New methods towards the prediction of the structure of transmembrane proteins and the simulation of helix-dynamics on large timescales

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich Physik
der Johann Wolfgang Goethe - Universität
in Frankfurt am Main

von
René Staritzbichler
aus Johannesburg

Frankfurt 2004
D F 1

Vom Fachbereich Physik der Johann Wolfgang Goethe-Universität
als Dissertation angenommen.

Dekan:	Prof. Dr. Wolf Aßmus
Gutachter:	Prof. Dr. Werner Mäntele Prof. Dr. Volkhard Helms
Datum der Disputation:	

Dedicated to my lovely son Leo Valentin (3.3.2004).

”Reports that say that something hasn’t happened are always interesting to me because, as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we don’t know.”

Donald H. Rumsfeld, US Secretary of Defense

„Da alles nicht mehr als eine bloße Erscheinung ist, perfekt, so wie es ist, nichts zu tun hat mit „gut“ oder „böse“, mit „annehmen“ oder „ablehnen“, kann man genausogut in Lachen ausbrechen.“

Logchen Rabjampa Rinpoche

Zusammenfassung

In dieser Arbeit werden neue Methoden zur dreidimensionalen Strukturvorhersage von Transmembranproteinen vorgestellt. Die Methoden setzen die Kenntnis der Primär- und der Sekundärstruktur der Proteine voraus.

Transmembranproteine spielen eine wesentliche Rolle in einer Vielzahl biologischer Prozesse in ihrer Funktion als aktive und passive Kanäle und als Rezeptoren. Fehlfaltungen und andere Störungen ihrer Funktion können am Entstehen von Krankheiten wie z.B. Alzheimer beteiligt sein. Einige Krankheitserreger docken an Transmembranproteine an. Zudem sind sie das Ziel von ca. 50% aller Medikamente. Das Wissen über die Struktur der Transmembranproteine bietet die Grundlage für die weitere Verarbeitung, z.B. in den Methoden des Drug-Designs. Experimentell konnten trotz ihrer Bedeutung und trotz großer Anstrengungen bisher nur wenige Strukturen bestimmt werden. Dies ist bedingt durch technische Schwierigkeiten in der Erzeugung zwei- oder dreidimensionaler Protein-Kristalle. Die Kristalle können mittels Röntgenbeugung bzw. Elektronenmikroskopie Aufschluß über die Struktur geben. Die Struktur kleinerer Transmembranproteine wie z.B. Glycophorin A konnten mit Hilfe der Kernspinresonanz-Spektroskopie gelöst werden, bei der keine Kristallisation der Proteine benötigt wird. Andere spektroskopische Methoden wie Infrarot-Spektroskopie lieferten weitere strukturell-funktionale Informationen.

Aufgrund der experimentellen Schwierigkeiten wären verlässliche, theoretische Vorhersagen von hohem Wert. Statistische und Homologie-Methoden, welche für lösliche Proteine sehr erfolgreich sind, können aufgrund der geringen Anzahl gelöster Strukturen noch nicht angewendet werden.

Eine Ab-initio-Strukturvorhersage ist die Suche nach derjenigen Struktur mit der geringsten freien Enthalpie, die mit höchster Wahrscheinlichkeit in der Natur zu finden ist. Diese Art der Strukturvorhersage beinhaltet die Berechnung der freien Enthalpie einer gegebenen Struktur und dem Suchen im Raum der möglichen Strukturen nach derjenigen mit der minimalen Enthalpie.

Bei den hier vorgestellten Methoden handelt es sich um ein Multi-Skalen-Verfahren, das es erlaubt, die Suche nach der nativen Struktur in verschiedenen Genauigkeiten bzw. Geschwindigkeiten durchzuführen. Eine erste Eingrenzung möglicher Strukturen wird mittels der Reduktion auf ein Residuenmodell und der Beschränkung auf ideale Helices erreicht. Hierdurch werden Freiheitsgrade und die Anzahl der zu berechnenden Wechselwirkungen wesentlich verringert. Die Ergebnisse dieser reduzierten Betrachtung dienen dann als Ausgangspunkt für die Verfeinerung durch atomistische MD-Simulationen der gesamten Proteine und ihrer Umgebung.

Die zu berechnende Energie setzt sich zusammen aus den Wechselwirkungen der Aminosäuren untereinander und den Wechselwirkungen der Aminosäuren mit ihrer Umgebung. Umgebung können die drei Phasen der wässrigen Lösung, der

Lipidkopfgruppen und die der Lipidketten sein. Die Grundidee ist, die Wechselwirkungsenergien der Aminosäuren im atomistischen Detail zu berechnen und die Ergebnisse mit einer mathematischen Funktion anzunähern. Die Verwendung dieser in der Anwendung weit weniger zeitaufwendigen, angenäherten Funktionen ist der Übergang zu der Residuenskala.

Insbesondere bei der Residuen-Residuen-Wechselwirkung werden im wesentlichen Beiträge zur inneren Energie berechnet. Der entropische Anteil ist schwierig vorherzusehen, da er eigentlich erst aus der Gesamtsituation heraus berechenbar wird. Die Entropie ist eine Funktion der Gesamtzahl der für ein System bei gegebenem Energieintervall erreichbaren Zustände. Einige entropische Beiträge können abgeschätzt werden, andere sind relativ konstant. Es bleiben jedoch Beiträge, die einem Residuenmodell nicht zugänglich sind.

Um die Wechselwirkung der Aminosäuren im atomistischen Detail zu berechnen, verwenden wir kurze MD-Simulationen von jeweils zwei Aminosäuren. Diese werden für eine ausreichende Anzahl verschiedener Abstände und Orientierungen der betreffenden Aminosäuren durchgeführt. [1] Es gibt verschiede Möglichkeiten, die Energiewerte durch Fit-Funktionen anzunähern. In dieser Arbeit wird ein Satz von einfachen, eindimensionalen und winkelparametrisierten Polynomfunktionen verwendet, die ohne Aufwand differenziert werden können. Anstelle einer Funktion je Residuenpaar, welche die Residuen in allen Orientierungen und Abständen beschreibt, sind es mehrere, rein abstandsabhängige Funktionen, welche die Residuen in jeweils einer bestimmten Orientierung beschreiben. Nach Einführung eines Maximalwertes für die in Betracht gezogenen Energien ist die Fitprozedur stabil und kann automatisiert werden. Die Daten der Fitfunktionen sind im atomistischen Detail berechnet worden, beim Durchsuchen des Konformationsraumes ist die Verwendung der Fitfunktionen jedoch wesentlich schneller als explizite Berechnungen.

Um die Wechselwirkung eines Proteins mit seiner Umgebung in einem Residuum-Modell zu berechnen, werden zwei Informationen benötigt. Erstens muss die Wechselwirkung der Residuen mit der jeweiligen Umgebung bekannt sein. Hierfür werden Literaturwerte für die freien Lösungs-Enthalpien der Residuen in Wasser und Chloroform verwendet. Diese wurden in unserer Arbeitsgruppe von Gu *et al.*[38] durch MD-Simulationen und Methoden der Multiconfiguration-Thermodynamic-Integration berechnet und sind in guter Übereinstimmung mit den existierenden experimentellen Werten. Der Übergang in der Kopfgruppenregion wird durch eine Funktion der hierfür üblichen Form angenähert.

Zweitens müssen Residuen, welche ihrer Umgebung ausgesetzt sind, von denen unterschieden werden, die zwischen den Helices „begraben“ sind, von denen also keine Wechselwirkung mit der Umgebung zu erwarten ist. Diese Unterscheidung wird insbesondere bei größeren Systemen wesentlich. Hierfür verwenden wir einen einfachen Kugelalgorithmus. Jedem Residuum wird eine Kugel zugeordnet,

[1] Aus technischen Gründen werden Tripeptide in helikaler Konformation (G-X-G) simuliert, von denen die äußeren beiden Glycine als Attrappen mitgeführt werden und keine Auswirkung auf die Energiewerte haben.

mit dem Mittelpunkt an der Position des C_α-Atoms. Für die Radien werden ebenfalls Literaturwerte genommen. Die Kugeln von Residuen, welche vergraben sind, werden vollständig von anderen Kugeln überlappt. Kugeln von Residuen an der Oberfläche der Proteine haben freie Oberflächen, welche nicht von anderen Kugeln überlappt werden - umso größere, je mehr sie der Umgebung ausgesetzt sind. Wir betrachten daher die Wechselwirkung mit der Umgebung als proportional zu der freien Oberfläche, die einem Residuum zugeordnet werden kann. Die Implementierung des Kugelalgorithmus basiert auf einer vektoriellen Integration in Kugelkoordinaten. Die ideale Größe der Kugeln hängt mit der Additivität der Lipophilizitäten zusammen. Diese wird zur Zeit von Gu *et al.* untersucht, indem die oben erwähnten Methoden zur Bestimmung der Lipophilizitäten auf Peptide mit mehreren Aminosäuren angewendet werden. Der Vergleich mit diesen Ergebnissen sollte eine Optimierung der Kugelradien ermöglichen.

Es hat sich herausgestellt, dass der Kugelalgorithmus in leicht modifizierter Form weitere sinnvolle Anwendungen gestattet. Wir bezeichnen einen Überlapp zweier Kugeln als Überlapp erster Ordnung usw. Je größer der Überlapp einer höheren Ordnung ist, desto dichter sind die Residuen gepackt. Eine höhere Dichte schränkt die Zahl möglicher Konformationen für einen bestimmten Energiebereich ein, d.h. sie führt zu einer geringeren Entropie und damit zu einer höheren freien Enthalpie. Daher ermöglicht die Berechnung des Überlapps höherer Ordnung eine Abschätzung der Seitenkettenentropie. Wird ein bestimmter Grenzwert der Dichte überschritten, wird die freie Enthalpie der Helices stark zunehmen, da nicht mehr ausreichend Raum für die Residuen vorhanden ist.

Nach der notwendigen Skalierung ist hiermit auch bei Überschreiten eines Grenzwertes ein Indikator für zu dichte Packungen gegeben, welche bei der üblichen paarweisen Berechnung des Residuenpotentials nicht erfasst werden können. Skaliert werden könnte der Algorithmus durch den Vergleich mit MD-Simulationen ganzer Helices. Für die Abschätzung der Seitenkettenentropie müssen Überlapp und Seitenkettenbeweglichkeit in den Simulationen miteinander verglichen werden. Des Weiteren muß der Grenzwert des Überlapps bestimmt werden, ab dem mit einem drastischen Energieanstieg aufgrund zu hoher Dichte der Seitenketten zu rechnen ist. Schliesslich bedarf es einer Skalierung des Überschreitens dieses Grenzwertes. Diese Skalierungen stehen noch aus. Trotz seiner konzeptuellen Einfachheit stellt der Kugelalgorithmus ein vielseitiges Hilfsmittel auf der Ebene des Residuenmodells dar.

Das aus zwei identischen Transmembranhelices gebildete Glycophorin A ist ein ideales Testsystem mit bekannter Struktur um die bislang vorgestellten Methoden zu prüfen. Wenn man Glycophorin A als starren Körper mit den ihm eigenen Symmetrien betrachtet, hat man einen fünfdimensionalen Konformationsraum, den man vollständig durchsuchen kann, ohne die im folgenden dargestellten Suchmethoden verwenden zu müssen. Die Residuen-Umgebung-Wechselwirkung als auch der Dichte-Packung-Algorithmus favorisieren die Umgebung der nativen Struktur. Die Residuen-Residuen-Wechselwirkung dagegen identifiziert eine andere Region als absolutes Minimum. Gründe hierfür können u.a. sein, dass die

Residuen-Residuen-Funktionen bislang nur für parallele Helices berechnet wurden und die Verkippung der Helices nicht ausreichend gut beschrieben wurde. Ein weiterer Faktor ist, dass die Funktionen Vakuumwechselwirkungen ohne polarisierbares Medium dazwischen beschreiben. Die Verwendung von im Vakuum berechneten Werten ist nicht immer sinnvoll. Zum Einen sind Teile der Helices durch Lipide voneinander getrennt, zum Anderen werden die Residuen häufig durch andere Residuen abgeschirmt. Hinzu kommt noch der Faktor, dass die Energien paarweise berechnet wurden und eine Abweichung in der Seitenkettenmobilität zu erwarten ist, wenn die Residuen in helikale Strukturen eingebettet sind. Dies sollte ebenfalls die Energiewerte beeinflussen. Park *et al.*[43] konnten durch eine Verschiebung der Residuenpositionen von den C_α-Positionen zu den geometrischen Zentren der Seitenketten die Qualität der Energiefunktionen deutlich verbessern. Ein ähnliches Vorgehen führte auch in dieser Arbeit zu einer wesentlichen Verbesserung der Ergebnisse. Dieser Schritt ist analog einer Kombination aus Dämpfungsterm und einer Beschränkung der Seitenkettenmobilität. Die Einführung eines Entfernungs-Grenzwertes, der alle Wechselwirkungen gleich Null setzt wenn die Entfernung oberhalb des Grenzwertes liegt, führte zu einer weiteren Verbesserung. Obwohl aufgrund dieser Modifikationen die Qualität der Energiefunktionen deutlich verbessert werden konnte, ist es zweifelhaft, ob zu dem aktuellen Stand für größere Systeme wie Bakteriorhodopsin verlässliche Ergebnisse zu erwarten wären. Insbesondere die Verkippung bis hin zu antiparallelen Helices spielt bei Bakteriorhodopsin eine noch wesentlichere Rolle. Die weiteren Methoden werden daher nur grundsätzlich beschrieben und einige einfache Tests duchgeführt.

Die ursprüngliche Idee für die Suche nach dem absoluten Minimum war eine Kombination aus einer Monte-Carlo Suche und einem genetischen Algorithmus. Es stellte sich heraus, dass dieses Vorgehen nicht der Komplexität der Energielandschaft gewachsen ist. Auch Methoden, die auf einer Gradientenbildung beruhen, sind dieser Aufgabe allein nicht gewachsen. Um in der weitläufigen und zerklüfteten Energielandschaft nicht in einem lokalen Minimum hängen zu bleiben, bedarf es einer Hilfestellung. Eine erfolgreiche Anwendung existierender Minimierungsalgorithmen ist nur dann zu erwarten, wenn die Anfangskonformation so geschickt gewählt wird, dass keine groben Hindernissse den Algorithmus stören können. Den Konformationsraum vollständig zu durchsuchen, ist zu aufwendig; andererseits dürfen keine Regionen unabgetastet bleiben. Wir betrachten die Energielandschaft zunächst aus der Vogelperspektive, indem wir die Helices auf einem Equidistanzgitter plazieren, deren Abstand größer ist als der zu erwartende Abstand im Protein. Die Helices auf diesem Gitter zu optimieren, um sie dann loszulassen und dem Minimierungsalgorithmus zu übergeben, ist längst nicht so aufwendig wie das Durchsuchen des gesamten Konformationsraumes und gleichzeitig ausreichend vollständig. Um die Helices auf dem Gitter zu optimieren, werden die in Frage kommenden Helices in Dreier-Kombinationen systematisch abgetastet. Aus diesen Dreier-Scans lassen sich dann die minimalen Konformationen des gesamten Proteins rekombinieren.

Praktisch bedeutet dies für die Untersuchung eines Proteins, dessen Struktur

gänzlich unbekannt ist, dass zunächst die möglichen Gitteranordnungen bestimmt werden müssen. Die möglichen Gittertypen hängen von der Anzahl der Helices ab. Die Helices werden in allen Permutationen den Gitterpunkten zugeordnet. Ausscheidungskriterium ist die Länge der Peptidkette zwischen den Helices. Für die in Frage kommenden Helix-Dreier-Kombinationen werden dann systematische Energieberechnungen auf der Basis des entwickelten Residuenpotentials durchgeführt und in „Karten“ abgespeichert. Die Wahl von Helix-Trios für das systematische Durchsuchen erlaubt, ausgehend von den Konformationen minimaler Energie des Ausgangstrios, die darauffolgenden Helices schnell und zuverlässig in Konformationen ebenfalls minimaler Energie hinzuzufügen. Da die Berechnung der Karten und die Rekombination nur auf der Residuen-Residuen-Wechselwirkung beruht, wird dem Minimierungsalgorithmus ein Ensemble möglicher Konformationen übergeben[2].

Der letzte Schritt der Strukturbestimmung wäre das Hinzufügen der zuvor vernachlässigten Residuen und der Seitenketten und die Übergabe der resultierenden Protein-Konformationen an ein MD-Programm, um die Struktur im atomistischen Detail zu verfeinern.

Abschließend werden einige Ansätze für eine Helix-Dynamik auf größeren Zeitskalen vorgestellt, die es ermöglichen sollen, die Dynamik von Proteinen über die zeitlichen Begrenzungen von MD-Simulationen hinaus zu beschreiben. Die Zeiträume, über die gegenwärtig mit MD ein System von ungefähr 100 000 Atomen simuliert werden kann, bewegen sich in der Größenordnung von einigen 10 Nanosekunden.

Die Funktion der Proteine ist häufig mit einer teilweise großräumigen Konformationsänderung verbunden. Bei der Verschiebung von Helices kann beispielsweise zwischen offenen und geschlossenen Zustand von Kanälen hin- und hergeschaltet werden, wie bei dem spannungsgesteuertem K^+ Kanal KvAP. Andere Proteine, wie Bakteriorhodopsin und Ca^2+ATPase, schalten durch Strukturänderung zwischen aktivem und ruhenden Zustand hin und her. Bakteriorhodopsin durchläuft durch die Absorption eines Photons einen Zyklus, bei dem ein Proton entgegen eines äußeren Gradienten durch das Protein gepumpt wird.

Unser Ansatz besteht aus einem Zwei-Schritt-Algorithmus. Im ersten Schritt werden die Helices als starre Körper betrachtet und für ein bestimmtes Zeitintervall werden die Bewegungsgleichungen starrer Körper gelöst. Die Berechnung der internen Dynamik, welche sich innerhalb des Residuenmodells auf die Verbiegung der Helices beschränkt, wird im zweiten Schritt berechnet. Hierfür müssen die Kräfte, welche auf die Residuen wirken, in Drehmomente umgerechnet werden, welche auf die Bindungen zwischen den Rückgrat-Atomen wirken. Die wesentlichen Freiheitsgrade der Peptidgeometrie sind jeweils die beiden Torsionwinkel der Bindungen des Protein-Rückgrats an die C_α-Atome. Für deren Auslenkung von der idealen Helix-Konformation werden die Bewegungsgleichungen gelöst - unter der Annahme, dass die Gleichungen entkoppeln.

[2]Im Prinzip können die gleichen Methoden auch auf der atomistischen Skala angewendet werden.

Abstract

Transmembrane proteins play crucial roles in biological systems as active or passive channels and receptors. Experimentally only few structures could be determined so far. Gaining structural insights enables besides a general understanding of biological mechanisms also further processing such as in drug design. Due to the lack of experimental data, reliable theoretical predictions would be of high value. However, for the same reason, missing data, the knowledge-based class of prediction methods that is well established for soluble proteins can not be applied. The goal of predicting transmembrane protein structures with ab initio methods demands locating the free energy minimum. Main difficulties here are, first, the computational costs of explicitly calculating all involved interactions and, second, providing an algorithm that is capable of finding the minimum within an extremely complex and rugged energy landscape. We have developed promising energy functions that describe the interactions of amino acids on a residue level, reducing computational costs while still containing most information on the atomistic level. We have also found a way to describe the interaction of the residues with its surrounding in a realistic manner by distinguishing residues exposed to the environment from those buried within helices using a sphere algorithm. The sphere algorithm can also be applied for a different purpose: one can measure how densely sidechains are packed for certain helical conformations, and thereby get an estimate of the sidechain entropy. In addition, overcrowding effects can be identified which are not well-described by the energy functions due to the pairwise calculation. To determine the absolute free energy minimum, we assume the helices to be located on an equidistance grid with slightly larger distances than to be expected. Optimizing the helices on the grid provides a starting point that should enable common minimizing algorithms, gradient-based or not, to find the absolute minimum beyond the grid.

To simulate the dynamics of the helices on large time scales, we split them into rigid body dynamics and internal dynamics in terms of the dihedrals. The former one is well-known with its inherent problem of numerical drift and plenty of approaches to it, among which we have chosen the quaternions to represent the rotation of the rigid bodies. The latter one requires a detailed analysis of the torque size exerted on the dihedrals caused by the forces acting on the residues.

Contents

List of Figures

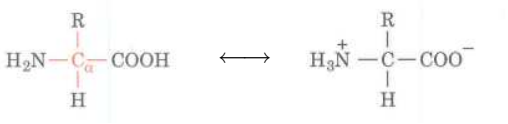
List of Tables

Chapter 1

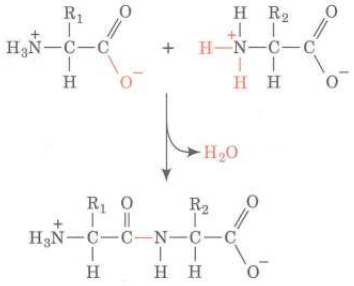
Introduction

1.1 About proteins

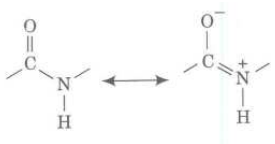
Proteins are made of one or more linear chains of amino acids, folded together. There are 20 natural amino acids that have the same headgroups



and differ by their sidechains. The differences in the sidechains lead to their very different chemical and physical behaviour. Amino acids and their derrivatives also have independent biological roles, like poisons and neurotransmitters. In proteins, they are connected via the peptide bond.



The peptide bond plays a major role for the protein characteristics. It has 40% double bond character:



This results in a rigid, planar structure for the peptide bond. The backbone, the peptide-bonded main chain, has two kinds of angles that are free within a broad range[1]. One belongs to the C_α-N bond (Φ) and another belongs to the C_α-C bond (Ψ). The other angles as well as the lengths alter on small scales only. Hence, the structure of a peptide with *n* residues is defined by a set of *2n-2* dihedrals (rotational angles)[2]. See the left of figure 1.1.

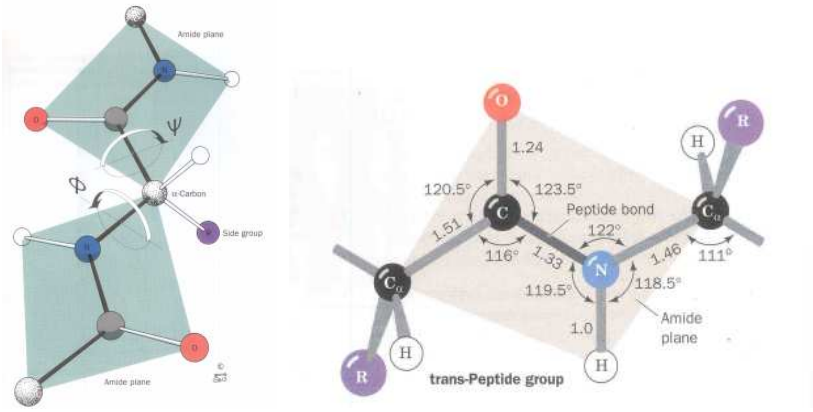


Figure 1.1: Angles and lenghts of the backbone-bonds. Taken from [1].

The way the peptides fold is not random but leads to the formation of highly ordered and energetically favoured secondary structure elements, where the most common are α-helices and β-strands. The dihedrals have identical values throughout the secondary structure element. Figure 1.2 shows them for an α-helix. Not all peptides form secondary structure elements. Some amino acids support the formation and some break them. In proteins, parts will remain uncoiled and parts will form helices or strands or both. These then assemble into the tertiary and quarternary structure. In which way they will lay together depends again on the properties of the involved amino acids. The high diversity in structure and function of proteins is solely grounded in the sequence, that is, in the combination of the amino acids.

Many proteins fold spontaneously into their native state, but some need assistance to fold properly by other proteins, called chaperons. The protein sequence specifies if it is a soluble protein within the cell or if it is located in the membrane. Many proteins include additional molecules, cofactors, which are crucial for the function of proteins, like the retinal chromophore in bacteriorhodopsin. The retinal absorbs light, undergoes isomerization and thereby induces the process of proton transport [2].

[1] The possible values of the dihedrals are typically plotted in Ramachandran diagrams.

[2] The conformations of the sidechains are of course not determined by the backbone dihedrals. They might be restricted by other sidechains, but still have some freedom to move.

1.1.1 Importance of membrane proteins

The human body contains nearly 100,000 different types of proteins, which can be either soluble or membrane proteins. The ratio of the proteins in the cell that are membrane proteins vary in the literature from 15-30% [3]. Transmembrane (TM) proteins are crucial for numerous biological processes. The photosynthetic reaction centers I and II are large protein complexes that belong to those membrane proteins involved in photosynthesis. The cytochrome bc_1 complex is involved in the respiratory chain. Rhodopsin is the protein that absorbs the light in the rod cells of the retina and creates an electric impulse by pumping a proton through its structure. Listing all their tasks would fill a book by itself.

In order to understand in detail the function of membrane proteins, knowledge about the structure is required. In this way, structural information enhances the understanding of the mechanisms of life. Also the fact that about half of all drugs are binding to membrane proteins is underlining the importance of determining the structures of transmembrane proteins.

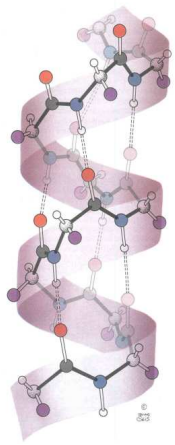


Figure 1.2: The peptide backbone in right-handed α-helical conformation ($\Phi = -57^\circ$, $\Psi = -47^\circ$). The dashed lines indicate the hydrogen bonds between the N-H group of the n^{th} and the C=O group of the $(n+4)^{th}$ residue. Taken from [1].

Thus, the determination of their structure is not a philosophical question, but it is of direct benefit, such as in the field of drug design, where structural information is essential. By blocking or activating a certain membrane protein, it is possible to bias the process it is involved in. Knowing the structure of the protein, it is possible to develop models of the appearance of the drug that shall dock to the protein and inhibit or activate certain processes. This can speed up the classical trial and error process enormously. Another good reason to deal with solving protein structures is that in many diseases misfolding or misfunctioning of proteins are

involved. Alzheimer, Parkinson, cystic fibrosis and atherosclerosis are some of the diseases which are caused by the misfolding of proteins [4, 5]. Because many of these proteins are membrane-associated, understanding the interaction of amino acids with bilayer and intermediate should promote progress in this field. There are indications that the interaction of proteins and membranes may play a role in folding or misfolding. First, there are peptides that are unfolded in solution and that fold in the interface. Second, some soluble, correctly folded proteins can be destabilized by interactions with the membrane [6, 7]. Also the misfolding of integral membrane proteins may be involved in disease [8]. One of the relatively few membrane proteins that are related to these diseases for which studies of the misfolded state were performed is diacylglycerol kinase [9].

Utilising structural information for the design of artificial flavours would present an application of doubtable use.

There are three main types of membrane proteins:

- receptors
- active transporters
- passive transporters

These types can be further categorized into more than 600 families according to their number of helices and sequence similarities [10]. These genomic analyses are based on the fact that the secondary structure prediction is reliable for the transmembrane part of membrane proteins (more than 90% accuracy [11, 12]), whereas the accuracy for soluble proteins is only about 75%. The large number of different membrane proteins reflects their specificity.

1.1.2 3D-structures of membrane proteins

Determining the structures of membrane proteins is a catchy business and requires artful procedures in cristallography and spectroscopy. Most transmembrane proteins do not keep their structure outside of the membrane. When within their native membrane there are too many different proteins and lipids to distinguish between input from the considered protein and noise. Due to these problems only 52 out of appoximately 17,000 structures have been determined so far [13].

There are two main strategies to solve structures of membrane proteins. The first is based on the extraction of proteins with detergent that wrap the lipophilic core of the proteins with the lipophilic part of the detergent. In some cases individual proteins are tagged together with antibody-fragments that dock to the protein. The antibodies are connected via bridging molecules. Through this, the proteins form 3D crystals, whose structures are determined by X-ray diffraction using synchrotron radiation [14]

The second strategy collects sufficient proteins of interest in a lipid-bilayer. These 2D crystals are investigated via electron microscopy [15].

Structures of single TM-helices or of TM-helix dimers have been determined by solid-state nuclear magnetic resonance (NMR) as well, like the one of glycophorin A [16]. The limiting factor for NMR is the relative molecular mass of the molecules. In the case of soluble proteins, the maximum mass of a structure solved is currently 900 kDa [17], due to the high internal symmetry of the protein. For TM-proteins, the state of the art is around 50 kDa.

But even structures that are solved do not reveal all information. Pathways for e.g. proton pumping require further investigations like mutagenesis experiments. By exchanging specific residues one can test whether a residue is involved in the pathway or not.

1.1.3 Energetics of membrane protein stability

The unfolding of membrane proteins in aqueous solution indicates that there are some inherent structural differences between soluble and membrane proteins. Disulfide bridges that can have crucial effects in soluble proteins [18] are not found in membrane embedded regions. One could expect that they try to bury hydrophilic residues that are located in the helices into the interior of the protein. Even though the thermodynamical costs of inserting charged or highly polar compounds into lipophilic environments are high, it does not seem to be a reliable rule that hydrophilic residues are buried in the interior [19]. Investigation of known structures have shown no evidence that membrane proteins are ”inside-out” proteins with a polar core and an apolar exterior. Instead, their interiors are about as lipophilic as those of soluble proteins [20]. There are also single transmembrane helices that contain a limited but present number of polar or charged groups. Moreover, the interactions between transmembrane helices can be very stable in the absence of any hydrogen-bond or salt-bridge. This would suggest that the packing of α-helices in the membrane is better understood by means of van der Waals interaction. On the other hand, interhelical hydrogen bonds cause strong interactions [21, 22].

While in the case of interhelical interactions the role of hydrogen bonds might vary, it is predominant for the stablility of transmembrane helices. The highly polar peptide bonds of the backbone must participate in hydrogen bonds, otherwise the costs for the insertion into the membrane are much too high. Helices as well as β-sheets are built according to this principle. The forming of the helices has to take place either in solution or in the interface. Inside the lipophilic core no uncoiled peptides can be found. This is a backbone effect that has the same validity for all residues.

As mentioned before, for the prediction of membrane helices within the sequence using a lipophilicity scale combined with accessible surface information is very reliable [11].

Nevertheless, certain regular patterns were found over-represented in the sequence of transmembrane proteins. The genomic analysis reveals that the GxxxG and the GxxxxxxG motifs are among the most prevalent. It also shows that there

are certain tendencies for some of the residues to be conserved among the proteins from the same family [10]. Furthermore, six amino acids (Leu, Ile, Val, Phe, Ala and Gly) account for two-thirds of all transmembrane residues [23].

The folding of α-helical integral membrane proteins has first been postulated to follow a two-stage path, involving the formation of individually stable α-helices and their association giving rise to tertiary and quaternary structures [24]. The idea for this model came from a series of experiments that demonstrated that isolated fragments of bacteriorhodopsin in lipid bilayers can reassemble spontanously into a fully functional form, consistent with the native protein residing in a free energy minimum. Next, a four-step model was proposed, containing partitioning, folding, insertion and association, which can be combined in different ways [25], (see figure 3.3 on page 30).

1.1.4 Conformational dynamics of membrane proteins

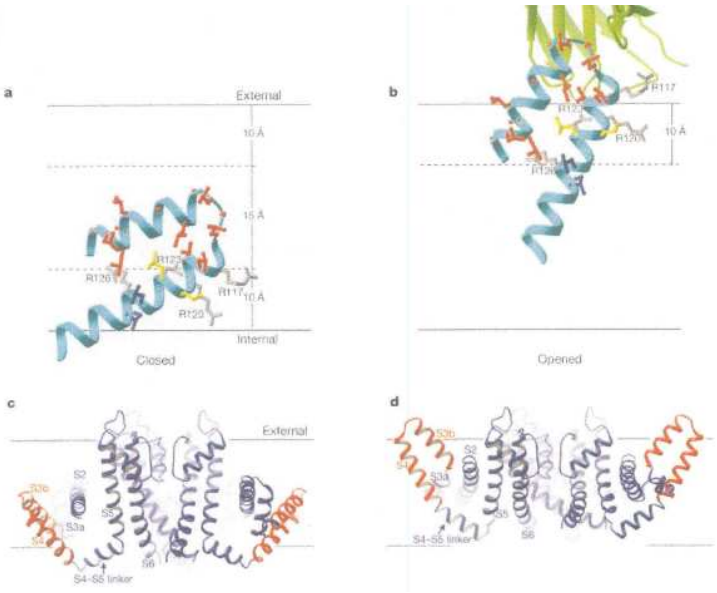


Figure 1.3: The different conformations of KvAP. The two images on the left show the protein in the closed state. The two on the right in the open state. Both lower images show the whole protein. The upper show in detail the conformations of the helices belonging to the selectivity filter, that undergo large conformational displacements. Taken from [26].

The functions of membrane proteins often involve conformational changes. Sometimes they are regulated by switching between active and inactive or open and closed states. In both transitions of large extent can occur. The best characterised

systems for the first scenario are bacteriorhodopsin and Ca ATPase. The former performs a proton pumping cycle during which helices E and F tilt together away from the center of the molecule, causing a shift of $3-4\mathring{A}$ at the loop between them [27]. This tilt enables one of the transfer-steps of a proton within the molecule. The latter undergoes a large rotation of an external domain, accompanied by pumping of Ca^{2+}-ions [28].

The voltage dependent K^+-channel KvAP is a good example for the second scenario [26], as plotted in figure 1.3. First, it has two distinct states that are sensitive to small changes in the applied voltage. It works like a switch, controlled by voltage. The two conformations are connected to the "open" and "close" state of the channel. The selectivity filter consists of 4 helices that can be tilted by more than 45° to the perpendicular of the membrane and are in a symmetrical disposition at the gate of the pore. The helices acts as a selectivity filter by means of a charged amino acid at the top of the four helices. (In other K^+-channels the selectivity filters work mainly through the dipoles of the helices [29].) Through the voltage change the helices move and actually break at certain position, enabling far opening. From the upper figures *a* and *b*, one can see how far the filter-helices move through the membrane. Clamp experiments investigate the voltage-dependency of the function of proteins [30].

1.2 About structure prediction

The difficulties in solving structures experimentally make theoretical predictions valuable. For soluble proteins about 20 000 structures are known. There is a solid base for knowledge based methods such as homology modeling and statistical potentials. The former seeks and finds analogous sequence elements in solved structures and models the unknown proteins using these similarities. The latter describes the total interaction between residues without detailed knowledge about what interactions are involved in which way. Because different energetic principles apply in membrane proteins the results are not transferable. On the other hand, there are not enough transmembrane protein structures solved to apply these methods in a reliable manner.

Besides their striking success for the soluble proteins, these approaches are not very satisfying from an analytical viewpoint. The possibility to learn about the underlying mechanisms is limited.

Starting from the opposite direction, one, in principle, can always perform explicit calculations of the interactions involved. Except for some entropical contributions, the interactions themselves are well known. The limiting factor for this approach are the computer capacities. Proteins may contain some thousand atoms and, additionally, it is necessary to include some thousand atoms from their surrounding. A well established method is Molecular Dynamics (MD) simulation, a powerful tool to do this in a classical approximation. With the present state of com-

putational capacities, systems of up to 100 000 atoms can be simulated for about 10 ns [31, 32]. Many mechanisms which are out of reach for experimental methods can be investigated with MD, but often biological processes where membrane proteins are involved occur on much longer timescales. The folding of proteins, for example, is out of reach - except for a few small ones, the so called "downhill folders" that fold within a few microseconds. Additionally, MD-simulations do not easily provide information on the thermodynamics of the folding. Lastly, MD-simulations are not capable of describing processes ruled by quantum mechanics - such as transitions between distinct quantum mechanical states, e.g. forming or breaking bonds. Dynamics based on quantum mechanics are much more time consuming and are not yet appliable to systems of comparable size or similar timescales. Several approaches are trying to overcome the quantum mechanical costs and the limitations of classical MD by combining them [33, 34]. Effective treatments of coupling one quantum degree of freedom with a classical system are also used to describe biological processes [35].

It is thus necessary to explore the systems beyond the limitations of MD simplifications. Some models were developed, starting from simple contact potentials [36] for residues (amino acids) to more refined energy functions, like continuum electrostatics or implicit models [37]. Most of them are on the residue level.

Another difficulty most search strategies have to face is that they are gradient-based and therefore are always endangered getting stuck in a local minimum of the energy landscape, instead of finding their goal, the global minimum.

1.3 About thermodynamics of biological systems

Molecules interact via various kinds of interactions like electrostatics, van-der-Waals, hydrogen-bonds and disulfide-bridges. The interactions alone are not sufficient to describe the behaviour of chemical and biological processes. Other information is necessary. The order of the system that includes geometrical conditions and the interactions has to be known. This is an information which cannot be obtained by simply looking at a system at a certain time, but is of statistical nature, i.e. it is necessary to know how a large number of systems in similar conditions will behave. The order of a system is described by its entropy:

$$S = k \ln \Omega \tag{1.1}$$

where Ω is the number of states a system can reach for a certain energy-range, and k the Boltzmann constant. How Ω exactly looks like depends on the conditions the system is in.

Thermodynamical systems can be described by a couple of different so-called thermodynamical potentials. According to the conditions, one of the potentials is always most suited for the description of the system. The conditions that determine which is the most favourable ensemble are mostly pressure p, volume V,

temperature T and entropy.

When pressure, number of particles N and temperature are constant, a closed system will try to reach the state where its free energy[3],

$$G(T,p,N) = U + pV - TS \tag{1.2}$$

with its differential,

$$dG = -SdT + Vdp + \mu dN, \qquad \mu \text{ chemical potential}$$

has a minimum. This is the case for most biological processes. Therefore those biological systems will try to minimize their free energy that thereby can be considered as the driving force. U is the inner energy as defined in the first law of thermodynamics

$$dU = \delta Q + \delta A$$

where δQ is the amount of heat and δA the amount of work.

From equation 1.2, it can be seen that an increase of the entropy will lead to lower free energy and conversely. Equation 1.1 shows that the entropy increases when the number of states increases, which the considered system can reach for a certain energy-range. If a degree of freedom of a molecule diminishes, for example, the sidechain-mobility, the entropy will shrink and the free energy will increase - what contradicts the aim of the system to reduce its free energy. In some cases, the interactions might favour a certain conformational change, but when the entropy opposes this change and prevails, then this change will not take place.

1.4 About this thesis

In this thesis, a novel approach to predict the structure of a membrane protein based on knowledge about the sequence and the location of secondary structure elements is presented. We have tried to find a way to keep as much atomistic information while making it faster and easier to apply than atomistic calculations. Structure prediction is the determination of the minimum of the free energy [4] in the conformational space. The dimensions of the conformational space are the degrees of freedom of the molecule whose structure shall be predicted. The determination contains two main parts:

- calculation of the free energy
- search through conformational space.

[3] dt.: freie Enthalpie

[4] see explanation for this in section 1.3 on page 8

The free energy has two main contributions:

$$G_{tot} = G_{res-res} + G_{res-env}$$

the residue-residue interaction $G_{res-res}$ in vacuum and the interaction of the residues with the environment $G_{res-env}$. The residues can be exposed to aqueous solution, to the lipophilic core or to the intermediate. Membranes that are themselves complex systems can only be treated implicitly in a residue model. In figure 1.4, the common shape of a function treating the membrane implicitly is shown. There are

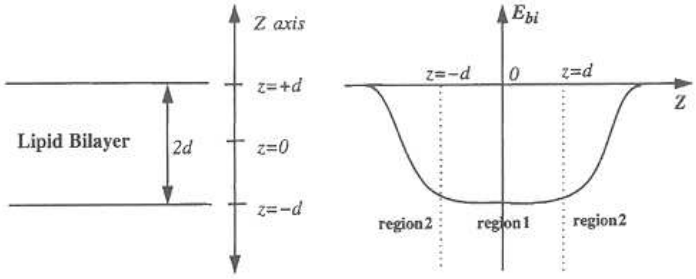


Figure 1.4: The z dependence of the interaction E_{bl} of a residue in a membrane. The lipophilic core has a thickness of 2d, the headgroups of 1d each.

experimental as well as calculated values for the interaction of each residue with lipophilic and hydrophilic environment: the lipophilicity scales [38, 25]. The interaction of a residue with the headgroup area is usually approximated, as in figure 1.4. Beside the different media the residue might be exposed to, it also happens that some residues are buried between others, without exposure to the environment. Especially in large systems this effect should be regarded. The interaction of the residues with the environment is considered as being proportional to their exposure. To measure the exposure of each residue, the sphere-algorithm was developed. Each residues gets a sphere assigned with a radius related to its own size. Those that are buried will be completely overlapped by neighbouring spheres, and the exposed ones will have non-overlapped, free surface. The free energy, $G_{res-env}$, is considered as being proportional to the free surface.

The sphere algorithm can also be used for measuring tight packing by using overlaps of higher order as an indicator for high densities. On one hand, the entropical terms of the free energy are proportional to solvent exposed surface, while on the other, they depend on the mobility of e.g. the sidechains. Large overlap of higher order means higher density, which results in fewer states that the protein can reach, and this means that the entropy is smaller. By determining both accessible surface and density, the sphere algorithm is a promising tool for both measuring the exposure and some entropical terms.

Atomistic MD-simulations are used to calculate a realistic residue-residue interaction field. For a pair of residues in a certain spatial orientation and distance

a short MD-simulation is performed, giving $G_{res-res}$. This has to be done for a sufficient number of distances and orientations as well as for all residue pairs. It is the computationally most expensive part. The resulting energy values can either be used as a look-up-table or are approximated by various kinds of fit-functions in the search of conformational space. It must be pointed out that this procedure keeps most of the atomistic information in a fast applicable fit function. The most expensive part, running hundred thousands of short MD-simulations, is done beforehand.

Results obtained at the residue level are the base for further refinment using all atom calculations.

The ideas of the multiscaling and the design of the residue-residue energy functions are described in chapter 2. Chapter 3 includes descriptions of the basic ideas of the sphere algorithm as well as of applications on a couple of test systems. In chapter 4 the methods developed in chapter 2 and 3 are applied to glycophorin A. Also, some further refinements of the residue-residue energy functions are explained and applied there.

For the search through conformational space, we first developed three methods:

- Monte-Carlo method
- genetic pair algorithm
- fine tuning

It turned out that this approach is not capable of locating the absolute minimum of the exeedingly complex energy landscape within reasonable time. To evercome this insufficiency, another set of methods was developed:

- optimize the orientation of the helices on an equidistant grid
- use resulting conformations as starting points for an energy minimizer

This is explained in detail in chapter 5.

Another topic of this work is to overcome the time limitations of MD-simulations that can only be extended by growing computational capacities. When using the atomistic level, the integration timesteps have to be rather short to preserve the constraints like bondlengths, etc. Based on the tools for structure prediction, we propose to perform helix-dynamics in a 2-step-model:

- rigid body dynamics
- backbone dynamics

For a certain conformation of the helices at a certain time, the total force that acts on each residue is calculated. Two types of kinetics, considered as decoupled, are calculated. In one step the helices are regarded as rigid bodies and the equation of

motions are solved. This is done using quaternions instead of rotation-matrices to overcome the inherent problem of numerical drift.

Next, the internal changes are calculated. The dihedrals Φ_i and Ψ_i are the only degrees of freedom. The equations of motion are solved for the deviation of the dihedrals from their values when being located in the ideal helix. The transformation of the spatial forces into functions of Φ_i and Ψ_i is the largest effort in the dynamics. The rigid body dynamics are calculated in an absolute space, while the internal dynamics are calculated in the space of the dihedrals. The timesteps have to be taken small enough to justify the decoupling of the internal and the inter-helical kinetics. The dynamics are the topic of chapter 6.

Chapter 7.1 contains plenty of suggestions what should be done next. Chapter 8 asserts the conclusions, and the appendix lists some of the methods and tools that were used here. Among them, especially, the vector transformations and decompositions were used extensively. The appendix also contains some more investigations about the sphere algorithm and applications, for the real enthusiasts.

Except for the atomistic MD simulations, all calculations of this work were performed with a new class library in $C++$ that was developed and implemented as part of this thesis.

Chapter 2

Residue-residue energy functions

2.1 Multi-scaling

The three-dimensional structure of a molecule that is populated in nature is the one with the lowest free energy. More correctly, the structure of a molecule at room temperature is best described by an ensemble of structures within the basin of the free energy minimum if we take into account the constant dynamics of molecular systems. Predicting the structure of a molecule means finding the conformation where it has its lowest free energy. Searching for this minimum requires calculating the energies for many conformations, especially when the system is large and no constraints are known.

In the all-atom representation of macromolecules, a large number of interactions must be calculated. The number N of pair interactions for n atoms is:

$$N = \frac{n(n-1)}{2}.$$

Bacteriorhodopsin has 1349 atoms[1]. Therefore, $1349 \cdot 1348/2 = 909226$ pair interactions have to be calculated at the atomistic level, plus 1349 interactions of the atoms with the environment. Introducing a residue-residue energy function leads to a significant reduction of interactions. The number of residues in bacteriorhodopsin is 222. Therefore, only $222 \cdot 221/2 = 24531 (+222 = 24753)$ interactions have to be calculated at residue level. The residues can be represented in various ways. For one dimensional energy functions, they can be simply represented by their C_α-atoms. For energy functions that are not purely distance dependent but also include the orientation of the residues, one or two more fixing-points are needed. When representing the residues by two atoms, the C_β-atom or the center of mass of the sidechains are useful choices for the second atom. For a full representation of the orientation of the residues three atoms are needed. Here the three backbone atoms, C_α, C, and N are most useful.

[1] According to 1C3W.pdb.

The introduction of the residue scale reduces the number of interactions significantly. Cutting off the loops is the next logical step, because focusing on the relative orientation of the helices leads to another reduction of the degrees of freedom. The structure of a peptide is determined by its dihedral angles. For n residues there are $N = 2n - 2$ dihedrals. While in the helices, the dihedrals vary only a little, they can adopt a wide range in the loops. The structure of a helical protein is basically determined by the dihedrals in the loops between the helices. Neglecting the terminal-loops the number of degrees of freedom in such a protein is

$$N = \sum_i 2n_i + 2,$$

where i is the number of loops and n_i the number of residues within that loop. Cutting off the loops reduces the degrees of freedom to

$$N = 6j$$

where j is the number of helices. Each helix has three degrees of freedom for the position of its center of mass and the three Euler-angles that define its orientation.

Why use a multiscale method? For large systems containing e.g. 7 helices like bacteriorhodopsin the conformational space is enormous. Performing costly all atom calculations while scanning an unconstrained conformation space would slow down the search beyond managability. Im & Brooks attempt ab initio predictions of helical TM proteins (F_o ATPase) using 20 multi-replica all-atom MD simulations at temperatures between 300 and 600K with a Generalised Born model [39] for the solvent and for the membrane portion (unpublished). These are still extremely expensive calculations. On the other hand, all simplifications will influence the quality of the prediction. Combining both simplified and all-atom model is a way out. For the first screen, where all conformations are equally probable, the simplest and fastest of the available scales should be choosen. That is the level of helices in a residue representation here. After the space of likely structures is reduced sufficiently for the helices on the residue scale, they can be passed to the full atomistic representation performing MD-simulations for the refinement.

2.2 Calculation of energy values

The main motivation to use a residue-residue potential is to reduce the computational complexity. Fewer interactions have to be calculated, no sampling of side-chain interactions is necessary and, finally, smoother potential functions allow using larger time-steps in dynamic simulations. Even with increasing computational capacities, the potentials would not become needless in the future but would allow to progress to larger systems.

How should one obtain such potentials? The first idea was to use contact potentials, but they would have been pretty rough, neglecting any distance- and

orientation-dependency. It was more tempting to find a way to calculate the energies atomistically beforehand and store the values in a fit-function. This would be very fast when applying it during the search of the conformational space but would carry the atomistic information in an implicit way. Although this seems obvious, it has not been described in the literature so far for the step from atomistic to residue potential. The atomistic energies have to be calculated for a large number of distances and relative orientations for all residue pairs to obtain reliable fit-functions. After suggesting this procedure to my supervisor Volkhard Helms, he suggested the usage of short MD-simulations to calculate the interaction energies of the residue pairs. This project was given to Markus Elsner as a diploma-thesis and was jointly supervised by Volkhard and myself [40]. Markus wrote the scripts that automatised the procedure of creating residues in various conformations and running MD-simulations. Later, after Markus's work was finished, the project was continued by a now graduate student, Yungki Park. He gave the project a slightly new orientation by primarily focussing on the van der Waals interaction energies.

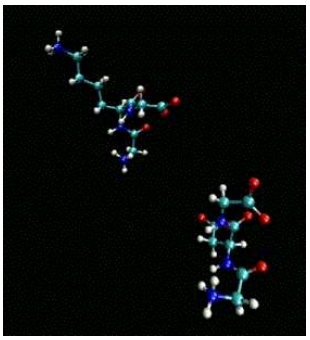


Figure 2.1: Calculation of the residue-residue interaction using tripeptides in MD-simulation. This picture is a snapshot from the simulation, done by Markus Elsner.

The MD simulations were performed using the NWChem-package [41]. Tripeptides were used for the simulation to construct residues without N-terminal and C-terminal blocking groups. For a given residue pair X and Y G-X-G and G-Y-G were used in helical conformation. The glycines were used as dummies, which means that they would not interact with the other residues of the system. These were generated using InsightII (Accelrys, USA). The defining variables are the distance between the helix-axis, the two rotational angles around the axis and the tilt angle between the helices - see figure 2.2. The van der Waals and the electrostatic term are listed separately and summed up. According to the stepsizes used, the calculations took between 4 and 10 days per residue-pair.

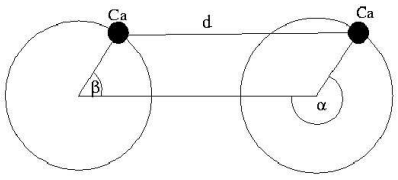


Figure 2.2: The variables in which the residue-residue energies were calculated. Taken from [40].

2.3 Fitting of the data

One of the major difficulties of the project turned out to be the fitting of the data. The reasons were that both the noise of the simulation data and the functional form of the potential are not known.

Markus tried a triple-fit. First he fitted the data as a function of the distance which was the easiest part. The fit parameters themselves should be functions of the angle, so he first fitted them over one of the angles and the resulting parameters again as a function of the second angle. This worked well, except that automatisation was not managable, and nobody would have liked to perform more than 200 fits by hand.

The next thing Markus tried was a fit algorithm based on genetical programming [42]. This was continued by Yungki. But here also the automatising caused serious problems for the triple fit.

The first step that led to results was a purely distance dependent function. This was rather easy business, but important information was lost - the attractive part of the interaction. Without any attraction one could hardly expect the helices to fold at all. Why was the fitting so bad when placing the helices at a certain distance and averaging over all angles? For each angle-combination, these functions had a repulsive and an attractive part (if there was any attraction, of course). But the distance at which the steep repulsive part of the energy-function becomes dominant depends on the spatial orientation. Averaging over these functions has the effect of averaging nearly all attraction out.

Nonetheless, Yunkgi managed to get a useful ranking of conformations from these fits. Even though the attractive part was missing, he could identify the native structure of glycophorin A [43] through the ranking. Yungki then introduced two modifications in his calculations that were very successfull. He averaged only over the central part of the angular range, receiving attracting fit functions. By using the C_β positions, instead of the C_α positions as the position of the residues, Yungki was able to increase the quality of the calculated energy landscape significantly, so that the native structure of glycophorin A could be identified [43]. This will be further discussed in detail in chapter 4.

As a last opportunity to include the angle dependency into the energy functions, one could always use the data as a look-up table, but no great insight could be expected that way.

Then I found a way that made the fitting very easy and stable. Furthermore, it was exactly the kind of function that I needed for dynamics simulations. I took an angle-parameterised set of purely distance dependent functions

$$f_{\alpha_1,\alpha_2}(d) = \sum_{i=1}^{n} \frac{a_i}{d^i}, \quad n \text{ number of fitting terms} \tag{2.1}$$

Thus, the derivatives, the forces, are instantaneously calculated.

The tilt. At first, the data was compiled with parallel helices. But in most real systems the helices are tilted and also oriented antiparallel. For glycophorin A it is known that the helices are tilted about 40°. We did some simulations with tilted helices and tried to describe the effect of the tilt on the fit functions by a factor analogous to the multipole-expansion of the electrostatic interaction.

$$\begin{aligned} W &= W^{monopole} + W_I^{monopole-dipole} + W_{II}^{monopole-dipole} + W^{dipole} \\ &= \frac{q_1 q_2}{r} - q_1 \frac{\vec{e}_r \cdot \vec{p}_2}{r^2} + q_2 \frac{\vec{e}_r \cdot \vec{p}_1}{r^2} + \frac{\vec{p}_1 \cdot \vec{p}_2 - 3(\vec{p}_1 \cdot \vec{e}_r)(\vec{p}_2 \cdot \vec{e}_r)}{r^3}. \end{aligned} \tag{2.2}$$

But this turned out to be rather unsuccessful. Plotting the behaviour of both data and dipole-term (figure 2.3) revealed the cause. In the close distance region

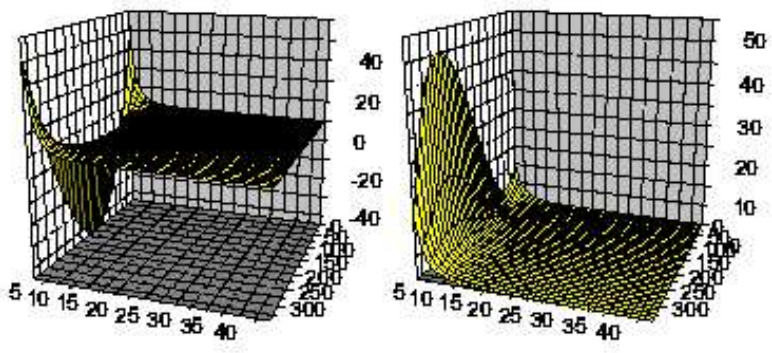


Figure 2.3: MD-data (right) vs. dipole formula (left) for glycine-glycine. The x-axis denotes the distance, the y-axis the rotation angle of the first helix and the z-axis the energy values. The tilt is 0° and the rotational angle of the second helix also 0°.

$r_{C_\alpha}[\AA]$	β_1 [°]	β_2 [°]	β_3 [°]	$E_{es}[kJ/mol]$	$E_{dipole}[kJ/mol]$
39.39	100.074	145.929	99.1139	0.00843853	-0.00983728
31.4202	100.074	145.321	99.9153	0.0114498	-0.0204456
23.4598	100.074	144.284	101.261	0.0219459	-0.053283
19.5001	100.074	143.437	102.341	0.0388623	-0.0983929
15.55	100.074	142.135	103.971	0.0687684	-0.210012
11.6202	100.074	139.896	106.708	0.138061	-0.562483
10.7598	100.074	139.171	107.578	0.113547	-0.730456
9.89998	100.074	138.31	108.603	0.102411	-0.969492
9.04995	100.074	137.284	109.812	0.191769	-1.31538
8.20003	100.074	136.025	111.28	0.235845	-1.83819
7.36993	100.074	134.486	113.054	0.106769	-2.63675
6.54	100.074	132.511	115.301	0.177956	-3.93651
6.06986	100.074	131.122	116.865	0.174636	-5.04297
5.60983	100.074	129.505	118.668	0.0984048	-6.53176
5.14993	100.074	127.555	120.823	0.12998	-8.6081
4.70001	100.074	125.209	123.387	0.535543	-11.4761
4.26983	100.074	122.404	126.418	0.957708	-15.3413
3.84994	100.074	118.901	130.155	1.89806	-20.5679
3.46001	100.074	114.627	134.648	3.28958	-26.8933

Table 2.1: Change of the relative orientation of the dipoles with decreasing distance for isoleucine-threonine, tilt = 0°, $\alpha_1 = 0°$, $\alpha_2 = 80°$, $\beta_1 = \arccos \frac{\vec{p}_1 \cdot \vec{p}_2}{|\vec{p}_1||\vec{p}_2|}$, $\beta_{2,3} = \arccos \frac{\vec{p}_{1,2} \cdot \vec{e}_r}{|\vec{p}_{1,2}|}$

the behaviour is opposing. This means that the multipole expansion that is only defined for distances larger than the size of the charge distribution gets to its limits at small distances and the use of a dipole term is insufficient in explaining the interaction of the residues at chosen distances where the discrete charge distributions and the shape of the residues have a prominent influence.

Looking at the data from the dipole fit reveals another problem connected to the choice of variables. The frame is parallel helices in fixed orientation. But varying the distance does not preserve the relative orientation of the residues. In table 2.1, one can see that for different distances the angles between the dipoles of the residues are changing. The choice of variables is not the most suitable for keeping the orientation of the residues. This may have contributed to the behaviour of the dipole fits in the plots 2.3.

Because an analytical expression for the influence of the tilt is missing at this point of the project, we plotted the MD-data for different tilts in figure 2.4, to estimate the influence of the tilt. The influence of the tilt is not negligible, but using the untilted values should enable a good approximation of the residue-residue interaction, especially for the van der Waals interactions.

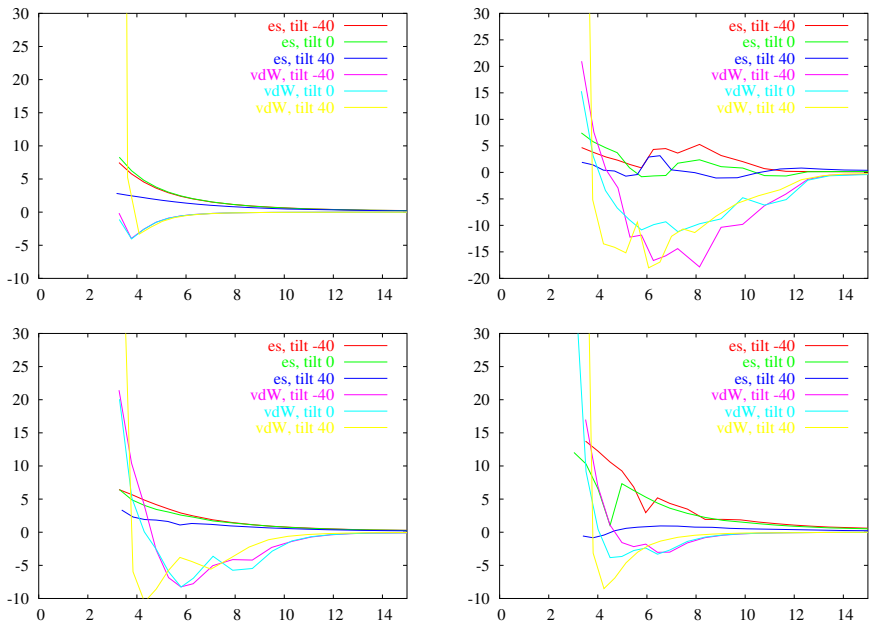


Figure 2.4: Influence of tilt on the energy values. Top left: glycine-glycine, top right: phenylalanine-phenylalanine, bottom left: leucine-valine, bottom right: serine-threonine, all $\alpha_1 = 0°, \alpha_2 = 0°$.

Problem cases of the fitting. The fitting with the polynomial function 2.1 was facing a problem that can be seen in figure 2.5. The large scale oscillations that are shown on the left are caused by an energy value of $\sim 250000 kJ/mol$ at a distance of 3.44 $\AA$.

Figure 2.6 shows the importance of a proper choice of the number of fitting parameters or polonomials that are used for fitting. The plots for 9 parameters often show the tendency of overfitting while with 7 parameters the fits are poor. We have chosen 8 parameters. The final point that has to be taken into account is that fitting functions can not guarantee proper behaviour beyond the interval where data is available. As one can see in figure 2.6, most of the plots have a steep positive gradient below the smallest energy value - opposing physical reality. We solved this problem by substituting the fit-function by another function with a large negative gradient below the smallest data point. That means there is an additional element in the vector where the fit parameters are stored, telling the minimum distance until which the fit-function is valid. Taking these things into account, the fit procedure becomes stable and the automised fit procedure reliable.

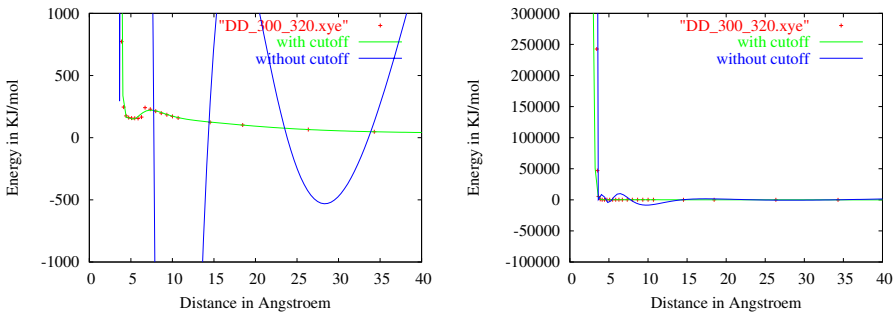


Figure 2.5: Necessity to introduce a cutoff. Both images show at different scales the distance dependency of the interaction of two aspartic acids with rotation angles $\alpha_1 = 300°, \alpha_2 = 320°$.

2.3.1 Levenberg-Marquardt method

To fit a set of datapoints (x_i, y_i) with a function $y = y(x;\vec{a})$ that depends nonlinearly on a set of n unknown parameters a_k, $k = 1,2,\ldots,n$ a merit function χ^2 can be defined

$$\chi^2(\vec{a}) = \sum_{i=1}^{n} \left[\frac{y_i - y(x_i;\vec{a})}{\sigma_i} \right]^2$$

and the best-fit parameters are determined by the minimization of this function. σ_i is the measurement error (standard deviation) of the i-th data point. The derivatives[2]

$$\beta_k \equiv -\frac{1}{2}\frac{\partial \chi^2}{\partial a_k} \qquad \alpha_{kl} \equiv \frac{1}{2}\frac{\partial^2 \chi^2}{\partial a_k \partial a_l}$$

can be used to minimize χ^2 in two ways. The first is the *inverse Hessian method*.

$$\sum_{l=1}^{n} \alpha_{kl}\delta a_l = \beta_k \tag{2.3}$$

This set of linear equation is solved by variation of the δa_l. The matrix α, that is one-half times the Hessian matrix, is called *curvature matrix*. The second is the *steepest descend method*, given by

$$\delta a_l = constant \times \beta_l. \tag{2.4}$$

The Levenberg-Marquardt method switches smoothly between the two methods. Far away from the minimum, the steepest descend method is used. As the minimum is approached, the inverse-Hessian is smoothly switched on. Marquardt introduced

[2] The factor 1/2 is convention.

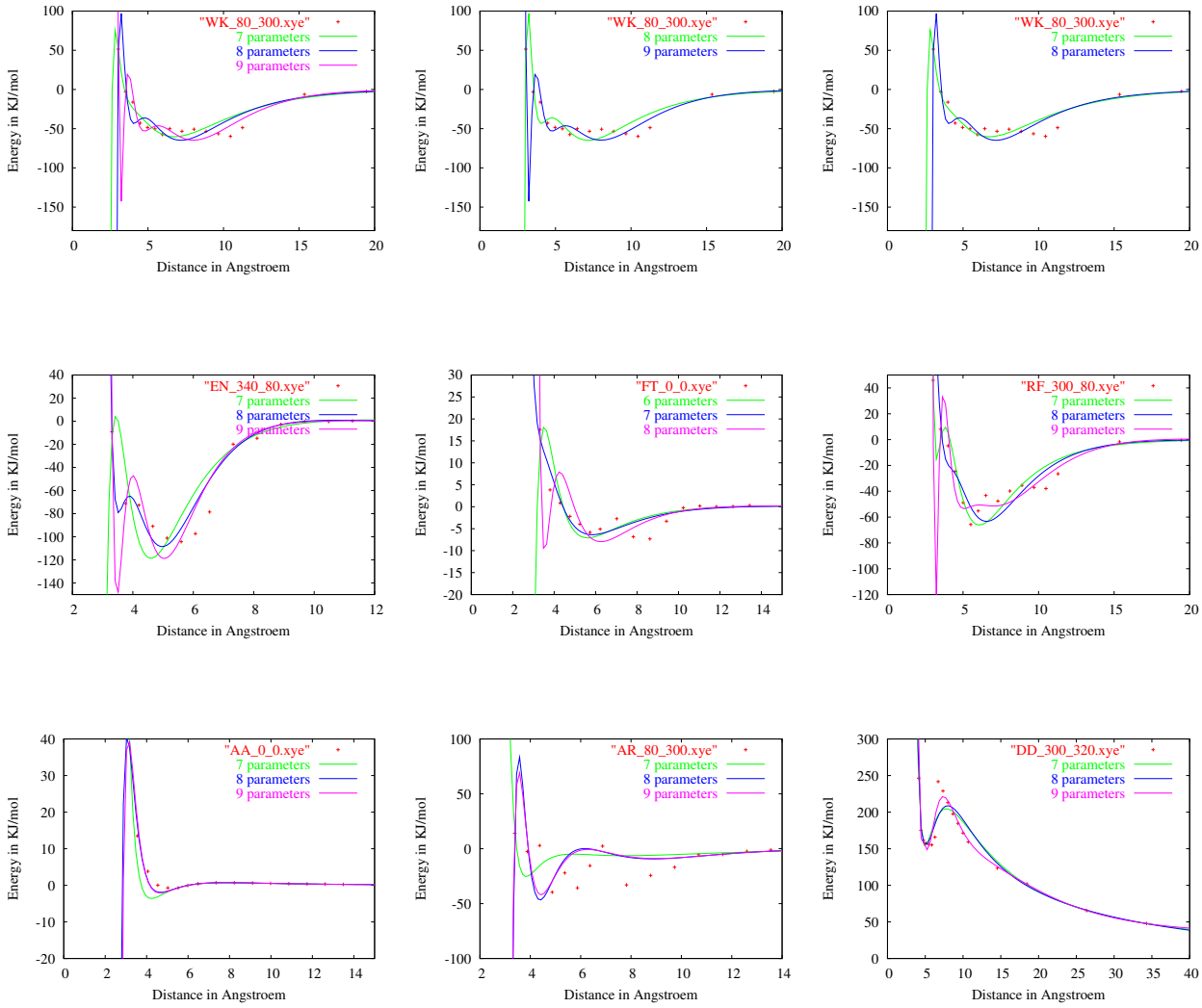


Figure 2.6: Influence of the number of fit parameters (order of polonomials). The code at the top of each image follows the pattern: residue 1, residue 2, rotation angle of first residue, rotation angle of second residue.

two modifications on equations 2.4 and 2.3. To find the right order of magnitude of the constant in equation 2.4 he modified the equation

$$\delta a_l = \frac{1}{\lambda \alpha_{ll}} \beta_l, \qquad \lambda \gg 1 \tag{2.5}$$

and combined it with 2.3 by defining a new matrix

$$\begin{aligned} \alpha'_{jj} &\equiv \alpha_{jj}(1+\lambda) \\ \alpha'_{jj} &\equiv \alpha_{jk} \quad (j \neq k) \end{aligned}$$

and replaced 2.5 and 2.3 by

$$\sum_{l=1}^{n} \alpha'_{kl} \delta a_l = \beta_k. \tag{2.6}$$

When λ is very large, equation 2.6 is identical to 2.5, and when it is close to zero 2.6 metamorphoses to 2.3.

An implementation of the GNU Scientific Library (GSL) [44] was used with the robust and efficient Levenberg-Marquardt method to perform the fit.

Chapter 3

The sphere algorithm for the description of molecular surfaces

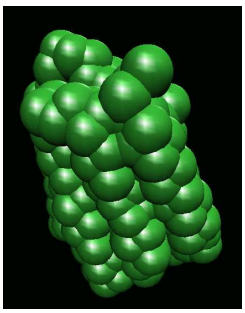


Figure 3.1: Bacteriorhodopsin with spheres around each C_α atom.

3.1 Motivation

Knowledge about how much a surface assigned to an atom or a subunit of a molecule is overlapped by others or how much is free (meaning it has no overlap) can be valuable in several cases.

Besides purely distance dependent interactions, like the electrostatics, there are also some that are surface dependent, like van der Waals. Statistical properties might also require information about the surface. One can distinguish two kinds of statistical properties. First, those like pressure or other thermodynamical properties that are statistical by their own nature. Second, those where one does not want to or simply can not calculate all interactions explicitly.

When modelling larger systems statistically, essential information can include the exposure of a molecule or a part of it to a certain surrounding, especially when it

is embedded in different or changing environments. In large proteins, for example, some of the residues might be buried between other residues in the interior of the protein while others at the surface are highly exposed, either to solvent or to membrane. In this case, the amount of surface that has no overlap with others can be used as a measure of the exposure to the surroundings.

A typical surface used to describe the building blocks of molecules, whether they be atoms or residues, are spheres. The total surface of a molecule could be described by adding the spheres of its constituents and summing the surfaces that have no overlaps with others. To create a precise image, one would have to put a sphere around each atom of a molecule. For large molecules it could be sufficient to construct a sphere for each subunit like amino acids. Of course, the surface dependent properties are also distance dependent, but only when they overlap with other surfaces will their values vary with the distance. In molecules that consist of a couple of spheres close together, internal conformational changes lead to different surfaces and to a change in the surface dependent properties.

Surface dependent properties also can be much more complex than the purely distance dependent - for the latter one only has to sum over all possible pair interactions as a first but far-reaching approximation. The summability means that the interactions do not influence each other - the interaction between a and b is not influenced by the interaction of c and d or any other. This is not the case for surface dependent interactions, due to the overlap of the surfaces which is not at all linear and, of course, where the overlap of two surfaces is influenced by a third sphere. This is also the reason why it is not yet possible to derive an analytical formula that tells the surface of an arbitrary number of spheres for any distance matrix. The complexity of the distance dependence of the surface increases drastically with higher number of spheres.

We are therefore looking for an algorithm that is able to calculate total surfaces of molecules as well as free surfaces of single atoms or residues in the molecule.

3.2 Methods

3.2.1 Basic ideas

The free surface of a sphere that is overlapping with a variable number of other spheres shall be computed. The method used is based on numerical integration in spherical coordinates. We calculate in an arbitrarily chosen but absolute space.

The spheres are ordered as an array. The first step is to create an overlap-array o_a for each sphere a, which contains the array-indices of all spheres b that it has overlaps with, i.e. all spheres that obey the condition: $d_{ab} \leq r_a + r_b$, where d_{ab} is the distance and r_a, r_b are the radii of the spheres. We represent the sphere by a set of m vectors, each vector representing a certain surface, which is: $\frac{4\pi r^2}{m}$ in first order (this will be specified later on). The origins of the vectors are at the position of the sphere. They have the length of the radius of the sphere and should be more or less equaly distributed. It is not a trivial topic to distribute a variable number of vectors homogenously over the sphere, so we do not put too much effort in that, but instead use a sufficient number of vectors to keep the error small enough. How do we distribute them? In spherical coordinates a vector *v = (x,y,z)* is:

$$v = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} r\sin\theta\cos\phi \\ r\sin\theta\sin\phi \\ r\cos\theta \end{pmatrix}$$

By defining following intervals in first order $\delta_\phi = 2\pi/k$ and $\delta_\theta = 2\pi/l$ the position p_{ij} of point ij is obtained:

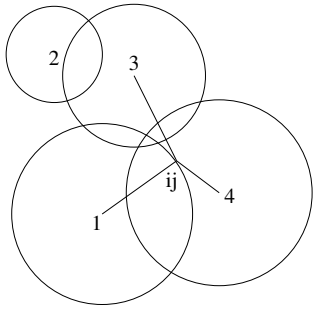
$$\vec{p}_{ij} = \vec{p}_0 + \begin{pmatrix} r\sin(i\cdot\delta_\theta + \frac{\delta_\theta}{2})\cos(j\cdot\delta_\phi + \frac{\delta_\phi}{2}) \\ r\sin(i\cdot\delta_\theta + \frac{\delta_\theta}{2})\sin(j\cdot\delta_\phi + \frac{\delta_\phi}{2}) \\ r\cos(i\cdot\delta_\theta + \frac{\delta_\theta}{2}) \end{pmatrix}, \quad \begin{array}{l} i = 0,...,l-1; j = 0,...,k-1; \\ \vec{p}_0 \text{ position of the center of the sphere.} \end{array} \tag{3.1}$$

Now the procedure that must be undergone to find the free surface of sphere a is like this:

- store the overlapping spheres b in an array o_a
- define a set of points ij on the sphere a as in equation 3.1
- check for each point if it is in an overlapped area or not:
  $|p_{ij,a} - p_{0,b}| < r_b$ for all spheres b from o_a,
  i.e. if the distance from point ij to sphere b is smaller than the radius of that sphere
  - if not: check next sphere b in the array o_a
  - if none at all: no overlap at this point
    $\Rightarrow$ add area represented by the vector to the free surface of the sphere a

– if yes: end search for this point, go to next point

Example:



- the free surface of sphere 1 is wanted
- sphere 1 has overlaps with spheres 3 and 4
- looking at point ij:
  * $|p_{ij,1} - p_{0,3}| > r_3$, no overlap at point ij with sphere 3
  * $|p_{ij,1} - p_{0,4}| < r_4$, overlap at point ij with sphere 4

The method can easily be modified for different frameworks. For example, it could be of interest which part of the sphere is embedded in which environment, like if it is located right at the edge of a phase transition of two media. The larger the sphere and the more distinct the two phases are, the more important this becomes. It could be done by checking pointwise the surrounding and summing it according to counters - one for lipids, one for solute in the case of the membrane.

3.2.2 Improving accuracy and performance

We spoke twice about first order. Why so? Taking fixed δ_ϕ and δ_θ will lead to significantly different surfaces at the poles than at the equator of the sphere. This would mean a significantly higher precision that is actually unusable at the poles than at the equator. This cannot be avoided completely, but the gap can be reduced. What follows now is one way of keeping the vector-surfaces as homogenous as possible. This means that we have to do a more refined definition of the δ_ϕ and δ_θ. We take δ_θ as constant and adjust δ_ϕ. As a second step, one could also vary δ_θ, but as the effort increases, so does the computational time.

We start from the equator, taking $\delta_\phi = \delta_\theta$ for the first interval $\theta = 0$ to $\theta = \pm\delta_\theta$. This results in the surface s_0:

$$s_0 = \int_0^{\delta_\theta} \int_{\frac{\pi}{2}-\delta_\theta}^{\frac{\pi}{2}} r^2 \sin\theta d\theta d\phi = r^2 \delta_\theta \sin\delta_\theta \tag{3.2}$$

The vector-surface s_0 is the surface that the vectors in this interval represent.

Coming to the next interval, one has to find the $l(k)$ that leads to a surface close to s_0. We do so by

$$s_0 \approx \int_0^{\delta'_\phi} \int_{\frac{\pi}{2}-(k+1)\delta_\theta}^{\frac{\pi}{2}-k\cdot\delta_\theta} r^2 \sin\theta d\theta d\phi = r^2 \delta'_\phi (\sin((k+1)\delta_\theta) - \sin(k\cdot\delta_\theta)). \tag{3.3}$$

Since δ_θ and s_0 are constant, we can solve for δ'_ϕ:

$$\delta'_\phi(k) = \frac{s_0}{r^2(\sin((k+1)\delta_\theta) - \sin(k\cdot\delta_\theta))},$$

what leads to a value $l'(k) = \frac{2\pi}{\delta_\phi(k)}$, which will be no integer, so it is rounded off to the next higher integer $l(k) = int(l'(k)) + 1$ and becomes $\delta_\phi(k) = \frac{2\pi}{l(k)}$, and finally the vector-surface will be

$$s(k) = r^2 \delta_\phi(k)(\sin((k+1)\delta_\theta) - \sin(k\cdot\delta_\theta))$$

One can introduce a common offset to the $l(k)$ to ensure a sufficient accuracy at the poles. The vectors are not equidistant.

The radii of the spheres are not determined but should fulfill the following criteria:

- buried parts of large proteins may have no free surface
- residues of a pore that point inward (e.g. in the water pore Aquaporin) should have a free surface comparable to the real size of the pore
- also cavities that contain water or other molecules could be used to find a useful definition of the radii

3.2.3 Free surface and nth-order overlap

The algorithm can be used in different ways:

- determining the free surface of a residue as described before
- summing them up resulting in the total surface of a molecule
- slightly modifying the integration-loops in order to calculate the overlap in different orders

By calling the overlap of two spheres an overlap of first order; overlaps of higher order will be exploited later as a tool for obtaining more geometrical information about modelled molecular conformations.

3.2.4 Interaction of amino acids with their environment

Residues can be exposed to either membrane or solvent. Due to the polar headgroups of the lipids at the surface of the membrane, the membrane is not constantly lipophilic but has a lipophilic core and an intermediate area where the interaction of the residues with the membrane can be modelled as in figure 3.2. The thickness of the two interface regions together is the same as the size of the core.

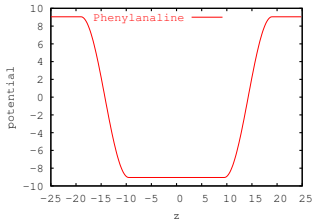


Figure 3.2: Insertion of a phenylalanine into a membrane ($|z| < 15$), y axis is the transfer free energy as calculated from Gu *et al.* [38] in kJ/mol.

The interaction of residue and membrane is described in terms of the free energy. Vital is the total free-energy difference between residues in the hydophilic core and in the aqueous solution (ΔG_{tot}).

ΔG_{tot} has many contributing parts and can be decomposed in the following manner [45],[46],[47]:

$$\Delta G_{tot} = \Delta G_{es} + \Delta G_{np} + \Delta G_{lip} + \Delta G_{imm} + \Delta G_{con}.$$

ΔG_{es} is the difference in the electrostatic contributions, ΔG_{np} nonpolar contributions, ΔG_{lip} lipid pertubation effects[1], ΔG_{imm} peptide immobilization effects[2], and ΔG_{con} are contributions from peptide conformational changes such as the formation of α-helices. The advantage of this decomposition is the possibility to calculate the contributions separately. As a note of caution, free energies may not be decomposed in a strict sense. Ben-Tal *et al.* [48] suggested to introduce a solvation free energy

$$\Delta G_{solv} = \Delta G_{np} + \Delta G_{es}.$$

which ought to be the dominant contribution. We use the values for the solvation free energy of amino acids in water or in chloroform calculated by Gu, Rahi and Helms using MD-simulations with multi-configuration thermodynamic integration [38]. For comparison we use the experimentally determined values from White *et al.* [25] and theoretically calculated values from Lazaridis [37]. White *et al.*

[1] includes altered ordering of lipids and entropic effects on lipid dynamics

[2] includes entropic effects for reducing the translational and rotational degrees of freedom of the peptide

measured the transfer free energy for the transfer of the whole residue from water to the interface as well as from water to octanol. We neglect the fact that the ratio between the interface scale and the octanol scale, the free energies for the transfer from water to the interface and to octanol, is not constant at all and models all residues with a similar function in the interface.

Attention must be paid for what experimental free energy values were measured exactly. To reproduce the experimental values, exact knowledge about the involved terms is crucial. White pointed out that in some experiments it is not clear which term they exactly measured.

3.2.5 Virtual charges

Transfer free energies alone are not sufficient to describe the interaction of peptides with their environment correctly. One has to consider two other contributions. First, there are the charged amino and carboxyl group at the ends of the peptides. Second, the helices build up a dipole moment. The hydrogen bonds that stabilize the helical conformation are all directed parallel to the helical axis. The same holds for the peptide bonds. Even though the individual dipole momenta are small they sum up to a total dipole moment of the helices that corresponds for a helix of average size to charges of 0.5 e magnitude at both ends [49]. This charge is not located at a defined position but is distributed over a delocalized volume. When the helix ends are exposed at the membrane surface the dipolarisation effect is damped by solvation and delocalisation effects [50].

To take them into consideration one can introduce virtual charges, summing dipole and termini charges, placed at the ends of the helix that can be placed at the projections of the first and the last C_α-atom on the z-axis of the helix. While located outside of the membrane their influence should be zero, but as soon as they enter they should add positive values to the interaction plots. Throughout the hydrophilic core this value should be a constant that is proportional to the charge. In the interface region, it should follow an analogous transition as the residue-membrane interaction.

The virtual charges are not implemented in the program yet. This will be a future step of the project. The numerical tests that were done for the sphere algorithm can be found in appendix B.1 on page 103.

3.3 Applications

3.3.1 Additivity of residue hydrophilicity

The solvation properties of single residues are not sufficient to describe large proteins, where some of the residues might be partially or completely buried. How do the values for single residues add up in the case of larger peptides? We use the amount of exposed surface obtained by the sphere-algorithm to measure how much a single residue contributes to the total peptide value.

If the interaction is proportional to the exposure of the residue, the solvation free energy of residue *i* should follow the relation

$$\Delta G_i^{solv} = \frac{free\ surface}{total\ surface} \Delta G_{i,total}^{solv}$$

We refer to the quotient above as the relative free energy. If not assigned differently we use the radii given in table B.3 right column on page 106, the lipophilicity scale from Gu *et al.* [38], and a membrane thickness of 38$\AA$.

Only for a few peptides can one find experimental values for the total transfer free energy of the whole peptide. The values we calculated are taken from the insertion profiles described in the following section.

A peptide where experimental data is available is **melittin** [51]. Melittin unfolds outside of the membrane. That means that the free energy ΔG_{conf} of forming or breaking an α-helical conformation is contributing. White and Whimley [25] found a transfer free energy for the unfolded peptide of -2.6 kcal/mol from water to interface and -7.6 kcal/mol for the transfer of the unfolded peptide in solution to the helix in the interface.

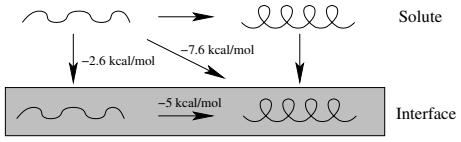


Figure 3.3: Different insertion paths for melittin.

It is unlikely that the alternative insertion path of first forming an α-helix and then inserting it should split the total energy of -7.6 kcal/mol in the same proportions as the experimentally observed path. However, a detailed analysis is out of the scope of this thesis. We simply calculate the transfer free energy for the insertion of the α-helix. The results depend on the rotational orientation of the helix, as can be seen from the inverse symmetry of the horizontal plot in figure 3.7. Note that the figures are not necessarily showing the curves with minimum energy values. The values vary from 0 to -5,95 kJ/mol = -1.42 kcal/mol. Lazaridis [37] calculated a value of -5 kcal/mol. A larger negative value for the helix insertion than the peptide insertion is reasonable because helix-formation in solution is not observed.

The next test case is the **c-helix of bacteriorhodopsin** [52]. Here the investigation is not as easy as for melittin. We looked at two different cases, without loops and with the loops as taken from Hunt *et al.* [52]. Including loops would make it necessary to sample over many loop conformations to get a complete image. Instead we assume, as a simple test, the conformation as being helical throughout the peptide. Figure 3.9 and table B.9 show that the loops are highly hydrophilic, and thus burying them in the interface is energetically unfavourable. Since the loops are taken to be in helical conformation, the plot reveals a large disaim of the helix to insert in the interface. When the loops can move freely they will bend in such a way that they remain in the hydrophilic area, while the helix is inserted into the interface. As one can see from figure 3.9 the effect for a horizontal helix is extremely unfavourable compared to the helix without loops, the reason being the burying of some strongly polar residues, which normally would have the chance to bend away from the lipophilic core towards the aqueous solute. Therefore, the values for the helix with loops are not relevant, but one may conclude that their effect will be an increase of the value for the helix without loops.

We can point out that our value for the insertion into the interface, -19,67 kJ/mol = -4.71 kcal/mol (without loops), is not far away from the one calculated by Lazaridis [37], which is -12 kcal/mol.

For the transmembrane conformation Lazaridis calculated -7 kcal/mol for a thickness of 23$\AA$ and +1 for a thickness of 26$\AA$, while we have -8,62 kJ/mol = -2.06 kcal/mol for 23$\AA$ and 0,44 kJ/mol = 0.11 kcal/mol for a thickness of 26$\AA$. Here the influence of the loops will not be that strong because none of them are buried into any lipophilic surrounding.

The last value is somehow arbitrary as one can see on figure B.2 on page 114. We set z = -4, as this is the most likely to occur since it is the closest to the center of the membrane with a smooth gradient.

While the last lines of tables B.4 to B.9 unveil that the sum $\sum_i \Delta G_i$ is far away from the experimental values, our results indicate that the free surfaces are a useful measure to reproduce the experimental values. No extensive assilimation of the radii was performed yet, what should lead to values even closer to experiment.

3.3.2 Membrane insertion

Single helices After the additivity was checked in 3.3.1, we take a look at the insertion profiles of different helices for one-dimensional insertions. Starting outside of the membrane for a certain orientation of the helices, and, while keeping the orientation fixed, we insert them, plotting the ΔG as a function of z only.

Figures 3.4 and 3.5 show the insertion behaviour of **polyalanine** and **polyleucine**, consisting of 20 residues each, which is the typical thickness of a membrane bilayer.

Due to the costs of inserting the ends of the helices, the transmembrane orientation will be the favoured one, but for polyalanine the gain of inserting is about

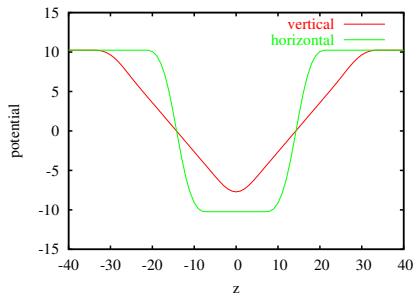


Figure 3.4: Insertion of a polyalanine helix, 20 residues [y axis kJ/mol]

-17,95 kJ/mol = -4.3 kcal/mol, while for polyleucine it will be -107.89 kJ/mol = -25.81 kcal/mol. Minding the costs for burying a virtual charge, one can distinguish polyleucine as being a good transmembrane helix former from polyalanine which should be a comparably poor one. That is in agreement with experiments[53], [54].

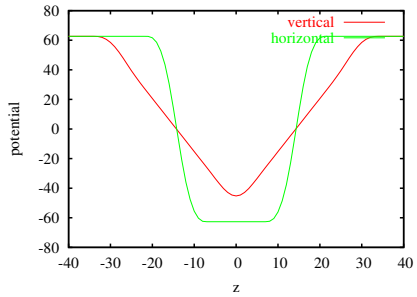


Figure 3.5: Insertion of a polyleucine helix, 20 residues [y axis kJ/mol]

The curves in figure 3.6 show the insertion profiles for a glycophorin A helix with constant orientation during the insertion. The shape of the curves for the vertical and the slanted helix reflect the different insertion behaviour of the individual residues. A detailed listing of the single residue energy values can be found in appendix B.3 tables B.4, B.5 and B.6. The curves reveal that from the interactions included so far the helices would have the aim to be completely buried within the lipophilic core instead of the observed transmembrane orientation. The crucial missing point are the virtual charges at the helix termini (see 3.2.5). The minimum of the plot at 45° and the one for horizontal orientation are very close together. Therefore, it would not require a very high energy value for burying a charge in the membrane to let the latter become the more likely conformation.

Figure 3.7 shows the insertion behaviour of **melittin**, and figures 3.9 and 3.8 show that of the *c*-helix of bacteriorhodopsin with and without some part of the

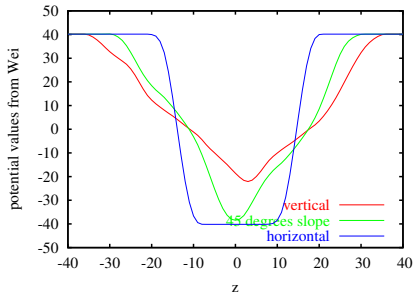


Figure 3.6: Insertion of a single glycophorin A helix for different tilts, z is the position of the cms of the helix, when z = 0 the cms is in the middle of the helix, y-axis is $\Delta G^{transfer}$ in kJ/mol, the orientation of the helix is kept fixed during insertion.

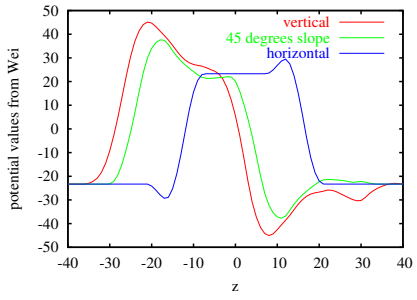


Figure 3.7: Insertion of a single melittin helix [y axis kJ/mol]

loop respectively. Additional insights into the effect of different membrane thicknesses on the bacteriorhodopsin-*c*-helix can be found in figure B.2 on page 114. In the case of melittin it is obvious that the helix cannot insert into the core in the horizontal orientation but only to the interface. The two peaks at side of the main peak indicate that the insertion behaviour depends on the rotational orientation. Furthermore it can only insert with one end in front, since the other is strongly rejected. The depth the helix can reach is limited as well. Here, with a membrane of 38$\AA$ thickness, the cms of the helix can only reach a distance of $\approx 6\AA$ to the middle of the membrane; but taking lower values of the thickness leads to a minimum at z = 0.

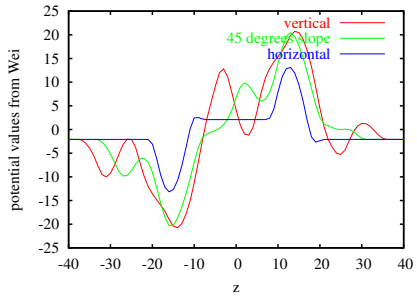


Figure 3.8: Insertion of a single bacteriorhodopsin C helix [y axis kJ/mol]

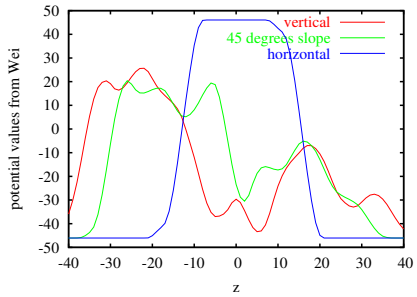


Figure 3.9: Insertion of a single bacteriorhodopsin C helix, with loops [y axis kJ/mol]

Figures 3.8 and 3.9 evince a more complex insertion behaviour of the *c*-helix of bacteriorhodopsin, with and without loops respectively. The symmetry of the horizontal case in figure 3.8 exhibits a strong rotation-dependency around the z-axis. The plot for the vertical insertion reveals a very different lipophilicity-symmetry in the sequence as compared to that seen in melittin. In melittin there is one half clearly dominated by lipophilic residues, while the other is dominated by hy-

drophilic. Bacteriorhodopsin-c has double frequency in the distribution of hydro- and lipophilic residues, which reflects the formation of a hydrophilic and a lipophilic side due to its nature as a part of a transmembrane protein that buries hydrophilic residues in its interior. The frequency in the plot is in agreement with the 3.6 residues per turn in an α-helix. This is a very useful result concerning the structure-prediction of larger systems like the whole bacteriorhodopsin protein. The "vertical"plot in figure 3.9 shows that the helix, once inserted, is very stable in the transmembrane conformation.

(See appendix B.3 for detailed tables of the distribution of the potential over the helices.)

Helix assembling. Another thing one should be able to see, using the algorithm, are the preferential assemblies of helices. For a complete analysis, one needs, of course, to also consider the residue-residue interaction (see chapter 4 on page 42). What one can see at this point is the reduction of unfavourable residue-environment interactions through the burying of hydrophilic residues, which are located in the lipophilic core, between the helices. Because this is a crucial point for the design of membrane proteins, such an isolated examination is useful.

We use glycophorin A as a testsystem for the helix-assembling. Due to its symmetries it has 5 degrees of freedom. For the definition of the variables, see figure 4.2 on page 43 [55].

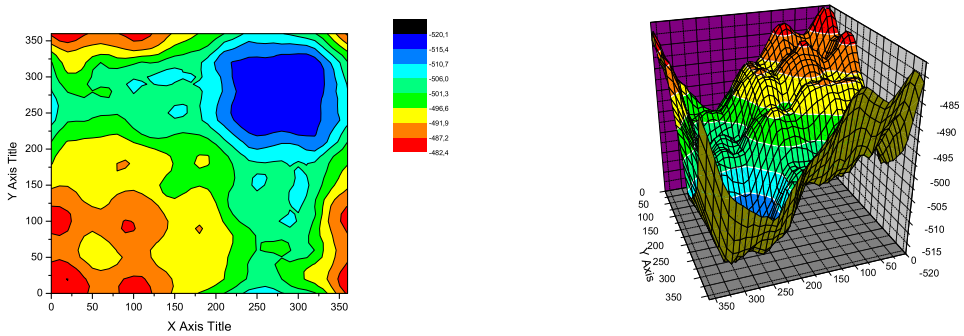


Figure 3.10: Rotation of the two glycophorin A helices around their z-axes with otherwise fixed orientation: distance 6Å, tilt 40°, l = 14Å. x and y axis are α_1 and α_2 respectively, z is the $\Delta G^{transfer}$ in kJ/mol

The 3d-plot in figure 3.10 reveals a strong preference for assembling the two glycophorin A helices in the conformation of the NMR-structure (see equation 4.1 on page 44 for location of native conformation). The plot was generated by

twisting the GpA helices around their z-axes, while the other orientations were kept fixed with the values written below the figure. It reveals that both have pronounced amphipathic sides.

Next we test whether the helices assemble at all. Using the favoured α_1 and α_2 values from figure 3.10 a one dimensional plot for three different tilts is shown in figure 3.11. The energy values of Gu were compared with values of White. The values of Gu showed an attraction only for a tilt of 40°. Both sets show an increasing attraction for larger tilt. Figure 3.12 plots the energy in the 0° tilt case for different membrane thicknesses. Below a thickness of $\approx 34\mathring{A}$, also the non-tilted helices attract each other. The next step will be a detailed search through the whole 5 dimensional conformation space, which will be performed together with the yet missing residue-residue interaction (see chapter 4).

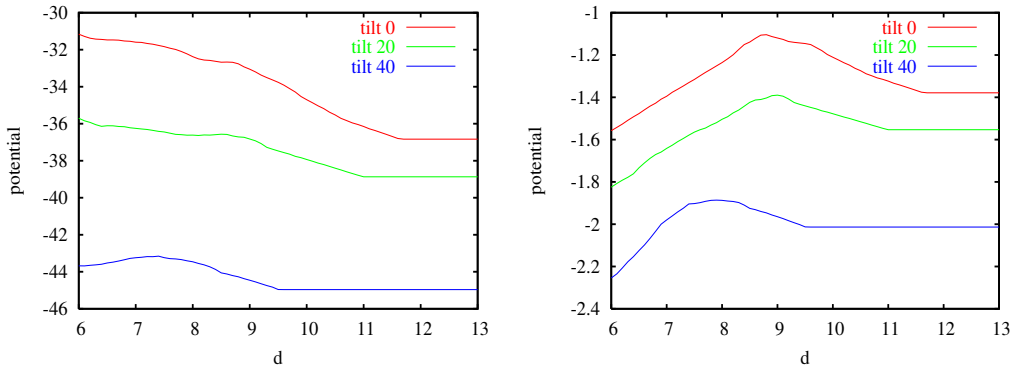


Figure 3.11: Assembling of two glycophorin A helices with $\alpha_1 = \alpha_2 = 275°$. $\Delta G^{transfer}$ as a function of the distance d is plotted. The left figure is calculated with values from Gu [kJ/mol] [38], the right with values from White [kcal/mol] [25].

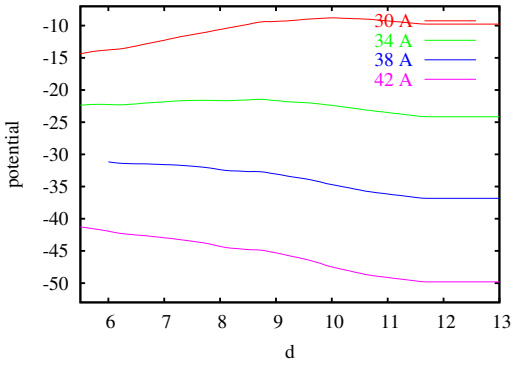


Figure 3.12: Assembling of two glycophorin A helices with different membrane thicknesses, for tilt = 0° and $\alpha_1 = \alpha_2 = 275^\circ$ [y axis kJ/mol]

3.3.3 Tight packing of residues

In the potential function, developed by Park, Elsner, Staritzbichler, Helms [43], tightly packed residues are well described until a certain density. The potential was described in chapter 2. Going beyond this limit will lead to an underestimation of the potential. The energy values were calculated with two residues which had the freedom to bend away in case they got closer than the sum of the two sidechain-lenghts.

$$d_{ab} \leq l_a^{sidechain} + l_b^{sidechain}$$

This bending could even lead to lower energy, due to a larger van-der-Waals interacting surface. However, the freedom to bend away might be limited when there are more residues close by, especially large ones. Too close contacts of residues should lead to highly repulsive van-der-Waals energies. Also, the number of accessible side-chain conformations is reduced what means a loss of entropy, i.e. higher free energy values. Our energy functions cannot distinguish between isolated or crowded residues. This missing geometrical constraint might result in too low (i.e. favourable) energy values and thereby might favour unreasonably packed conformations. One can expect a sterical clash especially when there are a couple of residues with long sidechains close together at the interface of two helices. One way out is to punish overlaps of higher order, using the sphere-algorithm. If we call the overlap of two spheres an overlap of first order, overlaps of fourth and higher order indicate a high spatial density of sidechains. The higher the order and the magnitude of the overlap the higher the density. Using the sum of these higher order overlaps one can perform a rough screening of the surface of the molecules and exclude unfavourable conformations. The energy term punishing tight packing could look like this:

$$\Delta G_{punish} = c_1 \cdot S^{(4)} + c_2 \cdot S^{(5)} + \cdots \quad , S^{(i)} \text{ overlapped surface of order } i$$

In this manner, the term could include enthalpic and entropic components, and is therefore marked as an effective free energy. The two terms in the sum will differ in the sharpness of the peaks in the screening landscape. The higher order term will be not as sensitive to smaller changes in the density but will be a good indication for the real nasty cases. One has to choose the radii of the spheres in such a way that they represent as well as possible the space the residues fill, unlike previous applications where the radii had to meet other requirements. We use the values for the radii calculated from Gu, Rahi, Helms [38] - see table B.3 on page 106. These values may actually underestimate the true size of the residues.

We use the glycophorin A homodimer to test this approach. We plotted

$$S^n = \sum_i S_i^n, \qquad \text{with } n \text{ the order of the overlap}$$

as function of the two rotational angles of the helices for different distances, tilt-angles and orders.

Figure 3.13 illustrates a high symmetry indicating a screw-like surface of the helices, which is in agreement with previous theoretical analysis [55]. The symmetrical pattern is changing with the tilt angle, which can be explained by different behaviour of fitting ridges into grooves for different tilts. Beside this universal helix-geometry, the figures provide some insights into specific structure. The peaks and the valleys have a periodicity due to the sidechains.

For comparison we used a different set of values for the radii, given in table B.3 right column. See figure 3.14, B.6, and B.7. Figure 3.14 shows a peak that cannot be seen in figure 3.13. Either the first set of radii leads to a sensitivity that is too low or the second to one that is too large. Still, the information obtained by the two sets is equivalent, even though the values in the latter set are much higher, because the radii are larger. The landscape in figure 3.14 is very similar to the one in figure B.5 on page 117, where the 3^{rd} order overlaps with the previously used radii are plotted. To decide about the most suitable set of radii and the order of overlaps, it would be best to look at a spacefilling model of the molecule in that configuration.

When the algorithm is used to avoid sterical clashes instead of a detailed search of the whole landscape, one can introduce a filter, taking into account only the overlaps above a certain size. Also looking at the fifth order turns out to be a sharper indicator for too high sidechain density. Another possibility that attaches more importance on larger overlaps is to sum over the exponentials of the overlaps:

$$\tilde{S} = \sum_i e^{S_i^n}$$

Especially when the helices are parallel, one might find a large number of small contributions summing up to the same magnitude that another conformation might have, due to only a few contributors that are pressing each other hard. In that case, the exponential sum should raise the value of the latter.

The higher order overlaps provide a good estimate for the tight packing of residues. The tighter the packing the lower the sidechain mobility and the lower the entropy (see section 1.3). That means the sphere-algorithm can be used as well to estimate the change of sidechain-entropy.

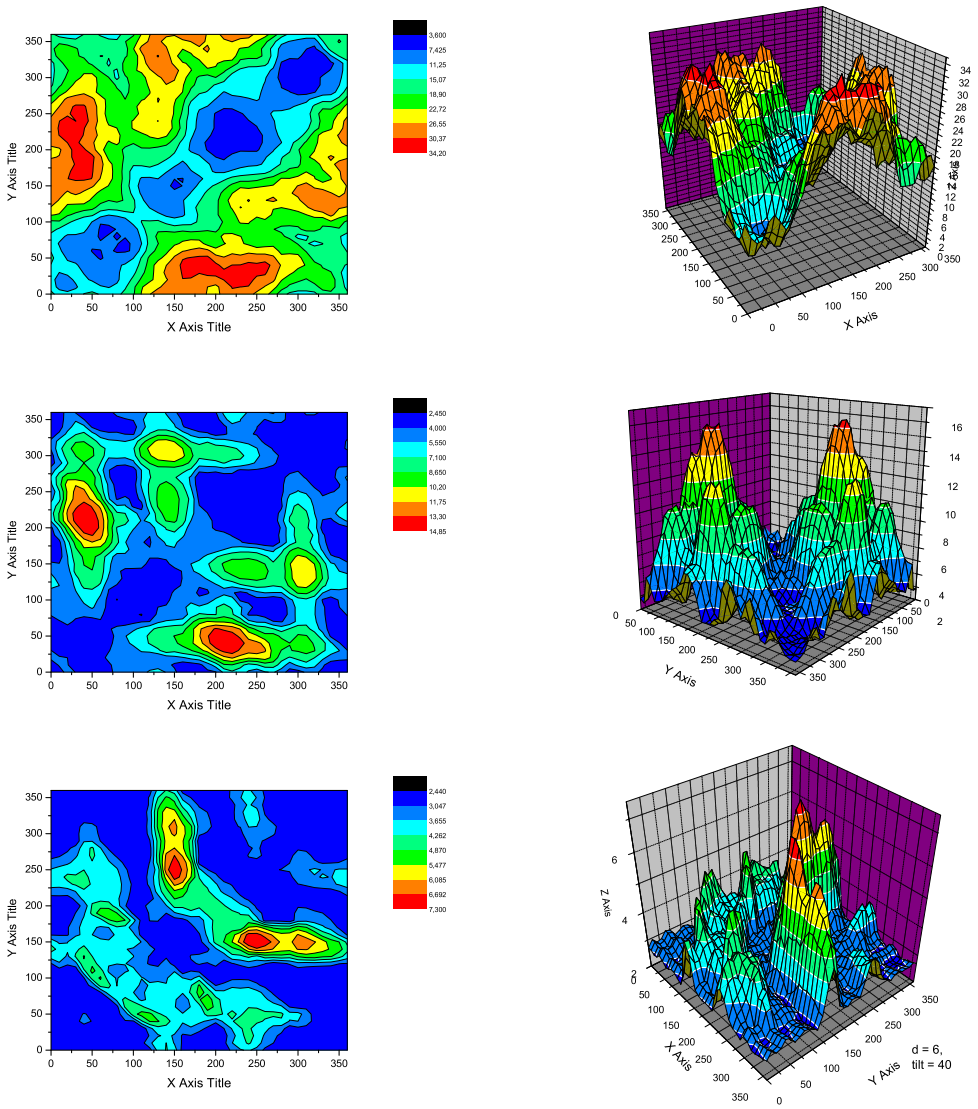


Figure 3.13: The magnitude of the 4^{th} order overlap as function of α_1 and α_2 for different tilts, left as contour map, on the right as 3d-plot, all at a distance of 6$\mathring{A}$. The upper two images show the overlap for the tilt θ = 0°, the middle two for $\theta = 20°$ and the lower two for $\theta = 40°$ [x,y in degrees, z in $\mathring{A}^2$].

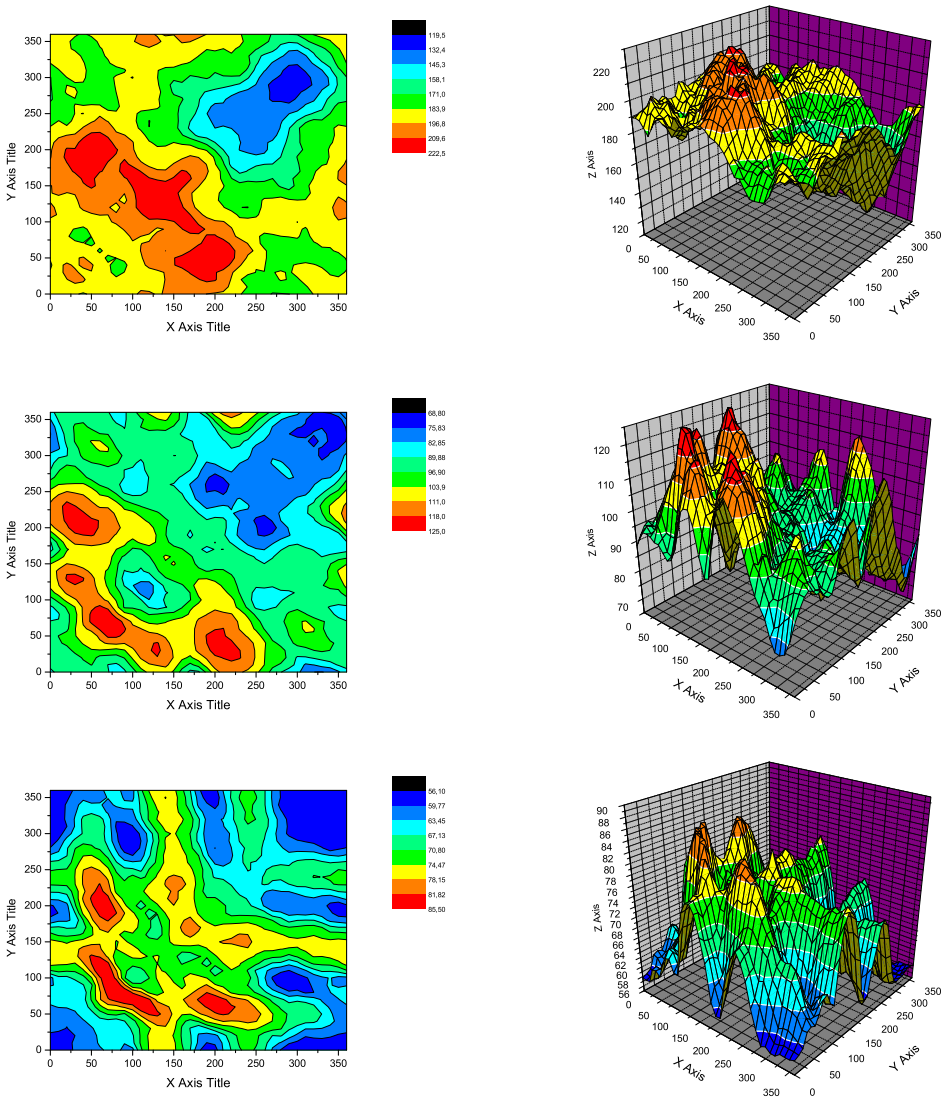


Figure 3.14: The same as in figure 3.13 but with a different set of radii. See table B.3 right column on page 106. The upper two images show the overlap for $\theta = 0^\circ$, the middle two for $\theta = 20^\circ$ and the lower two for $\theta = 40^\circ$ [x,y in degrees, z in $\AA^2$].

Chapter 4

Testing of the energy functions

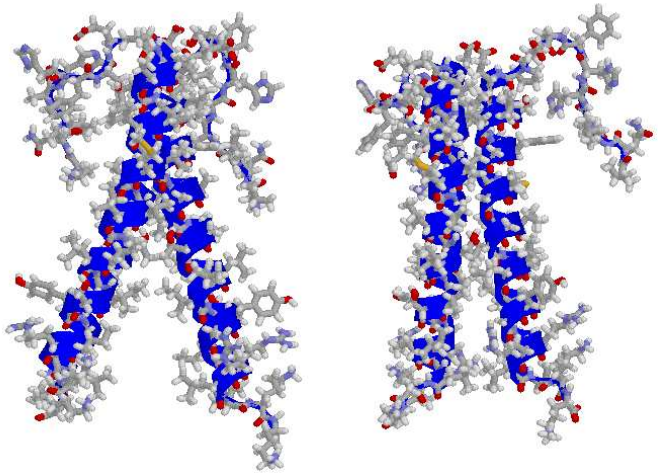


Figure 4.1: The GpA dimer from two perspectives.

4.1 Glycophorin A

For two reasons, glycophorin A (GpA) is an ideal test-system to commonly investigate the quality of residue-residue energy functions and the sphere algorithm in its both applications to determine the environment interaction and the crowding of the residues. First, its structure is known. Second, when regarding the helices as rigid bodies without internal dynamics, it is possible to systematically scan the whole conformational space of this protein. Minimum search methods presented in chapter 5 are not required and the quality of the energy calculation can be tested isolated.

Glycophorin was the first membrane proteine that had its entire amino acid sequence being determined. The structure was determined with NMR [56]. It is one of the most common proteins in the red blood cells (RBC), a rather simple kind of cell that contains only a single (plasma-)membrane. In the cells one finds mainly glycophorin A, and its relatives B, C, D and E only in lower concentration. It consists of three domains. First, the amino-terminus that binds about 100 sugars in 16 units on the extracellular side, giving glycophorin its name. Only 40% of the mass of glycophorin is protein, the other 60% are sugars. Most of the mass is located on the outside surface of the membrane. Each RBC contains around a million glycophorins - which means a large amount of carbohydrates on its surface. The MN blood group of the RBCs, one of 15 genetically distinct blood group systems, depends on the constitution of the oligosaccaride groups [1]. Many of the sugars are the negatively charged sialic acids. The negative charges on the surfaces have a repulsive effect on the RBCs, what reduces their likelihood of clumping [57].

The second domain is a single α-helix crossing the membrane. Last, the carboxy-terminus, that points into the cytosol, carrying plenty of polar and ionised sidechains. This is of importance for the cytosceleton, which binds to the membrane via GpA and another TM-protein, Band 3 [58, 59].

GpA works as a receptor for the influenza virus [60] and for the malaria parasite [61]. Humans that lack glycophorin C are relatively resistant to malaria [62].

Because the two helices of glycophorin A form a pretty symmetric dimer, one ends up with 5 dimensions in which the scan of the conformational space has to be performed. These are the distance d between the z-axes of the helices, the rotation-angles α_1 and α_2 around their z-axes, the tilt-angle ϕ between them and the slide l that is the length from the top of the helices to their crossing point where the z-axes have their shortest distance.

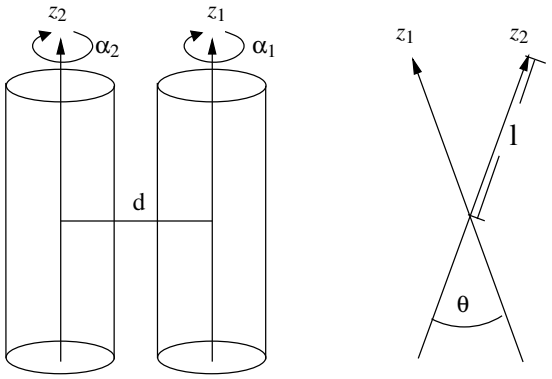


Figure 4.2: The 5D variables of GpA.

The symmetry is reflected in the equal slides for both helices and the same tilt

angle relative to the absolute z-axis perpendicular to the membrane. Regarding them as rigid bodies enables an explicit search through all 5 dimensions of the conformational space.

The native structure (model 1 of the pdb-structure 1AFO.pdb [63]) has the following values for these variables:

$$\begin{aligned} d^{(native)} &= 7.53\mathring{A} \\ l^{(native)} &= 22.35\mathring{A} \\ \phi^{(native)} &= 38.52^\circ \\ \alpha_1^{(native)} &= 229.59^\circ \\ \alpha_2^{(native)} &= 239.14 \end{aligned} \tag{4.1}$$

For the reason of completeness it should be mentioned that the PDB-file contains 20 structures. Figure 4.3 shows models 1 to 7. The helical structure of GpA should

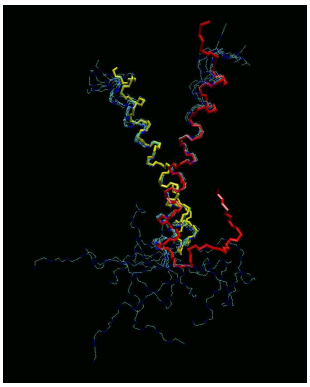


Figure 4.3: The backbones of model 1 - 7 from 1AFO.pdb for GpA. The thick drawn backbone belongs to model 1.

be rather flexible in the part that is further away from the interface of the dimer. Because plotting the 6D energy-conformation space is impossible, we try different approaches to understand the properties of the system by 3D plots.

4.1.1 The contribution of the residue-residue interaction

The dependency of the residue-residue interaction-energy on the two rotation-angles α_1, α_2 for the native values for $d^{(native)}, l^{(native)}, \phi^{(native)}$ from 4.1 is shown in figure 4.4 from different perspectives. The energies were calculated here and in all following plots with a step-size of $\Delta\alpha_1 = \Delta\alpha_2 = 10^\circ$. We refer to the native conformation as being located at $(\alpha_1, \alpha_2) = (230^\circ, 240^\circ)$ from here on, instead of the values in 4.1.

These results reveal that at this point the native structure is not in the absolute minimum of the residue-residue energy-landscape, which is at $(60^\circ, 60^\circ)$. This will

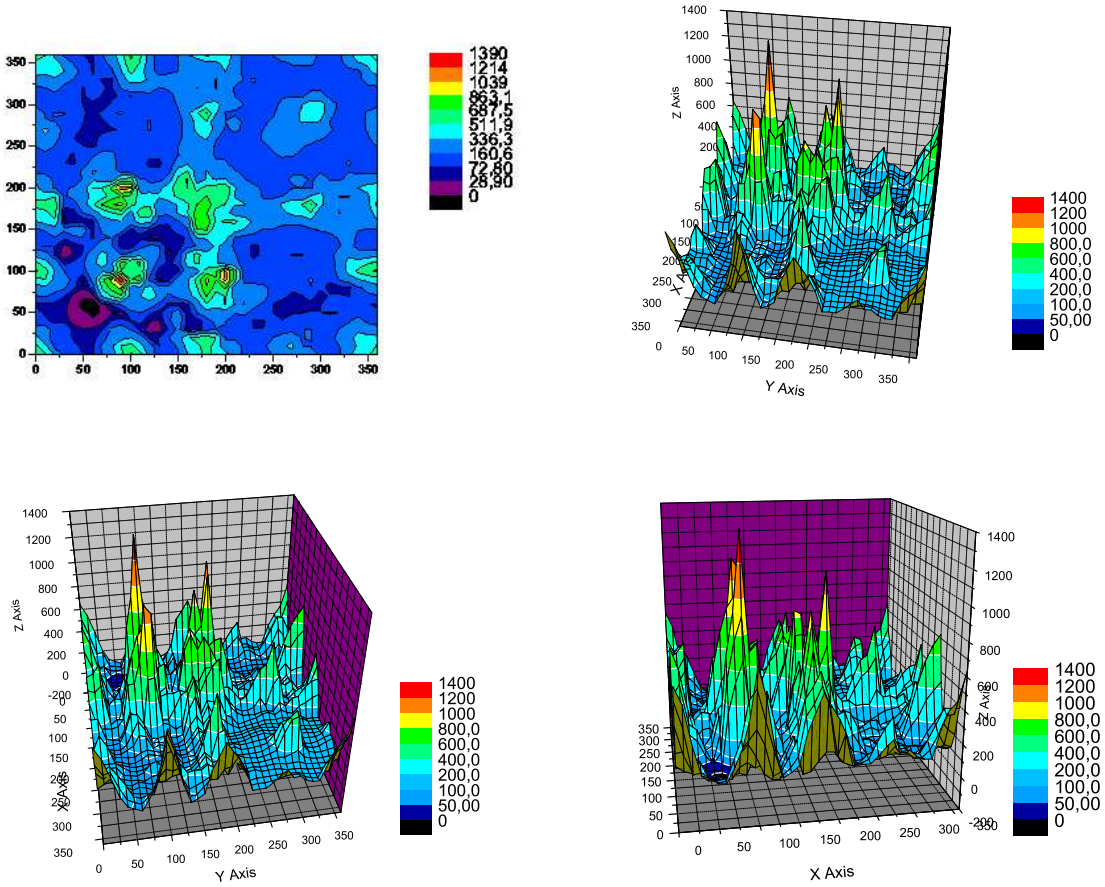


Figure 4.4: The energy landscape of glycophorin A as a function of its two rotational angles. Distance, tilt and slide have the values of the native conformation. Top left in 2D, all others in 3D - from different perspectives!

be refined in section 4.1.7 and 4.1.8. The plot does not show how large the energy gap between absolute minimum and the native conformation actually is. Table 4.1 lists the exact energy values for these two conformations.

Note the crosslike symmetry in the plot, showing that there are largely repulsive residues at the interface for a rotation of about 180°.

4.1.2 The contribution of the residue-environment interaction

For the same variables the residue-environment interaction is plotted in figure 4.5. The radii we used here are based on the calculations of Gu *et al.* [38]. They are a useful measure for distinguishing the different expansions of the residues. To find the ideal radii for this application we introduced a common factor to the set of radii. The influence of this factor will be investigated in section 4.1.5. Here a factor of 1.25 was used. Even if the absolute values are not sure yet, the influence

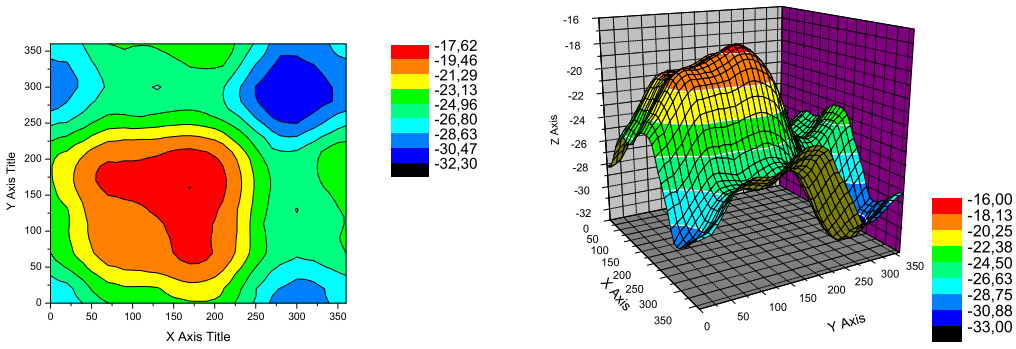


Figure 4.5: The residue-environment interaction of GpA as a function of the rotational angles α_1, α_2. All other variables have the values of the native structure. Radii from Gu *et al.* [38] scaled by a factor 1.25.

of the environment can be seen from the plots as a relative upshift of some of the bothersome areas favoured by the residue-residue interaction and a distinct highlighting of the plateau next to the native values. Note that the effect on both values $(60°, 60°)$ and $(230°, 240°)$ is rather similar. See table 4.1 for the exact values. Especially the minimum around $(150°, 150°)$ in the energy landscape of the residue-residue interaction gets a relative uplift by the environment interaction.

4.1.3 Influence of the 5^{th}-order overlap

Figure 4.6 shows a third contribution: the overlap of 5^{th} order, that is a measure for the entropical contribution from the sidechain mobility together with overcrowding effects. It has a similar shape as the residue-environment interaction - favouring the same plateau. Also here the scaling is still to be found.

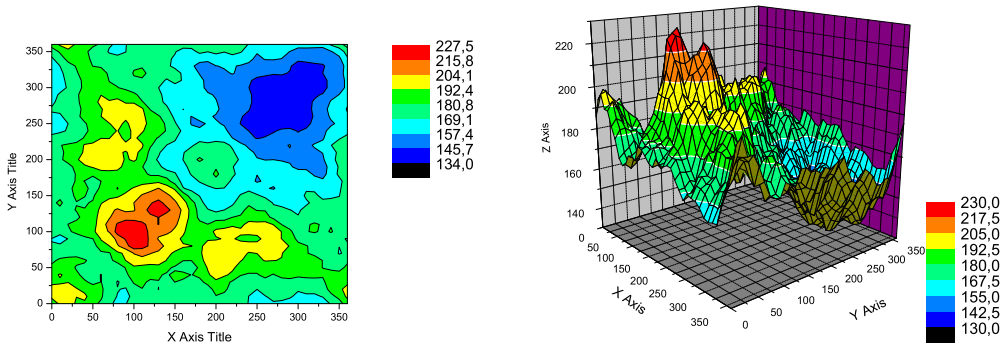


Figure 4.6: The overlap of 5^{th} order for GpA as a function of the rotational angles α_1, α_2. All other variables have the values of the native structure.

Remarkable is that both plots for environmental interaction and for the overlap have a wide minimum next to the native conformation, while the native conformation itself is not within that minimum. It is like a snapshot of a rather flexible structure at the border to the region that is accessible for it - or in more physical terms, it seems that the native structure is located right on the side of the potential well.

α_1	α_2	E_{env} [KJ/mol]	E_{res} [KJ/mol]	Overlap [$\AA^2$]
60°	60°	-21.38	-14.17	177.17
230°	240°	-24.94	124.65	149.33
Δ		3.56	138.65	27.84

Table 4.1: The values of the absolute minimum compared to the native conformation. The values for the overlap and E_{res} are unscaled.

4.1.4 Other perspectives

So far we have looked only at the influence of the rotation angles α_1, α_2 at $d^{(native)}$, $l^{(native)}, \phi^{(native)}$. The next step is to vary d and l for fixed $\alpha_1^{(native)}, \alpha_2^{(native)}, \phi^{(native)}$, as can be seen in figure 4.7. Three features are noteworthy. First, around the native value is a rather smooth valley, which is thermodynamically favorable. Second, there is a steep wall around that valley that has two branches. One that is slide- and tilt-dependent in a nearly parabolic way. The other approximately constant one branches at a slide of about 30$\AA$. Third, the almost homogeneous wall near the regime of parallel-helices. For a distance of 7.5$\AA$ it is obvious that the parallel case will lead to highly repulsive energies.

It is reasonable to assume that tilt angles somewhere above $\approx 40°$ will have larger energies than calculated. As soon as the ends of the helices are buried in

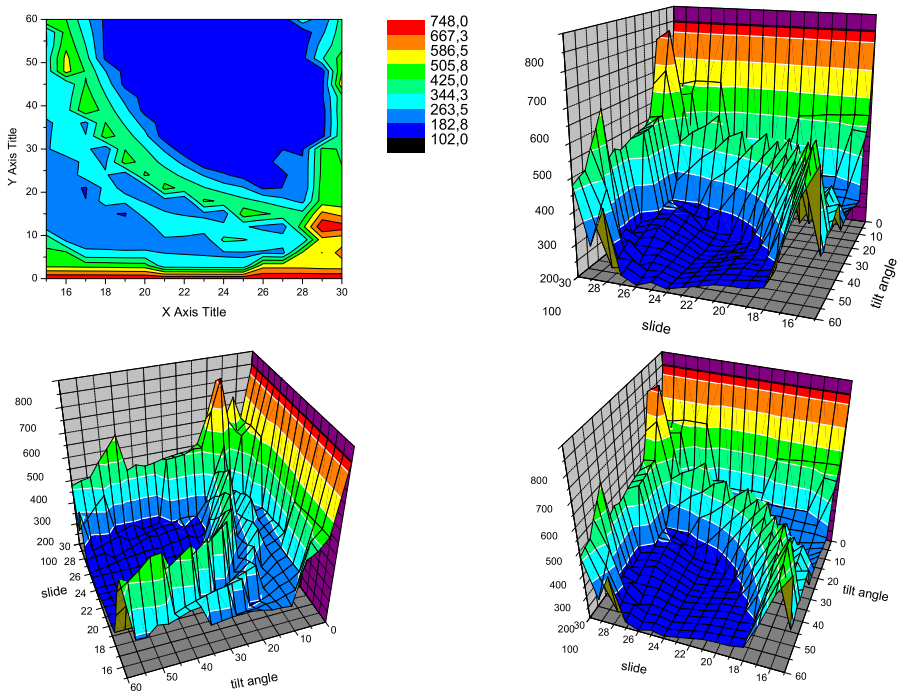


Figure 4.7: The residue-residue interaction as a function of tilt and slide.

the membrane, two contributions become important that are still missing. The total dipole moment of the helix can be regarded as located at the ends of the helices carrying a partial charge $\pm\delta e$. In our model we chopped off the parts of the protein which are not in a α-helical conformation. Some of the residues close to the helix would be buried in the interface for larger tilt angles.

The behaviour of the environment-interaction and the overlap with the same variables are somewhat simpler, as shown in figure 4.8. The residue-environment interaction, plotted in the left image, shows a weak dependence on the slide variable, e.g. in the given native rotational orientation all residues have a similar interaction with the environment, but are favouring the crossing point of the helices to be located in the middle of the helices rather than at the ends. The tilt angle shows a strong preference for burying the helix that will avoid parallel helices but is opposed by the previously explained, not included effects that do not allow the helix to be buried.

The overlap of the residues naturally reduces with larger tilt. The overlap is increasing at the ends and has a smooth valley in the middle of the helices - which is a similar behaviour to the one in the plots for the residue-residue interaction.

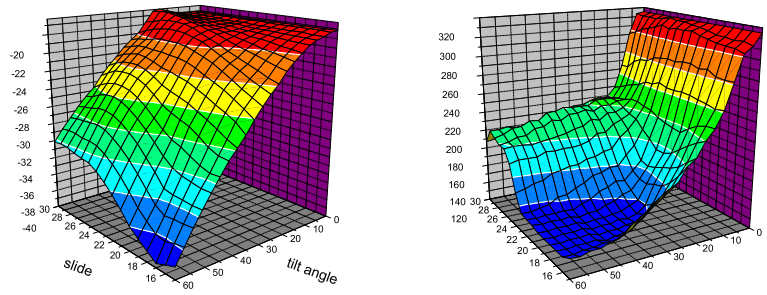


Figure 4.8: Tilt and slide dependence of residue-environment interaction (left) and overlap (right).

4.1.5 Influence of the radii on the residue-environment interaction

As described in 4.1.2 we introduced a common factor to the set of radii that we took from Gu *et al.* [38]. Figure 4.9 shows the residue-environment interaction for factors 0.6, 0.8, 1, 1.25, 1.5, 1.75, 2. The two features of interest are shape and z-range of the plots. The major change in the shape occurs between the factors 0.6, 0.8 and 1. It is reasonable to assume that for the factor = 0.6, a contribution is missing that starts to have an effect at a factor of 0.8, while it seems to be fully developed when the radii remain unchanged for a factor of 1. When the factor gets above 1.75 the shape is flattening, indicating that a limit is reached, above which information will be lost.

The dependency of the z range on the radii is shown in table 4.2. First, the z values continuously and strongly decrease with increasing radii. This emphasises the importance of an exact determination of the radii to get a proper scaling of the algorithm. Looking at the gap between maximum and minimum in each plot uncovers that the gap does not decrease continuously in the same manner as the absolute values, but has a maximum around the radius-factor 1. This will be discussed in more detail in the following section 4.1.6.

factor	0.6	0.8	1	1.25	1.5	1.75	2
maximum [KJ/mol]	-102.9	-60.9	-30.2	-17.6	-12.3	-9.4	-8.1
minimum [KJ/mol]	-113.9	-72.4	-46.6	-32.3	-23.5	-17.7	-13.7
Δ [KJ/mol]	11	11.5	16.4	14.7	11.2	8.3	5.6

Table 4.2: The z-ranges of the results of the residue-environment mode of the sphere-algorithm for different factors of the radii.

Table 4.3 shows that while the choice of the radii affects the energy landscape strongly, the influence of different radii on the two main points of interest remains small.

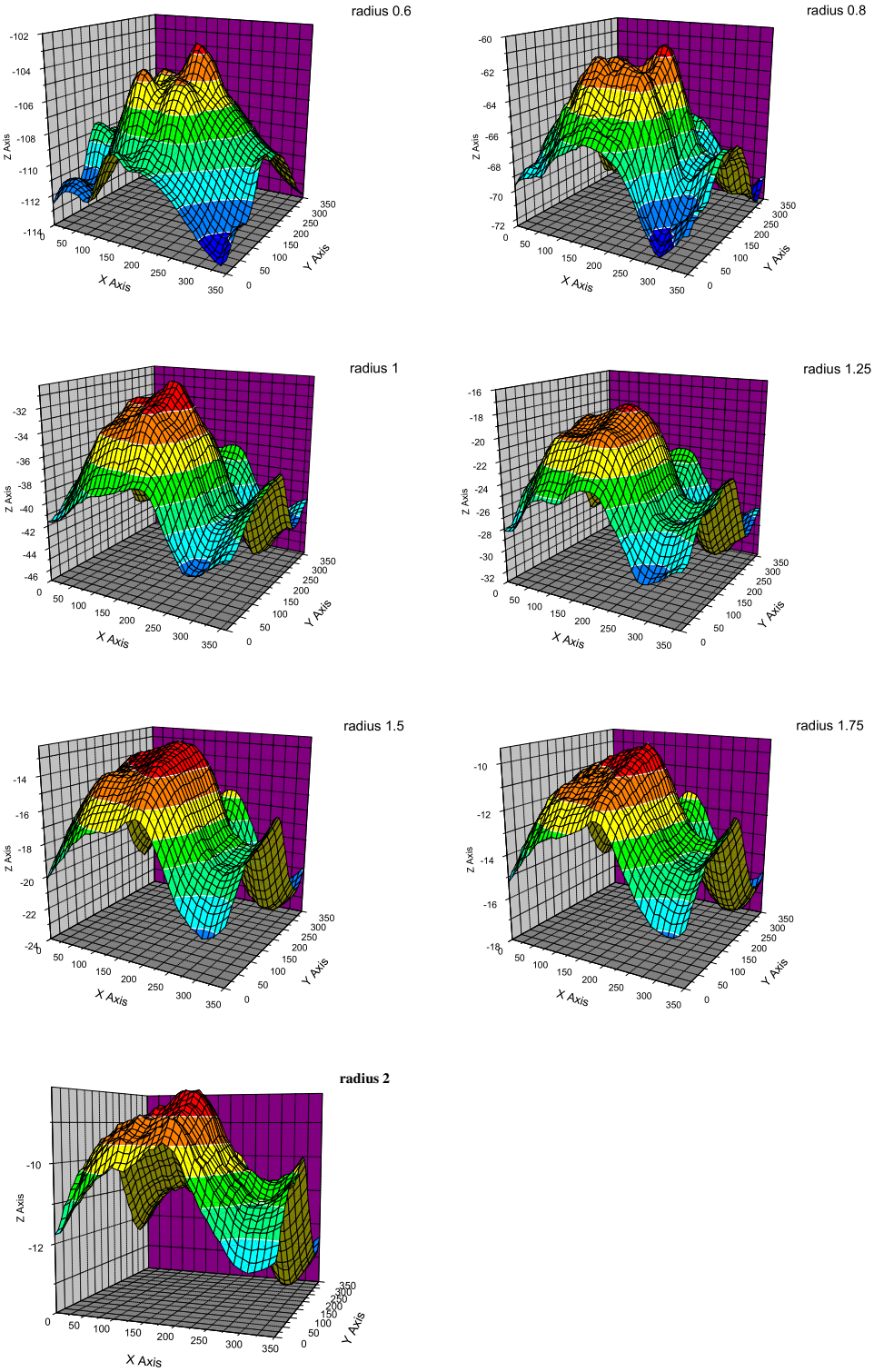


Figure 4.9: Different radii and their effect on the environmental interaction. Mind the different scales of the z axes.

factor	E(60°,60°)	E(230°,240°)	ΔE
0.6	-111.527	-106.491	5.036
0.8	-66.1719	-66.0963	0.0756
1.0	-35.7376	-38.0178	-2.2802
1.25	-21.3817	-24.9463	-3.5646
1.5	-14.9908	-17.496	-2.5052
1.75	-11.6517	-13.4919	-1.8402
2	-9.56057	-11.25	-1.68943

Table 4.3: Influence of the radius on the residue-environment interaction for the native and for the minimum conformation from the residue-residue plot.

4.1.6 Splitting of the environmental plots

As discussed in 4.1.5, there seem to be two contributions combined in the plots for the interaction of the residues with their surroundings. These become visible by moving the helices far away from each other, such that no overlap between the helices can occur. The separation allows to distinguish the change in the residue-environment that is caused by twisting a single and tilted helix in the membrane, from the contribution of the interface of the helices. Moving the helices sufficiently far away from each other, here 50$\AA$ leads to the top left plots of figure 4.10 and 4.11, for radius-factors 1 and 2 respectively. The shape indicates that some residues that are in the interface, or close to it, are moved along the z-direction. This is an effect of the tilt. For a perpendicular helix there would be no difference in the interaction due to the rotation. The difference plots show two things. The interface region clearly prefers the helices to be oriented around (280°,280°). While the plot with a radius factor 1 is very sharp and refined, the plot for a factor 2 seems to have lost some information. Table 4.4 is an extension of table 4.2, containing values not only for the plots with a 7.5$\AA$ distance, but also for the plots with 50$\AA$ and for the difference plots. The high quality of the difference plot for radius-factor 1 (fig. 4.10) goes together with the largest Δ-value. It is therefore reasonable to assume that a radius-factor of 1 or 1.25 should lead to the best results in the frame of the environmental interaction.

4.1.7 Influence of a distance-cutoff

The tilted helices of glycophorin A do not cross at the middle of the helices. This leads to a rather large separation of the ends of the helices. It is likely that the helices will have lipids or headgroups between them. Consequently, using a vacuum interaction between residues may be problematic. This is related to the pending formulation of a damping term for the interaction of residues in media. Instead of attempting an exact determination of the damping term, here we only perform a simple test by introducing a distance cutoff to see in which way the long-range

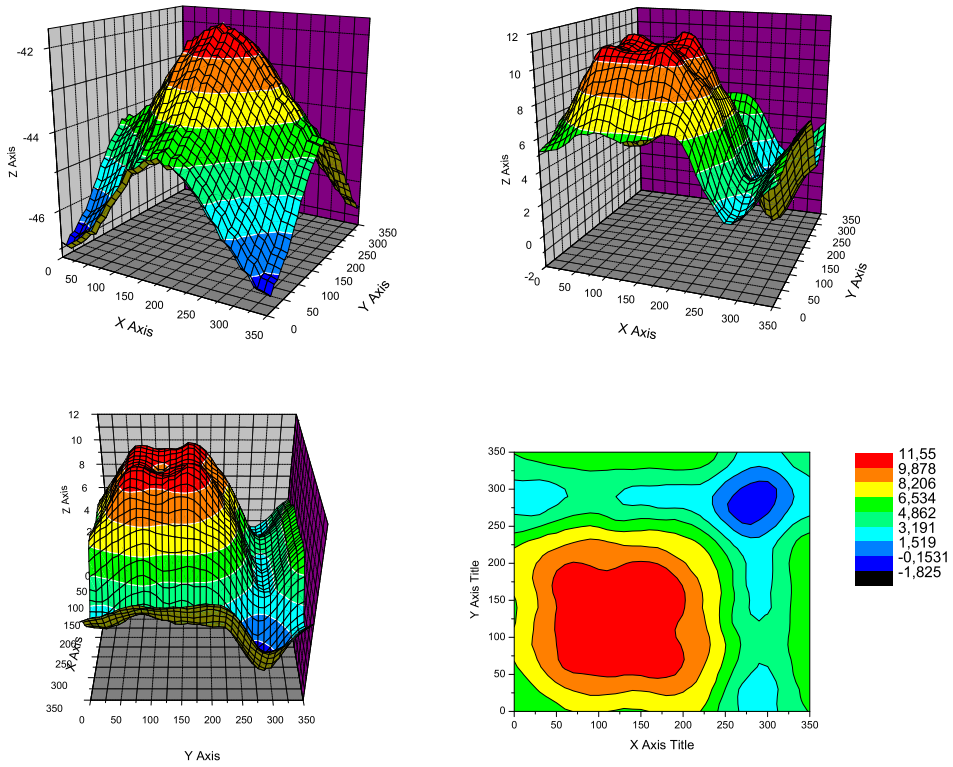


Figure 4.10: Splitting of environmental interaction into general change due to the rotation of the helices and the specific interaction in the interface for a radius-factor of 1. Top left image shows the interaction of the helices with the environment at a distance of 15 Angstroem. The other three show the same data, namely the difference between the plots for 15 and 7.5 Angstroem.

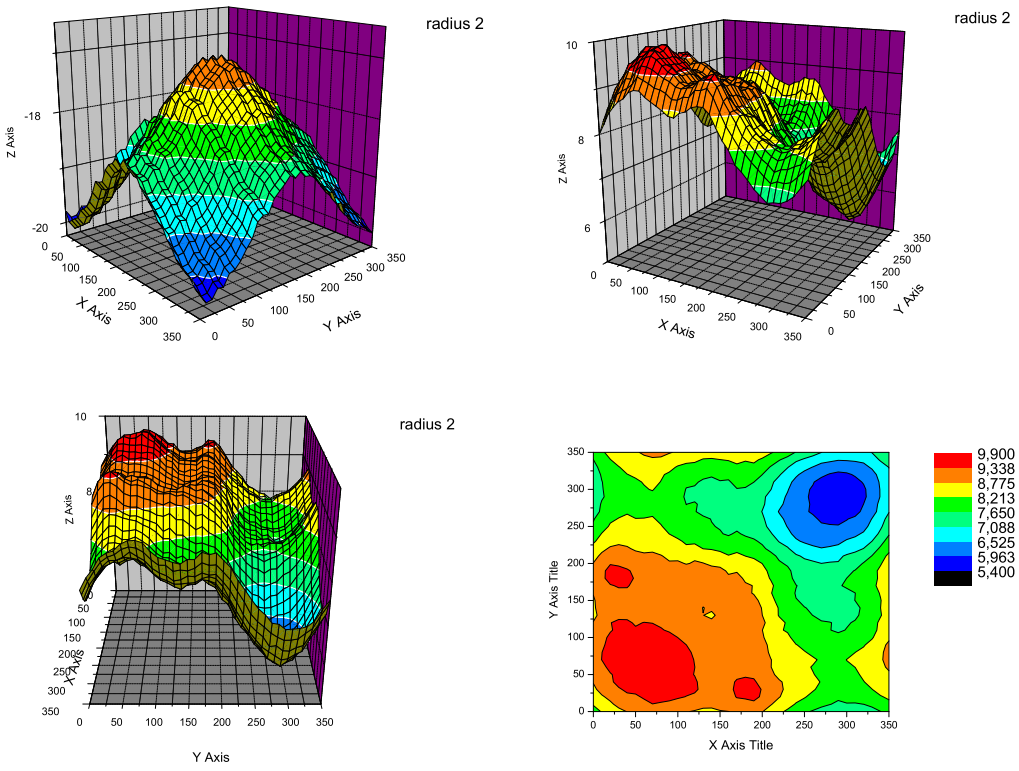


Figure 4.11: Splitting of environmental interaction for a radius-factor of 2. Top left image shows the residue-environment interaction for a distance of 50 Angstroem. The others show the difference between the plots for 50 and 7.5 Angstroem.

distance	factor	0.6	0.8	1	1.25	1.5	1.75	2
7.5Å	maximum	-102.9	-60.9	-30.2	-17.6	-12.3	-9.4	-8.1
	minimum	-113.9	-72.4	-46.6	-32.3	-23.5	-17.7	-13.7
	Δ	11	11.5	16.4	14.7	11.2	8.3	5.6
50Å	maximum	-104.7	-66.3	-41.6	-29.9	-23.7	-19.7	-17.0
	minimum	-113.9	-73.2	-47.0	-34.2	-27.4	-23.0	-20.1
	Δ	9.2	6.9	5.5	4.3	3.7	3.3	3.1
7.5Å − 50Å difference-plot	maximum	2.7	6.9	11.5	12.5	11.8	10.8	9.9
	minimum	0	-1.0	-1.8	0.2	2.6	4.3	5.4
	Δ	2.7	7.9	13.3	12.3	9.1	6.5	4.5

Table 4.4: The z-ranges for different radius-factors.

interaction influences the energy landscape. Different cutoffs at 20, 15, 10 and 8 Angstroem were tested. The second type of interactions that this cutoff will influence are the long range residue-residue interactions within the protein, which also are overestimated by vacuum values. As soon as there are other residues between the considered residues their interaction will be damped due to polarisation effects.

Figure 4.13 shows the resulting energy landscape for the cutoff of 8Å, which has the largest impact. Because it is difficult to see the changes in the energy landscape with these plots, a detailed investigation in terms of difference plots was performed.

The difference-plots are shown in figure 4.14. The first plot at the top left is the difference plot between the plots for no cutoff and a cutoff of 8Å, which should show the largest change. A cutoff of 8Å will definitely lead to an underestimate of the interactions, but is still regarded here to investigate in which way it influences the energy values. More reasonable for further usage seems a cutoff of 10Å. The difference-plot between the plot for no cutoff and the one for 10Å is shown on top right. Both plots have in common that they offset the whole energy landscape to a lower level and that they decrease especially the area around the native conformation. This is a step in the right direction, because it influences the primordial [1] energy landscape in such way that the native conformation becomes more likely to be predictable with these functions.

After introducing the total effect of the cutoff we now go step by step through the cutoffs to its corresponding influence. The middle left plot of 4.14 shows the

[1] Throughout this section we refer to the residue-residue energy-landscape without shift and cutoff as the primordial.

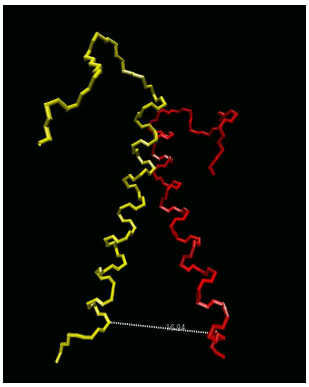


Figure 4.12: The backbone of glycophorin A. The white line connects the C_α atoms of Leu 98(yellow chain) and Lys 100 (red chain). The distance between them is 16.94 $\AA$.

difference between having no cutoff and one of 20$\AA$. The plot has a different gradient principle than the others.

The plot on the middle right shows the change which occurs when the cutoff is changed from 20$\AA$ to 15$\AA$. Cutting the tail of the energy functions in this region has a different effect. The gradient pattern of this plot inverts the previous.

This investigation is of course approximate, but the main effects offsetting the whole energy landscape and preferring the area of the native structure are apparent. Therefore we postulate a similar behavior also resulting from more refined damping terms.

4.1.8 Influence of a distance-shift

The results for the applications of the sphere algorithm are very promising, even when they are not yet properly scaled. The fit functions that describe the residue-residue interaction do not yield the native conformation as the absolute minimum of the energy landscape. Even if the introduction of a more realistic damping factor should reduce the gap, as was indicated in 4.1.7, it will still not disappear. Yungki Park *et al.* [43], who used the same dataset, were able to identify the native structure. Restricting on the Van-der-Waals part of the calculated energies, Yungki used one dimensional fit functions that were averaged over a range of angles.

The crucial step he made was in using positions close to the geometric center of the sidechains as the location of the residues instead of the C_α positions - the variable in which the data was calculated. This might appear unjustified on first sight, but leads to some useful insights. In this way, the energy function becomes more specific and penalizes overlapping side chain conformations stronger than before.

For residues with parallel orientation, the shift will have no consequence. We

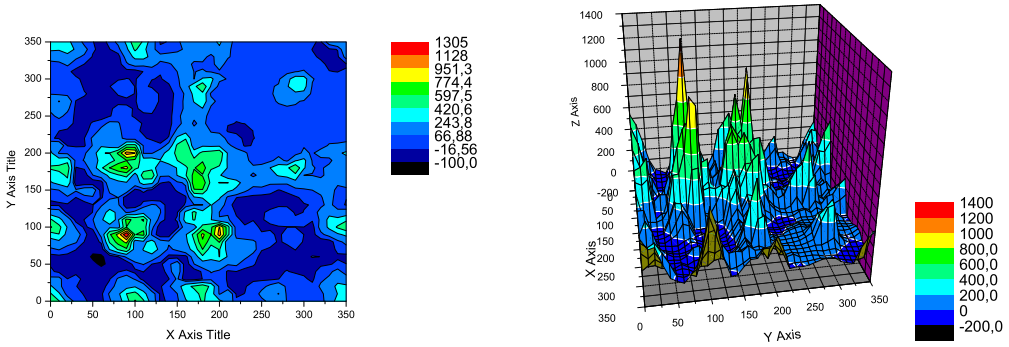


Figure 4.13: Influence of a distance cutoff of 8$\AA$ on the energy landscape of glycophorin A as function of the two rotational angles of the helices, with the other variables having native values.

present now some tests on this to see the effect of different kinds of shifts on our energy functions. Besides discussing a few full energy landscapes and difference plots we also look more explicitly at the effect on two reference points that can be used as indicators for the quality of the energy functions. These are the native conformation $(230°, 240°)$ and the conformation belonging to the absolute minimum $(60°, 60°)$ found in figure 4.4 on page 45. Table 4.5 shows the energy values for all kinds of shifts for these two conformations. In this table the results from the shift are as well combined with the cutoff, introduced in section 4.1.7.

A simple test was done first by shifting all distances about 0.84 $\AA$ (the reason for such an odd value will be justified in the following step). This will have a comparable effect on the residues pointing towards each other as shifting the location of all residues away from the helix axis. The residues that are not effected by a location shift will be affected here. Residues pointing towards each other will experience an opposite effect, e.g. they will move closer instead of away from each other.

Figure 4.15 reveals a large change resulting from the simple difference shift, which is easier to see in terms of a difference plot. Figure 4.16 shows the difference between the energy landscapes with and without shift. All values are upraised, while relative to the rest of the energy landscape the area around the native conformation is favored.

Next, we shift the location of the residue stepwise outwards, and thereby approach the area where the geometrical center can be assumed. We use the neighbouring backbone atoms of the C_α to define two vectors, pointing to the C_α. Summing these two vectors we get our reference vector with the magnitude s. We simply multiply this vector with an increasing factor to see which factor leads to

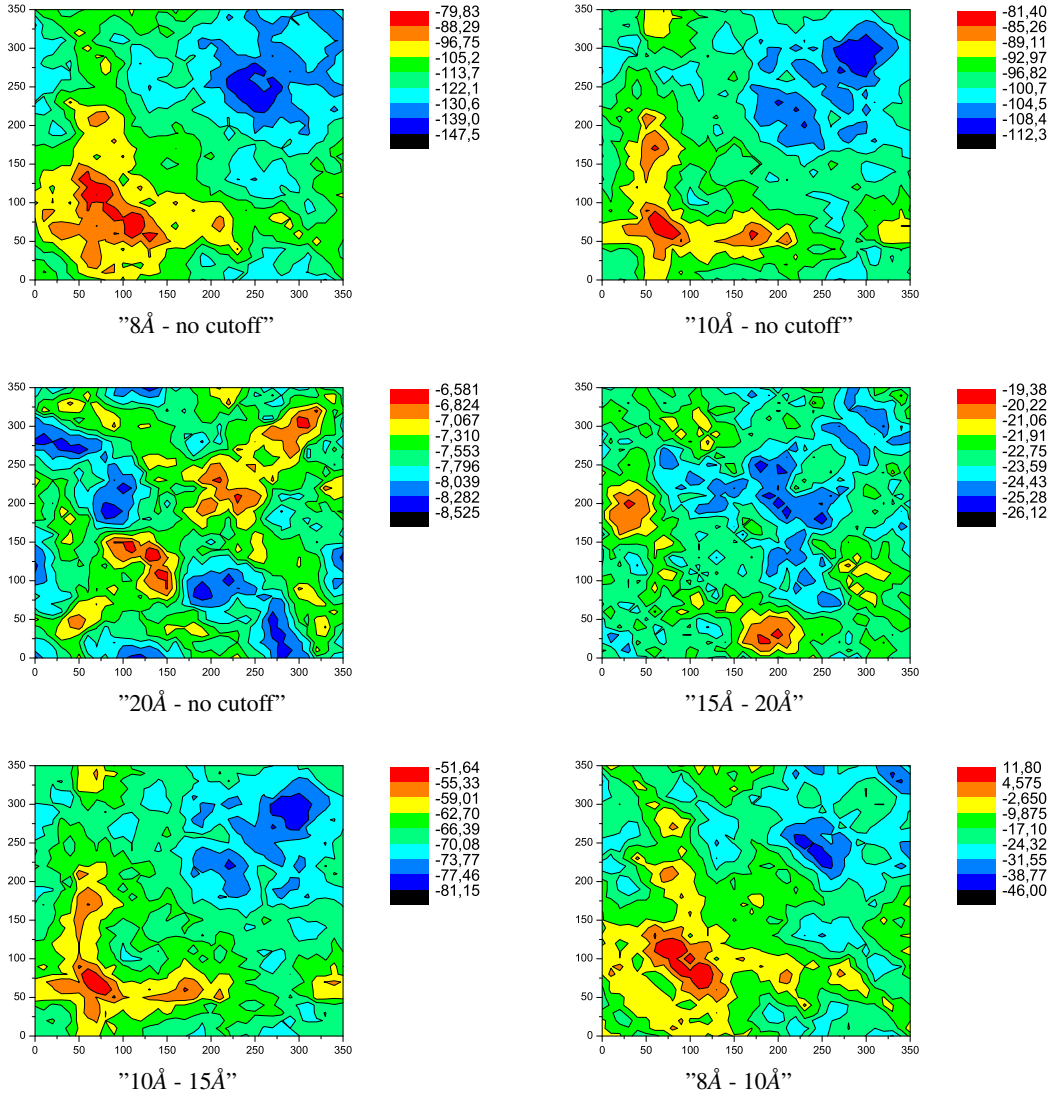


Figure 4.14: Difference plots for different cutoffs. Top left: the difference plot for the cutoff of 8Å versus the primordial (native without cutoff) conformation. Top right: with cutoff 10Å versus the primordial. Middle left: cutoff 20Å vs. primordial. Middle right: cutoff 15Å *vs.* cutoff 20Å. Bottom left: cutoff 10Å *vs.* 15Å. Bottom right: cutoff 8Å *vs.* 10Å.

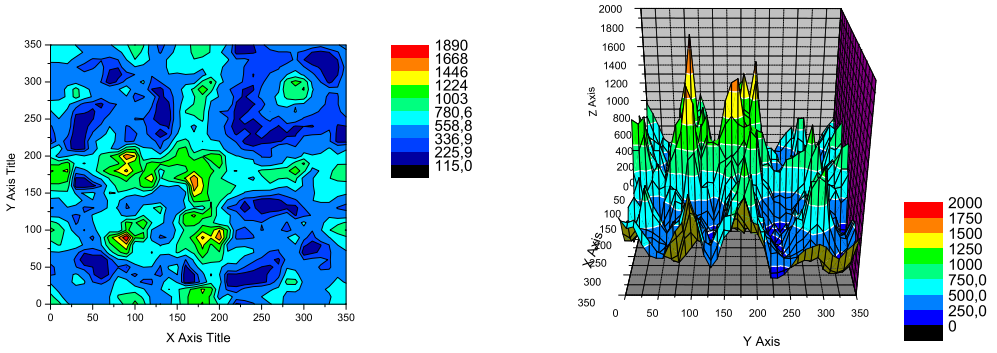


Figure 4.15: A simple distance-shift of 0.84$\mathring{A}$.

the best results. Figure 4.17 shows the results, figure 4.18 the results in terms of difference plots. This testing is more of qualitative nature - to get the right refinement one should start from the damping term and the whole helix simulation. The length s is appoximately 1.68$\mathring{A}$. Therefore, the previously introduced simple shift of 0.84$\mathring{A}$ corresponds to $0.5 \cdot s$. This choice allows a direct comparison of the effects of the simple shift and the shifting of the residue location. As can be seen in the bottom right image of figure 4.18 the simple shift and the location shift have largely differing effects. It is interesting that the difference plot shows such high symmetry. What can be seen from figure 4.17 and table 4.5 is that the gap between the native structure and the one where the energy landscape had the min-

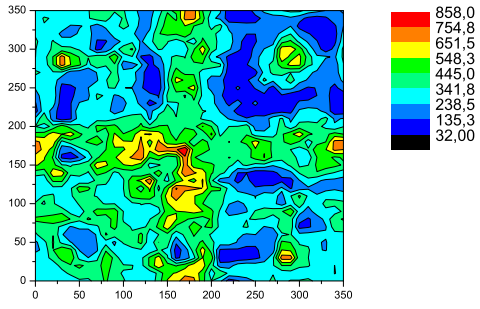


Figure 4.16: The difference between the energy landscapes with and without the simple difference-shift of 0.84$\mathring{A}$.

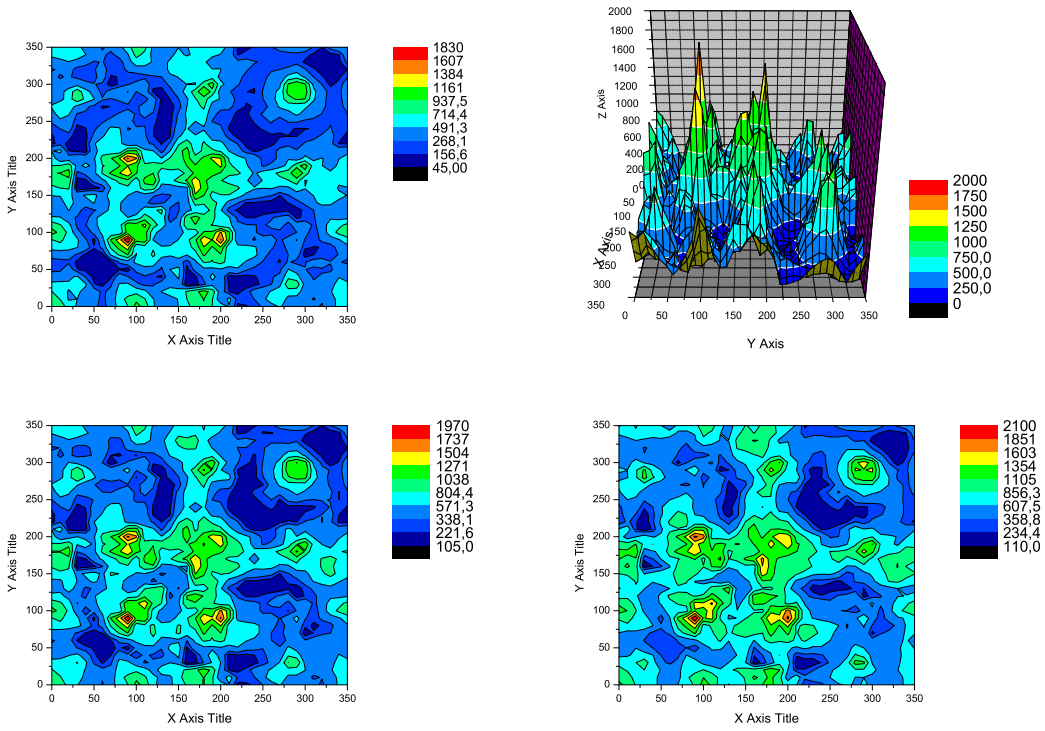


Figure 4.17: The influence of the distance shift on the energy landscapes. Top for a shift of $0.3 \cdot s$, left in 2D, right in 3D. Bottom left for a shift of $0.4 \cdot s$, bottom right for a shift of $0.5 \cdot s$.

imum before introducing the shift is getting smaller already in the first picture[2]. The difference plots in figure 4.18 show this effect clearly. Increasing the shift increases this effect. At the same time the absolute values are increased. It is obvious that with increasing shift a part of the helices are moved closer together and repulsive contributions will rise the strongest. Combining these results with the

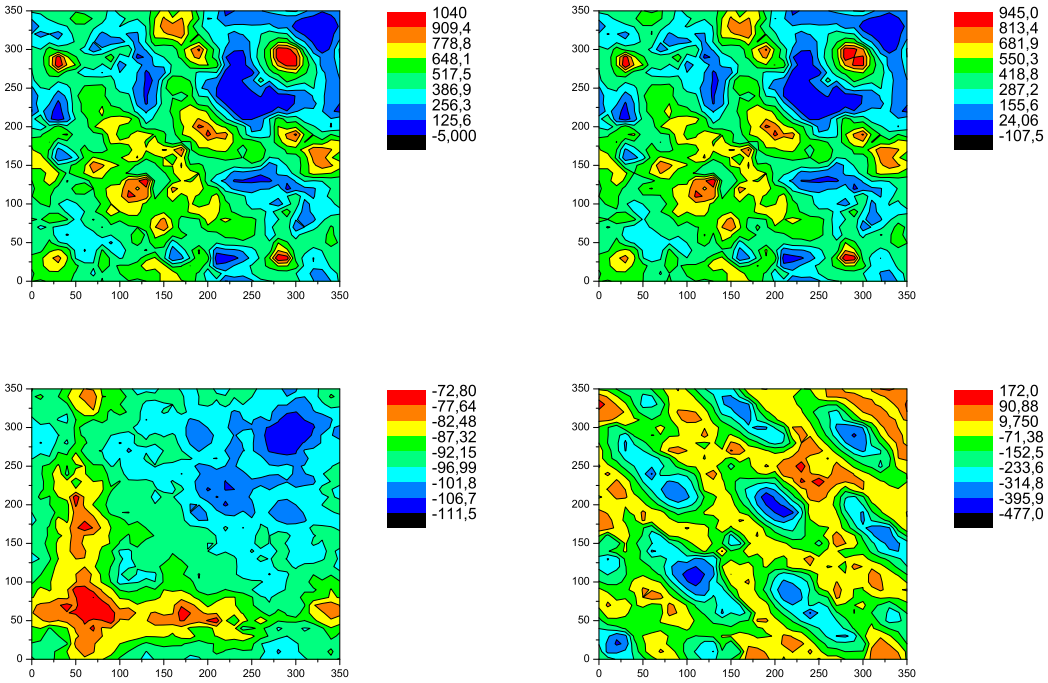


Figure 4.18: Difference plots for distance shift. Top left the $0.5 \cdot s$ shifted minus the primordial. Top right same shift with cutoff of $10\mathring{A}$ minus the primordial. Bottom left the the shifted with cutoff minus the shifted without cutoff. Bottom right the simpleshift of $0.84\mathring{A}$ minus the shift of $0.5 \cdot s$.

cutoff of the previous section leads to the most striking results. Not only does the gap between the native and minimal-energy conformation inverse, but they are also close to attracting energies. The lower part of table 4.5 shows the common effect of shift and cut. Note that the optimal shift seems to be around $0.4 \cdot s$ because the gap is clearly inverse and at the same time the absolute value appears reasonable.

How should we understand the change? We can assume that it is necessary to introduce shift and cutoff together. The effect of the shift is twofold. First,

[2]Recall table 4.1 on page 47 as reference to this.

α_1	α_2	cutoff	$0.2 \cdot s$	$0.3 \cdot s$	$0.4 \cdot s$	$0.5 \cdot s$	$t = 0.84\AA$
60°	60°	none	7.85	45.65	144.83	257.53	124
230°	240°	none	100.64	107.06	111.19	126.63	199.16
60°	60°	10Å	-77.57	-33.37	68.84	183.9	34.62
230°	240°	10Å	6.3	3.69	7.99	22.69	80.38

Table 4.5: The influence of different shifts and a cutoff of 10Å on the values of native and absolute minimum. $s \approx 1.7\AA$.

for the interaction of residues nearby the distance is getting smaller - as explained before, this can be interpreted by a too large flexibility of the sidechains to bend during simulation - due to the 'unrealistic' pair model. The residues pointing away from each other are assigned larger distances, which might be explained by internal shielding. The comparison to the simple shift data shows that they are not leading to the same result which should be the case if mainly the nearest residues would play a role in this effect. Internal shielding should also play a role. The cutoff is affecting mainly long range interactions that are definitely shielded by other molecules or other residues. A distance dependent shielding term ought to be introduced with a minimum distance, below which it has no effect.

Inspired from the success of the shift/cutoff-introduction, we rescanned the conformational space with the modified potentials for the following ranges:

$$\begin{aligned} d &= 7, 7.25, ..., 8 \\ l &= 12.35, 14.35, ..., 26.35 \\ \theta &= 23.5, 28.5, ..., 43.5 \end{aligned} \tag{4.2}$$

and the usual step sizes for the rotation around the helical axes. This means 259,200 conformations. 334 conformations have lower energy values than the native structure, but most of them for larger distances, where lower values become anyhow more likely for the residue-residue interaction. For a distance of $7.5\AA$ there are 71 conformations below the native. That is of some orders of magnitude lower than before the introduction of the new positions of the residues. The influence of the residue-environment interaction of the tight-packing and of the tilt are not included here. Also, the manipulation of the energy functions was pretty rough yet. There is enough evidence to suppose that after proper scaling of the sphere algorithm and a more refined treatment of the potentials, the native conformation can be identified within the whole conformational space or that there are only few rivals that will be eliminated after switching to atomistic scale.

Chapter 5

Search strategies in large systems

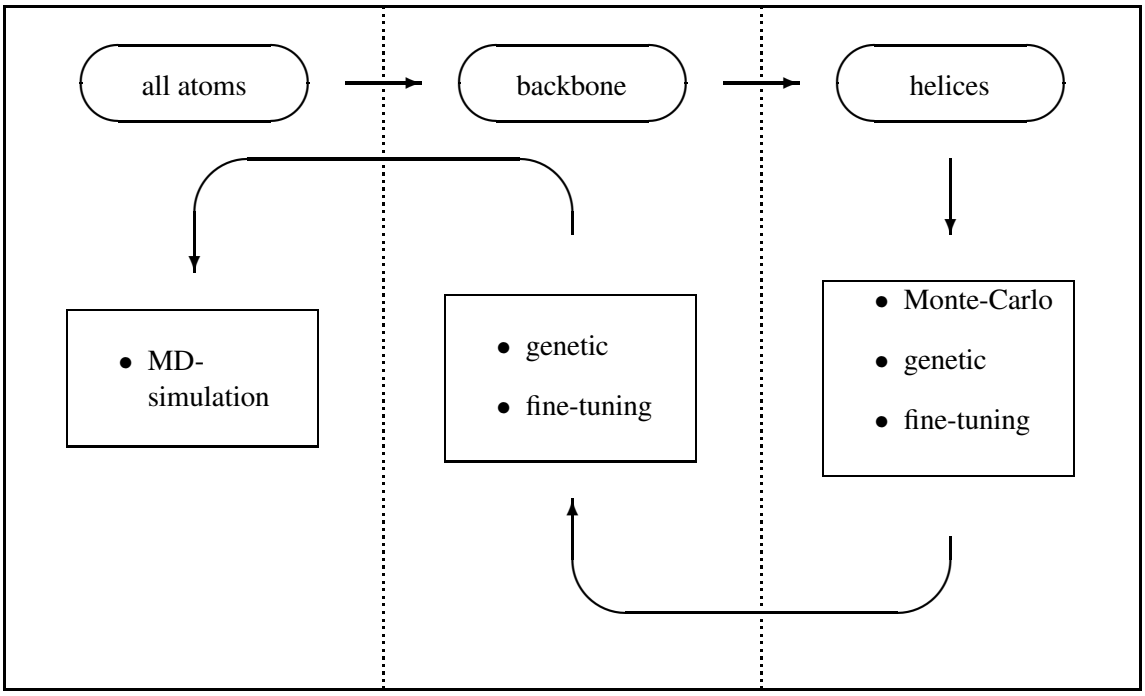


Figure 5.1: Overview of scales and methods of first approach for searching conformational space.

5.1 Minima-search through complex energy landscapes - the first approach

The figure 5.1 gives an overview of the scales and the methods that are applied within each scale. By introducing the residue model, the number of interactions is reduced drastically, since, by cutting off the loops, the number of degrees of freedom is reduced. The former step accelerates the energy calculation while the latter simplifies the search through conformational space. In our first approach, we started with a Monte-Carlo search using helices on the residue level. That was followed by a genetic algorithm and fine-tuning. The idea was then to go up one level by adding the sidechains, using an algorithm that first adds the loop to one of the helices and then minimizes the distance and, when close enough, also the angle of the loop end and the helix end. On this level, only the application of the genetic algorithm and the fine-tuning makes sense. The Monte-Carlo method is not necessary, because the rough screening was done already. Note that the search is performed in different variables in these two levels. Regarding helices, the variables are position and orientation of the helices. Adding the loops, the variables are the dihedrals. Finally, the sidechain-atoms can be added and MD-simulations can be performed for the best structures obtained previously. The addition of the sidechains can be done conveniently by storing average sidechain conformations in relative coordinates of the backbone geometry. The relative coordinates remain the same, independently from the actual orientation of the backbone, and the sidechains can be added in a very simple fashion. To meet the requirements of the conformational conditions a sidechain optimizing algorithm could be applied before passing the structure to MD-simulations. This scheme is similar to the one recently used by Goddard *et al*. [64] to predict structures of G-coupled receptors.

5.1.1 Genetic algorithm

Because this approach is not the final one, and since there is nothing special about the Monte-Carlo search and since the fine-tuning is close to what is done in the final approach, we concentrate on explaining the concept of the genetic algorithm. After running the Monte-Carlo search for some time one will see that while the total picture is not very exciting yet, one can find frequently helix pairs within the full structure that seem to have found relative orientations that are promising. Collecting a sufficient number of structures with promising helix pairs one can accelerate the search by learning from them. By calculating the energy pairwise and storing their relative position and orientation in a gene, the structure of a protein can be determined pairwise using relative positions and orientation of a sequence of helix-pairs. The sequence of helices does not necessarily has to be the one given by the connecting loops. The advantage of the native one is that the constraint of loop-length is conserved more easily. One way to do this is to express the connecting vector $\vec{v}_c$ between two helices in terms of the relative coordinate system of the first

helix:

$$\begin{aligned} \vec{v}_c^{(ij)} &= \vec{r}_{cms}^{(j)} - \vec{r}_{cms}^{(i)} = d_x^{(ij)}\vec{e}_x^{(i)} + d_y^{(ij)}\vec{e}_y^{(i)} + d_z^{(ij)}\vec{e}_z^{(i)} \\ \vec{d}^{(ij)} &:= (d_x^{(ij)}, d_y^{(ij)}, d_z^{(ij)}) \end{aligned} \tag{5.1}$$

and the orientation of the second helix in terms of the Euler-angles that transform the relative coordinate system of the first helix into the one of the second:

$$\begin{aligned} \mathbf{A}(\phi^{(ij)}, \theta^{(ij)}, \psi^{(ij)}) \begin{pmatrix} \vec{e}_x^{(i)} \\ \vec{e}_y^{(i)} \\ \vec{e}_z^{(i)} \end{pmatrix} &= \begin{pmatrix} \vec{e}_x^{(j)} \\ \vec{e}_y^{(j)} \\ \vec{e}_z^{(j)} \end{pmatrix} \quad , \mathbf{A}(\phi^{(ij)}, \theta^{(ij)}, \psi^{(ij)}) \in \mathbb{R}^3 \times \mathbb{R}^3 \\ \vec{\Phi}^{(ij)} &:= (\phi^{(ij)}, \theta^{(ij)}, \psi^{(ij)}) \end{aligned} \tag{5.2}$$

One gets the absolute position of the helix-bundle by including the absolute position of one of the helices - most conveniently that of the first in the sequence. Note that the Euler-angles in the genes are relative orientations, which allows to calculate the orientation of the second helix within the coordinate system of the first helix. To calculate the absolute orientation a second rotation has to be performed, which is equivalent to rotating the absolute coordinate system to the one of the first helix. Using the definitions of $\vec{d}^{(ij)}$ and $\vec{\Phi}^{(ij)}$ from equation 5.1 and 5.2 the relative conformation of a protein with 5 helices can be expressed as the following set of values:

$\vec{d}^{(12)}$	$\vec{\Phi}^{(12)}$	$\vec{d}^{(23)}$	$\vec{\Phi}^{(23)}$	$\vec{d}^{(34)}$	$\vec{\Phi}^{(34)}$	$\vec{d}^{(45)}$	$\vec{\Phi}^{(45)}$

Running the random search for a certain time will lead to conformations where parts of the protein are in a favourable orientation, while others are still crap. For example a certain run might have a very low energy value for helices 2 and 3:

$\vec{d}^{(23)}$	$\vec{\Phi}^{(23)}$

In another run it might be helices 4 and 5 that found a nice relative orientation:

$\vec{d}^{(45)}$	$\vec{\Phi}^{(45)}$

By collecting the low energy helix pairs in this way, one can build up a gene-pool, that can be ordered [1] and then recombined in numerous ways. Figure 5.2 shows

[1] That requires the storage of the energy value in the gene.

a couple of results of the genetic algorithm for an 'unrealistic' test. It is based on a simple Lennard-Jones-like potential favoring any compact states, that gave the helices the tendency to cluster in any kind of orientation. Obviously these examples have no biological relevance. The images just show that the algorithm worked well in a rather short time - even when it was done for the entire protein at residue detail, meaning a large number of degrees of freedom.

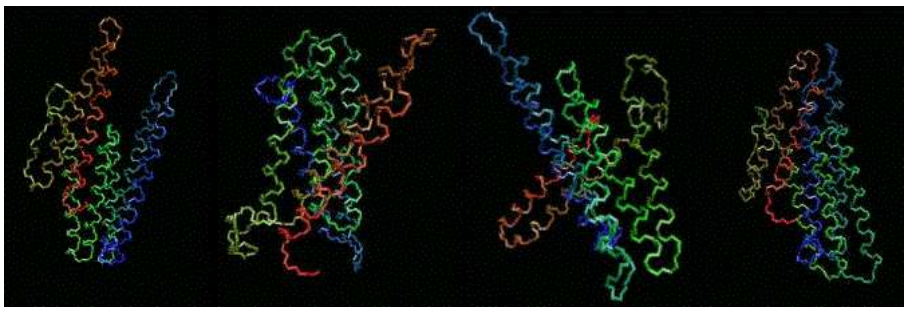


Figure 5.2: 4 different conformations of the 7 helices of bacteriorhodopsin with loops after short run time of genetic algorithm with arbitrary potentials.

Remark For some proteins it might be not the native sequence that has to be minimized. Bacteriorhodopsin for example has one very long loop, that would allow the two helices that are connected by that loop to move away from each other in that way that these two helices are not direct neighbours anymore. Trying to optimise the pair potential between these two helices wouldn't lead to the right result. If one can not exclude this possibility for a protein of interest it is necessary to permutate the sequence of helices in which the protein is built.

5.2 The final approach

Why another approach? The results of testing the previous approach on bacteriorhodopsin were not satisfactory and a more detailed analysis of the problem was necessary. The reasons for the insufficiency were

- the long loops of bacteriorhodopsin create a huge possible conformational space, while most of it is very unlikely
- because the multidimensional energy landscape is extremely rugged one cannot expect a random/genetic algorithm to find the absolute minimum in reasonable time.

The algorithm was allowed to create any possible conformation and a long loop allows many unlikely conformations, where the protein is split into two parts. Picture

5.3 shows the results after approximately one day. The problem of the long loops could be solved by using a random number distribution that favours shorter helix-distances. But also the helices that are connected by shorter loops were not folded properly. There is no reason to expect the proper folding to occur in reasonable time.

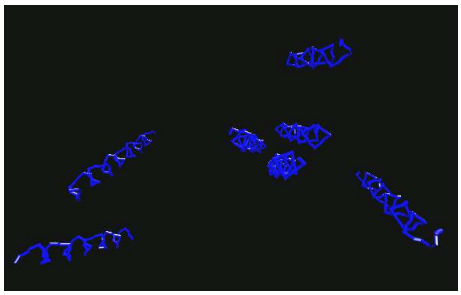


Figure 5.3: The helices of bacteriorhodopsin after about one day runtime of Monte-Carlo and genetic algorithm.

The long loops in principle would allow the helices to assemble in such way that the helices that are connected by a loop would not have to be next neighbours.

The basic idea. The first aim in searching a conformational space is to be smart and fast. One would like to apply methods that do not have to search too long, but follow a gradient or learn quickly during the search. As mentioned before the energy landscape for proteins is huge and rugged. Too huge to search through all of it. Too rugged for the smart approaches. Reduction can be achieved by either simplifying the interaction calculation or in restricting conformational space. The ruggedness requires completeness - limiting the ability to confine. In order to find an optimal compromise we begin our search from a bird's-eye view. By scaning the helices at a fixed distance a complete scan is within reach. Using the same distance for all helices implies that they are placed on an equidistance-grid. It is reasonable to assume that the real structures are not too far apart from the ideal grids. To recombine the single scans most easily, the scans are performed for helix-triplets. The interaction is already simplyfied by introducing the residue modell. For scanning the triplets we only compute the residue-residue interaction, because it makes no sense to calculate any free surfaces before the protein is totally assembled.

The resulting minima are then the starting points for applying an off-grid minimization algorithm, that can move the helices freely after 'release' and should be able to find its way to the absolute minimum by comparing the different minima found for different starting positions. There should be no major hindrances that could let even a good minimizer get stuck. We use available minimizers, gradient-

based and non-gradient-based, from the GNU scientific library [44]. For the resulting conformations atomistic MD-simulations should then be performed.

Figure 5.4 shows all possible grids for 7 helices. Alternative numbers of helices will of course lead to other sets of grids.

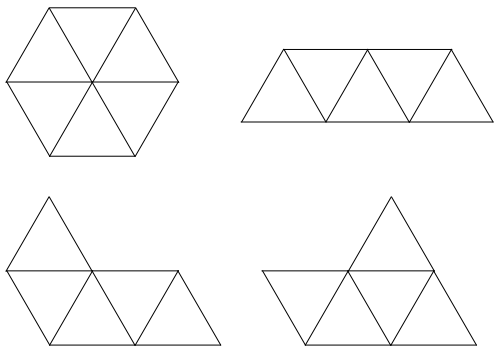


Figure 5.4: The 4 possible equidistant grids for 7 helices.

For a system without initial information one has to test all four grids and fit the sequence onto the grids in all possible ways. This can be done by assigning the helices in all permutations to the grid and checking if the loop-constraint on the maximally allowed distances between consecutive helices is fulfilled. Any structural evidence from experiments will significantly reduce the number of possible permutations and the number of needed helix-triplet scans. In figure 5.5 the cms of the helices of bacteriorhodopsin are shown. The cms are projected on a plane going through the cms of the first three helices a, b and c.

In this picture the centers of mass of the helices are plotted, surrounded by a sphere that would connect the C_α-atoms in an ideal helix from top view. This is only a rough ilustration using the x,y-coordinates of the helices and neglecting the z-coordinate, with the effect that the positions of the last four helices are not very precisely drawn. But it shows that the helices are indeed in a conformation close to the bottom-right as well as to the top-right grid in figure 5.4. Starting from one of the two grids, in optimized orientations, one may expect a minimizer to find the right arrangement of the helices.

Design of the grid-assignment. For each number of helices the possible grids have to be found. Given a certain grid, a construction path has to be defined. First, the grid-points are numbered, e.g. as in figure 5.6. With this numbering,

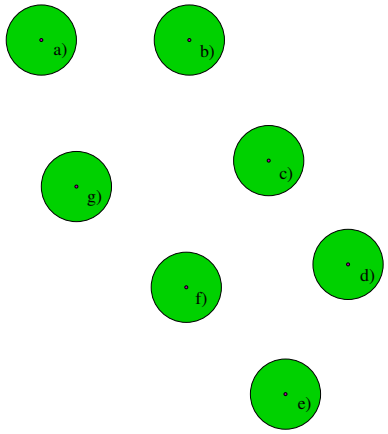


Figure 5.5: The helices from the native structure of bacteriorhodopsin (1C3W.pdb). The cms are projected on a plane. The spheres that are assigned to the cms have a radius that represents the average distance of the C_α's to the axis.

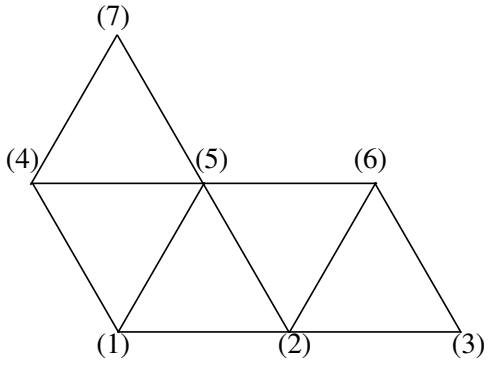


Figure 5.6: One of the grids with assigned numbering.

construction paths can be defined for the grid as an array of triples:

$$\begin{aligned} gridpath[0] &= (1,2,5), \qquad \phi[0] = 0^\circ \\ gridpath[1] &= (1,5,4), \qquad \phi[1] = 300^\circ \\ gridpath[2] &= (4,5,7), \qquad \phi[2] = 0^\circ \\ gridpath[3] &= (5,2,6), \qquad \phi[3] = 60^\circ \\ gridpath[4] &= (6,2,3), \qquad \phi[4] = 120^\circ \end{aligned}$$

The construction directions begin with the first three grid-points (1,2,5) given by $gridpath[0]$. They are located in the same manner as the helix-triplet-scan was performed with the line $\overline{12}$ between the first two helices pointing in x-direction and anti-clockwise ordered. Next, grid-point 4 is added in anti-clockwise manner relative to grid-point 1 and 5. All other grid-points are added in the same way. This choice is arbitrary and does not tell in which order the helices are assigned to the grid. This part is so far not automated, ehich means it has to be provided to the algorithm.

The assignment of the helices is the next step:

$$\begin{aligned} h_1 &\rightarrow g_1 \\ &\vdots \\ h_7 &\rightarrow g_7, \end{aligned} \tag{5.3}$$

where h_i is helix i and g_i the grid-point i. The helices can be ordered on the grid in many different ways, what means to permute the assignment in 5.3:

$$\begin{array}{lccccccc} perm_1: & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ perm_2: & 1 & 2 & 3 & 4 & 5 & 7 & 6 \\ perm_3: & 1 & 2 & 3 & 4 & 6 & 5 & 7 \\ perm_4: & 1 & 2 & 3 & 4 & 6 & 7 & 5 \\ perm_5: & 1 & 2 & 3 & 4 & 7 & 5 & 6 \\ perm_6: & 1 & 2 & 3 & 4 & 7 & 6 & 5 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{array} \tag{5.4}$$

Regarding the permutations as arrays one can perform the following assignment to the elements of the permutation-array:

$$perm_i(j) = k: \quad h_k \rightarrow g_j, \qquad j,k = 1,\ldots,7 \tag{5.5}$$

For example:

$$perm_3(6) = 5: \quad h_5 \rightarrow g_6 \tag{5.6}$$

There is one constraint that has to be checked and that reduces the number of allowed permutations drastically - the loop-length. For 7 numbers one has 5040 permutations. The relation between the position $\mathcal{G}(i)$ of grid-point i and the position $\mathcal{HP}(j)$ of helix j is:

$$\mathcal{HP}(perm(j)) = \mathcal{G}(j).$$

With this the constraint can be written as:

$$\mathcal{L}(i) > ||\mathcal{HP}(i) - \mathcal{HP}(i+1)||, \quad i = 1,\ldots,6 \tag{5.7}$$

with $\mathcal{L}(j)$ representing the loop-length between helix j and j+1.

Now, the actual construction path in terms of the helices has to be determined.

$$act(i,j) = perm(gridpath(i,j)) \tag{5.8}$$

The elements of the construction paths are referring to rather large interaction maps of the three helices involved in each step respectively.

The helix-triplet scans. The helices are arranged in anti-clockwise order. The variables used for the scans are given by the Euler angles of the three helices and additionally a z-shift. Looking at known structures like the one of bacteriorhodopsin reveals that the cms of the helices are not in plane but are shifted. That a minimizer would overcome the well in z-direction, coming from the ridges of the helical conformation, is unlikely and therefore the maps should contain the z-shift. We allowed a maximum difference of the helices of $2 \cdot \Delta z$, that leads to 19 possible combinations of the relative z-location. We have chosen $2 \cdot \Delta z = 5.5\mathring{A}$, what is close to the pitch of the α-helix. The other variables had the following ranges in our first tests:

$$\begin{aligned} \phi_i &= 0^\circ, 60^\circ, \ldots, 300^\circ \\ \theta_i &= 0^\circ, 15^\circ \\ \psi_i &= 0^\circ, 20^\circ, \ldots, 340^\circ; \quad i = 1,2 \end{aligned}$$

For $\theta = 0^\circ$ ϕ is 0°. With these ranges we ended up with $\sim 3.8 \cdot 10^7$ elements per map and a runtime of $150h - 200h$ (depending on the cutoff). For $\theta = 0^\circ$ there is no variation of $\phi = 0^\circ$. Therefore ϕ and θ will be combined into one variable in further use.

What there is to learn from glycophorin. Before one can apply the methods on larger systems one should investigate how our widely exploited test-system glycophorin behaves under conditions similar to those used for the triple scans. Therefore we look at the energy landscape at a distance larger than the native one, namely $9\mathring{A}$. Using the previously optimized energy functions reveals that the optimizations that were worked out are obviously valid only for the conditions of the native conformation. The top-left image of figure 5.7 shows that the native conformation is not in an absolute minimum or close to it. It looks more like the situation where we started the optimization. As the next step we increased the shift-factor s up to 1.5. The top-right image shows that this leads to a more favourable shape of the energy landscape but a shift factor of 1.5 is not so easy to justify anymore and it seems to be reasonable to wait till the whole helix simulations were performed.

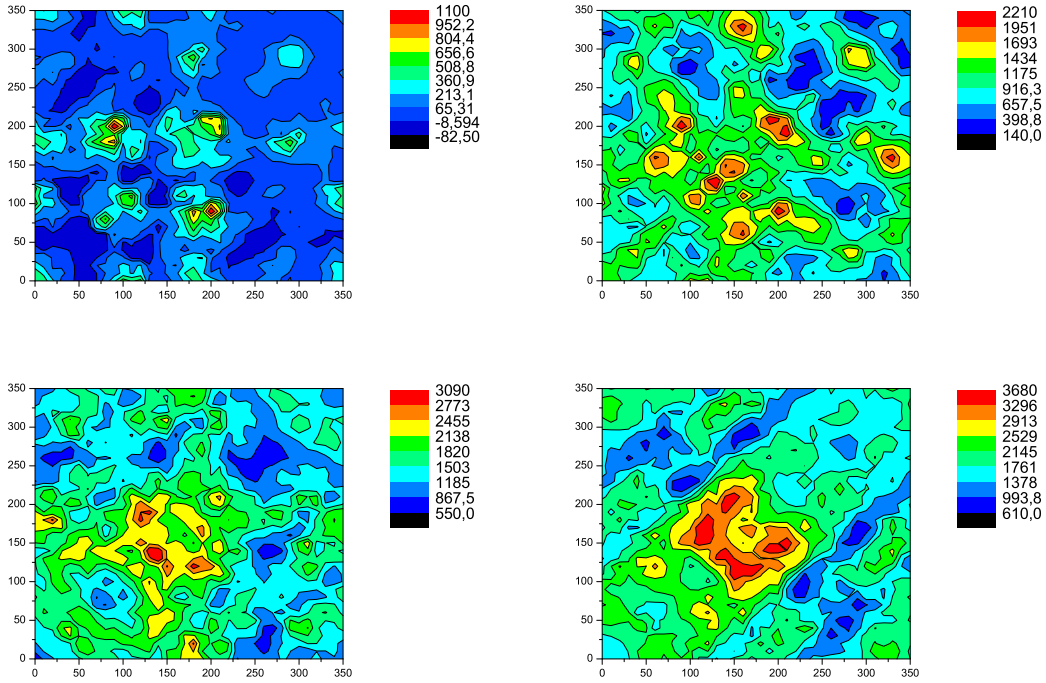


Figure 5.7: Glycophorin A at distance of 9$\mathring{A}$. Both helices are twisted around their z-axes. Top-left: tilt 40°, shift 0.5. Top-right: tilt 40°, shift 1.5. Bottom-left: tilt 20°, shift 1.5. Bottom-right: tilt 0°, shift 1.5. All have a cutoff of 12$\mathring{A}$ and the native slide.

It would be most convenient if it would be possible to identify the ideal rotational angles of the tilted helices already from the untilted case. Then the triple helix scans could be performed with untilted helices and this would decrease the size of the maps enormously. The image on the bottom-left shows the energy landscape for a tilt of 20° and the bottom-right for a tilt of 0°. As far as one can judge at the present state of the energy functions, it does not seem to be possible to identify in this way the ideal rotation angles of the tilted helices from the untilted at larger distances.

The actual grid construction. To construct the grids from the maps of the triple-helix-scan requires taking several steps that are now explained one by one. Taking the grid as in figure 5.6 and filling it with the helices so that one obtains the grid closest to the native one, leads to the triangle shown in figure 5.8.

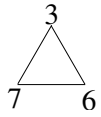


Figure 5.8: The first helix triple in the construction path. The numbers refer to the helix numbers, not the grid numbers.

This triangle is exactly in the orientation the related map was calculated - no rotation is necessary. In the general case the triangles do not have to be in the same orientation. The values obtained from the maps then have to be modified. In the first construction step, the search for minima in the map is comparably easy. Since none of the helices has to be in a specific orientation, all elements of the map are equally treated. The algorithm probes for each element of the map whether it is candidate to be one of the searched minima. If yes, the elements of the map and the conformation of the helices that is connected to the location of the element within the map, are stored in a vector.

The matrix $\mathcal{M}_0$ carries the information of the first construction step in vectors (its lines) like this:

$$\mathcal{M}_{0,i} = (E^{(i)},\, c_+^{(i)},\, \Phi_7^{(i)},\, \psi_7^{(i)},\, \Phi_6^{(i)},\, \psi_6^{(i)},\, \Phi_3^{(i)},\, \psi_3^{(i)},\, z_{763}^{(i)}). \tag{5.9}$$

The variable Φ combines ϕ and θ. The meaning of c_+ has to do with the recombination of the matrices $\mathcal{M}_l$ and will be explained in the following sections.

After the minima for helices 7-6-3 are found, helix 1 is added.
The values for $\Phi_{7,3}$, and $\psi_{7,3}$ have to be taken from $\mathcal{M}_0$ and then rotated about 60° to move it to the conformation used in the map for helices 7-3-1 (shown in the right of figure 5.9). From the z_{763} the z_{73} has to be extracted.
With these values the search through *map731* is performed. All values of Φ_1, ψ_1, z_{731} that are in conformance with the given set of $\Phi_{7,3}$, $\psi_{7,3}$, z_{73} are located within the

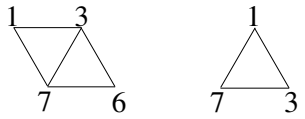


Figure 5.9: The second step of the grid construction. Helix 1 is added to the first triple, shown on the left side. On the left the second helix triple is shown in the way its map was calculated.

map and those identified as minima are rotated back about -60° and then stored in $\mathcal{M}_1$. If $\mathcal{M}_0$ contains n sets of values, and for each set of $\mathcal{M}_0$ also n minima are stored in $\mathcal{M}_1$, then $\mathcal{M}_1$ is of size n^2. The information contained in $\mathcal{M}_1$ is slightly different from the information in $\mathcal{M}_0$:

$$\mathcal{M}_{1,j} = (E^{(j)},\, c_-^{(j)},\, \Phi_1^{(j)},\, \psi_1^{(j)},\, z_{731}^{(j)}). \tag{5.10}$$

The value c_- is the connection to $\mathcal{M}_{0,c_-}$, to which $\mathcal{M}_{1,j}$ belongs. The energy $E^{(j)}$ is the total energy of the four helices 7-6-3-1 in the given conformation.

In a similar way helix 2 is added. $\mathcal{M}_2$ is defined analogously:

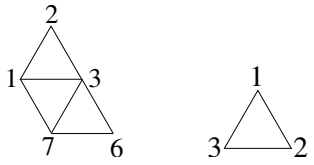


Figure 5.10: Grid after adding helix 1 (left), and the new helix triple in the orientation its map was calculated (right).

$$\mathcal{M}_{2,k} = (E^{(k)},\, c_-^{(k)},\, \Phi_2^{(k)},\, \psi_2^{(k)},\, z_{321}^{(k)}), \tag{5.11}$$

where c_- is the connection to $\mathcal{M}_{1,c_-}$. By starting from any line of $\mathcal{M}_2$, one will immediately find the according lines in $\mathcal{M}_1$ and $\mathcal{M}_0$. Storing the connectivities in this way has the advantage that it is conserved even when the order of matrix $\mathcal{M}_2$ is changed[2]. A ranking of the different construction paths is done by simply ordering $\mathcal{M}_2$ for ascending energies. Here the values of $\mathcal{M}_1$ have to twisted around 120°, to be in agreement with the values of *map321*. Then again the minima can be located in *map321* and then rotated back and stored in $\mathcal{M}_2$.

The next step, adding helix 3, reveals a new feature. $\mathcal{M}_3$ is not connected to $\mathcal{M}_2$, as it would be done in the previous manner, but to $\mathcal{M}_0$. This means there are

[2]One has to take care when to change the order in the matrix. When, for example, the order of $\mathcal{M}_1$ is changed, after $\mathcal{M}_2$ is already calculated the connectivity is lost. But in this frame this does not appear.

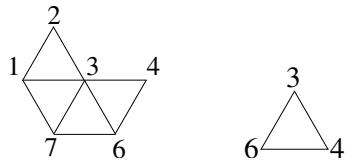


Figure 5.11: The 4^{th} step in the grid construction. Helix 4 is added.

two independent branches starting from $\mathcal{M}_0$. To construct the total grid efficiently, it has to be known which lines of $\mathcal{M}_1$ and $\mathcal{M}_3$ belong to the same minima of $\mathcal{M}_0$. Otherwise an expensive trial and error procedure would be needed. To be able to start with one member of $\mathcal{M}_2$ at one end of the branch finding the according values in $\mathcal{M}_1$ and $\mathcal{M}_0$ and then to go through all suitable conformations in $\mathcal{M}_3$ and $\mathcal{M}_4$ another way of connection is introduced. While the construction from $\mathcal{M}_2$ till $\mathcal{M}_0$ is performed in a backward manner, represented by the c_- notation, the construction from $\mathcal{M}_0$ over $\mathcal{M}_3$ till $\mathcal{M}_4$ is done in a forward manner - recalling the previously introduced c_+ in $\mathcal{M}_0$.

A major difference between c_+ and c_- is that c_- points to exactly one vector in the previous matrix, while c_+ defines a set of vectors in the following matrix, suiting to the condition of the given $\mathcal{M}_0, i$. For a given i in equation 5.9 the following range of vectors within $\mathcal{M}_3$ is obtained:

$$\mathcal{M}_{3,j(i)}, \qquad j(i) = \sum_{z=0}^{i-1} c_+^{(z)}, \ \ldots\ , \sum_{z=0}^{i-1} c_+^{(z)} + c_+^{(i)} - 1, \quad j \in \mathbb{N}. \tag{5.12}$$

The rest is analogous to the described steps. The recombined energies are not the total energies, but the sum of helix triplet energies. The total energy can be only calculated after the triplets are recombined. The same applies to the interaction with the environment. Due to these limitations the method of scanning helix-triplets and recombining them is just a preselection of an ensemble of conformations that have to be examined in more detail.

The minimizer. The results from the previous section are now the starting point for the off-grid minimum search including all interactions and tools on the residue-level. Determining the structure of a molecule is equivalent to minimizing its multi-dimensional energy function. There are many algorithms available to find the minimum of a multi-dimensional function. So far we implementated a minimizer from the GNU Scientific Library [44], based on a non-gradient based Simplex algorithm of Nelder and Mead [65]. After implementing the forces as well many other algorithms will become applicable.

The grand final will be atomistic MD simulations for the results of the minimizer. By storing averaged sidechain conformations as relative coordinates in

terms of the backbone atoms the side chains can easily be constructed for arbitrary backbone conformations.

Remarks. The data structure presented here is a simplified tree-structure. Especially for a more general application than needed at this state, it will be useful to provide a complete tree-class. That way the symmetry-break in the treatment of the two branches will be avoided.

Note that the ideas presented in this section could be, if sufficient computational capacities are available, applied to MD-simulations using the helices in atomistic detail instead of the residue model. Keeping the backbone fixed for the triple helix scan on the equidistant grid, would be a well-defined system that could be easily computed in a parallel manner.

5.3 Bacteriorhodopsin

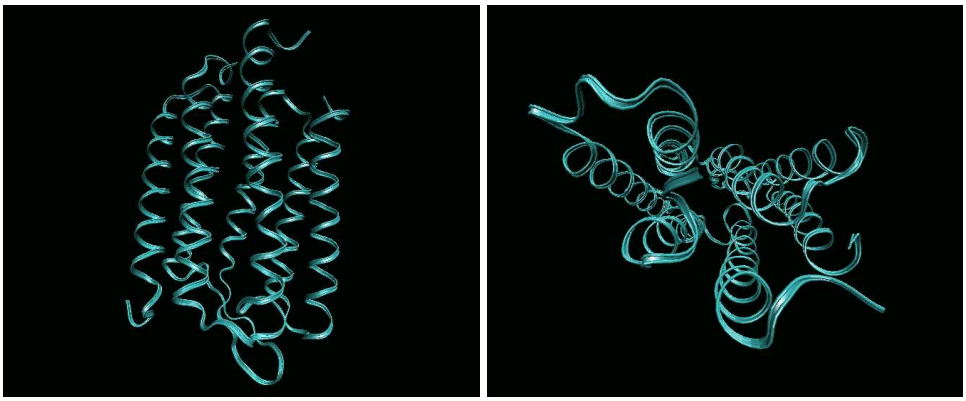


Figure 5.12: The native structure of bacteriorhodopsin (1C3W.pdb) from top (right image) and from the side (left image). Note that the apparent opening of the helices in the right image is caused by the perspective drawn by the VMD package.

Bacteriorhodopsin is a 7-helical transmembrane-protein. As its brother in law, rhodopsin, it contains a retinal as an active group, that is covalently bound to Lys 216. When retinal absorbs a photon it changes its conformation. This starts the process of proton-pumping, that also includes conformational changes of the helices [66]. Through the help of water molecules within the protein the proton moves along two possible pathways. The resulting proton gradient is then used for ATP-synthesis. Since the structure of bacteriorhodopsin is one of the best determined, we use it as a reference system for the methods that we present here.

First, we superimpose ideal helices onto the native structure of bacteriorhodopsin. Approximating the real structure with ideal helices is limited by the fact that the

native structure is not in the ideal conformation. The approximated helices as well as the native structure are shown in figure 5.13. The mean distance between the residues is 1.082$\AA$, as can be seen in table 5.1.

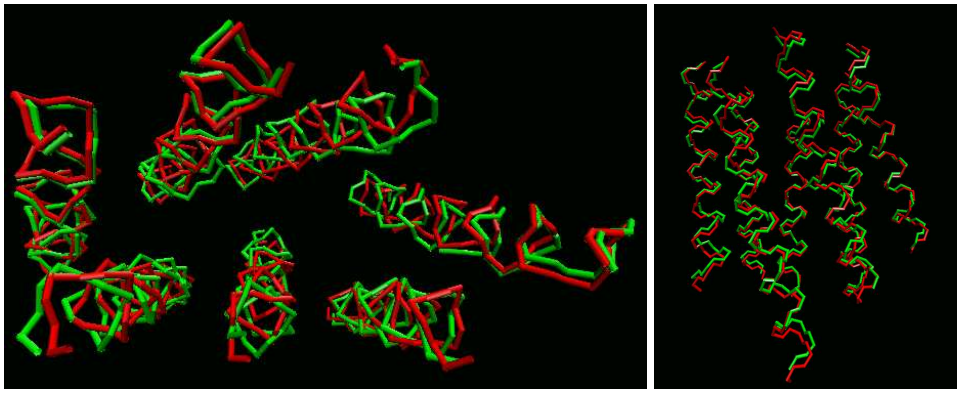


Figure 5.13: Coordinates of 1C3W.pdb (green) and the ideal helices approximating it (red).

Test on the minimizer. The first test was done using the unrefined residue-residue energy functions. To test the stability of the native conformation, the helices in the close-to-native conformation were passed to the minimizing algorithm. For the results see figure 5.14. While some helices are rather conserved, some had a distinct drift towards the center of the protein. That is reasonable, because the retinal molecule and the internal waters are missing in our model.

Performing the same investigations with the modified interactions led to different conformations. The left image of 5.14 has as a third plot the one for a cutoff of 15$\AA$. The right image shows the resulting conformation for three different manipulations of the energy functions. They all show similar results in that sense that the same helices experience a drift for different manipulations while the others are conserved. Unfortunately one could not conclude any reliable improvements of the results due to the manipulations yet.

Tests on the triple helix scans. As a final method to investigate the quality of the maps a similar approach as for glycophorin is used. All helices are put as close to the native conformation as possible and then two variables are varied, while the others are kept constant. This has to face several limitations in the case of bacteriorhodopsin, beside the unsufficient quality of the energy functions. While in the case of glycophorin there were 5 dimensions, there are 9 dimensions in the case of helix triplets on an equidistant grid. The cut-out is even less convincing

	native	$c = \infty, s = 0$	c = 12	c = 15	s = 0.25	s = 0.5	c = 12, s = 0.5
1C3W.pdb	1.082	1.295	1.288	1.229	1.264	1.287	1.284
native	-	0.505	0.542	0.486	0.470	0.541	0.613
c = ∞, s = 0	-	-	0.437	0.453	0.416	0.541	0.557
c = 12	-	-	-	0.570	0.565	0.654	0.629
c = 15	-	-	-	-	0.493	0.633	0.554
s = 0.25	-	-	-	-	-	0.466	0.456
s = 0.5	-	-	-	-	-	-	0.445

Table 5.1: The mean distances of the C_α's for different cutoffs c and shifts s using the minimizer.

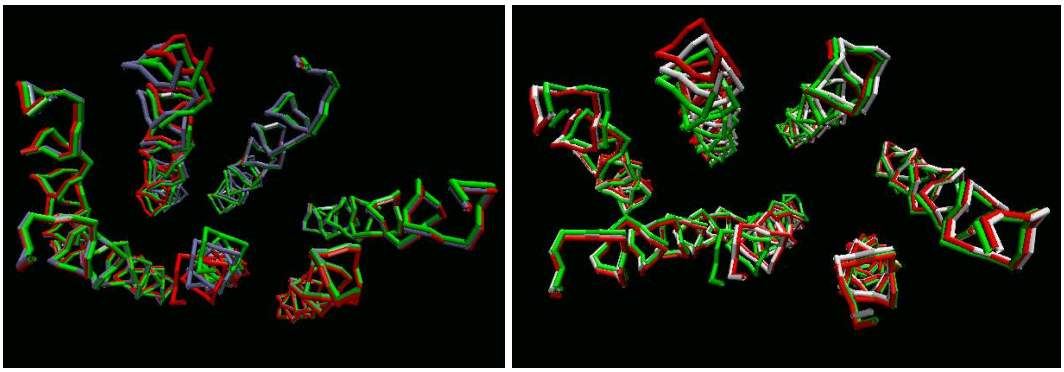


Figure 5.14: Bacteriorhopopsin after minimization. Left: as a reference the ideal helices in close-to-native conformation are plotted in red, the minimized conformation using energy functions with no cutoff and shift are plotted in green, and for a cutoff of 15$\AA$ in blue. Right: green: cutoff 12$\AA$, no shift; red: no cutoff, shift 0.5; grey: cutoff 12$\AA$, shift 0.5.

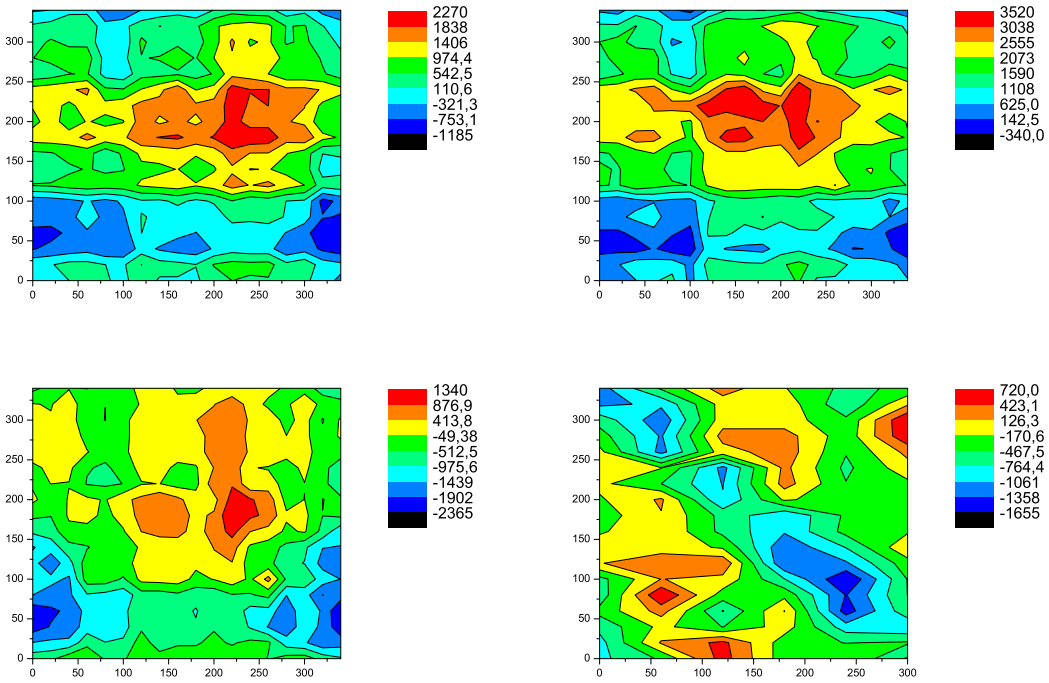


Figure 5.15: Some 2D extracts from the 9D triple maps of helices 763 of bR. In the first three images helices 7 and 6 are twisted around their axes, while the helices are as close to native conformation as the choice of Euler angles in the maps allow. The top-left is for the values closest to the native with unmanipulated energy functions. Top-right for a cutoff of 13$\AA$ and a shift of 0.5. Bottom-left shows a variation in the z-directions (see text). In the last image (bottom-right) helices 6 and 3 are fixed and helix 7 varies in ϕ and ψ with constant $\theta = 15^\circ$. The x-axis is assigned to ϕ with $\Delta\phi = 60^\circ$, the y-axis to ψ with $\Delta\psi = 20^\circ$ (bottom-right). Note the different scales!

than for 5 dimensions.
Then, the maps were calculated with rather rough divisions. Recalling that $\Delta\phi = 60°, \theta = 0°, 15°, \Delta\psi = 15°$ it becomes obvious that finding the native conformation within the maps can only be approximative.

Figure 5.15 shows some 2D energy landscapes for helices 7,6 and 3 of bacteriorhodopsin. There are two useful ways to investigate the maps. First, keeping one helix fixed and rotating two helices around their body-z-axes. This is performed in the two top and the bottom-right image. Second, keeping two helices fixed and let the third vary ϕ and ψ for $\theta = 15°$ - as can be seen for helix 7 in the bottom-right image.

The close-to-native conformation should be around $(200°, 60°)$ and $(200°, 80°)$ for the first three plots. In the top-right image helices 6 and 7 are twisted around their body z-axis using the unmanipulated energy functions. The close-to-native conformation is in the attracting area of this energy landscape but not in the minimum. Introducing cutoff and shift of 15$\mathring{A}$ and 0.5 respectively leads to hardly any change in the shape of the energy landscape, but to an offset. Please note that in native bR, individual helix triples don't need to adopt their relative minima. It is the minimum of the entire protein that matters.

The bottom-left image plots the energy landscape for following shifts in the z-direction: the cms of helix 6 is -5.5$\mathring{A}$ displaced relative to the cms of helix 7, and the cms of helix 3 has the same z-value as the one of helix 7. In all other images the cms of helix 6 is shifted 5.5$\mathring{A}$ relative to the one of helix 7 and the one of helix 6 2.5$\mathring{A}$. Eventhough the difference in displacement between the two plots is rather large, it also does not change the shape of the energy landscape. The offset here leads to an even larger attraction. Other variations of the z-displacements that are not shown here, revealed similar behaviour with different offsets. This might indicate that the z-displacement might not have to be regarded in the helix triple scan. This would mean a factor 19 in the size of the maps, which are now around 380 MB each and that took 7-10 days each on PentiumIII Xeon processors (700 MHz).

Chapter 6

Helix-dynamics on large timescales

6.1 The big plan

The function of proteins is mostly correlated with dynamic changes. These can consist of large scale conformational changes of domains or small scale local changes. On the other hand, not all TM-proteins function via conformational changes. Docking ligands to receptors mainly leads to an energetical change in the molecule that causes a reaction on the other side of the membrane and is not necessarily connected to conformational changes within the receptor. Also, passive transporters do not have to undergo certain changes to fulfil their function.

There are also continuous conformational changes that are not related to any function. Completely rigid and stiff molecules would be in contradiction to quantum mechanics. At any moment molecules oscillate around all bond lengths and angles. The oscillations differ largely on the timescale. Bond length oscillations have small amplitudes and high frequencies. The largest oscillations and lowest frequencies are, in principle, allowed for the dihedral angles between 4 atoms. Interjacent are the bond angles, in frequency as well as in amplitude.

One can split the motions in a protein into those of the sidechains and those of the backbone. In the backbone, the peptide bond has a rather rigid and planar structure which is caused by its partial double bond character of about 40%. The remaining dihedrals define the global structure of the protein.

When a secondary structure element like β-sheet or α-helix is formed, the dihedrals oscillate around the value of the ideal conformation of the helix or sheet.

Atomistic MD-simmulations consider all non-electronic degrees of freedom. The smallest timescale becomes the limiting factor to the capacities of MD. Even if one is interested in global changes only, one has to consider all vibration modes of the atoms, which differ in orders of magnitude on the timescale.

Nowadays, MD-simlutations are capable to reach timescales of 10-100 ns for

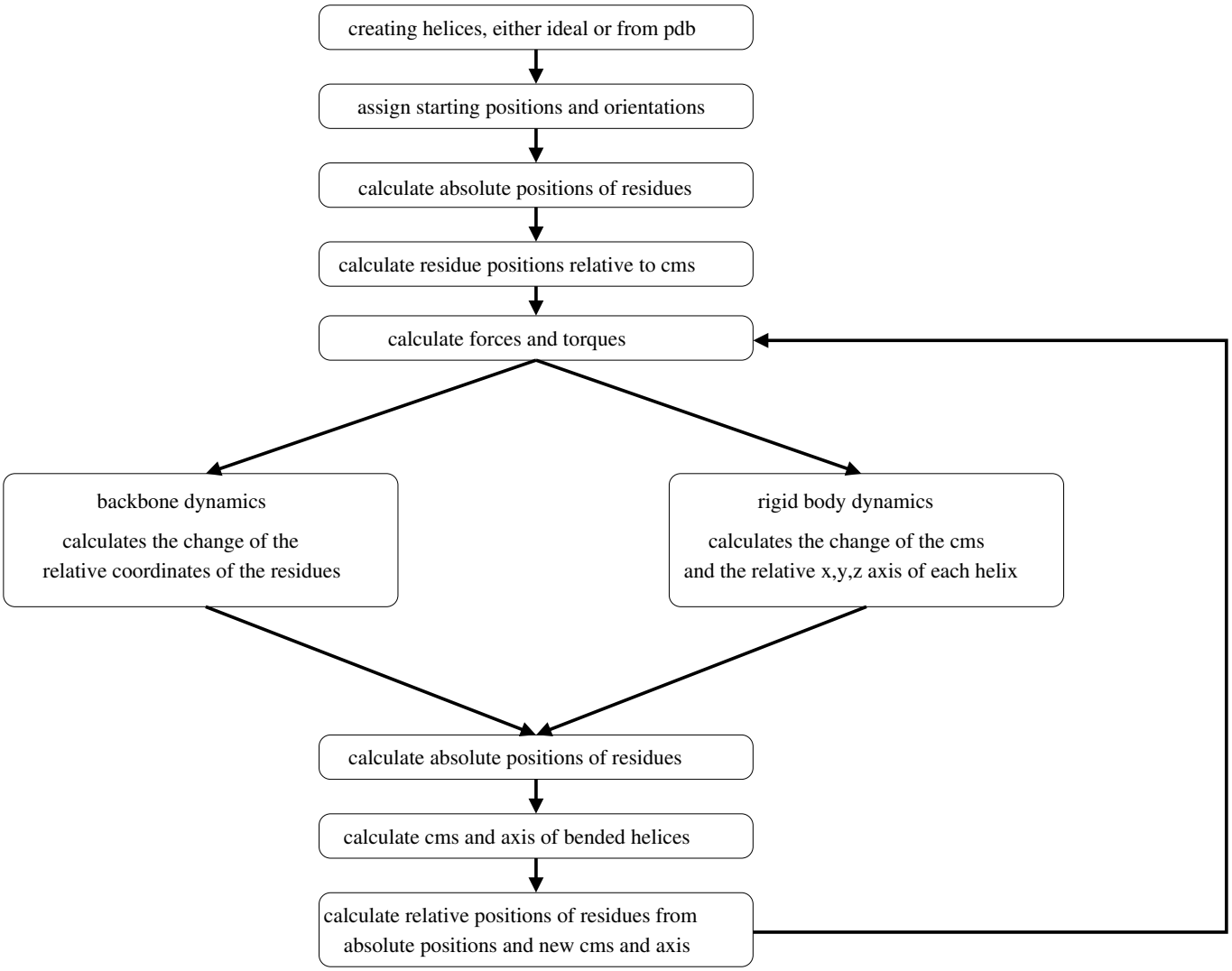


Figure 6.1: Flow-chart of the dynamics on residue level.

proteins that contain around 10 000 atoms, including solute and lipid atoms of their neighbourhood.

Using the residue model introduced in previous chapters, it is possible to develop dynamic tools that consider only the dihedrals as degrees of freedom. Cutting off the loops and considering only helices for the dynamics is another major reduction of the degrees of freedom.

This step is justified by the experimentally shown fact that cutting single loops of TM-proteins conserves both structure and function, while cutting off all does not [67]. That single loops may be removed without disturbing the function indicates that their individuality plays a crucial role. On the other hand, they do have an effect. Otherwise, all loops could be chopped off and structure and function would be conserved. The influence should be sufficiently described by a distance constraint or a simple potential term for the endpoints of the helices.

After cutting off the loops, new variables are needed to describe the dynamical behaviour of the relative positions and orientations of the helices. This is the point where it is useful to split the internal dynamics of the helices from their relative movements and to introduce a two step model, that is sketched in figure 6.1.

The internal bending and twisting is calculated in terms of the backbone dihedrals. However, the helices are considered as being rigid bodies when calculating the relative motion. The rigid body dynamics that determines the change of cms and body-axes, is a well-known physical-mathematical method described briefly in section 6.4. The backbone dynamics requires a much larger mathematical effort. The forces that act on the residues and cause the dynamical changes must be converted into the torques acting on the dihedrals. (See section 6.3.)

The procedure Initially we need a starting conformation. This can be provided by either a pdb-file for any kind of helix or a set of conformation-specifying values like position and orientation for ideal helices. Additionally, velocities and angular velocities for the starting time t_0 must be chosen. Knowing the positions of all residues allows one to quantify the over-all forces that are acting on each residue. From this one can calculate the total forces and torques acting on the helices. If not given from the beginning, like in the case of using a pdb-file as donator of an inital conformation, the dihedrals must be determined. The final things needed are the relative positions of the residues to the cms. Now the actual calculation of the dynamics can be executed.

The rigid body dynamics fixes the location of the cms and the orientation of the helices after the set timestep. The backbone dynamics alters the relative positions of the residues that were not affected by the rigid body dynamics. Recombining them contains another little challenge. The backbone dynamics starts from a set of dihedrals and also leads to a set of dihedrals. What is needed for the recombination are the new relative positions. From the new cms, axes and relative residue-positions the new absolute positions are easily obtained. At this point all

information needed for the next pass is present that starts with calculating the overall forces acting on the residues.

The interplay of the two types of dynamics demands excessive use of coordinate transformations.

Subject of discussion is still the point if it would be better to perform the two dynamics simultaneously or successively. Testing this will be necessarily performed, by varying the timesteps.

6.2 Forces caused by the environment

The forces between the residues can be computed either directly from the MD-simulations or as derivatives of the energy functions. The influence of the surroundings on the dynamics are more difficult to obtain. We start from:

$$\vec{F} = -\vec{\nabla}U, \tag{6.1}$$

We consider the potential energy of a residue as proportional to its free surface. Therefore the force is proportional to the change of the free surface. The change of the surface is not measurable analytically, due to the complexity of the overlaps.

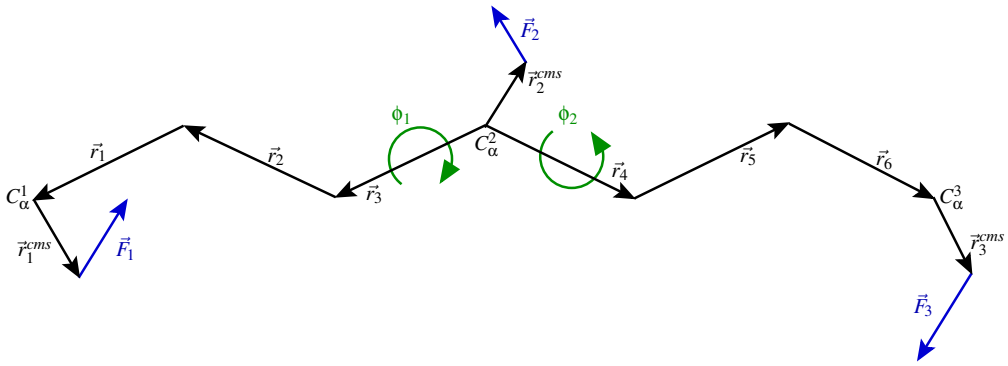
However, one can calculate the free surfaces of the residues numerically using the sphere-algorithm. The environmental forces can be approximated most simply by the following gradient:

$$\vec{F}_i^{env} = -\vec{\nabla}U_i^{env} = -\begin{pmatrix} \frac{U_i^{env}(t_0+\Delta t)-U_i^{env}(t_0)}{x_i(t_0+\Delta t)-x_i(t_0)} \\ \frac{U_i^{env}(t_0+\Delta t)-U_i^{env}(t_0)}{y_i(t_0+\Delta t)-y_i(t_0)} \\ \frac{U_i^{env}(t_0+\Delta t)-U_i^{env}(t_0)}{z_i(t_0+\Delta t)-z_i(t_0)} \end{pmatrix}. \tag{6.2}$$

Note that the forces $\vec{F}_i^{env}$ are not the real forces acting on the residues, but are rather the contribution of the total force which is parallel or antiparallel to the displacement vector $\vec{d}_i = \vec{r}_i(t_0+\Delta t) - \vec{r}_i(t_0)$, where the $\vec{r}_i$ are the locations of the residues. One could use equation 6.2 to obtain the total force for the protein at time t_0 by performing three linearly independent translation steps. A simpler estimation of the residue-environment force can be obtained by regarding the residue-residue contribution as the driving force, and the residue-environment part as reacting to it. Following for a timestep the dynamics caused by residue-residue force only allows us to calculate equation 6.2. This leads to an estimate of how much the environment supports or opposes the change caused by the residue-residue interactions. With this information the dynamics can be recalculated, starting again at t_0, but including the environment-residue approximation. This procedure requires only one step instead of three for the total force and is well justified.

To obtain the total force without performing three intermediate steps for each timestep, one could use the knowledge from previous timesteps $t_0 - i\Delta t$. This is also justified, because the change in the environmental forces can be expected to be less sensitive to small movements than the residue-residue forces.

6.3 Backbone dynamics



Because the global structures of proteins are determined mainly by the backbone torsional angles, they are the coordinates in which the internal dynamics of the helices must be calculated. A way to describe how the forces that are acting on each residue are affecting the torsional angles needs to determined. An approximation must be made because it is difficult to analytically calculate the effect of n (number of residues) forces on $2 \cdot n - 4$ angles instantaneously. Therefore a set of $4 \cdot n - 8$ uncoupled ordinary differential equations of first order will be used.

To find the differential equations that describe the change of the rotational angles, one can start with the general equation for a rotating solid body:

$$\frac{d\vec{L}}{dt} = \vec{\tau}$$

where $\vec{L}$ is the angular momentum and $\vec{\tau} = \sum_i \vec{r_i} \times \vec{F_i}$ the torque.
This is equal to:

$$\frac{d\vec{L}}{dt} = \frac{d\mathbf{I}\vec{\omega}}{dt} = \vec{\tau}, \tag{6.3}$$

with $\mathbf{I}$ being the tensor of inertia and $\vec{\omega}$ the angular velocity.
The inertia tensor is defined as:

$$\mathbf{I} = \sum_{i=1}^{n} m_i \begin{pmatrix} y_i^2 + z_i^2 & -x_i y_i & -x_i z_i \\ -y_i x_i & x_i^2 + z_i^2 & -y_i z_i \\ -z_i x_i & -z_i y_i & x_i^2 + y_i^2 \end{pmatrix}. \tag{6.4}$$

From this, the moment of inertia I is calculated by:

$$I = \vec{e}_\omega \cdot \mathbf{I} \cdot \vec{e}_\omega = \sum_i m_i (r_i^2 - (\vec{r}_i \cdot \vec{e}_\omega)^2), \tag{6.5}$$

with $\vec{r}_i = (x_i, y_i, z_i)$. The rotation axis is constant and can be chosen as z.
This leads to:

$$I = \vec{e}_z \cdot \mathbf{I} \cdot \vec{e}_z = \vec{e}_z \cdot \sum_i m_i \begin{pmatrix} -x_i z_i \\ -y_i z_i \\ x_i^2 + y_i^2 \end{pmatrix} = \sum_i m_i (x_i^2 + y_i^2). \tag{6.6}$$

Taking into account only the next-neighbour sidechains and using the center of mass of the neighbouring sidechains, we will get two terms I_1, I_2. One has to be aware of which atoms to include into the cms, since it is not the cms of the whole residue that are used in other chapters. With $\Phi = \Phi' - \Phi_0$ being the deviation from the ideal helix value and $\tau_{bond} = \kappa\Phi$ we get:

$$(I_1 + I_2)\frac{d^2\Phi_i}{dt^2} = -\kappa\Phi_i + \tau_{ext}, \quad \tau_{ext} = \tau(\vec{F}_j) + \tau(\vec{F}_{j+1}). \tag{6.7}$$

The $\overrightarrow{F_j} = \sum \overrightarrow{F_{jk}}$ are the total interhelical forces exerted by residues k acting on residue j. There are two ways of defining κ. Either by an arbitrary value or by perfoming MD-simulations with a couple of residues, starting in ideal helical conformation in vacuum and measuring frequencies and amplitudes of the rotation angles and through that calculating κ. One could also apply extension-forces or compression-forces on the helices during the MD-simulation. Actually, κ should be a nonlinear function of Φ: $\kappa = \kappa(\Phi)$. A set of $4 \cdot n - 8$ uncoupled differential equations, given by equation 6.7 will need to be solved. To do so, the way the forces act on the rotational degree of freedom, and what torque they cause then must be calculated. In the remaining part of this section we will determine

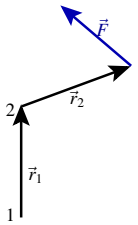
$$\ddot{\Phi}_i = \ddot{\Phi}_i(\vec{F}_j, \vec{F}_{j+1}), \tag{6.8}$$

the functional relation between the forces and their effect on the diheadrals. The general expression for a torque that acts on a point 0, caused by the force $\overrightarrow{F}$ via a lever arm $\overrightarrow{r}$ is:

$$\overrightarrow{\tau} = \overrightarrow{r} \times \overrightarrow{F}$$

But before the actual needed torque can be calculated, three effects must be considered:

I. If the direction of the torque vector is not the same as the one of the rotational axis, the force that acts on the axis will have two components: one is the torque that makes the axis twist, and one is a lever force that would cause the axis to change its orientation if it would not be fixed, as in our case.



With definitions as in the picture above, $\vec{F}$ can be expressed in terms of the vectors $\vec{e}_{r_1}, \vec{e}_{r_2}$ and $\vec{e}_{r_{12}} = \frac{\vec{r}_1 \times \vec{r}_2}{|\vec{r}_1 \times \vec{r}_2|}$:

$$\vec{F} = F_{12}\vec{e}_{r_{12}} + F_1\vec{e}_{r_1} + F_2\vec{e}_{r_2} \equiv \vec{F}_{12} + \vec{F}_1 + \vec{F}_2. \tag{6.9}$$

This is a general expression for the decomposition of a vector into 3 others. How to express a vector in terms of others is explained in appendix A.1.1. The torque that acts on point 2 is:

$$\vec{\tau}_2 = \vec{r}_2 \times \vec{F}_{12}. \tag{6.10}$$

The other part of $\vec{F}$ causes a torque on point 1:

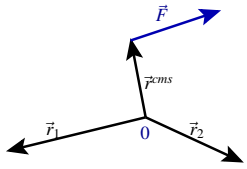
$$\vec{\tau}_1 = (\vec{r}_1 + \vec{r}_2) \times (\vec{F}_1 + \vec{F}_2) \tag{6.11}$$

This expression holds as long as the angle between $\vec{r}_1$ and $\vec{r}_2$ is fixed.

II. Another case to be considered is if the torque-vector is not pointing in the same direction as the vector representing the rotation axis. The magnitude of the torque τ_a acting on an axis $\vec{a}$ is equal to the projection of the torque vector $\vec{\tau}$ on the axis:

$$\tau_a = \tau \cos \alpha_{\vec{a},\vec{\tau}} = \vec{\tau} \cdot \vec{e}_a. \tag{6.12}$$

III. The third effect is that the force that acts on the cms of the sidechains acts on two rotational axis, e.g. the effect of the force is split onto the two degrees of freedom.



The torque acting on point 0 is:

$$\begin{aligned}\vec{\tau} &= \vec{r}^{cms} \times \vec{F} \\ \vec{\tau} &= \tau_{r_1}\vec{e}_{r_1} + \tau_{r_2}\vec{e}_{r_2} + \tau_{r_{12}}\vec{e}_{r_{12}}. \end{aligned} \tag{6.13}$$

In the same manner as in (6.9) one can calculate the factors τ_{r_1} and τ_{r_2}, which define the proportion of the distribution of the force.

Now all requirements are fulfilled to describe the backbone dynamics.
As an example $\ddot{\Phi}_1$ is calculated:

$$(I_1 + I_2)\ddot{\Phi}_1 = \kappa\Phi_1 + \tau_{ext}, \quad \tau_{ext} = \tau_1(\vec{F}_1) + \tau_2(\vec{F}_2). \tag{6.14}$$

The calculation of τ_2 is straightforward. First the torque acting on C_α^2 has to be calculated. By splitting the torque vector into the two rotation axis vectors and the vector perpendicular to them, the searched value is found by taking the projection on the $\vec{r}_3$.

$$\begin{aligned}\vec{\tau}_2' &= \vec{r}_2^{cms} \times \vec{F}_2 \\ \vec{\tau}_2' &= \tau_3'\vec{e}_{r_3} + \tau_4'\vec{e}_{r_4} + \tau_{34}'\vec{e}_{r_{34}}, \quad \vec{e}_{r_{34}} = \frac{\vec{r}_3 \times \vec{r}_4}{|\vec{r}_3 \times \vec{r}_4|} \\ \tau_2 &= \vec{\tau}_3'. \end{aligned} \tag{6.15}$$

Calculating τ_1 would be straightforward as well, if $\vec{F}_1$ wouldn't split on two axis as well, but it does. That means an addition step must be made. What is left to be determined is the fraction of the force which acts on $\vec{r}_1$ not as a torque, but rather as a lever.

$$\begin{aligned}\vec{\tau}_1' &= \vec{r}_1^{cms} \times \vec{F}_1 \\ \vec{\tau}_1' &= \tau_0'\vec{e}_{r_0} + \tau_1'\vec{e}_{r_1} + \tau_{01}'\vec{e}_{r_{01}}, \quad \vec{e}_{r_{01}} = \frac{\vec{r}_1 \times \vec{r}_1}{|\vec{r}_0 \times \vec{r}_1|}\,. \end{aligned} \tag{6.16}$$

τ_0' and τ_1' are causing torque, τ_{01}' is the lever part. τ_{01}' is acting on both axis, but it is perpendicular to both, so it is equaly distributed on them. That means $\frac{\tau_{01}'}{2}$ is the part that is needed. But what force is causing this torque? This force will be called $\tilde{\vec{F}}_1$:

$$\begin{aligned}\frac{\tau_{01}'}{2}\vec{e}_{r_{01}} &= \vec{r}_1^{cms} \times \tilde{\vec{F}}_1, \\ &= \left|\vec{r}_1^{cms} \times \tilde{\vec{F}}_1\right| \vec{e}_{r_{01}}, \\ &= r_1^{cms}\tilde{F}_1 \sin\alpha\, \vec{e}_{r_{01}} \end{aligned} \tag{6.17}$$

We can construct $\tilde{\vec{F}}_1$ the following way:

$$\begin{aligned}
\tilde{\vec{F}}_1 &= \tilde{F}_1 \vec{e}_{\tilde{F}_1}, \\
\vec{e}_{\tilde{F}_1} &= \frac{\vec{e}_{r_{01}} \times \vec{e}_{\left(r_1^{cms}\right)}}{\left|\vec{e}_{r_{01}} \times \vec{e}_{\left(r_1^{cms}\right)}\right|}, \\
\tilde{F}_1 &= \frac{\tau'_{01}}{2} \frac{1}{\left|\vec{r}_1^{cms} \times \vec{e}_{\tilde{F}_1}\right|} = \frac{\tau'_{01}}{2 r_1^{cms} \sin\alpha}.
\end{aligned} \tag{6.18}$$

That leads to:

$$\tilde{\vec{F}}_1 = \frac{\tau'_{01}}{2} \frac{\vec{e}_{r_{01}} \times \vec{e}_{\left(r_1^{cms}\right)}}{\left|\vec{r}_1^{cms} \times \vec{e}_{\tilde{F}_1}\right|} \tag{6.19}$$

And finally the torque τ_1 that it is acting on $\vec{r}_3$:

$$\vec{\tau}_1 = (\vec{r}_1^{cms} + \vec{r}_1 + \vec{r}_2) \times \tilde{\vec{F}}_1 \tag{6.20}$$

that causes:

$$\tau_1 = \vec{\tau}_1 \cdot \vec{e}_{r_3} \tag{6.21}$$

the torque we were looking for.

6.4 Rigid body dynamics

This part is common knowledge and is thus explained only briefly. The state of a rigid body can be described by a state vector $\vec{Y}(t)$:

$$\vec{Y}(t) = \begin{pmatrix} \vec{X}(t) \\ \mathbf{R}(t) \\ \vec{P}(t) \\ \vec{L}(t) \end{pmatrix}, \tag{6.22}$$

where $\vec{X}$ is the position of the rigid body, the rotation matrix $\mathbf{R}(t)$ the orientation, $\vec{P}(t)$ the linear, and $\vec{L}(t)$ the angular momentum of the rigid body.

The equations of motion of the rigid body are given by the derivative $\frac{d}{dt}\vec{Y}(t)$:

$$\frac{d}{dt}\vec{Y}(t) = \frac{d}{dt}\begin{pmatrix} \vec{X}(t) \\ \mathbf{R}(t) \\ \vec{P}(t) \\ \vec{L}(t) \end{pmatrix} = \begin{pmatrix} \vec{V}(t) \\ \vec{\omega}(t) * \mathbf{R}(t) \\ \vec{F}(t) \\ \vec{\tau}(t) \end{pmatrix}. \tag{6.23}$$

with $\vec{\tau}(t)$ the torque, $\vec{F}(t)$ the force that acts on the rigid body, $\vec{V}(t)$ its velocity and $\vec{\omega}(t)$ its angular velocity. The '*'-notation has the following meaning:

$$\vec{\omega}(t) * \mathbf{R}(t) = \left(\vec{\omega}(t) \times \begin{pmatrix} r_{xx} \\ r_{xy} \\ r_{xz} \end{pmatrix} \quad \vec{\omega}(t) \times \begin{pmatrix} r_{yx} \\ r_{yy} \\ r_{yz} \end{pmatrix} \quad \vec{\omega}(t) \times \begin{pmatrix} r_{zx} \\ r_{zy} \\ r_{zz} \end{pmatrix} \right).$$

The angular velocity $\vec{\omega}(t)$ is connected to the inertia tensor $\mathbf{I}$ by:

$$\vec{\omega}(t) = \mathbf{I}(t)\vec{L}(t). \tag{6.24}$$

The intertia tensor $\mathbf{I}(t)$ can be expressed in terms of the body-space intertia tensor $\mathbf{I}_{body}$:

$$\mathbf{I}(t) = \mathbf{R}(t)\, \mathbf{I}_{body}\, \mathbf{R}^T(t). \tag{6.25}$$

The body-space inertia tensor $\mathbf{I}_{body}$ is a constant and can be calculated before the simulation.

Using these equations has the inherent problem of numerical drift that is caused by the usage of 3×3 matrix to describe the orientation of the rigid body, which is a 3 dimensional property. Several alternatives are discussed in the literature [68, 69, 70]. One in particular is the usage of quaternions instead of rotation matrices. (See section A.2 on page 99 for a detailed definition of quaternions.) They are

a 4 dimensional generalisation of the complex numbers, introduced by Hamilton. Still they have one more dimension than the rotation matrices. However, their drift can be easily corrected at any time by a simple normalisation. This makes them a useful tool in the frame of rigid body dynamics. In equation 6.23

$$\dot{\mathbf{R}}(t) = \vec{\omega}(t) * \mathbf{R}$$

becomes

$$\dot{\mathbf{q}}(t) = \frac{1}{2}\vec{\omega}(t)\mathbf{q}(t).$$

Chapter 7

Outlook and suggestions

7.1 Aiming at the most general residue-level force field

We started designing the residue force field, aiming at parallel helices. The choice of the variables in which the calculations were performed were in accordance with the peptides being in parallel helical conformations. To apply the force field in arbitrary orientations, may the residues be in helices, β-sheets, uncoiled loops or termini, other variables are more suitable. As mentioned before, the most common variables to describe the spatial orientation of bodies are the Euler angles. A useful choice for defining the coordinate system of the residues are the backbone atoms.

This choice leads to fit functions that depend on 7 variables:

$$f(\phi_1, \theta_1, \psi_1, \phi_2, \theta_2, \psi_2, d) : \mathbb{R}^7 \rightarrow \mathbb{R}.$$

It is left open in this relation which distance exactly is meant by d. The positions of the residues could be defined by the positions of the C_α's, the C_β's or the centers of mass of the residues. The C_α's are a well defined and the simplest choice. The C_β's and the *cms* are less well defined in the sense that they are not as fixed as the C_α's, but will oscillate and can be used only as mean values. Nevertheless, taking the *cms* as reference point is a promising choice - in a residue model it should be the point of highest symmetry, interactionwise. While attempting, one of them should turn out to become the best choice concerning the fitting, e.g. leading to the simplest fitting functions that are able to identify the native structure. Data collection in terms of Euler angles would require much more computational time and storage. To reduce the required data to a reasonable amount and at the same time to make the fitting easier and more stable, it is useful to not only store the energy values in the data files, but also their derivatives, the forces. Fitting energy and force simultaneously makes it much easier to find a stable automated fitting algorithm, even for 7 dimensions. Especially for the forces this will lead to much better fits, instead of constructing them by taking the derivatives of the energy functions. Also, the minimizer algorithms that are needed for the structure prediction will be more

efficient, working simoultaneously with energy and forces. In order to allow the most flexible way of treating and investigating the data, the data files should then contain the most general and complete information: the positions of the C_α's, C_β's and the *cms*, the Euler angles, the dipole vectors, the different energy terms, and the forces.

Another way to fit the data in such a complex way is the usage of neural networks. Splitting the huge amount of data and passing it to different networks will stabilize the learning of the networks.

If one has the most general data set available, one can try different fitting approaches that might simplify the fitting. A set of three angles together with the distance might define each residue sufficiently. These are the angle between the dipoles of the residues and the angles between the dipoles and the connecting vector, given here in terms of their scalar products:

$$\vec{p}_1 \cdot \vec{p}_2, \quad \vec{p}_1 \cdot \vec{e}_r, \quad \vec{p}_2 \cdot \vec{e}_r$$

These are familiar from the formula for the interaction energy of two dipoles.

As stressed in the discussion of the results for glycophorin A and bacteriorhodopsin, the calculations done so far need some refinements. Yungki Park [43] was able to identify the native structure of glycophorin A by using the center points to locate the residues, instead of the C_α positions - the variable in which the data was calculated. This might be surprising at first hand, but indicates that in the way the data was calculated some modifications are necessary. As discussed in section 4, the distance shift may account for two contributions. The first is that the sidechains have more freedom to bend during pairwise simulations than they will have being integrated in a helix. The second, a missing damping term of the residue pairs seperated by other molecules like lipids, waters or other residues.

The data for the energy functions was calculated using tripeptides, where the two outer residues were dummies. During the simulation of the two amino acids they can bend within a certain range. When part of different helices, they are surrounded by other residues that, especially when tightly packed, will decrease the accessible conformational space, which should lead to different energy values and a different entropy. One way to estimate this effect is to simulate entire helices in MD and compare the results to those one obtains by using the energy functions. This is also necessary to scale the overlap mode of the sphere algorithm. The statistical refinement based on the entire helix simulations can be combined with knowledge based methods, exploiting solved structures.

By simulating two or more entire helices in vacuum, the additivity of the energy functions and the damping that is caused by intermediate residues should also be checked. So far, only the vacuum energies were calculated, but, when the residues are separated by other residues or lipids or solution, polarisation will lead to a decrease of the residue-residue interaction that is not included yet. This should lead to a distinct effect, especially for the interaction of charged residues that now have a strong contribution even when quite separated.

The residue-residue energy functions will have two regions, for small distances the undamped vacuum functions and after a certain limit the damping term included. If a separation of the residues by media is possible, an extension of the sphere algorithm can be introduced, which measures if the connecting vector of the residues crosses free surface. This is the case for systems like glycophorin A or when the energy functions are used for docking proteins. By simulating entire peptides or helices with varying distances within the membrane, an idea about the damping terms in the three phases can be developed.

To further increase the quality of the fit functions, more expensive simulations on a selection of conformations could be used to see how much the results are affected. If the deviations are small, it is a justification of the previously calculated values. In case of larger deviations, one can hope that the systematics behind the deviation will be understandable and that they can be included into the fit functions.

7.2 What's left to do

- the sphere algorithm can be easily improved by implementing code that either has a faster and more precise way to integrate over the surface of the spheres or is based on a library (the latter is in progress).
- the absolute scale for both applications of the sphere algorithm has to be determined by simulating entire helices. Also, the choice of the radii should be verified or improved.
- so far, the minimizer was based on the calculation of energies only. As a further information the forces could be used. This should make the algorithm both faster and more reliable.
- different types of available minimizers should be tested.
- implementation of the virtual charges that are superimposing the raise in the potential due to the burial of the total backbone dipole and from burying loop-residues next to the helix into the membrane. Both will contribute as smooth 1D step functions, as in figure 1.4 on page 10, the latter as a multistep function.
- a pressure term should be introduced that is dependent on the total free surface of the protein.

Chapter 8

Conclusions

In this thesis, a set of novel approaches towards the goal of predicting the structures of transmembrane proteins was developed and implemented in an object-oriented program package (30 000 lines of code). It contains mainly four parts, namely the residue potential, the sphere algorithm, search algorithms and the helix dynamics.

Despite its conceptual simplicity, the sphere algorithm in its two applications which determine free surfaces and higher order overlap, turned out to be a very useful tool in the frame of molecular modelling. The free surface was used to measure the exposure of a residue to the surrounding, the overlap as an indicator for the sidechain entropy and overcrowding effects. For some systems, a direct comparison with experiments or other theoretical calculations was possible. Insights were gained about the insertion behaviour of single-helix systems like melittin and many others. Also, some detailed maps were plotted for tight packing. Still lacking a proper scaling, the results shown here are of qualitatively nature but are nonetheless very promising. Scaling would be best done by simulating entire helices.

The residue-residue energy functions were applied on a simple and well-defined test system, glycophorin A. At first hand, the native structure of glycophorin A could not be identified as the absolute minimum of the energy landscape. However, by introducing a distance shift and cutoff, the quality of the prediction was increased significantly. Effects of the pair-nature of the energy functions and a missing damping term seem to justify this. To prove this in detail, whole-helix MD-simulations would be required. The quality of a residue model is limited due to the deviation from pairwise energy calculations that depend on the specificities of all the neighbours in whole helices. These deviations cannot be integrated sensibly into a residue model, however, using it as a pre-filter for atomistic MD is highly promising.

Applying the energy functions on helix triplets from bacteriorhodopsin located on the grid revealed that while the conformation closest to the native one is in a highly attracting orientation, it is not in the orientation of minimal energy. This indicates that the simple modifications that were sufficient in the case of glycophorin

A are insufficient for bacteriorhodopsin. The scans with different distances, shifts and cutoffs were then performed. One cannot expect to identify the native conformation in the very large conformational space of bacteriorhodopsin before a whole-helix analysis is performed. The quality of the energy functions will increase by understanding the deviations from the pairwise calculation of the energy functions and introducing more realistic damping terms. It then becomes more likely to identify the native conformation for bacteriorhodopsin as well.

The search methods described here do not depend purely on the residue model, but can be applied to atomistic search as well. These methods are based on the idea to perform the first scan of the energy landscape from a bird's-eye view on a certain height by rotating the helices at a fixed distance. Applying the same distance to all helices leads to a grid arrangement. In order to recombine the helices the scans were performed as triple scans. The application of minimizing algorithms off-grid was then performed.

The dynamic tools were presented here only conceptually, without any applications. At this stage of development of the energy functions it would be premature to apply them for dynamic calculations. Deriving the forces from the energy functions will conserve the error that the energy fits contain. Obtaining them directly from MD would, in contrast, reduce the error of both energy and force fit through simultaneous fitting. Manipulating NWchem in this fashion was hindered due to major technical problems so far.

After the presented refinements are performed the methods, developed here, can be expected to be a powerful tool in the field of molecular modelling.

Appendix A

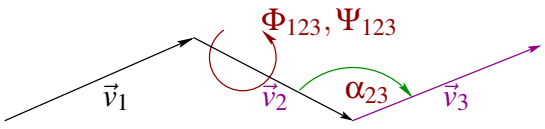
Methods and tools

A.1 Matrix based methods

A.1.1 Vector transformation

Two kinds of transformations are heavily exploitet in this package.
The change of the coordinate system, that is not only a rotation can be calculated via vector decomposition (refdecomposition). An example are the coordinates of sidechain atoms, which e.g. are known in the absolute, orthonormal space and shall be calculated in terms of the backbone atoms. This has the advantage that the coordinates of the sidechain atoms are constant [1] relative to the backbone atoms. Once the relative positions are known, it is very easy and fast to calculate their absolute position after the helices have moved.

The second transformation is the rotation of a vector in a given coordinate system or the twist of the coordinate system using the Euler transformation. One important case of this is the calculation of the backbone atom positions from a given set of diheadrals. The structure of a peptide is determined by the diheadrals. The backbone can be seen as a chain of vectors, connecting the backbone atoms. Knowing two vectors v_1, v_2, the length l_3 of the third v_3, the angle α_{23} between the second and the third and the diheadral Ψ_{123} or Φ_{123} allows the calculation of the third, via a double Euler transformation.



The first step is the rotation of a vector $\vec{x}_3 = l_3\vec{e}_x,\ \ \vec{e}_x = (1,0,0)$ around the

[1] In this framework we consider the sidechains as stiff.

angles α and Ψ or Φ.

$$\mathbf{A}(0,\alpha,\Psi/\Phi)\cdot\vec{x}_3 \equiv \vec{t}_3.$$

This puts the vector $\vec{t}_3$ in the right orientation, but relative to $\vec{x}_3$ not to $\vec{v}_2$. To get the real orientation a second Euler transformation has to be performed:

$$\mathbf{A}(\phi,\theta,\psi)\cdot\vec{t}_3 = \vec{v}_3,$$

where the angles ϕ,θ,ψ are defined as those that fulfil

$$\mathbf{A}(\phi,\theta,\psi)\cdot\vec{x}_2 = \vec{v}_2$$

In this manner the whole backbone can be constructed recursively.

A.1.2 Vector decomposition

Changing a coordinate system means to decompose a vector v in terms of its new coordinate axis $\vec{\chi},\vec{\eta},\vec{\xi}$:

$$\vec{v} = a\vec{\chi} + b\vec{\eta} + c\vec{\xi}.$$

so v becomes:

$$\vec{v} = \begin{pmatrix} v_x \\ v_y \\ v_z \end{pmatrix} \longrightarrow \vec{v}' = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

Determining the a, b, c is equivalent to solving:

$$\mathbf{A}\cdot\vec{v}' = \vec{v}$$

where

$$\mathbf{A} = \begin{pmatrix} \chi_x & \eta_x & \xi_x \\ \chi_y & \eta_y & \xi_y \\ \chi_z & \eta_z & \xi_z \end{pmatrix}$$

We use the LU-decomposition described below to solve this.

A.1.3 Inverse matrices

A quadratic $n \times n$ matrix $\mathbf{A}$ is invertible if there is a matrix $\mathbf{A}^{-1}$ that obeys:

$$\mathbf{A}\cdot\mathbf{A}^{-1} = \mathbf{A}^{-1}\cdot\mathbf{A} = \mathbf{1}. \tag{A.1}$$

To obtain the inverse matrix $\mathbf{A}^{-1}$ we solve the equation:

$$\mathbf{A}\cdot\vec{x}_j = \vec{e}_j, \qquad j = 1,\ldots,n \tag{A.2}$$

where the $\vec{e}_j$ are n unit vectors. The corresponding $\vec{x}$'s are the columns of the inverse matrix $\mathbf{A}^{-1}$. The method we use to calculate the $\vec{x}$'s is again the LU decomposition, described in the following section.

A.1.4 LU-decomposition

This section follows the describtion in [71].

A quadratic $n \times n$ matrix $\mathbf{A}$ can be written as the product:

$$\mathbf{A} = \mathbf{L} \cdot \mathbf{U} \tag{A.3}$$

of the upper triangular $\mathbf{U}$ and the lower triangular matrix $\mathbf{L}$. With this decomposition we can solve the linear set:

$$\mathbf{A} \cdot \vec{x} = (\mathbf{L} \cdot \mathbf{U}) \cdot \vec{x} = \mathbf{L} \cdot (\mathbf{U} \cdot \vec{x}) = \vec{e} \tag{A.4}$$

by first solving for the vector $\vec{y}$ such that

$$\mathbf{L} \cdot \vec{y} = \vec{e} \tag{A.5}$$

and then solving

$$\mathbf{U} \cdot \vec{x} = \vec{y} \tag{A.6}$$

Solving for the components of $\vec{x}$ and $\vec{y}$ is straightforward. These triangular sets of equations are rather easier to solve.

We use an implementation of the GNU Scientific Library to perform the LU decomposition [44]

A.2 Quaternions

Quaternions are a generalisation of the complex numbers. Hamilton was the first to show in 1833 that complex numbers form an algebra, e.g. that it is possible to built up consistent rules to calculate with pairs of numbers. For ten years he tried in vain to extend this concept to number-triplets. It is said that the idea to use four numbers came during a walk with his wife in 1853. He was so excited that he carved it into the pillars of the Broom Bridge (now Hamilton Bridge) in Dublin. The Quaternion is an extension of the normal complex numbers and is defined as:

$$\mathbf{q} = q_0 + \mathfrak{i} q_x + \mathfrak{j} q_y + \mathfrak{k} q_z, \quad q_0, q_x, q_y, q_z \in \mathbb{R}. \tag{A.7}$$

They obey the following basic rules:

$$\mathfrak{i}^2 = \mathfrak{j}^2 = \mathfrak{k}^2 = -1, \qquad \mathfrak{ij} = \mathfrak{k}, \mathfrak{ji} = -\mathfrak{k}. \tag{A.8}$$

With $\vec{\mathfrak{i}} = (\mathfrak{i}, \mathfrak{j}, \mathfrak{k})^T$ and $\vec{q} = (q_x, q_y, q_z)^T$ one may write:

$$\mathbf{q} = q_0 + \vec{q} \cdot \vec{\mathfrak{i}} \quad \text{and the conjugate as:} \quad \mathbf{q}^* = q_0 - \vec{q} \cdot \vec{\mathfrak{i}}. \tag{A.9}$$

The multiplication of two quaternions is defined as:

$$\mathbf{q}_1 \mathbf{q}_2 = (q_{10} q_{20} - \vec{q}_1 \cdot \vec{q}_2) + (q_{10} \vec{q}_2 + q_{20} \vec{q}_1 + \vec{q}_1 \times \vec{q}_2) \cdot \vec{\mathfrak{i}}. \tag{A.10}$$

The magnitude of the quaternion is:

$$\sqrt{\mathbf{q}\mathbf{q}^*} \quad \text{where} \quad \mathbf{q}\mathbf{q}^* = q_0^2 + \vec{q}^2. \tag{A.11}$$

Quaternions with the magnitude 1 are called unit quaternions. To each quaternion $\mathbf{q}$ different from zero there is a inverse quaternion $\mathbf{q}^{-1}$, so that: $\mathbf{q}\mathbf{q}^{-1} = 1$. The reason why the quaternions are useful in the description of rotations is that they can be written as matrices. There are two ways to write quaternions as a matrix. The first is a complex 2×2 matrix:

$$\mathbf{Q} = \begin{pmatrix} q_0 + q_x\mathrm{i} & q_y + q_z\mathrm{i} \\ -q_y + q_z\mathrm{i} & q_0 - q_x\mathrm{i} \end{pmatrix} \tag{A.12}$$

The second is a real 4×4 matrix:

$$\mathbf{Q} = \begin{pmatrix} q_0 & q_x & -q_y & -q_z \\ -q_x & q_0 & -q_z & q_y \\ q_y & q_z & q_0 & q_x \\ q_z & -q_y & -q_x & q_0 \end{pmatrix} \tag{A.13}$$

The conjugate $\mathbf{q}^*$ is in the matrix-representation the transposed matrix $\mathbf{Q}^T$. For a unit quaternion is:

$$\mathbf{Q}\mathbf{Q}^T = \mathbb{1} \qquad \text{or} \qquad \det(\mathbf{Q}) = 1. \tag{A.14}$$

From this directly follows that for the case of a unit quaternion the corresponding matrix $\mathbf{Q}$ is a rotation matrix. This is not yet very useful, but one can derive from the real 4×4 matrix also a matrix in $\mathbb{R}^3$, which is also fullfilling the requirements of beeing a rotation matrix:

$$\mathbf{Q} = \begin{pmatrix} (q_0^2 + q_x^2 - q_y^2 - q_z^2) & 2(q_x q_y + q_z q_0) & 2(q_x q_z + q_y q_0) \\ 2(q_x q_y + q_z q_0) & (q_0^2 - q_x^2 + q_y^2 - q_z^2) & 2(q_y q_z - q_x q_0) \\ 2(q_x q_z - q_y q_0) & 2(q_z q_y + q_x q_0) & (q_0^2 - q_x^2 - q_y^2 + q_z^2) \end{pmatrix} \tag{A.15}$$

This is the matrix we were looking for. Each unit quaternion corresponds with exactly one 3×3 rotation matrix.

A.2.1 How to describe rotations with quaternions

While using rotation matrices to describe the orientation of a rigid body, this formula describes the change of the orientation:

$$\dot{R}(t) = \omega(t)R(t). \tag{A.16}$$

With quaternions instead of matrices it becomes:

$$\dot{\mathbf{q}}(t) = \omega(t)\mathbf{q}(t) \tag{A.17}$$

The rotation θ about an unit(!) axis $\vec{u}$ is described by the unit quaternion:

$$\mathbf{q} = [q_0, \vec{q}] = \left[\cos\left(\frac{\theta}{2}\right), \sin\left(\frac{\theta}{2}\right)\vec{u}\right]. \tag{A.18}$$

A rotation $\mathbf{q}_1$ followed by $\mathbf{q}_2$ is represented by $\mathbf{q}_1\mathbf{q}_2$ and therefore all possible rotations can be described as combinations of unit quaternions like (A.18).A point $P = (x_0, y_0, z_0)$ is represented by the quaternion $\mathbf{p} = (0, x_0, y_0, z_0)$. The point P' which is rotated around the axis $\vec{u}$ is recieved through:

$$\mathbf{p}' = \mathbf{q} \cdot \mathbf{p} \cdot \mathbf{q}^* \tag{A.19}$$

A.3 Electrostatic energy

The electrostatic potential energy W of a charge distribution $\rho(\vec{x})$ in an outer field with the potential $\Phi(\vec{x})$ is[72]:

$$W = \int \rho(\vec{x})\Phi(\vec{x})d\vec{x}. \tag{A.20}$$

The potential is defined as:

$$\Phi(\vec{x}) = \int \frac{\rho(\vec{x}')}{|\vec{x} - \vec{x}'|}d\vec{x}'. \tag{A.21}$$

The charge distribution in a residue can be written in this way:

$$\rho(\vec{x}) = \sum_i q_i \delta(\vec{x} - \vec{x}_i), \tag{A.22}$$

where the q_i are the partial charges of the atoms in the residue. With this the potential becomes:

$$\Phi(\vec{x}) = \sum_i \frac{q_i}{|\vec{x} - \vec{x}_i|}. \tag{A.23}$$

This leads to the general expression of the energy:

$$W = \sum_{ij} \frac{q_i q_j}{|\vec{x}_i - \vec{x}_j|}. \tag{A.24}$$

If one is interested in the interaction of two molecules in varying conformations it is not the most useful way to use this formula, simply because it is expensive on the computational time on large systems to calculate all atom-atom interactions. Instead one can calculate the electrostatic energy between two charge distributions through their total charges, dipole momenta and spatial orientation. Total charges and the relative dipole moment of a molecule remain more or less the same, even if position and orientation change. To do so we write the potential as a Taylor-expansion:

$$\begin{aligned}\Phi(\vec{x}) &= \Phi(\vec{x}_{coc}) - \vec{x}\cdot\vec{E}(\vec{x}_{coc}) - \cdots \\ &= \frac{q_t}{r} + \frac{\vec{p}\cdot\vec{x}}{r^3} + \cdots \end{aligned} \tag{A.25}$$

regarding only the monopole- and the dipole-terms here, with q_t as the total charge. The point at which one wants to expand the function can be choosen and is taken as the "center of charge":

$$\vec{x}_{coc} = \frac{\int \vec{x}|\rho(\vec{x})|d\vec{x}}{\int |\rho(\vec{x})|d\vec{x}}. \tag{A.26}$$

The field at a point $\vec{x}$, that is caused by an dipole located at $\vec{x}_0$ is:

$$\vec{E}(\vec{x}) = \frac{3\vec{n}(\vec{p}\cdot\vec{n}) - \vec{p}}{|\vec{x}-\vec{x}_0|^3}, \quad \vec{n} = \frac{\vec{x}-\vec{x}_0}{|\vec{x}-\vec{x}_0|}. \tag{A.27}$$

The dipole moment $\vec{p}$ is defined as:

$$\vec{p} = \int \rho(\vec{x})\vec{x}d\vec{x} = \int \sum_i q_i \delta(\vec{x}-\vec{x}_i)\vec{x}d\vec{x} = \sum_i q_i\vec{x}_i. \tag{A.28}$$

The sum depends on the choice of the origin, when the total charge $q_t = \sum_i q_i$ is not zero. Using (A.22) in (A.20) results in:

$$W = \int \sum_i q_i \delta(\vec{x}-\vec{x}_i)\Phi(\vec{x})d\vec{x} = \sum_i q_i \Phi(\vec{x}_i). \tag{A.29}$$

Now, with the expansion of the potential (A.25) we get:

$$W = \sum_i q_i(\Phi(\vec{x}_{coc}) - \vec{x}_i\cdot\vec{E}(\vec{x}_{coc})). \tag{A.30}$$

This describes the interaction of a charge distribution with a second charge distribution which is represented by its the monopole and the dipole term of the multipole expansion. The next step is to represent also the "first" distribution as a monopole and a dipole. Considering two charge distributions as consisting of a monopole and a dipole each the potential energy can be written as:

$$\begin{aligned} W &= W^{monopol} + W_I^{monopol-dipol} + W_{II}^{monopol-dipol} + W^{dipol} \\ &= \frac{q_1 q_2}{r} - q_1\frac{\vec{e}_r\cdot\vec{p}_2}{r^2} + q_2\frac{\vec{e}_r\cdot\vec{p}_1}{r^2} + \frac{\vec{p}_1\cdot\vec{p}_2 - 3(\vec{p}_1\cdot\vec{e}_r)(\vec{p}_2\cdot\vec{e}_r)}{r^3}. \end{aligned} \tag{A.31}$$

This is the general expression for two charged residues, representing (A.24).

Appendix B

Plots, figures and tables on sphere-algorithm

B.1 Tests

B.1.1 Basic test on the free-surface method

To ensure a sufficient accuracy of the sphere-algorithm we compare the numerical result with an analytical formula, which is available for the overlap of two spheres. Through this we will determine an appropriate value of δ_θ for our further investigations. The δ_θ can be chosen freely, and should be taken in such way that it leads to a satisfying compromise betweeen precision and computation time. The smaller δ_θ the higher the precision and the longer the calculation-time. For each of a number of different values for δ_θ we perform 2000 tests with 2 spheres of random radii and positions. The analytical formula for the overlap of sphere a in terms of the radii and the distance is:

$$S_{a/b}^{overlap} = 2\pi r_a^2 \left(1 - \frac{r_a^2 + d_{ab}^2 - r_b^2}{2 r_a d_{ab}}\right)$$

The results are listed in the table below. In the first column are the values for δ_θ, in the second the mean-value of the difference between numerical and analytical result, in the third the standard deviation, and in the last the time needed to perform 2000 runs on a PC.

The table provides evidence that the algorithm is trustworthy. It probably could be easily optimized, by using a different scheme for the distribution of the vector-surfaces. For the investigations performed here we have chosen a δ_θ of 3^o.

δ_θ	mean-difference	rmsd	time / 2000 runs
2^o	-0.000496541	0.0681082	1 m 40.082 s
4^o	0.0056022	0.200258	0 m 24.9 s
5^o	0.00546776	0.249301	0 m 16.156 s
8^o	-0.0457886	0.514143	0 m 6.432 s
15^o	0.0107227	1.28436	0 m 1.991 s

Table B.1: Accuracy and performance of sphere-algorithm with different δ_θ

B.1.2 Basic test on the nth-order-overlap method

Following a simple strategy we tested the consistency of the higher order overlap method. Our reference is again the formula for the overlap of two spheres. For the overlap of more spheres a little trick is used. If one takes three spheres *a, b, c* and two of them (*b* and *c*) with the same radii, one can use the first order formula as a test for the cases that *b* and *c* do not overlap and overlap completely. One can do a little scan with *a* and *b* fixed and *c* as a probe. We performed the test with four spheres *a, b, c, d*, where *b, c, d* have the same radii and *a, b, c* are fixed.

Sphere *a* has radius 10, the others radius 4. *a* is placed in the origin, *b* and *c* on the z-axis at z = 10, i.e. with their center on the surface of *a*. Sphere *d* moves as a probe on the surface of *a*, starting with its center on the z-axis at z = 10 ($\Phi = 0°$), until there is no overlap with *b* and *c* anymore ($\Phi = 60°$). Φ is the angle between the z-axis and the connecting vector between the centers of *a* and *d* .

Results from plotting the overlaps in different orders can be seen in table B.2. The consistency can be seen from the upper and lowest block.

ϕ	sphere	first order	second order	third order
0°	a	54.321	54.321	54.321
	b	201.062	201.062	79.6294
	c	201.062	201.062	79.6294
	d	201.062	201.062	79.6294
20°	a	77.9067	54.321	24.3737
	b	201.062	110.481	26.076
	c	201.062	110.481	26.076
	d	111.49	57.1409	26.2737
40°	a	101.382	54.321	3.18207
	b	201.062	84.8181	9.45921
	c	201.062	84.8181	9.45921
	d	85.4284	14.7345	9.59984
60°	a	103.971	54.321	0
	b	201.062	79.6294	0
	c	201.062	79.6294	0
	d	80.3415	0	0

Table B.2: Basic test on overlaps up to third order

B.2 Tables

Amino acid		r_{cavity} [1]	r^2
Glycine	G	2.46	2.585
Alanine	A	2.73	2.904
Valine	V	3.18	3.362
Leucine	L	3.38	3.613
Isoleucine	I	3.39	3.667
Methionine	M	3.47	3.868
Proline	P	3.02	3.290
Phenylalanine	F	3.62	3.959
Tryptophan	W	3.94	4.250
Serine	S	2.83	3.216
Threonine	T	3.05	3.362
Asparagine	N	3.17	3.535
Glutamine	Q	3.36	3.969
Tyrosine	Y	3.70	4.203
Cysteine	C	3.00	3.278
Aspartic Acid, protonated	1	3.11	3.602
Glutamic Acid, protonated	2	3.32	3.929
Lysine, deprotonated	3	3.49	4.039
Histidine, protonated at δ 1/π	6	3.46	3.827
Histidine, protonates at ε 2/τ	5	3.44	3.827

Table B.3: The radii taken for the sphere-algorithm for seeking tight-packing in conformational space. Radii are listed in $\mathring{A}$.

[1] as calculated in [38].

[2] taken from preliminary work of a summer student in our group, Sahand Jamal Rahi.

B.3 Residuewise list of relative free surfaces and resulting energies

residue	typ	ΔG kJ/mol	free surface $\AA^2$	relative free surface	$\Delta G \times$ relative free surface kJ/mol
0	I	15.8	109.8	0.65	5.13
1	T	-4.7	61.7	0.435	-1.02
2	L	15.5	72	0.439	3.4
3	I	15.8	67.3	0.398	3.15
4	I	15.8	59.8	0.354	2.8
5	F	18.1	79.1	0.402	3.63
6	G	-5.1	31.9	0.379	-0.97
7	V	12.5	61.6	0.434	2.71
8	M	11.3	72.3	0.384	2.17
9	A	1.8	46.8	0.441	0.4
10	G	-5.1	42.0	0.5	-1.28
11	V	12.5	57.1	0.402	2.51
12	I	15.8	72.2	0.427	3.38
13	G	-5.1	36.2	0.431	-1.1
14	T	-4.7	59.2	0.417	-0.98
15	I	15.8	63.2	0.374	2.95
16	L	15.5	63.5	0.387	3
17	L	15.5	58.2	0.358	2.75
18	I	15.8	67.9	0.402	3.18
19	S	-9.7	42	0.323	-1.57
20	Y	5.2	114.9	0.518	1.35
21	G	-5.1	29.7	0.354	-0.9
22	I	15.8	117.2	0.694	5.48
Σ		179	1485.38		40.17

Table B.4: Free surface and lipophilicity of a single glycophorin A helix, based on the transfer free energies from Gu *et al* [38].

			helix 1		helix 2	
	typ	ΔG kJ/mol	rel. free surf.	$\Delta G \times$ rel. free surf. kJ/mol	rel. free surf.	$\Delta G \times$ rel. free surf. kJ/mol
0	I	15.8	0.651	5.14	0.65	5.13
1	T	-4.7	0.434	-1.02	0.433	-1.02
2	L	15.5	0.439	3.4	0.440	3.41
3	I	15.8	0.399	3.15	0.400	3.16
4	I	15.8	0.354	2.8	0.355	2.8
5	F	18.1	0.402	3.64	0.402	3.64
6	G	-5.1	0.38	-0.97	0.38	-0.97
7	V	12.5	0.432	2.7	0.433	2.7
8	M	11.3	0.384	2.17	0.385	2.17
9	A	1.8	0.393	0.35	0.393	0.35
10	G	-5.1	0.271	-0.69	0.273	-0.67
11	V	12.5	0.395	2.47	0.394	2.46
12	I	15.8	0.426	3.37	0.426	3.37
13	G	-5.1	0.207	-0.53	0.207	-0.53
14	T	-4.7	0.107	-0.25	0.106	-0.25
15	I	15.8	0.371	2.93	0.371	2.93
16	L	15.5	0.382	2.96	0.382	2.96
17	L	15.5	0.154	1.19	0.153	1.19
18	I	15.8	0.286	2.26	0.286	2.26
19	S	-9.7	0.324	-1.57	0.324	-1.57
20	Y	5.2	0.518	1.35	0.517	1.35
21	G	-5.1	0.354	-0.9	0.355	-0.9
22	I	15.8	0.692	5.47	0.693	5.47
$\sum$						78.84

Table B.5: Free surfaces and lipophilicities of two glycophorin A helices at 6$\mathring{A}$ distance and 40° tilt. When compared to table B.4, the only noticable differences occur at positions 10, 13, 14.

			helix 1		helix 2	
	typ	ΔG kJ/mol	rel. free surf.	$\Delta G \times$ rel. free surf. kJ/mol	rel. free surf.	$\Delta G \times$ rel. free surf. kJ/mol
0	I	15.8	0.617	4.88	0.618	4.88
1	T	-4.7	0.433	-1.02	0.433	-1.02
2	L	15.5	0.27	2.09	0.269	2.09
3	I	15.8	0.102	0.8	0.102	0.81
4	I	15.8	0.355	2.8	0.354	2.8
5	F	18.1	0.395	3.58	0.394	3.56
6	G	-5.1	0.054	-0.14	0.056	-0.14
7	V	12.5	0.342	2.14	0.343	2.14
8	M	11.3	0.383	2.16	0.384	2.17
9	A	1.8	0.401	0.36	0.400	0.36
10	G	-5.1	0.157	-0.4	0.155	-0.39
11	V	12.5	0.402	2.51	0.401	2.51
12	I	15.8	0.427	3.37	0.426	3.37
13	G	-5.1	0.227	-0.58	0.227	-0.58
14	T	-4.7	0.204	-0.48	0.205	-0.48
15	I	15.8	0.372	2.94	0.373	2.95
16	L	15.5	0.374	2.9	0.374	2.9
17	L	15.5	0.027	0.21	0.028	0.21
18	I	15.8	0.315	2.49	0.316	2.49
19	S	-9.7	0.323	-1.57	0.323	-1.57
20	Y	5.2	0.401	1.04	0.401	1.04
21	G	-5.1	0.059	-0.15	0.06	-0.15
22	I	15.8	0.692	5.47	0.693	5.47
Σ						70.85

Table B.6: Free surfaces and lipophilicities of two glycophorin A helices at 6$\AA$ distance and 0° tilt.

residue	typ	ΔG kJ/mol	free surface $\AA^2$	relative free surface	$\Delta G \times$ relative free surface kJ/mol
0	G	-5.1	56.4	0.671	-1.71
1	I	15.8	102.8	0.608	4.8
2	G	-5.1	38.4	0.458	-1.17
3	A	1.8	48.3	0.456	0.41
4	V	12.5	56.8	0.4vv	2.5
5	L	15.5	62.2	0.379	2.94
6	K	-47.8	87.5	0.427	-10.12
7	V	12.5	51.9	0.365	2.28
8	L	15.5	68	0.414	3.21
9	T	-4.7	50.1	0.353	-0.83
10	T	-4.7	61.8	0.435	-1.02
11	G	-5.1	34.7	0.413	-1.05
12	L	15.5	67.8	0.413	3.2
13	P	8.3	55.0	0.404	1.68
14	A	1.8	45.7	0.431	0.39
15	L	15.5	59.8	0.365	2.82
16	I	15.8	62.9	0.372	2.94
17	S	-9.7	42.4	0.326	-1.58
18	W	12.2	81.1	0.357	2.18
19	I	15.8	49.5	0.293	2.31
20	K	-47.8	64.2	0.313	-7.48
21	R	-45.2	80.3	0.324	-7.32
22	K	-47.8	60.1	0.293	-7
23	R	-45.2	104.5	0.421	-9.53
24	Q	-12.8	74.3	0.376	-2.4
25	Q	-12.8	113.3	0.572	-3.66
Σ		-135.3	1679.8		-23.29

Table B.7: Free surfaces and lipophilicities of a single melittin helix.

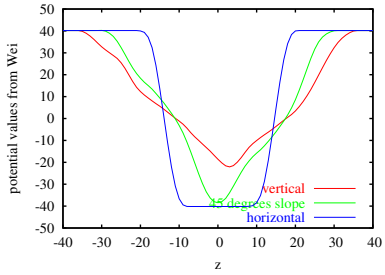
residue	typ	ΔG kJ/mol	free surface $\AA^2$	relative free surface	$\Delta G \times$ relative free surface kJ/mol
0	N	-15.1	96.4	0.614	-4.63
1	P	8.3	56.9	0.418	1.74
2	I	15.8	73.8	0.436	3.45
3	Y	5.2	73.9	0.333	0.86
4	W	12.2	86.2	0.38	2.32
5	A	1.8	30.7	0.289	0.26
6	R	-45.2	87.9	0.354	-8.01
7	Y	5.2	70.1	0.316	0.82
8	A	1.8	35.8	0.338	0.30
9	D	-82.9	52.2	0.320	-13.27
10	W	12.2	79.3	0.349	2.13
11	L	15.5	56.0	0.342	2.65
12	F	18.1	77.0	0.391	3.54
13	T	-4.7	45.8	0.323	-0.76
14	T	-4.7	51.5	0.362	-0.85
15	P	8.3	47.9	0.352	1.46
16	L	15.5	60.6	0.369	2.86
17	L	15.5	60.9	0.371	2.88
18	L	15.5	62.5	0.381	2.95
19	L	15.5	63	0.384	2.98
20	D	-82.9	59.5	0.365	-15.13
21	L	15.5	78.17	0.476	3.69
22	A	1.8	45.29	0.427	0.38
23	L	15.5	111.8	0.681	5.28
Σ		-36.3	1563.4		-2.09

Table B.8: Free surface and lipophilicity of a single bR C helix, without loops.

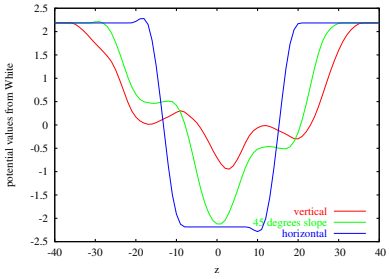
residue	typ	ΔG kJ/mol	free surface $\AA^2$	relative free surface	$\Delta G \times$ relative free surface kJ/mol
0	G	-5.1	57.7	0.687	-1.75
1	G	-5.1	40	0.476	-1.21
2	Q	-12.8	102.4	0.517	-3.31
3	Q	-12.8	77.1	0.39	-2.49
4	N	-15.1	57.7	0.367	-2.77
5	P	8.3	43	0.316	1.31
6	I	15.8	56.7	0.336	2.65
7	Y	5.2	68.7	0.31	0.80
8	W	12.2	86.2	0.38	2.32
9	A	1.8	30.7	0.29	0.26
10	R	-45.2	88.2	0.355	-8.04
11	Y	5.2	70.2	0.316	0.82
12	A	1.8	35.9	0.338	0.30
13	D	-82.9	51.9	0.318	-13.20
14	W	12.2	79.5	0.350	2.14
15	L	15.5	55.9	0.341	2.64
16	F	18.1	77.0	0.391	3.54
17	T	-4.7	45.8	0.323	-0.76
18	T	-4.7	51.7	0.364	-0.85
19	P	8.3	47.8	0.351	1.46
20	L	15.5	60.6	0.369	2.86
21	L	15.5	60.7	0.37	2.87
22	L	15.5	62.4	0.38	2.95
23	L	15.5	63.2	0.385	2.98
24	D	-82.9	58.3	0.357	-14.81
25	L	15.5	63.3	0.386	2.99
26	A	1.8	38.9	0.367	0.33
27	L	15.5	64	0.39	3.02
28	L	15.5	64.1	0.391	3.03
29	V	12.5	53.4	0.376	2.35
30	D	-82.9	64.5	0.395	-16.39
31	A	1.8	43	0.406	0.36
32	D	-82.9	64.4	0.395	-16.37
33	Q	-12.8	101.4	0.512	-3.28
34	G	-5.1	37.3	0.444	-1.13
35	T	-4.7	99.8	0.703	-1.65
Σ		-230.7	2223.6		-46.05

Table B.9: Free surface and lipophilicity of a single bR C helix, with loops.

B.4 Insertion profiles for different values of ΔG



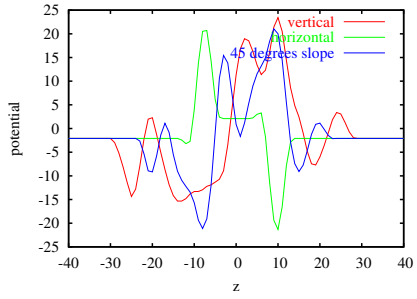
ΔΔ G from W. Gu



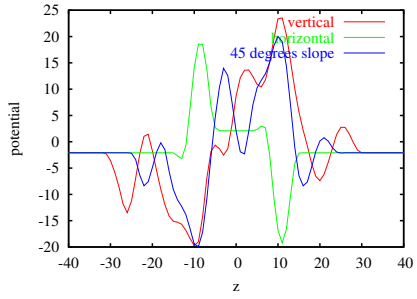
S. White

Figure B.1: Membrane-insertion of a single GpA helix.

B.5 Some more insertion profiles - for those who really like them



membrane thickness 23*Å*



membrane thickness 26*Å*

Figure B.2: Insertion of a c-helix of bR with different membane sizes.

B.6 Tests on n^{th}-order overlap method

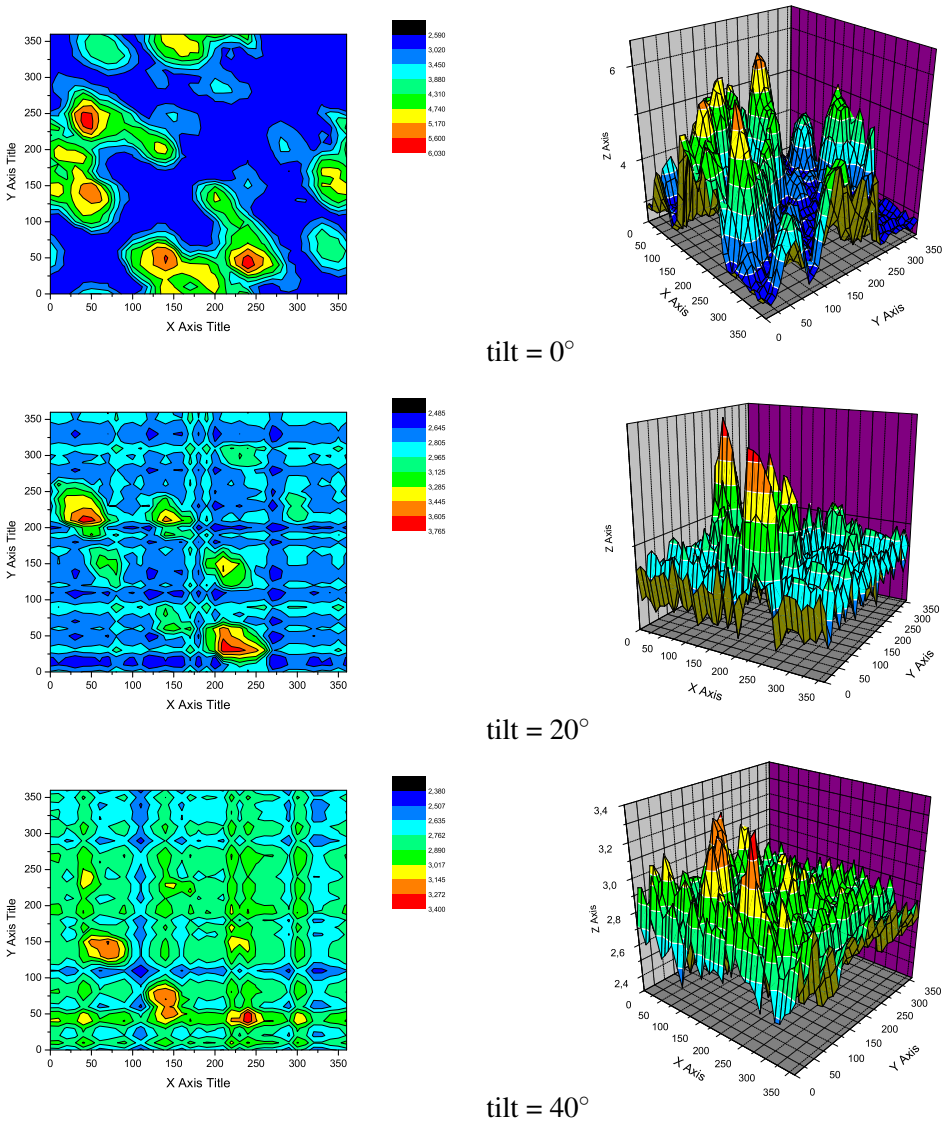
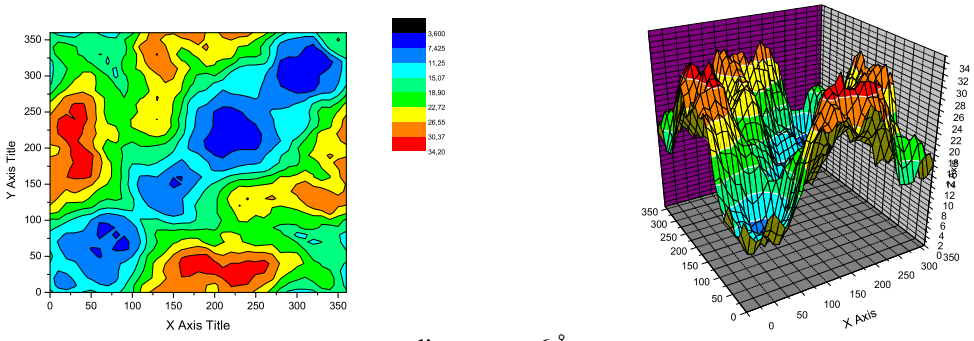
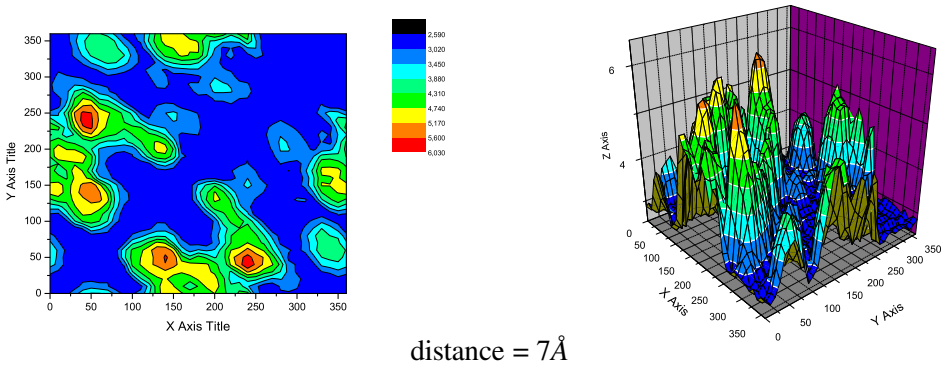


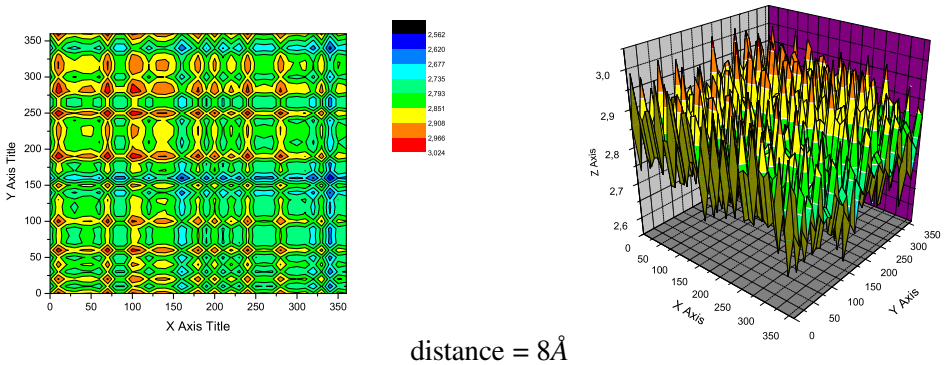
Figure B.3: 4th order overlap of glycophorin A, distance = 7$\mathring{A}$. The lowest set of figures has a very small energy range. The differences shown are merely noise.



distance = 6Å

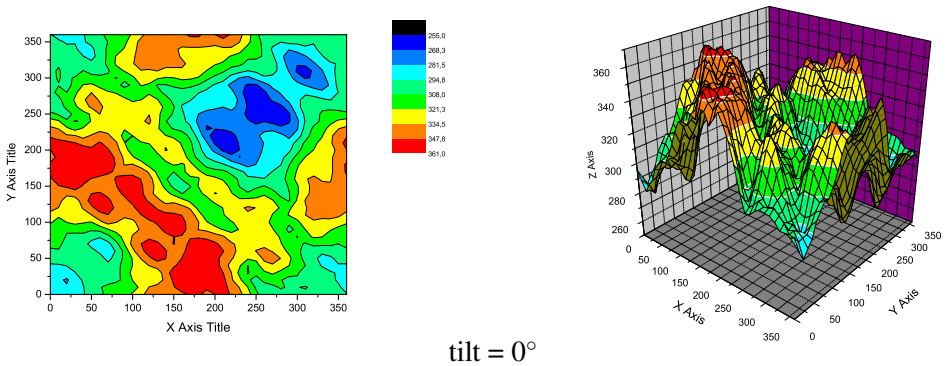


distance = 7Å

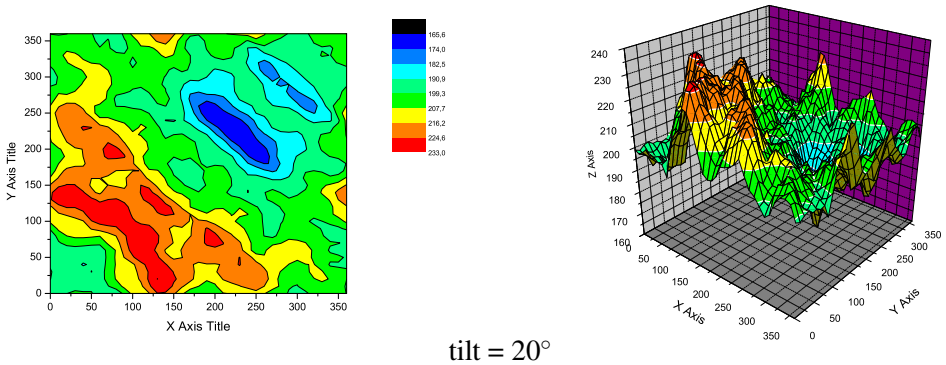


distance = 8Å

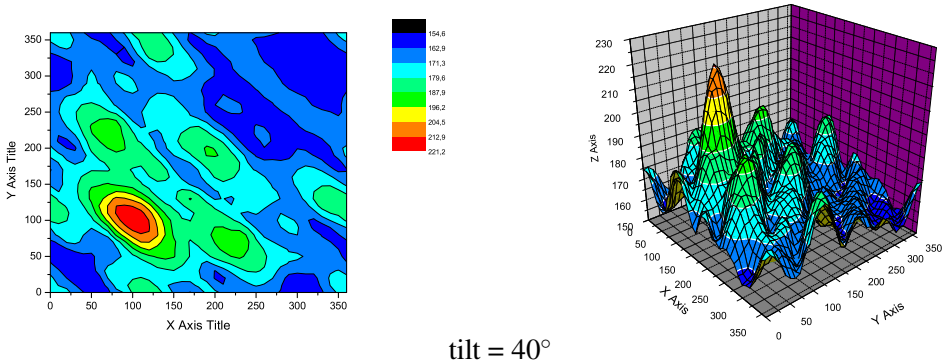
Figure B.4: 4th order overlap of GpA, tilt = 0°.



tilt = 0°

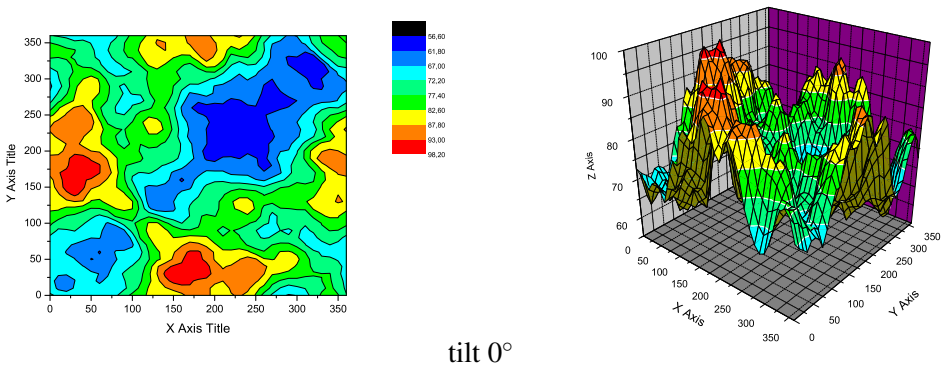


tilt = 20°

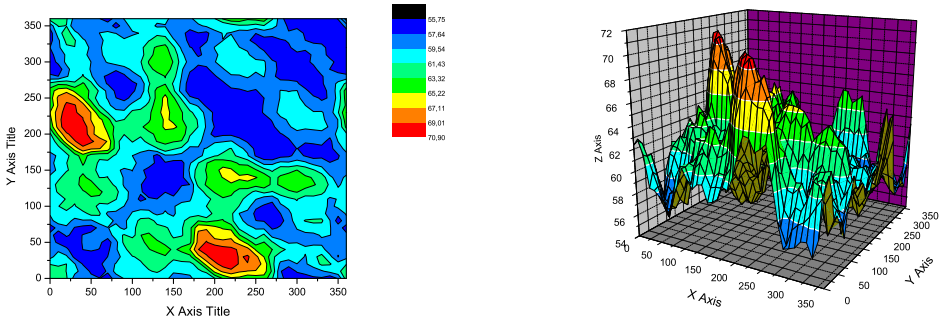


tilt = 40°

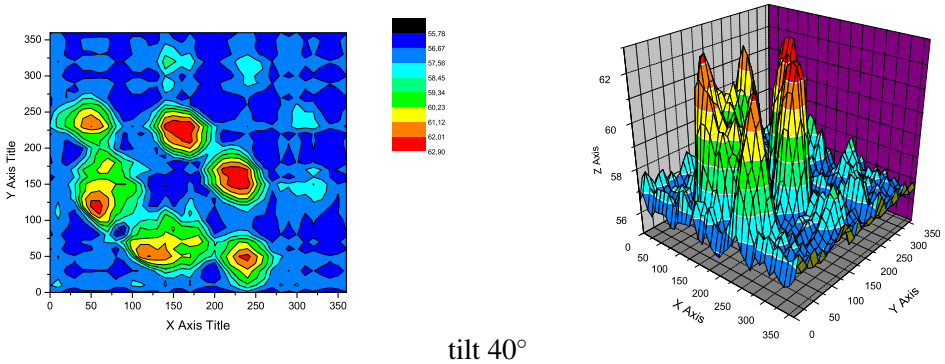
Figure B.5: 3^{rd} order overlap of GpA for distance = 6$\mathring{A}$.



tilt 0°



tilt 20°



tilt 40°

Figure B.6: The 4^{th} order overlap at a distance of $7\mathring{A}$ using alternative radii from Jamal. See table B.3.

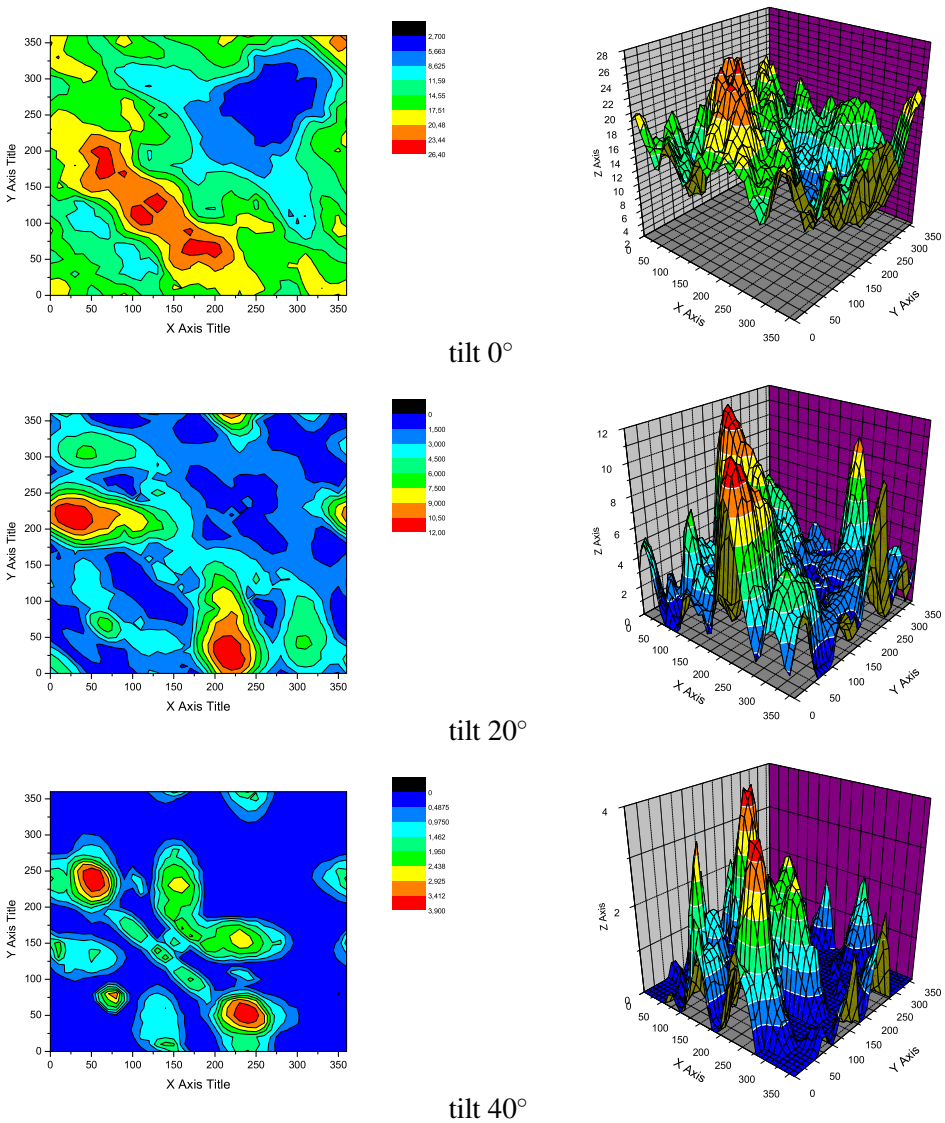


Figure B.7: The 5^{th} order overlap at a distance of 6$\mathring{A}$ using alternative radii from Jamal. See table B.3.

Bibliography

[1] Voet , Voet, *Biochemistry*

[2] Neutze R, Pebay-Peyroula E, Edman K, Royant A, Navarro J, Landau EM, *Bacteriorhodopsin: a high-resolution structural view of vectorial proton transport.* Biochim Biophys Acta 1565(2):144-67, 2002.

[3] Stevens TJ, Arkin IT. *Do more complex organisms have a greater proportion of membrane proteins in their genomes?* Proteins 39(4):417-420, 2000.

[4] Aguzzi A, Haass C, *Games played by rogue proteins in prion disorders and Alzheimer's disease [Review].* Science 302(5646):814-818, 2003.

[5] Dobson CM, *Protein misfolding, evolution and disease.* Trends Biochem Sci 24(9):329-332, 1999

[6] Shin I, Kreimer I, Weiner L, *Membrane-promoted unfolding of acetylcholinesterase: a possible mechanism for insertion into the lipid bilayer.* Proc Natl Acad Sci USA 94:2848-2852, 1997.

[7] Banuelos S, Muga A , *Binding of molten globule-like conformations to lipid bilayers: structure of native and partially folding lactalbumin bound to model membranes.* J Biol Chem 1995, 270:29910-29915.

[8] Sanders CJ, Nagy JK, *Misfolding of membrane proteins in health and disease: the lady or the tiger?* Curr Opin Struct Biol 2000, 10:438 442.

[9] Gorzelle BM, Nagy JK, Oxenoid K, Lonzer WL, Cafiso DS, Sanders CR, *Reconstitutive refolding of diacylglycerol kinase, an integral membrane protein.* Biochemistry 1999, 38:16373-16382.

[10] Liu Y, Engelman DM, Gerstein M, *Genomic analysis of membrane protein families: abundance and conserved motifs.* Genome Biol 2002 Sep 19;3(10).

[11] Jayasinghe S, Hristova K, White SH, *Energetics, stability, and prediction of transmembrane helices.* J Mol Biol 312(5):927-934, 2001.

[12] Chen CP, Kernytsky A, Rost B, *Transmembrane helix predictions revisited [Review].* Prot Sci 11(12):2774-2791, 2002

[13] www.mpibp-frankfurt.mpg.de/michel/public/memprotstruct.html updated on april 24, 2004

[14] Hunte C, Michel H, *Crystallisation of membrane proteins mediated by antibody fragments.* Curr Opin Struct Biol 12:503-508, 2002.

[15] Rhee KH, Morris EP, Barber J, Kühlbrandt W, *Three-dimensional structure of the plant photosystem II reaction centre at 8Å resolution.* Nature 396:283-286, 1998.

[16] Mackenzie KR, Prestegard JH, Engelman DM, *A transmembrane helix dimer - structure and implications.* Science 276(5309):131-133, 1997.

[17] Fiaux J, Bertelsen EB, Horwich AL, Wüthrich K, *NMR analysis of a 900K GroEL-GroES complex* Nature 418:207 - 211, 2002.

[18] Requena JR, Levine RL, *Thioredoxin converts the Syrian hamster (29-231) recombinant prion protein to an insoluble form.* Free Radical Biol Med 30(2):141-147, 2001.

[19] Stevens TJ, Arkin IT, *Are membrane proteins ïnside-outproteins?.* Proteins 36(1):135-143, 1999.

[20] Rees D, Komiya H, Yeates T, Allen J, Feher G, *The bacterial photosynthetic reaction center as a model for membrane proteins.* Annu Rev Biochem 58:607-633, 1989.

[21] Zhou FX, Cocco MJ, Russ WP, Brunger AT, Engelman DM, Nat Struct Biol 7:154-160, 2000.

[22] Choma C, Gratkowski H, Lear JD, DeGrado WF, Nat Struct Biol 7, 161-166(2000).

[23] Senes A, Gerstein M, Engelman DM, *Statistical analysis of amino acid patterns in transmembrane helices: The GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions.* J Mol Biol 296(3):921-936, 2000.

[24] Popot J-L, Engelman DM, *Membrane protein folding and oligomerization: the two-stage model.* Biochemistry 29:4031-37, 1990.

[25] White SH, Wimley WC, *Membrane protein folding and stability: physical principles.* Annu Rev Biophys Biomol Struct 28:319-365, 1999.

[26] Jiang YX, Ruta V, Chen JY, Lee A, MacKinnon R, *The principle of gating charge movement in a voltage-dependent K+ channel.* Nature 423(6935):42-48, 2003.

[27] Vonck J, *Structure of the bacteriorhodopsin mutant F219L N intermediate revealed by electron crystallography.* EMBO Journal 19(10):2152-2160, 2000 May 15.

[28] Toyoshima C, Nomura H, *Structural changes in the calcium pump accompanying the dissociation of calcium.* Nature, 418(6898):605-611, 2002.

[29] MacKinnon R, in a talk on the Gordon Conference on Ligand Recognition and Molecular Gatingïn Italy, May 2002

[30] Geibel S, Zimmermann D, Zifarelli G, Becker A, Koenderink JB, Hu Y-K, Kaplan JH, Friedrich T, Bamberg E, *Conformational Dynamics of Na+/K+- and H+/K+-ATPase Probed by Voltage Clamp Fluorometry.* Ann NY Acad Sci Volume 986 - Na+/K+-ATPase and Related Cation Pumps - Structure, Function, and Regulatory Mechanisms. S. 31-38 (2003).

[31] Shen T, Tai K, Henchman RH, McCammon JA, *Molecular Dynamics of Acetylcholinesterase.* Acc Chem Res; 2002; 35(6) pp 332 - 340;

[32] Böckmann RA, Grubmüller H, *Nanoseconds molecular dynamics simulation of primary mechanical energy transfer steps in F1-ATP synthase.* Nat Struct Biol 9:198-202 (2002).

[33] Devi-Kesavan LS, Garcia-Viloca M, Gao J, *Semiempirical QM/MM potential with simple valence bond (SVB) for enzyme reactions. Application to the nucleophilic addition reaction in haloalkane dehalogenase.* Theor Chem Acc 109(3):133-139, 2003.

[34] Rohrig UF, Frank I, Hutter J, Laio A, VandeVondele J, Rothlisberger U. *QM/MM Car-Parrinello molecular dynamics study of the solvent effects on the ground state and on the first excited singlet state of acetone in water.* Chemphyschem. 4(11):1177-82, 2003.

[35] Lill MA, Helms V, *Molecular dynamics simulation of proton transport with quantum mechanically derived proton hopping rates (Q-HOP MD).* J Chem Phys 115:7993-8005, 2001

[36] Fleishman SJ, Ben-Tal N, *A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane alpha-helices.* J Mol Biol 321(2):363-378, 2002.

[37] Lazaridis T, *Effective energy function for proteins in lipid membranes.* Proteins: Struc Func Gen 52:176-192, 2003.

[38] Gu W, Rahi SJ, Helms V, *Solvation Free Energies and Transfer Free Energies for Amino Acids from Hydrophobic Solution to Water Solution from a Very Simple Residue Model.* J Phys Chem B 108:5806-5814, 2004

[39] Im WP, Lee MS, Brooks CL, *Generalized born model with a simple smoothing function.*— J Comp Chem 24(14):1691-1702, 2003.

[40] Elsner M. *Berechnung der Wechselwirkung von Aminosaüren in benachbarten α-Helices und Parameterisierung eines effektiven Potentials.* Diplomarbeit im Fachbereich Chemische und Pharmazeutische Wissenschaften der Johann Wolfgang Goethe-Universitat Frankfurt am Main, 2002

[41] Straatsma TP, Apra E, Windus TL, Dupuis M, Bylaska EJ, de Jong W, Hirata S, Smith DMA, Hackler M, Pollack L, Harrison R, Nieplocha J, Tipparaju V, Krishnan M, Brown E, Cisneros G, Fann G, Fruchtl H, Garza J, Hirao K, Kendall R, Nichols J, Tsemekhman K, Valiev M, Wolinski K, Anchell J, Bernholdt D, Borowski P, Clark T, Clerc D, Dachsel H, Deegan M, Dyall K, Elwood D, Glendening E, Gutowski M, Hess A, Jaffe J, Johnson B, Ju J, Kobayashi R, Kutteh R, Lin Z, Littlefield R, Long X, Meng B, Nakajima T, Niu S, Rosing M, Sandrone G, Stave M, Taylor H, Thomas G, van Lenthe J, Wong A, Zhang Z, *NWChem, A Computational Chemistry Package for Parallel Computers, Version 4.5"* (2003), Pacific Northwest National Laboratory, Richland, Washington 99352-0999, USA.

[42] Makarov DE, Metiu H, *Fitting potential-energy surfaces: A search in the function space by directed genetic programming* J Chem Phys 1998 Volume 108(2) 590-598

[43] Park Y, Elsner M, Staritzbichler R, Helms V; *A Novel Scoring Function for Modeling Structures of Oligomers of Transmembrane alpha-helices.* 2004, Proteins, accepted for publication.

[44] GNU Scientific Library: www.gnu.org/software/gsl/

[45] Jacobs RE, White SH 1989, *The nature of the hydrophobic binding of small peptirdes at the bilayer interface: implications for the insertion of transbilayer helices.* Biochemistry 28: 3421-3437

[46] Engelmann DM, Steitz TA, *The spontanous insertion of proteins into and across membranes: the helical hairpin hypothesis.* Cell 23:411-422, 1981.

[47] Jähnig F, *Thermodynamics and kinetics of protein incorporation into membranes.* Proc Natl Acad Sci USA 80:3691-3695, 1983.

[48] Ben-Tal N, Ben-Shaul A, Nicholls A, Honig B, *Free-energy determinants of α-helix insertion into lipid bilayers.* Biophys J 70:1803-1812, 1996.

[49] Hol WG, Halie LM, Sander C, *Dipoles in alpha helices and beta sheets: their role in protein folding*. Nature 294:532-536, 1981.

[50] Ben-Tal N, Honig B, *Helix-helix interactions in lipid bilayers*. Biophys J 71(6):3046-50, 1996.

[51] Ladokhin AS, White SH, *Folding of amphipathic α-helices on membranes: energetics of helix formation by melittin*. J Mol Biol 285:1363-69, 1999.

[52] Hunt JF, Rath P, Rothschild KJ, Engelman DM, *Spontanous, pH-dependent membrane insertion of a transbilayer α-helix*. Biochemistry 36:15177-15192, 1997.

[53] Bechinger B, *Membrane insertion and orientation of polyalanine peptides: a 15N solid-stateNMRinvestigation*. Biophys J 81:2251-2256, 2001.

[54] Gurezka R, Laage R, Brosig B, Langosch D. *A heptad motif of eucine residues found in membrane proteins can drive selfassembly of artificial transmembrane segments*. J Biol Chem 274:9265-9270, 1999.

[55] Fleishman SJ, Ben-Tal N, *A novel scoring function for predicting the conformations of tightly packed pairs of transmembrane α-helices*. J Mol Biol 321:363-378, 2002.

[56] MacKenzie KR, Prestegard JH, Engelman DM, *A transmembrane helix dimer: structure and implications*. Science 276:131, 1997.

[57] Stryer L, *Biochemie*, Spektrum Verlag 1995

[58] Luna EJ, Hitt AL, *Cytoskeleton-plasma membrane interactions*. Science 258:955-964, 1992.

[59] Lux SE, *Dissecting the red cell membrane skeleton*. Nature, 281:426-429, 1997.

[60] Garett RH, Grisham CM, *Biochemistry*. Saunders, 1995.

[61] Wang HY, Tang H, Shen CKJ, Wu CI, *Rapidly evolving genes in human. I. The glycophorins and their possible role in evading malaria parasites*. Mol Biol Evol 20(11):1795-1804, 2003.

[62] Pasvol G, *How many pathways for invasion of the red blood cell by the malaria parasite?*. Trends in Parasitology. 19(10):430-432, 2003.

[63] The RCSB protein data bank (PDB): www.rcsb.org/pdb/

[64] Vaidehi N, Floriano WB, Trabanino R, Hall SE, Freddolino P, Choi EJ, Zamanakos G, Goddard WA, *Prediction of structure and function of G protein-coupled receptors.* Proc Natl Acad Sci USA 99(20):12622-12627, 2002.

[65] Nelder JA, Mead R, *A simplex method for function minimization.* Comp J 7:308-315, 1965.

[66] Lanyi JK, *Bacteriorhodopsin* Annu Rev Physiol 66:665-688, 2004

[67] Popot JL, personal discussion at Gordon Conference, 2002, Italy.

[68] Stiegelmeyr A, Pfeiffer F, *Simulation of rigid multibody systems with unilateral constraints.* Zeitschrift fur Angewandte Mathematik und Mechanik. 81(2):S407-S408, 2001.

[69] Hemami H, *Evolutionary trends in rigid body dynamics.* Comp Meth Appl Mech Engn 192(5-6):635-654, 2003.

[70] Dullweber A, Leimkuhlera B, McLachlan R, *Symplectic splitting methods for rigid body molecular dynamics.* J Chem Phys 107(15), 1997.

[71] *Numerical Recipes in Fortran 77 - Vol. 1* Second Edition, Cambridge University Press 1999

[72] Jackson JD, *Klassische Elektrodynamik, 2.Auflage*, Kapitel 4.1 und 4.2, de Gruyter, 1983 .

Danksagung

Mein besonderer Dank gilt meinem Betreuer Professor Dr. Volkhard Helms für die Ermöglichung meiner Arbeit an einem spannenden Thema, das viel Raum für die Entwicklung eigener Ideen und Ansätze gegeben hat. Auch für viele inspirierende und motivierende Gespräche möchte ich ihm danken.

Herrn Professor Dr. Mäntele möchte ich für die Betreuung von Seiten der Universität Frankfurt danken und dem Lesen dieser Arbeit vor der Abgabe.

Herrn Professor Dr. Bamberg danke ich für die Aufnahme nach dem Umzug meiner Arbeitsgruppe nach Saarbrücken und für die Bewilligung der finanziellen Unterstützung.

Für fachliche Diskussionen danke ich Dr. Michael Hutter, Dr. Markus Lill, Markus Elsner, Yungki Park, Wei Gu und Dr. Tihamer Geyer.

Für die Unterstützung in Computer-Fragen möchte ich Dr. Michael Hutter, Christian Gorba, Sam Ansari, Barbara Schiller, Lutz Kampmann, Johan Postma, Alexander Haas und Elena Herzog danken. Paolo Lastrico für die Hilfe bei der Poster-Erstellung.

Für das Korrekturlesen der Arbeit bin ich Volkhard, Lutz, Alex, Hildur Palsdottir und Laura Pretsch sehr dankbar.

Spezieller Dank gilt auch Tomaso Frigato für reichlich Spass und viele sinnlose Diskussionen. Hildur danke ich für eine sehr besondere Zeit. Dem italienischem Haufen für die vielen Feste, dem Genuß am Guten und für die Freude: Gianni, Lucia, Luana, Paolo, Emanuella, Mauro und Emiliano. Lena Olkhova für den besten Vodka den ich bisher getrunken habe und die schönsten traurigen russischen Lieder. Ana Bicho und Eva Lorinci für ihre Freundschaft.

Sehr dankbar bin ich meinen Eltern, für dass was sie mir ermöglicht und gezeigt haben, meiner Mutter für die Ausdauer, meinem Vater für die Offenheit.

Mehr als dankbar bin ich meiner Liebsten, Nadine, die sehr bald auch meine Braut sein wird, für ihr Verständniss, ihre Unterstützung und ihr riesengrosses Herz!

Lebenslauf

Geboren am 1. Februar 1969 in Johannesburg / Südafrika.

1975 - 1977 Gorch-Fock Grundschule, Hamburg.
1977 - 1979 Grundschule Müssenredder, Hamburg.
1979 - 1988 Gymnasium Müssenredder, Hamburg.

Oktober 1988 - Mai 1990 Zivildienst.

WS 90 / 91 Medizin- und Umwelttechnik, Fachhochschule Hamburg-Bergedorf.

SS 91 - WS 92 / 93 Physik-Diplom, Universität Hamburg.
SS 93 - WS 95 / 96 Physik / Mathe Lehramt Oberstufe, Universität Hamburg.
SS 96 - WS 99 / 00 Physik-Diplom, Universität Hamburg.

Januar bis Dezember 2000 Diplomarbeit bei Prof. Dr. J. Bartels am Institut für theoretische Physik, DESY, Hamburg.
Thema: „Einfluss der Charm-Quark-Massen auf Wirkungsquerschnitte des BFKL-Pomerons in $\gamma^*\gamma^*$ - Streuungen am LEP“

Seit September 2000 Doktorarbeit bei Prof. Dr. Volkhard Helms am Max-Planck-Institut für Biophysik, Frankfurt.