On the effects of structure and function on protein evolution

Kristoffer Illergård
List of Papers

This thesis is based on the following papers:


III Kristoffer Illergård, Anni Kauko, and Arne Elofsson Polar residues in the core of membrane proteins are conserved and directly involved in function. *Submitted*

IV Kristoffer Illergård, Simone Callegari and Arne Elofsson MPRAP: An accessibility predictor for α-helical transmembrane proteins. *Submitted*

Reprints were made with permission from the publishers.

Other publications:


¹ Authors contributed equally to the article.
Abstract

Many proteins can be described as working machines that make sure that everything functions in the cell. Their specific molecular functions are largely dependent on their three-dimensional structures, which in turn are mainly predetermined by their linear sequences of amino acid residues. Therefore, there is a relation between the sequence, structure and function of a protein, in which knowledge about the structure is crucial for understanding the functions. The structure is generally difficult to determine experimentally, but should in principle be possible to predict from the sequence by computational methods.

The instructions of how to build the linear proteins sequences are copied during cell division and are passed on to successive generations. Although the copying process is a very efficient and accurate system, it does not function correctly on every occasion. Sometimes errors, or mutations can result from the process. These mutations gradually accumulate over time, so that the sequences and thereby also the structures and functions of proteins evolve over time.

This thesis is based on four papers concerning the relationship between function, structure and sequence and how it changes during the evolution of proteins. Paper I shows that the structural change is linearly related to sequence change and that structures are 3 to 10 times more conserved than sequences. In Paper II and Paper III we investigated non-helical structures and polar residues, respectively, positioned in the nonpolar membrane core environment of α-helical membrane proteins. Both types were found to be evolutionarily conserved and functionally important. Paper IV includes the development of a method to predict the residues in α-helical membrane proteins that after folding become exposed to the solvent environment.
Sammanfattning

Många proteiner kan beskrivas som arbetsmaskiner som ser till att allt fungerar i cellen. Deras specifika molekylära funktioner är till stor del beroende av deras tredimensionella strukturer, vilken i sin tur huvudsakligen är förutbestämd av deras linjära sekvenser av aminosyror. Därför finns det ett samband mellan sekvens, struktur och funktion hos ett protein, där kunskap om strukturen är avgörande för förståelse av funktionen. Strukturen är i allmänhet svår att bestämma experimentellt, men bör i princip vara möjlig att förutsäga från sekvensen med hjälp av beräkningsmetoder. Instruktionerna om hur man bygger de linjära proteinsekvenserna kopieras under celldelningen och förs vidare till efterföljande generationer. Även om kopieringen är ett mycket effektivt och exakt system, fungerar det inte korrekt vid varje tillfälle. Ibland kan fel eller mutationer resultera från processen. Dessa mutationer ackumuleras successivt med tiden, så att de sekvenser och därmed också deras strukturer och funktioner hos proteiner förändras med tiden.

Denna avhandling bygger på fyra artiklar om förhållandet mellan funktion, struktur och sekvens i proteiner och hur det förändras under evolutionens gång. Artikel I visar att den strukturella förändringen är linjär i relation till sekvensförändring och att strukturer är 3 till 10 gånger mer bevarade jämfört med sekvenser. I artikel II och artikel III undersökt strukturer i icke-helix struktur och polära aminosyror, respektive, placerade i den centrala opolära membranmiljön i α-helix membranproteiner. Båda typerna upptäcktes vara evolutionärt bevarade och funktionellt viktiga. Artikel IV omfattar utveckling av en metod för att förutsäga aminosyrorna i α-helikala membranproteiner som efter veckning förblir exponerade mot lösningsmedlet.
## Contents

1 Introduction

2 The mapping between sequence and structure
   2.1 Protein folding ........................................... 5
   2.2 Structure determination .................................. 6
   2.3 Structure prediction .................................... 7

3 Evolution of protein structures
   3.1 Detecting homology ........................................ 9
   3.2 Evolutionary distance ....................................... 10
   3.3 Natural selection and genetic drift ......................... 11
   3.4 How do structure and function affect evolution? ........... 12
   3.5 How do structures respond to sequence changes? ............. 14

4 Membrane proteins - introduction
   4.1 Biological membranes ...................................... 15
   4.2 What do membrane proteins look like? ....................... 16
   4.3 Which functions do membrane proteins have? .................. 17

5 Membrane protein folding
   5.1 Are membrane proteins inside-out of water-soluble proteins? 19
   5.2 Translocon-mediated folding ................................ 20
   5.3 Membrane insertion of transmembrane segments ............... 21
   5.4 Measuring free energy of insertion .......................... 22
   5.5 Assembly of transmembrane segments ......................... 22

6 Membrane protein structure prediction
   6.1 Predicting membrane protein topology ......................... 25
   6.2 Predicting solvent accessibility and residue contacts ........ 27
   6.3 Predicting full-atom structure ............................... 27

7 Membrane proteins with complex structure
   7.1 Recent structures are more complex ........................ 29
   7.2 Repositioning of transmembrane regions ....................... 30
7.3 Extending topology models in structure prediction 32

8 Methodology 33
8.1 Creation of dataset 33
8.2 Evolutionary considerations 33
8.3 Assigning features from structure 34
8.4 Predicting structural features 35

9 Summary of papers 37
9.1 Paper I 37
9.2 Paper II 38
9.3 Paper III 38
9.4 Paper IV 39
9.5 Final thoughts 39

10 Acknowledgements 41

Bibliography 43
Chapter 1

Introduction

The basic unit of all known living organisms is the cell [1]. In each cell and at every moment a large number of biochemical reactions and processes are taking place, which are responsible for keeping the cells alive. The main actors to carry out these duties within the cell are proteins of different kinds.

Enzymes are proteins catalyzing chemical reactions. They are usually highly specific and accelerate one or a few chemical reactions thousand-fold. Some proteins are involved in the process of cell signaling and signal transduction. Some proteins, such as insulin, are sent to the outside of the cells to transmit a signal from the cell in which they were synthesized to other cells in distant tissues. Others are positioned in the membrane and act as receptors whose main function is to bind a signal molecule and to induce a biochemical response in the cell. Some proteins in the membrane transport ions, metabolites and larger molecules such as proteins and RNA across their membranes. Antibodies are protein components of adaptive immune system whose main function is to bind foreign substances in the body, and target them for destruction. Transport proteins, like hemoglobin, bind usually small bio-molecules and often transport them to other locations in the body of a multicellular organism [1]. To summarize, many proteins can be described as the working machines that make sure that everything functions in the cell.

All instructions of how to build proteins and how to put together all parts into livings organisms are stored in DNA. In this way genetic information flows from DNA, via an intermediate level, RNA, to protein in a process that has been called the central dogma of molecular biology, Figure 1.1. In this process the linear sequence of nucleotides in DNA is transcribed into RNA, which in turn is translated into a linear sequence of amino acid residues [1]. The translated linear protein sequence folds, by a process that is not completely understood, into shapes that are predetermined by their sequences and the interaction with the solvent. In the folding, some parts become buried in the protein interior while other remain exposed to the solvent environment and locally contiguous segments form recurring secondary structure elements in the form of $\alpha$-helices or $\beta$-strands. The elements often pack into a com-
Figure 1.1: DNA is replicated during cell division to successive generations. Parts of the linear sequence of nucleotides in DNA are transcribed into RNA, which in turn are translated into linear sequences of amino acid residues. The translated linear protein sequences fold into structures that are mainly determined by their sequences. These structures determine the functions of the proteins.

The instructions of how to build the proteins, which is stored in the DNA, are replicated (copied) during cell division so that these instructions are passed on to successive generations. Although DNA replication is a very efficient and accurate system, it does not function correctly on every occasion. Sometimes errors, or mutations can occur in the process. These mutations gradually accumulate over time, so that species that have diverged recently usually differ by fewer mutations than those that went separate ways in the distant past, Figure 1.2. Moreover, the genetic changes in DNA are transferred via the flow of information to protein sequences, and via protein folding to protein structures and protein function. With knowledge about how these processes occur it is therefore possible to reconstruct the molecular history by comparing evolutionarily related proteins or genes.

Knowledge about the structure of proteins in the cell is important for characterizing molecular functions in e.g. complex diseases. However, experimental determination of protein structures is time consuming and represents a bottleneck in biology today. The difficulty of determining structures motivates development of computational methods to predict the structure (as well as function) from sequence. In addition, it motivates to seek an increased understanding of the relationship between sequence, structure and function in proteins and also how this relation changes during evolution.

This thesis represent the main research results and conclusions from my PhD-studies. The main focus is the relationship between sequence, structure and function during evolution of proteins. It is based on four research articles. Paper I concerns how structures respond to evolutionary changes in protein sequences and whether structures change slower than sequences or not. In Paper II, Paper III and Paper IV we investigated one special type of proteins,
Figure 1.2: Mutations are introduced in DNA and gradually accumulate over time, so that species that have diverged recently usually differ by fewer mutations than those which went their separate ways in the distant past. The molecular history can be reconstructed by studying evolutionarily related proteins or genes. This figure show a simplified phylogenetic tree of how donkey, zebra, and two different kinds of horses have evolved from a common ancestor. The phylogenetic information is extracted from data in [4].

called α-helical membrane proteins. These proteins span the whole membrane and interact with the distinct environments both inside and outside the membrane. More specifically Paper II and Paper III study non-helical structures and polar residues, respectively, positioned in the nonpolar membrane core environment, and their relation with structure and function. In Paper IV an analysis of how the properties of protein regions that after folding become exposed to distinct environments differ is presented. In addition, it includes a separate method for predicting from sequence which residues are exposed to the solvent environment.

The first two chapters of this thesis describe some of the theoretical background on how proteins fold into structures and how proteins evolve. These are necessary for understanding all four papers in general, but Paper I in particular. Thereafter, four chapters are devoted to describe the relation between structure, sequence and environment in the evolution of membrane proteins, constituting essential background knowledge for Paper II, Paper III and Paper IV. The following chapters contain a short overview of the general methodology, a brief summary of results and a discussion of my contributions to the research field.
Chapter 2

The mapping between sequence and structure

For some decades, much effort has been directed to predict protein structure from sequence by computational methods. This should in principle be possible, since the three-dimensional structures of proteins are mainly determined by their linear sequences of amino acid residues [6]. A strong motivation for development of computational methods devoted for this purpose is that the structure of a known sequence is functionally informative but difficult to determine experimentally. Although, there has been a lot of progress, the protein structure prediction problem is far from solved.

This chapter describes the process by which sequences fold to structures, how structures are experimentally determined and how they can be predicted from sequence.

2.1 Protein folding

In a series of conformational changes a fully extended and unfolded water-soluble protein often folds by lowering its free energy into a single, compact, biologically active fold. The final native state is most often believed to be the one with the lowest free energy. Its thermodynamic stability is, however, usually small and depends on the differences in entropy and enthalpy between the native state and the unfolded state [2]. For a protein molecule to adopt all possible conformations one at a time in a random fashion would take astronomically long time. However, measurements have shown that in reality a protein typically folds to the native structure in less than a second. This means that the folding must be directed in some way through a kinetic pathway of unstable intermediates to escape sampling a large number of irrelevant conformations [2].

Early in the folding process of water-soluble proteins hydrophobic residues tend to be buried in the interior so that hydrophobic side chains are brought out
Protein sequences fold into shapes that are predetermined by their sequences and the interaction with the solvent. This means that there is a “mapping” between sequence and structure.

of contact with water and into contact with each other. This hydrophobic effect, which minimizes unfavorable interactions between hydrophobic residues and water, is a strong driving force of folding of this type of protein. The burying of hydrophobic residues greatly restricts the number of possible conformations the molecule can assume and allows proteins to fold in seconds rather than years [2].

Some native structures are completely made up of α-helices or of β-sheets, whereas others consist of a mix of both types. Many of these secondary structure elements are often formed early on in the folding process of water-soluble proteins. This formation of secondary structure early in the folding process can be regarded as a consequence of burying hydrophobic sidechains and not as the driving force for the formation of a folding intermediate [2]. Finally, the native interactions are formed throughout the protein, including packing of the interior as well as the fixation of surface loops. How the folding of membrane proteins differs is discussed in later chapters.

### 2.2 Structure determination

Determination of the structure is crucial for detailed functional mapping of a protein, but is experimentally difficult. Two common techniques are X-ray crystallography and nuclear magnetic resonance (NMR).

X-ray crystallography is a method of determining the arrangement of atoms within a crystal. For proteins it usually requires many different experiments to obtain well diffracting crystals. In the method, the interaction of X-rays with electrons in the crystal is used to obtain an electron-density map of the molecule, which can be interpreted in terms of an atomic model. The quality of
the derived electron-density map depends on the resolution of the diffraction data, which in turn depends on how well-ordered the crystals are [2].

In NMR, the magnetic-spin properties of atomic nuclei within a molecule are used to obtain a list of distance constraints between these atoms, from which a three-dimensional structure of the protein can be obtained. The method does not require protein crystals and can be used on protein molecules in concentrated solutions. Historically it has, however, been restricted in its use to small proteins [2].

Experimentally determined protein structures are deposited in Protein Data Bank [7] as coordinate files. These files represent models that best explain experimental data. As with all models, these structures have uncertainty associated with them owing in part to variations in experimental conditions. An unavoidable fact is that crystallized proteins represent only a snapshot of truly flexible structures. Two identical protein sequences crystallized under different conditions may e.g. show substantial structural differences [8]. This could be the case if they e.g. are caught in different functional states, such as pre-sense and absence of bound ligands, or open or closed conformation of an ion channel.

2.3 Structure prediction

The three-dimensional structure of a protein is guided by two distinct sets of principles operating at vastly different time scales: the laws of physics and the theory of evolution [9]. From a physical point of view the protein molecule is a consequence of a variety of electrostatic interactions that under native conditions usually fold a protein into a stable well-defined structure in less than a second by minimizing its free energy. From an evolutionary point of view the protein molecule can be seen as a result of gradual changes in sequence and structure over millions of years. These two view points give rise to two different approaches to the protein structure prediction problem [10]: de novo and comparative modeling.

De novo methods predict the structure from the sequence alone, without relying on similarity between the modeled sequence and any known structure. These methods assume that the native structure corresponds to the global free energy minimum of the protein and attempts to find this minimum by an exploration of many possible protein conformations. Two key steps in de novo methods are the procedure for efficiently carrying out the conformational search and the free energy function used for evaluating possible conformations. Success in water-soluble protein structure prediction has so far been restricted to rather small proteins.

The aim of all comparative modeling methods is to build a model for a target protein of unknown structure using similarity to a template protein of known structure. In principle it consists of four steps: i) finding a template of known
structure to the target sequence to be modelled, ii) aligning the sequence to the template, iii) building a model and iv) assessing the model. The quality of the model is largely dependent on the similarity in sequence between the target and the template, and therefore also on how much information can be transferred from template to target in the model building process [9].

Despite recent progress in the de novo prediction field, the comparative modeling approach is by far more accurate than de novo predictions. Since the quality of the structures determines the information that can be extracted from them, comparative modeling is therefore always the recommended choice when a close homolog of known structure is known [9].
Chapter 3

Evolution of protein structures

The genetic changes in DNA are transferred via the flow of information to protein sequences, and via protein folding to protein structures and protein function. With knowledge about the process, by which evolutionary changes occur, it is possible to reconstruct the molecular history by comparing evolutionarily related proteins or genes [3].

This chapter describes i) the processes of detection of homology between proteins, ii) measurement of evolutionary distances between them, iii) how structure and function affect evolution and iv) finally estimation of how structures respond to evolutionary changes in sequence.

3.1 Detecting homology

Since proteins gradually change, two proteins can be evolutionarily related and have a common ancestor, Figure 3.1. Per definition, two proteins are said to be homologous if they share a common ancestor. The concept of homology is central in molecular biology and is often inferred between proteins if the similarity in sequence or structure is high. Often, the publically available repositories of known protein sequences and structures are scanned in automatic ways to detect homologous proteins [5].

In the protein homology detection process it is important to create alignments, which are sets of protein sequences that specifies the correspondence among the residues of those sequences. An evolutionary alignment is needed e.g. to be able to reconstruct ancient evolutionary changes. In this kind of alignment every pair of characters at a site (meaning within a column of the alignment) are to be interpreted as having evolved from an ancestral residue in an ancestral sequence. Regions of either sequence left unpaired in this alignment are interpresented as having been inserted or deleted since the divergence from their common ancestor [3].
Figure 3.1: Divergence of two proteins from a common ancestor. Changes in protein sequences are transferred to protein structures and protein function. Some changes in sequence result in major changes in structure and function (a), whereas other changes only result in minor changes in structure and function (b).

For proteins a distinction can be made between sequence alignments, which are derived from only protein sequences, and structural alignments, which are derived from protein structures.

A pairwise sequence alignment method often optimize the evolutionary alignment by optimizing over a molde of possible evolutionary histories of two sequences assumed to have diverged from a common ancestor. The alignment optimization is calculated with the assumption that different sites evolve independently of one another, and that the sequences have a certain common and specific composition and pattern of substitution [5].

Protein structural alignment methods assign a correspondence among the structurally similar parts in the structural models of two or more protein structures. In general, they optimize a measure of structural similarity of parts, rather than any evolutionary quantity. One structural similarity measure commonly included in the optimization score is root mean square deviation of the atomic coordinates of the superposable regions. The structural alignment might be represented in the same way as a sequence alignment, i.e. with a number of residue correspondences [11].

3.2 Evolutionary distance

When an alignment between two homologous proteins has been constructed it is often of interest to determine how much change that has occurred between these two sequences since divergence. The simplest measure of the distance between two sequences is to count the number of sites in the alignment at which the two sequences differ. Often, the fraction of identical or
Figure 3.2: The observed number or fraction of non-identical sites in a pairwise alignment is easy to calculate. However, since multiple substitutions can occur at the same site this measure underestimates the actual number of substitutions that have taken place since divergence from the common ancestor. The figure is partly adapted from [3].

non-identical sites is used. Although, this kind of measure is easy to calculate it is a poor measure of the actual number of evolutionary changes. The reason is that multiple substitutions can occur at the same site, which results in underestimation of the actual number of substitutions that have taken place, Figure 3.2. The sequence identity measure saturates as more substitutions are accumulated and severely underestimates the sequence divergence at larger evolutionary distances [3].

There has been a considerable amount of research on developing methods to estimate the actual number of evolutionary changes in DNA sequences. Most methods are based on four assumptions, which in many cases might be unrealistic: (i) all sites change independently, (ii) the substitution rate is constant over time and in different lineages, (iii) the sequence compositions is at equilibrium, (iv) the conditional probabilities of substitutions are the same for all sites and do not change over time. In addition, some methods model a variation in substitution rate among sites, most often by discretizationing a gamma distribution [3].

There are relatively few methods for computing distances for protein sequences, which is probably largely due to the greater complexity of the problem; instead of four states required for DNA data, protein sequences comprise some 20 different amino acids.

3.3 Natural selection and genetic drift

A central task of molecular evolution is to determine the processes by which gene sequences change over time. Today we know that the specific molecular functions associated with proteins are the result of randomized processes such as mutation and genetic drift and organizing processes like natural selection.
Natural selection is the process by which heritable traits that make it more likely for an organism to survive and successfully reproduce become more common in a population over successive generations. Mutations that are not subject to natural selection are called neutral and are lost (usually) or fixed (very occasionally) by the process of genetic drift. Non-neutral mutations can be of two types; advantageous for the organism, which are subject of positive selection, and disadvantageous, which are subject of negative selection [3].

Reconstructions of ancient mutations that have survived evolution and become fixed in populations have shown that most common types of changes within proteins are short (non-frame shifting) insertions or deletions and replacements of single amino acid residues that conserve the hydrophobicity [5]. Among these mutations, the majority can be believed to have been neutral or nearly neutral, whereas most of the deleterious ones have been selectively removed [3].

Although most evolutionary changes in proteins are believed to occur by small changes that are neutral for structure and function there is evidence that other types of changes also occur. It has e.g. been shown that a few non-neutral changes in protein sequences can cause dramatic changes to protein structures and functions with large evolutionary consequences [19, 20]. However, it is not known how common and important such events are.

### 3.4 How do structure and function affect evolution?

In proteins, some sites evolve faster than others. From multiple sequence alignments the types and rates of substitutions can be reconstructed. Earlier studies have claimed that there is an inverse correlation between the rate of substitution and the functional constraint acting on a site [3]. Indeed, it is well-established that the slowest evolving sites are most often directly functionally important and might be responsible for e.g. binding of ligands or catalysis.

If the structure of one of the proteins in an alignment is known it is possible to “map” the features from the structure to the alignment (or vice versa). By this kind of mapping it is e.g. possible to investigate whether local structures of distinct secondary structure or surface accessibility classes exhibit different amino acid substitution patterns and rates of substitution, Figure 3.3. The most evident such trend for water-soluble proteins seems to be that residues at exposed sites are often hydrophilic and replaced at a faster rate than the often hydrophobic and buried sites are. Similarly, residues at positions of non-regular secondary structures (loops and coils) are replaced at a faster rate than regular secondary structure (α-helices and β-sheets) [12, 13].

Similarly, if the function of one of the proteins in an alignment is known it is possible to “map” the functional features with “evolutionary features”. Functional annotations of specific residues in well-characterized proteins are, however, most often restricted to just a few sites that are responsible for e.g
binding of ligands, while the majority of sites lack information. One observation of the functionally important sites is, as previously mentioned, that they often are evolutionary conserved. Another observation, is that they often are found in rare, unfavorable or unusual energetic environments, such as left-handed helices [14], ionizable groups with perturbed titration curves [15], and polar residues engaged in unfavorable electrostatic interactions in enzyme active sites [16]. Therefore, it has been suggested that residues involved in function often contribute unfavorably to the stability of the native state and that evolution often optimizes functional properties at the expense of thermodynamic stability [17, 18]. However, the generality of the rule is not known.

Another application of mapping the structure or function of one protein to an alignment than correlating the structural and functional features with substitution patterns is to infer them to residues of unknown or less-characterized structure and function. This is done by assuming that the structural or functional features are similar in regions that correspond to each other in the alignment (i.e. being in the same column in the alignment). This type of inference by homology is commonly used in structure and function prediction and is in general more accurate the more similar the local protein sequence regions are. However, the local structure and function might occasionally differ drastically even at conserved sites.

Figure 3.3: If the structure and function of one of the proteins in an alignment is known it is possible to “map” the features from the structure and function to the alignment (or vice versa). Structural and functional features could be e.g. secondary structure and ligand binding sites. Features derived from a multiple alignment could be e.g. type and rate of substitution. Both type of features can be done at a residue level.
3.5 How do structures respond to sequence changes?

Since the three-dimensional structures of proteins is mainly determined by the amino acid sequences [6], the structure will change in response to changes in the sequence. The extent of structural perturbation in response to sequence evolution will depend on type and location of the changes. Some single mutations will completely disrupt the structure, while others that conserve the physico-chemical properties of the sequence will barely affect the structure at all [21].

One way of studying how much structures change for different numbers of sequence changes on average, is by aligning a large number of homologous protein pairs and compare there structural similarity at different sequence similarities. Proteins with closely related protein sequences will most often have very similar structures, whereas these distantly related sequence will differ more in ther structures. The relation between structural similarity and sequence similarity was first reported to be exponential [22], but has since been shown to depend on which similarity measures that are used and can also be linear [23]. Proteins that might be problematic to study in this way are those that partly or completely consist of regions that do not fold into well-defined structures or proteins with several functional states with quite different structures.

One claim that is often used in discussions about proteins is that “structure is more conserved than sequence.” However, what is meant by that statement is not often clearly stated. A change in sequence is discrete and represents a well-defined event, while a corresponding change in a native state of a flexible structure is continuous and depends on the particularities of the structural measure used. Therefore, comparing structural divergence and sequence divergence with their different dimensionalities, characteristics, and systematic sources of noise and measurement bias can be like comparing apples and oranges.
Chapter 4

Membrane proteins - introduction

There are two main types of solvent environments in which proteins perform their functions. The first environment is hydrophilic, with water as the most important component and the second is hydrophobic, with lipids as the main component. The proteins in these environments are in the thesis called water-soluble proteins and membrane proteins, respectively.

This chapter describes what a membrane is and what the main types of structures and functions of membrane proteins are. It serves as an introduction to membrane proteins.

4.1 Biological membranes

A biological membrane functions as a barrier between living cells and their environment, and to compartmentalize intracellular organelles within eukaryotes. Therefore, it needs to be both mechanistically strong and flexible. It also needs to be impermeable to compounds that are unwanted in the cell and to have mechanisms for passage of desired compounds into the cell [24].

The fundamental structural component of biological membranes is lipid. This lipid molecule often consists of a hydrophilic headgroup, and one or more hydrophobic fatty acid hydrocarbon tails. As a consequence of their amphipilic nature, lipids aggregate and spontaneously self-organize in water, generally into either micelles or bilayers. The biological membrane is a bilayer, where the hydrophobic fatty acid chains point away from the hydrophilic water molecules, arranged into two oppositely oriented fluidic leaflets [24], Figure 4.1.

The lipid composition varies between membranes and determines the thickness, fluidity and curvature of the bilayer. The actual distribution of the functional groups across a 1,2-Dioleoyl-sn-Glycero-3-phosphocholine (DOPC) lipid bilayer have been measured experimentally [25], and simulated
by molecular dynamics [26], Figure 4.1. The emerging picture is a highly heterogeneous environment, where the chemical environment across the membrane varies markedly over short distances.

4.2 What do membrane proteins look like?

In biological membranes, proteins are interspersed among the lipids and as much as half of the mass fraction might consist of proteins embedded within the lipids [24]. Broadly, membrane proteins can be classified as peripheral or integral, Figure 4.2 a. Peripheral membrane proteins are only loosely attached to the membrane through electrostatic or van der Waals interactions with the lipid headgroups or other membrane proteins, or through a covalently attached anchor [24]. Integral membrane proteins most often span the whole membrane, and can be divided into two type of structures; i) closed barrels of amphipathic β-strands and ii) as bundles of tightly packed α-helices, Figure 4.2 b.

β-sheets are used in the β-barrel type of transmembrane proteins. Residues in the β-sheets alternatively point outwards, facing the lipids, and inwards, facing the inside of the barrel, resulting in a sequence pattern in which the residues are typically alternatively polar and hydrophobic. The outcome is a polar channel through which water-soluble molecules can cross. It has been estimated that around 2–3 % of the genes in bacteria encodes β-barrel membrane proteins [28] and that they can be found in the outer membrane of mitochondria and chloroplasts.
Figure 4.2: Different types of membrane proteins. a) Membrane proteins can be divided into peripheral and integral. b) Integral membrane proteins can be divided into \( \alpha \)-helical bundle and \( \beta \)-barrel types.

The \( \alpha \)-helical bundle type of membrane proteins is the most abundant and well-studied. The \( \alpha \)-helices are in general longer than the \( \beta \)-strands of \( \beta \)-barrel proteins and are easier to predict from sequence. A recent study estimate that, in humans, this type of membrane protein constitute around 27\% of the genes that codes for membrane proteins [29].

Hereafter, the term “membrane protein” will refer to \( \alpha \)-helical membrane proteins.

4.3 Which functions do membrane proteins have?

One of the most effective way to increase the knowledge of specific functions of proteins is to determine their structure. In the case of membrane proteins, it is experimentally particularly difficult and have led to a 100-fold underrepresentation in structural databanks compared to non-membrane proteins [30]. Although, the number of experimentally determined structures grows exponentially, the number of available sequences grows faster. Current estimates predicts that it will take at least three decades before 80\% of the membrane superfamilies will have at least one determined structure, which can be used as template for comparative modeling [30]. This means that there is a huge need of structural information about membrane proteins at the moment and probably for a long time in the future.

Membranes contain a substantial amount of proteins with diverse functions, Figure 4.3. They transport ions, metabolites and larger molecules such as proteins and RNA across their membranes. Membrane proteins are also responsible for sending and receiving chemical signals, to propagate electrical impulses, attach to neighbouring cells, anchor other proteins to specific locations in the cell. Other functions include regulating intracellular vesicular transport, controlling membrane lipid composition, organizing and maintaining the shape of organelles and the cell itself [24].
Due to their important functions, membrane proteins are highly interesting from a medical point of view. Cystic fibrosis is caused by misfolding mutations in a chloride transporter protein in the lung. Neurological and cardiac diseases are caused by nonfunctional photoreceptors (rhodopsins) in the eye. Water-channel proteins (aquaporins) are critical for kidney function but also involved in stroke. Finally, mutations in membrane proteins involved in protein import into organelles such as peroxisomes and mitochondria underlie many serious (but fortunately rare) heritable disorders [31]. Due to this medical relevance, many membrane proteins are prime drug targets, and it has been estimated that more than half of all drugs currently on the market are directed against membrane proteins.
Chapter 5

Membrane protein folding

The solvent environments and the folding process of membrane proteins is different from water-soluble proteins.

The following chapter describes how linear amino acid sequences of membrane protein fold into three-dimensional structures. The process can be described according to a two-stage model with initial formation of transmembrane-helices followed by assembly of transmembrane regions into tertiary structure.

5.1 Are membrane proteins inside-out of water-soluble proteins?

In the folding of a protein some residues will remain exposed to the environment while others will become buried in the protein interior. For water-soluble proteins it is energetically favorable to bury hydrophobic residues and expose polar and charged residues to the surrounding water. However, transmembrane proteins face three distinct environments; a hydrophobic lipid environment inside the membrane, a hydrophilic water environment outside the membrane and an interface region rich in phospholipid headgroups. Because of the energetic conditions resulting from these distinct environments the accessible surfaces of transmembrane proteins need to expose different types of residues at different locations.

As previously explained, the dominant driving force behind folding of water-soluble proteins is the hydrophobic effect, that minimizes unfavorable interactions between hydrophobic residues and water [32]. Therefore, water soluble proteins generally have a hydrophobic interior and hydrophilic exterior. In contrast, the drive to bury polar residues from the solvent environment within the membrane is much weaker. Early studies of the bacteriorhodopsin structure suggest that membrane proteins are “inside-out” , i.e. that they consist of a hydrophilic interior and a hydrophobic exterior, Figure 5.1. However, later studies indicated that the “inside-out” rule is not
Figure 5.1: Water-soluble proteins (a) generally have a hydrophobic (orange) interior and hydrophilic (green) exterior. It has been suggested that membrane proteins (b) are inside-out of water-soluble proteins, with a hydrophilic interior and hydrophobic exterior. However, such a (inside-out) rule is generally not applicable to all membrane proteins.

generally applicable to all $\alpha$-helical membrane proteins [33, 34, 35]. Since membrane proteins are exposed to distinctly different chemical surrounding across the membrane that varies markedly over short distances, they will have to adapt their surface to the different environments. Therefore, the main driving forces of folding and stabilization are different and less understood for membrane proteins. However, it is likely that stably folded proteins in membrane regions also reside in a free energy minimum [35].

5.2 Translocon-mediated folding

Most $\alpha$-helical transmembrane proteins are targeted by an N-terminal signal sequence to the endoplasmatic reticula membrane. For such proteins, the ribosome, which is the unit that synthesizes the polypeptide sequence, binds to a channel complex known as the Sec translocon during translation. Thereafter, translocation is resumed and the nascent molecule is threaded directly through the channel as it is being synthesized [36]. In bacteria, the mechanism for insertion of proteins into the plasma membrane is similar, whereas insertion of proteins into the bacterial periplasm or the outer membranes of mitochondria differs. Since protein translation and translocation of membrane insertion occur simultaneously the process is termed co-translational [24].

The translocon apparatus can be seen as a switching station [36] that receives elongating peptide sequences from the ribosome, and directs them either across the membrane or into the membrane bilayer, depending upon the segment’s properties, Figure 5.2. Whether the molecule is folded or unfolded
Figure 5.2: The translocon apparatus can be seen as a switching station that receives elongating peptide sequences from the ribosome, and directs them either across the membrane or into the membrane bilayer, depending upon the segment’s properties. The figure is from [27] with ideas influenced by [37].

upon exit from the ribosome tunnel and the molecule is entering the translocon channel is not known [36].

5.3 Membrane insertion of transmembrane segments

Somehow the ribosome-translocon complex manages to recognize the incipient transmembrane helices in the nascent polypeptide and allows them to insert into the lipid bilayer rather than being fully translocated across the membrane into the lumen of the endoplasmatic reticulum. Most probably the transmembrane segments leave the channel through a lateral gate in the channel wall that opens sideways toward the bilayer. In a high-resolution structure of a translocon one can see a possible opening from the interior into the lipid bilayer, which is hypothesized to control the passage of nascent transmembrane helices into the bilayer from the interior of translocon [36].

It has been assumed that transmembrane helices move into the lipid membrane from the protein-conducting channel by a partitioning process, where sufficiently hydrophobic helices prefer the bilayer, whereas more polar helices favor the translocon. Therefore, the process would appear as an equilibrium process in which favourable interactions between lipids and amino acid side chains promote membrane integration rather than translocation [36]. The efficiency with which a transmembrane segment is integrated by the translocon should then be sensitive to the relative positions of amino acids within the segment. This trend is evident when looking at amino acid distributions in known membrane protein structures.
Figure 5.3: An experimental model to determine insertion efficiency of designed putative transmembrane segments in the membrane [38]. With the model, the requirements in terms of amino acid composition for a sequence segment to be recognized as a transmembrane segment and inserted into the membrane, as opposed to being translocated across can be investigated. The figure is adapted from [38].

5.4 Measuring free energy of insertion

Recently, von Heijne and co-workers have made progress in describing the requirements in terms of amino acid composition for a sequence segment to be recognized as a transmembrane segment and inserted into the membrane, as opposed to being translocated across [38]. The experimental data comes from insertion efficiency measurements of designed putative transmembrane segments in a model membrane protein, Figure 5.3. A large number of designed and natural test segments were systematically introduced in the experimental setup. From the measured relative amounts of inserted and translocated forms of the proteins, an apparent free energy of membrane insertion of the segment, \( \Delta G_{\text{app}} \), was calculated. A model for predicting the membrane insertion efficiency of transmembrane segments was derived, which takes amino acid position, length of the segment and hydrophobic moment into account [38].

The experiments showed that strongly polar residues give highly position-specific contributions to membrane insertion, and are much better tolerated towards the water-lipid interface regions than inside the hydrophobic core. Hydrophobic residues give favorable contributions and are less position specific [38].

5.5 Assembly of transmembrane segments

The selection of transmembrane segments is only the first step in the complex process of gathering the transmembrane segments together to form the native protein structure. After translocon-guided insertion, the transmembrane segments of multiscanning membrane proteins must condense to form the final folded three-dimensional structure.
For single-spanning proteins it seems that a “threshold hydrophobicity” can be defined with the experimentally derived $\Delta G_{app}$-scale that determines whether or not a given polypeptide segment will form a transmembrane segment. However, this kind of approach does not appear to work for multi-spanning membrane proteins. Actually, a surprisingly large fraction (25 %) of the transmembrane helices in the multi-spanning proteins of known three-dimensional structure have $\Delta G_{app} > 0$ kcal/mol, which suggest that they are not sufficiently hydrophobic to insert efficiently as single transmembrane segments [39].

The relatively large fraction of transmembrane segments that cannot insert efficiently into the membrane in the absence of other parts of the protein suggests that some transmembrane segments in multi-spanning proteins may depend on other parts of the same protein for efficient insertion. Possibly pairs or higher assemblages of transmembrane helices interact among themselves before partitioning into the bilayer [36].
Chapter 6

Membrane protein structure prediction

The need of computational methods for predicting structures from sequence is even more pronounced for membrane proteins than for water-soluble proteins [24, 31, 40]. As was described in the previous chapter, membrane proteins can conceptually be decomposed into two consecutive steps: folding of the individual hydrophobic segments into transmembrane segments, often corresponding to helices, followed by association of the different transmembrane segments. In this chapter it will be described that structure prediction of α-helical membrane proteins can, similarly, be broken down into two steps: (i) determining the topology of the proteins, which includes delineating the boundaries of the transmembrane segments and (ii) predicting the tertiary conformation of the protein, which includes determining how the transmembrane segments pack together [41].

6.1 Predicting membrane protein topology

The topology of membrane proteins refers to a description of the way in which a polypeptide chain weaves back and forth across the membrane [40]. The simplest description is a list of the different segments of the polypeptide chain that form the transmembrane helices and their orientation relative to the membrane. The concept of topology is widely used in membrane protein structure prediction. Its determination can be a crucial preliminary step to modeling the structure as it constrains the way individual transmembrane segments could associate within the membrane [40].

The first predictors for α-helical transmembrane proteins calculated segmental hydrophobicity averaged over 10 to 20 residues along the sequence. They used the fact that transmembrane helices are on average more hydrophobic than the connecting loop regions. First, only the so called topography was predicted, meaning that no attempt was made to predict the orientation of the
Figure 6.1: The membrane protein Bacteriorhodopsin represented as a topology (left) and as a full atom structure (right). Structure prediction of \( \alpha \)-helical membrane proteins can be broken down into two steps: (i) determining the topology of the proteins, which includes delineating the boundaries of the transmembrane segments and (ii) predicting the tertiary conformation of the protein, which includes determining how the transmembrane segments pack together.

protein. The topographic signal based on the segmental hydrophobicity seems to be a rather symmetric property and depends mainly on the distances from the membrane center [40, 41].

With the observation that the loops connecting the transmembrane helices differ in amino acid composition, depending on whether they face the inside or the outside of the cell, it became possible to extend predictions to include the orientation of the proteins [42]. This, so called positive inside rule [43], is still widely used although it is still unclear whether it is the result of properties of the translocon or the cytoplasmic membrane [41].

Later significant improvements of topology predictions is to include information from evolutionarily related sequences and the use of machine learning techniques, such as Hidden Markov Models, to extract the relevant sequence features [40, 41]. Today the best topology predictors can assign the correct number of transmembrane regions, their approximate locations and the correct side of the termini with an accuracy of 70–90 % [40]. Some problems that remain difficult is how to predict the exact location of transmembrane regions, to distinguish between signal peptides and transmembrane regions, to identify marginally hydrophobic transmembrane segments and to predict the topology of more complex structures.
6.2 Predicting solvent accessibility and residue contacts

Most efforts in structure prediction of α-helical membrane proteins have been focused at improving the predictions of the topology [40]. Although, a predicted topology might be a useful initial step, it does not give any information of how different segments in the membrane are packed together into bundles in the tertiary structure. Two common steps in structure prediction of water-soluble proteins to aid prediction of the packing of different segments are prediction of solvent accessibility and residue contacts. These structural features are in the case of membrane proteins complementary to the topology.

A solvent accessibility predictor assigns which residues are exposed to the solvent environment in the native structure, and which are buried in the protein interior. Residues responsible for catalysis or substrate binding, positioned within the membrane region are often partly buried in the protein interior, while temporary protein-protein-interactions naturally occur on solvent exposed sites. Therefore, a predictor of solvent accessibility is interesting for predicting functional relevance of individual residues. The attempts, hitherto, made to predict accessibility of membrane proteins [44] have been developed to predict the exposure within the membrane environment and, thus, require an initial prediction step for determining if a residue is located within the membrane or not. Such an initial prediction step is coupled with errors at the residue level, since the exact start and stop points of the transmembrane segment is hard to assign correctly.

A contact predictor assigns which residue pairs are in close contact in the structure. One common way to identify contacts is to look for sites that undergo similar or compensating substitutions. Such sites have been shown to be close in structure more than randomly picked pairs of sites. Since networks of correlated mutations have been found to appear near binding regions and active sites [45], identification of these could potentially be used to identify functionally important regions. Predicting contacts in water-soluble proteins has successfully been utilized to constrain structure prediction and to identify incorrect models generated by structure prediction methods. However, for membrane proteins this type of predictions have received little attention [46]. Recently, there have been a few attempts with varying success to predict contacts in membrane proteins with the aim of identifying interacting α-helices [46].

6.3 Predicting full-atom structure

Predicted structural features like topology, surface accessibility and residue contacts might be useful, but, the goal of structure prediction is to get a full-atom model. Structure prediction can, as previously described, be divided into
de novo prediction and comparative modeling. In the case of membrane proteins, both methods are quite difficult, which might explain the effort to develop methods for predicting structural information of low resolution, like topology.

In principle de novo prediction of membrane proteins could be done with the same methods and protocols as for water-soluble proteins. However, there are two main differences in which the predictions of membrane proteins differ from water-soluble proteins; i) the solvent environment is different and more complex and ii) many membrane proteins are large [40, 41]. As an example of the latter, G-protein coupled receptors contains 7-transmembrane regions and several hundred residues per chain. Another example is the photosynthetic reaction center which consist of many chains that constitute a very large complex. Historically, succesful de novo predictions has been limited to rather small proteins. Together, these difficulties with predictions of membrane proteins have effectively restricted the development so far. However, recently Rosetta, which has been successful for water-soluble proteins by using a fragment-based assembly algorithm, was adopted for transmembrane protein structures [47]. Another, interesting attempt is FILM [48], which similarly might be seen as a adaption for membrane proteins of a previous succesful method.

The most obvious problem for comparative modeling of membrane proteins is the lack of templates with known structures. Many methods have hitherto used quite similar protocols as for prediction of water-soluble proteins [49]. An interesting attempt to model the structures of most of the 900 human G-protein couple receptors was recently performed by Skolnick and coworkers using the TASSER algorithm [50], resulting in quite good accuracy of the predicted rhodopsin structure. However, as for other methods the correctness of the structures can not be verified until more structures are available [40, 41].

Topology, surface accessibility, residue contacts and other structural features can be interpreted as low-resolution data. Predictions of such features can be used as constraints for “complete” structure prediction of all atoms. Alternatively, low resolution data can be derived experimentally from e.g. site-specific mutagenesis, chemical crosslinking, NMR, massspectrometry, glycosylation or binding affinities, protein-protein interactions. Low-resolution data can also be used to evaluate agreement between full atom models and predicted features [40, 41]. Combining different types of low-resolution data in structure prediction might be a field that will grow in the future.
Chapter 7

Membrane proteins with complex structure

The folding of \( \alpha \)-helical membrane proteins is often described as a two-step process, where the insertion of individual transmembrane helices into the lipid bilayer in their energetically most favored position is followed by a folding process where the formed helices find their optimal packing interactions. While the two-stage model is still a useful first-order approximation of the folding process, both structural and biochemical studies have begun to unravel a more complex reality [51]. Many recently determined structures of \( \alpha \)-helical membrane proteins have a complex type of structure, possibly also more complex folding and reaction cycle than the first structures that were determined.

This chapter describes the discovery that has highlighted that many proteins do not fit in the old models for folding and structure prediction of \( \alpha \)-helical membrane proteins.

7.1 Recent structures are more complex

For a long time the general view was that membrane proteins form simple helix bundles, with their hydrophobic transmembrane helices crisscrossing the membrane in more or less perpendicular orientations. Indeed, many membrane proteins, abide by this principle. Bacteriorhodopsin is one of them, Figure 7.1. It is a seven-helix bundle and functions as a light-driven proton pump: small, light-induced movements in its transmembrane helices entice protons to translocate across the membrane against an electrochemical gradient. The helices lie almost straight in the membrane and pack with typical knobs-into-holes packing angles [36].

However, some more recently solved membrane protein structures show that reality is not always this simple. Membrane-embedded helices can be short, long, kinked or interrupted in the middle of the membrane, they can
Figure 7.1: Illustration of the concept of membrane topology, which is a description of the way in which a polypeptide chain weaves back and forth across the membrane [40]. Bacteriorhodopsin (right) represents a protein with simple membrane topology. Glutamate transporter homolog (left) represents a protein with a complex membrane topology. It contains e.g. kinked (yellow) and breaked (green) transmembrane helices, reentrant regions (red) and coils in the deep membrane core (blue).

be almost perpendicular to the membrane plane, be strongly tilted, lie flat on the surface of the membrane, or even span only a part of the membrane and then turn back. In addition, some do not pack to each other according to simple “knobs-into-holes” geometries [36]. The glutamate transporter homolog is one protein with many such deviations and with a complex topology, Figure 7.1. This protein contains long steeply inclined helices and short, closely spaced pairs of helices that penetrate only halfway through the membrane (forming so-called reentrant loops). There are also stretches of non-helical structure deep within the membrane that are largely buried between the transmembrane helices [40].

7.2 Repositioning of transmembrane regions

The underlying assumption in many topology predictions is that the most hydrophobic sequence segments are centered in the membrane core. However, the positioning in the membrane of transmembrane regions in the folded structure does not always correspond to the thermodynamically favored position in the membrane of the isolated region. This is e.g. the case for some of the transmembrane segments in glutamate homolog transporter, Figure 7.2. Long-range tertiary interactions might make it more energetically favorable for transmembrane regions to have an altered (or “shifted”) position in the membrane. Occurrence of such shifts suggest that transmembrane regions may undergo rather dramatic repositioning in the membrane during the folding and oligomerization process. From the point of view of structure prediction, repositioning of transmembrane regions may explain why topology pre-
Figure 7.2: The topology of Glutamate transporter homolog. The underlying assumption in many topology predictions is that the most hydrophobic sequence (green) segments of length 19-23 residues are centered in the membrane core. However, the positioning in the membrane of transmembrane regions in the folded structure does not always correspond to the thermodynamically favored position in the membrane of the isolated region.

Predictions are at best mediocre at predicting the exact length and location of transmembrane regions as they are found in high-resolution structures [40, 36].

Another expectation is that the transmembrane region is hydrophobic. However, as previously mentioned a large fraction of the transmembrane segments in known multispansing structures are marginally hydrophobic and contain many polar residues in the membrane core. How, and at which stage during folding, these regions are inserted into the membrane is not completely understood. Neither is it known at which stage during folding reentrant regions and residues in coil regions in membrane core are formed. A better understanding regarding these issues of the folding processes will probably increase structure prediction performance [40, 36].

Observations of transmembrane and reentrant regions that are in thermodynamically disfavored positions in the membrane in the folded structure suggest that membrane proteins are rather dynamic entities in the early folding stages and that repositioning of transmembrane regions might be common [36]. Larger repositioning might also occur after folding during the reaction cycle of the proteins. Small-molecule transporters must flip between drastically different conformations so that the protein are open either towards the external or the internal side of the cell. Large-scale conformational changes, including repositioning of transmembrane helices, have e.g. been documented by X-ray crystallography during the catalytic cycle of the Ca²⁺ ATPase [36]. Most often the structure of only one of the states is available.

Other membrane proteins form stable structures with little flexibility. Proteins involved in proton and electron transfer typically coordinate a range of cofactors that need to be positioned relative to each other with high precision and hence must be quite rigidly packed [36].
7.3 Extending topology models in structure prediction

As noted above, the simplest topology model with hydrophobic perpendicularly penetrating transmembrane helices is too simplified to describe the recently determined complex membrane protein structures. Most topology prediction methods are not suited to identify transmembrane regions that are marginally hydrophobic (rich in polar residues), strongly tilted, kinked or interrupted by coils, or identify reentrant regions or helices that lie flat on the membrane surface. Recently, however, methods have been developed to predict presence of reentrant regions [52, 53] and proline-induced kinks [54]. An interesting alternative to topology prediction is to predict the distance from the membrane center for each residue in a protein [55].
Chapter 8

Methodology

This chapter describes which databases and evolutionary considerations that were used to create the datasets in the investigations. Further, it explains how structural annotations were derived from known structures and predicted annotations from sequence-derived information.

8.1 Creation of dataset

The included analysis datasets of determined structures were SCOP 1.74 [56] (Paper I) and Orientation of Proteins in Membrane (OPM) [57] (Paper II, Paper III and Paper IV).

In SCOP, protein domains of known structures are divided into classes like “All-α”, “All-β”, “α-β (a/b)” and “α-β(a+b)” and fold descriptions like “layer”, “sandwich”, “barrel” and “bundle”. Based on evolutionary relatedness, the structures are also hierarchically divided into superfamilies and families.

OPM [57] is specialized on membrane protein structures. The database has classified the structures into superfamilies and families as well as included functional information. The database have, by energetic calculations, oriented the structures so that the predicted membrane center are located at the X-Y plane. This means that the absolute value of the Z-coordinate corresponds to the distance from the membrane center.

8.2 Evolutionary considerations

In Paper I, the homology among protein pairs was inferred from SCOP annotations if they belong to they same superfamily or family. Pairs of proteins were structurally aligned with STRUCTAL [58]. From the resulting alignments the evolutionary distance (average number of substitutions since divergence of common ancestor) were estimated using the maximum likelihood
estimates with the JTT+Γ substitution model [59], allowing site-rate heterogeneity, computed using Tree-Puzzle v.5.2 [60].

In Paper II, Paper III and Paper IV, the local sequence alignment program PSIBLAST was used to identify homologous proteins to the ones of known structure in OPM. In this process, proteins within the database Uniprot90, that contain many proteins of unknown structure, with high sequence similarity to the known structures were identified as evolutionary related. From the multiple alignments, amino acid substitution (replacement) rates within each protein were estimated using the program rate4site [61]. The program computes a phylogenetic tree from the alignment and uses the stochastic process underlying sequence evolution within protein families to calculate a conservation score using an empirical Bayesian method for each residue. In Rate4site, the rate (conservation score) of each site (residue) i in protein j is normalized by subtracting the average conservation score and dividing by the average standard deviation of the reference protein j.

To avoid overrepresentation of proteins from a few evolutionary families the data must often be homology reduced. This was done after some initial filtering based on structural quality. The main dataset in Paper I was reduced by taking one protein pair per family or superfamily, respectively. The homology reduction of the membrane proteins in OPM in Paper II, Paper III and Paper IV used the program CD-hit [62] to select representative structures with maximum 40 % sequence identity to each other. In addition, Paper III and Paper IV also took OPM superfamily annotations into account to make sure to only pick maximum one representative structure from each superfamily.

8.3 Assigning features from structure

The assignment of protein secondary structure and relative surface accessibility from coordinates can be done by fast and consistent automatic computational methods. Secondary structure assignment were, Figure 8.1, assigned with STRIDE (Paper I) and DSSP [63] (Paper II, Paper III and Paper IV). The assignment of the relative surface accessibility of residues in a protein can be calculated from the coordinate file by rolling a probe over the whole structure. The absolute calculated area per residue is normalized in order to get the relative surface accessibility, for example by an extended Ala-X-Ala tri-peptide conformation or the maximum surface accessibility. In Paper I, Paper II, Paper III and Paper IV the surface accessibility was calculated by Naccess 2.1.1 [64] with the probe size 1.4 Å, corresponding to the size of water. The output from all three programs gives structural states for each residue in the polypeptide sequence. The many secondary structural states given by STRIDE and DSSP were merged into three classes; helix, sheet and coil, and the accessibility values were classified into two groups, buried and exposed.
If the structure is known, features such as secondary structure can be assigned from coordinates by fast and consistent automatic computational methods. This or other types of features can also be predicted from sequence-derived information.

In Paper II, Paper III and Paper IV the Z-coordinates from OPM were used to classify all residues into different groups: membrane core region, water-lipid interface region and water-soluble region (outside the membrane). The topology was automatically classified from the structures according to a protocol as described in [65]. Here each residue is classified to belong to one out of four topological states: transmembrane, reentrant, inside or outside.

In all these methods one annotation per residue is obtained in an automatic and objective way. This made it possible to e.g. analyze the distribution of amino acids, surface accessibility, secondary structure as a function of Z-coordinate.

### 8.4 Predicting structural features

A few features were predicted by sequence-derived information, Figure 8.1. These features were Z-coordinate (by Zpred [55]), topology (by Octopus [52]) and free energy of membrane insertion (by ΔG-predictor [38]). The ability to predict the relative solvent accessibility and secondary structure was investigated using the machine learning technique called support vector machines, as implemented in the svmlight package [66]. During evaluation and optimization of the training, the homology reduced dataset of known structures was divided into five groups. All combinations of four groups were used to train while the fifth was used to evaluate the performance (cross-validated). By comparing the predicted with the known structural features the matthew correlation coefficient or the mean average error was calculated. A number of sequence derived parameters, such as amino acid scores from PSI-BLAST,
substitution rate, and predicted distance from membrane center, as well as SVM model parameters, were evaluated as input to the SVM.
Chapter 9

Summary of papers

Paper I concerns how structures respond to evolutionary changes in protein sequences and whether structures change slower than sequences or not. Paper II and Paper III concerns non-helical structures and polar residues, respectively, positioned in the non-polar membrane core environment, and their relation with structure and function within $\alpha$-helical membrane proteins. In Paper IV an analysis is presented of how the properties of membrane protein regions that after folding become exposed to distinct environments differ. In addition, it includes a separate method for predicting which residues are exposed to the solvent environment from sequence. Below you will find a short summary of each of the papers that are included in the thesis. At last, there are some final thoughts.

9.1 Paper I

We investigated how protein structures respond to sequence changes (structural response) by comparing the structural similarity between a large number of structurally aligned and evolutionarily related proteins at different distances of divergence. Initially, we used the commonly used root mean square deviation as a measure of structural similarity. The first finding was that by changing the sequence similarity measure from fraction identity to number of amino acid changes since divergence, a measure called evolutionary distance, the relationship between sequence and structure transforms from exponential to linear. Moreover, we developed three other measures of structure similarity at the residue level. The analysis of these measures showed that the number of changes that have occurred in secondary structure, relative surface accessibility and residue contacts are linearly related to the average number of changes that have occurred in amino acid sequence. According to these measures, structure is 15 to 30 times more conserved than sequence. Thereafter, we grouped the amino acid residues according to physico-chemical similarity. It was found that by grouping the amino acids into the same number of states
as the structural state sequences, structure is 3 to 10 times more conserved than sequence.

Although the structural response is linear on average, the variation is large. We tried to identify structural or functional features connected to proteins with high or low structural response. The only strong trend we identified was that protein domains with a high fraction of β-sheets change states in secondary structure sequence “faster” than proteins with high fraction of α-helices do for same amount of sequence change.

9.2 Paper II

In this paper, we systematically analyzed the 7% of residues in non-helical (coil) conformation in the membrane core in known α-helical membrane proteins. The coils can be found in transmembrane-helical kinks, as major breaks in transmembrane-helices and as parts of reentrant regions. On average, the sites in coil conformation are more evolutionary conserved than those of other sites, particularly within reentrants and breaks. The backbone around coil residues has a polar character and the residues are often either buried or located near aqueous channels. The residues are frequently found within channel and transporter proteins where they introduce the flexibility and polarity required for transport across the membrane. A literature study confirmed that coils are essential for the function of the proteins.

9.3 Paper III

In this paper, we characterized the 9% of strongly polar residues (D, E, K, R, H, N, P and Q) that after folding within α-helical membrane proteins end up in the middle of the membrane. These residues are more conserved than other residues. This is connected to their tendency to be buried in the protein interior and to their high functional importance. The environment that surrounds strongly polar residues in the membrane core resembles the environment that surrounds strongly polar residues in water-soluble proteins. Frequently, their polar groups form hydrogen bonds with water and line functionally important cavities. In addition, strongly polar residues in the membrane core are often directly involved in binding of small compounds in channels and transporters or long-term interactions to prosthetic groups within electron transporters. Further, it was predicted that in human membrane proteins polar residues are over-represented among transport proteins and G-Protein Coupled Receptors. Finally, we saw a trend towards more polar residues in proteins with several transmembrane regions. In particular, the first transmembrane region contains fewer polar residues than the others.
9.4 Paper IV

Here, we analyzed the properties of different classes of sites within a set of membrane proteins of known structures. We found that sites that face the solvent environments, are very different inside and outside the membrane. In contrast, residues buried within the interior of the protein are rather similar. As in water-soluble proteins, the non-membrane regions of α-helical membrane proteins have a hydrophilic exterior and a hydrophobic interior. Within the membrane, both the exposed surface and the buried surface of membrane proteins are hydrophobic, but the difference between the buried sites is much smaller.

We found that state-of-the-art predictors for surface area are optimized for one of the environments and therefore performs poorly in the other environment. To circumvent this problem we included complete transmembrane proteins in a training set and developed a new predictor, membrane protein residue surface accessibility predictor (MPRAP). The new predictor performs well both inside and outside the membrane, and it is the best predictor inside the membrane and in the water-lipid interface region. Further, it is better than the combination of previous specialized predictors. Potential uses of the predictor were demonstrated by identification of error-containing protein structures.

9.5 Final thoughts

In Paper I we investigated how structures respond to sequence changes by comparing four measures of structural similarity to two measures of sequence similarity, for a large number of protein pairs. We found a linear relation between evolutionary distance and the number of changes in discrete structural states in structural core. The finding made it possible to quantify how much more conserved structure is than sequence. By comparing the number of changes in discrete structural states with changes in amino acid sequence we got a simple way to test these features that normally occur in different dimensions. Although, structure previously have been stated to be more conserved than sequences and the sequence-structure similarity relationship have been investigated before, we believe this was the first attempt to quantify how much more conserved structure is compared to sequence.

Although, it is well known that evolutionary distance is a better measure of sequence divergence, it is mostly used for closely related proteins that can be aligned by sequence alignment methods. In this study, structural alignments were used, which made it possible to include both closely and distantly related proteins. This paper, therefore, in addition illustrate a new use of the measure.
Moreover, we found the variation in how much structures change in response to sequence to be large. What determines the large variation remains an open question.

Many recently solved membrane protein structures have structural features that deviate from the earliest and simplest topology model. Recent analysis of more complex structures showed that these structural deviations frequently are directly associated with the functionality of membrane proteins. Interface helices are common in TM-proteins performing electron transport \[67\], while reentrant regions are common in transporter proteins \[65\]. In Paper II and Paper III, we analyzed two additional features, coil regions and polar residues situated in the membrane core region. We showed that these features are also associated with the functionality and should not be ignored. More specifically, we found them too be common within transporters.

Much focus in the past has been directed at developing topology predictions of membrane proteins. Recent innovations include approaches to increase the accuracy by modeling reentrant regions. In addition, in Paper II and Paper III we developed methods to predict coils and polar residues in the membrane core. Identification of such regions might be important for finding regions enabling flexibility or binding ligands. In contrast to topology, few studies have been aimed to predict the assembly of different transmembrane regions. The development of the predictor MPRAP, which predicts the solvent accessibility of the residues in membrane proteins (Paper IV) represents one step in this direction. In contrast to previous methods this method does not ignore the fact that the membrane environment is highly heterogeneous and the hydrophobicity changes drastically over short distances. Therefore, the new predictor performs well both inside and outside the membrane.

The demand for structural and functional knowledge of membrane proteins will most likely sustain for the next few decades. The understanding of membrane protein structures increase for each new structure that is solved. The increased understanding of the structures has recently lead to a few more sophisticated prediction and modeling attempts, both from our lab as well as from other labs. It will be interesting to see how well these methods perform on new structures in the future and which modifications that are needed.
Chapter 10

Acknowledgements

During my work I have received assistance and encouragements from many people. I would like to thank all of you so much! Here are some persons that deserve some special thanks.

Arne, my supervisor, thank you for scientific contributions, inspiration, relaxed attitude and stimulating time of research. In short, it has been a great time!

Anni, my main collaborator, thank you for all loudly enthusiasm about “anomalies”, for being “open” in funny and unexpected moments and for letting me discover what “circular writing” means. Most of all, thank you for all nice time in the lab.

Simone and Yogesh, my best students, thank you for always being so interested and curious about all sorts of things. Simone, thank you for trying to convince me that imperial stouts taste better than belgium-style ales, no matter of how stupid it sounds. Yogesh, thank you for all stories about Nepal. I wish that you some day can show me your home country.

Linnea and Aron, thank you for scientific collaboration. Who would believe 9 years ago that we would publish a scientific paper together (not only us, but anyway...)? Thank you for helping me out with all sorts of work- and non-work related problems. You are the best friends!

Öjvind, Håkan, Andreas, Sara, Erik and Björn, my past and present room mates, thank you for all work-related help. Thank you for daily non-work related discussions about things such as “kvintcirkeln” and the philosophy of “att snitta ned”. Sara, thank you for exxcellent proofreading of an earlier version of this thesis.

Uno Carlsson, Lars-Göran Mårtensson, Per Hammarström, Bengt Persson and David Ardell, my former supervisors, thank you for your relaxed atmoshpere in your labs. Your enthusiasm for doing protein science research convinced me to stay in the “business” after graduation from Linköping University.
Pauwel Kwak, thank you for creating the best beer. Ever!

All not previously mentioned co-workers, administrators, and friends, thank you for making my time so pleasant at work. You are so many that I cannot acknowledge you all by name. I would like to thank you for help at work, nice company at the lunch tables, cake clubs, journal clubs, seminars and badminton-games. Thank you also for nice “after works”, poker evenings and social activities with the Pärk team Svedmyra 1.

Mamma, Pappa and Johanna, my family, thank you for all encouragement, support and for aways being so kind. You are the best family (I have ever had)!

Josefin, my wife and personal illustrator, thank you for creating the beatiful figures in this thesis. Thank you for all understanding, especially during some long working days during the writing of this thesis. Thank you for letting me finding out that everything in life is not about work. Thank you for making my life so enjoyable, partly because of your “kyckling med jordnötssås” and partly for making future so interesting and unpredictable. I love you so much!
Bibliography


