Prediction and analysis of protein structure

Siv Midtun Hollup

PhD thesis

Department of Informatics
University of Bergen
2010

Preface

I have no recollection of wanting to be a researcher as a child. I had no idea there was such a thing. Luckily I got a job at the Department of Molecular Biology more than 10 years ago and I was immediately hooked. There is something deeply satisfying about finding out how things work and carefully designing your experiment so that you can answer a specific question.

In science, there are so many unanswered questions and yet more ways to answer them. For me the combination of informatics and biology was immediately appealing, and I was drawn to protein research from day one. I do not think proteins will ever cease to amaze me with their complexity and ingenuity. This journey has been fun, educational, hard, and a lot of other things, and I would not want to be without this experience.

Bergen, 2009,

Siv Midtun Hollup

Acknowledgements

I would never have managed these four years alone, and there are many who deserve my most heartfelt thanks. Inge, who has supported me through my projects and manuscripts, thank you for being a very understanding boss and for letting me work in your group. Thank you for creating a great working environment at CBU, being positive and encouraging everyone to participate in social events. To Willie, thank you for sharing your knowledge of proteins, your incredible experience and wonderful ideas, and for the great time I had in London in your group. To Nathalie, thank you for teaching me about dynamics, for discussions and patience through my winding path to finding the right project. You have all been a great inspiration to me!

To the Department of Informatics, for hiring me and trusting me with teaching a course, I have never learnt so much in such a short space of time! To everyone at CBU, both past and present, thank you for a fantastic working place; the different people, cultures and research fields that come together, learning about other subjects, be it through seminars or over a cup of tea. The Department of Mathematical Biology in London was also a great place to work where I would have liked to spend more time. At Mathbio, there was a very nice tea break tradition, which has luckily also evolved at the CBU recently. Tea breaks are a life saver, so to all of you: thank you!

To Trond and Kristian, for rants about code, funny stories and lively pub nights. To Harald and Swati for sharing their office and time with me. To Anders and Gisle for Italy, a trip I will never forget! To Edvin, for a good collaboration and breaks with chats about everything and nothing.

I must thank my family, my friends and my husband, who have listened to all my ideas, complaints, tales of talks and cool research (to me), who have supported me when it's been rough and when it's been good, who have let me figure it all out for myself. To Ane, for your invaluable friendship and your interest in what I do. To Alexander, for being at the other end of the terminal, for music and tea and talks. To Ivar, for running, hiking, and Guitar Hero. To Anette and Christian, for proof reading and encouragements. Most of all, to Ove, for being here for me all the way through, for cooking dinner for me when I've slept on the sofa, and for absolutely everything else.

Summary

This thesis, which contains an introduction and four manuscripts, summarises my efforts during my the past four years to understand proteins, their structure and dynamics. The first manuscript presents a protocol that refines models as part of a protein structure prediction pipeline. To achieve this, we used spatial information from determined structures and sequence information from multiple alignments. The protocol was used to improve the quality of rough models containing only one point per residue.

In the second manuscript we investigated protein fold space. We compared models with known fold to determined structures and found that out models contained many folds that were not seen in the present pool of structures in the PDB. Comparison of structural features revealed no reason why the model folds could not exist.

We investigated how well geometric comparison methods distinguished fold in the third manuscript. We presented a novel measure of topological similarity and showed that geometric methods have trouble distinguishing fold differences between both models and PDB structures.

In the last manuscript we showed that the architecture is the most important factor for dynamics as measured by normal modes. Protein fold has some effect and cannot be discarded completely, but larger differences in fold does not necessarily correspond to larger differences in flexibility if the architecture is the same.

List of publications

Model refinement (I) – Structural Fragments in Protein Model Refinement, Hollup S.M., Taylor W.R. and Jonassen I., published in Protein and Peptide Letters in 2008.

Protein fold space (II) – Probing the 'dark matter' of protein fold space, Taylor W.R., Chelliah V., Hollup S.M., MacDonald J.T and Jonassen I. published in Structure in 2009.

Evaluation of comparison methods (III) – Protein Fold Discrimination: an Evaluation of Geometric and Topology Based Measures, Hollup S.M., Sadowski M.I., Jonassen I. and Taylor W.R., submitted to Journal of Molecular Biology in 2009.

Dynamics of protein folds (IV) – Exploring the factors determining the dynamics of different protein folds, Hollup S.M., Fuglebakk E., Taylor W.R. and Reuter N., unpublished manuscript.

Contents

Part I

Introduction

1

Concerning proteins

We are small but we are many, we are many we are small;
we were here before you rose, we will be here when you fall.

Neil Gaiman in Coraline, 2002

The word protein was first introduced by Jöns Jakob Berzelius in 1838 [1] to describe the primary constituent of animal nutrition. Nearly a hundred years later the enzyme urease was shown to be a protein [2], heightening the awareness of how important and prevalent proteins were. In the following three decades, scientists learned that proteins were built using amino acids and that proteins had different characteristics. The first actual structure of a protein was shown when the structure of myoglobin was determined using X-ray diffraction in 1958 [3]. Today, over 60 000 structures have been determined by experiment and these results are available through the Protein Data

Bank [4] (PDB)[1], a free repository of structural information.

Proteins are found in all living organisms, from the smallest bacterium to humans, from plants and algae to sharks and whales. A single protein molecule cannot be seen with the naked eye, yet a substantial amount of living tissue is protein. In fact, the reason we can see anything at all is due to proteins in our eyes and in our brain. Nutrients are processed by proteins, yielding molecules and energy to fuel the organism. Without proteins, we would simply not exist in the form we are today, as the development of the fetus and indeed any organism is regulated by proteins controlling DNA expression [5, 6].

Spider webs are made of proteins linked together [7] and our own bones, muscles and skin are largely proteins. In general, proteins can be divided into rough groups based on what they do. Proteins involved in structural tasks, like forming bones, sinew and muscle, are called structural, or fibrous proteins. Their function is to build and support the elements that give shape to different cells, rather than take part in specific reactions. Structural proteins are among the largest proteins and can have up to several thousand amino acids, often in repeated sequences that form regular 'building blocks' (see Section 1.1.2), like very long α (alpha) helices twisted together in collagen (skin) [8], in coiled coils in keratin [9] (hair), or in case of the spider thread, β (beta) sheets.

Another group of proteins have a compact form and are called globular proteins. They are usually smaller than the structural proteins are have on average a few hundred amino acids, or residues. Many enzymes, proteins that catalyse chemical reactions, are a part of this class of proteins. Most globular proteins are water soluble even though a large part of the residues

[1] http://www.pdb.org

making up the protein chain are not soluble in water. These residues are buried inside the protein, while the water-soluble residues are mostly at the surface. Examples of globular proteins include enzymes that take part in regulation and development of the organism, e.g. responding to external stimuli or to a change of conditions [10, 11].

Membrane proteins are often classified in a separate group. These proteins are situated in or at cell membranes and perform a variety of functions. Some of them pass signals through the membrane by changing their shape [12], while others allow whole molecules to pass through [13]. Some membrane proteins provide sign posts for proteins and molecules and are used in passing messages between different cells in an organism or even between different organisms.

Since proteins are essential for our existence and as most diseases involve protein malfunction, they have been studied extensively by many disciplines: physics, chemistry, molecular biology, statistics and informatics. Even through our best efforts our knowledge of proteins is still incomplete, as not all proteins are given the same amount of attention. There are different ways to investigate proteins, from functional studies aimed at understanding how a protein behaves under various conditions, to determining which shape a protein has. Determining the structure of a protein experimentally is a long and complicated process, and some proteins cannot be determined at all [14].

Proteins involved in the development of an organism, or a protein's role in disease, gives extra interest to proteins. This means that specific proteins from model species are studied by a great number of people, are tested in many ways and determined experimentally in different variants. While there are over 60 000 entries in the PDB today and the number of ways a chain can fold upon itself is limited [15], the PDB does not show more than a small part of the millions of protein structures that actually exist.

Determining the structure of a protein is not sufficient to understand its properties. In order to understand how a protein interacts with its surroundings and itself, its structure must be studied on many levels. Most proteins have some internal movement associated with their function. The study of the dynamics of proteins is important as the function of a protein is tied intimately with the motions it can undergo [16].

The structure of most proteins is not known at all, or only a portion is determined. By studying the relation between a protein's sequence and structure it is possible to develop methods that can predict protein structure if we know the sequence. This can be done because proteins, however diverse in function, share some basic characteristics. Predicting a protein structure can be greatly helped by the fact that proteins that share some sequence similarities are also likely to be even more similar in structure [17].

On a more challenging level, researchers try to create new proteins or optimise and tailor the behaviour of existing proteins [18]. This requires knowledge of both protein structure and dynamics, and is applied in the search for new medicines and materials [19, 20, 21].

1.1 Protein Characteristics

Proteins are complex, both in terms of their structure and function. This has led researchers to develop different ways of analysing these protein characteristics. On a high level we can describe proteins as structural, globular or membrane proteins. These descriptions relate to either the function or the placement of the protein, but they provide limited information on what the structure of a protein might be. The function of a protein is tied intimately with its structure and its dynamic properties [16], and we must therefore endeavour to understand the characteristics of protein structure. The remainder of this chapter will explore ways to describe proteins, introduce structural and dynamical principles as well as how proteins relate to each other through evolution.

1.1.1 Weak forces underlying protein fold

Every protein consists of at least one chain of amino acid residues. These residues have different characteristics, such as their solubility in water and whether or not they are charged. This section gives a short introduction to how the atoms and groups comprising the residues interact with each other.

In a single protein there are thousands of atoms. Even though many of the atoms do not interact, the number of interactions counts several millions. The covalent bonds between atoms in a single residue and the bonds between atoms in different residues are by far outnumbered by the weaker interactions between atoms, making it important to understand the weak forces occurring between atoms.

It is vital for a protein's stability to shelter non-polar atoms from water. The tendency of non-polar atoms to aggregate, called the hydrophobic effect, impacts protein structure as the hydrophobic side chains of residues are mostly buried in the core of the protein, away from the surface. The backbone of the protein, which is also buried in the core, is polar, and interactions between the polar atoms form so that these atoms can be in the core without destabilising the protein.

The compactness of a protein globule is also governed by the weak Van der Waals forces. These forces are active between atoms that are in immediate vicinity of each other, and they describe how atoms are attracted to each other until their masses are close enough for steric repulsion to take effect.

Both in the core and most particularly on the surface, electrostatic interactions are involved in the shape and integrity of the structure. Groups with opposite charges can form salt bridges. Polar groups on the surface also control the degree to which the protein is solvated in water. The hydrogen bond, an electrostatic interaction, occurs between atoms in side chains and in the backbone of the protein.

1.1.2 The structure of proteins

Globular proteins can be described in many ways depending on the characteristics we wish to analyse. A protein's role in disease can be hinted at by mutations in the sequence [22]. This leads to structural changes and thereby functional changes. In order to discuss these and other issues, we must have a clear understanding of how structure is described (see Figure 1.1 for an overview).

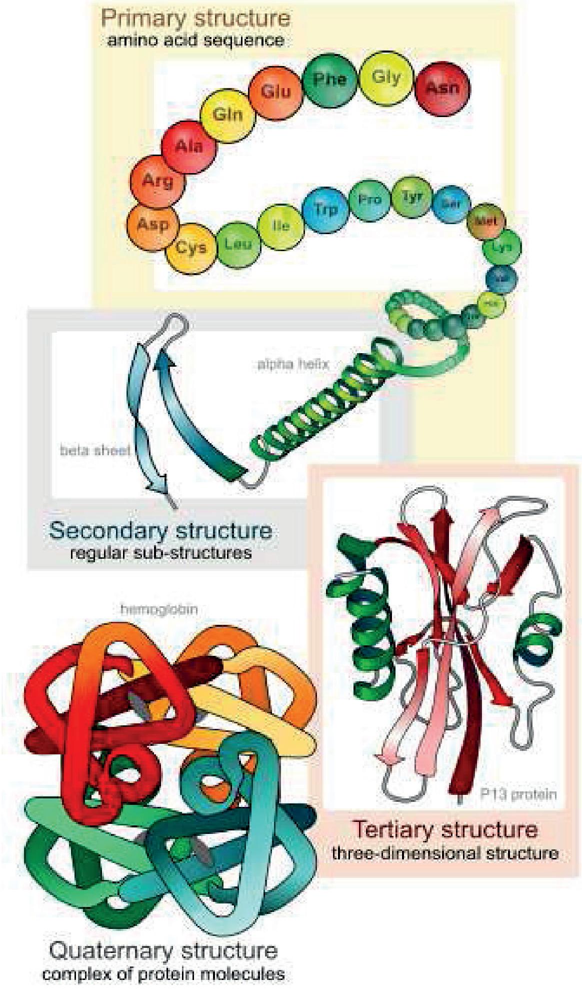


Figure 1.1: Descriptions of protein structure. Figure downloaded from Wikimedia Commons.

Primary structure

The simplest way to describe a protein is through its primary structure. The primary structure is the protein sequence represented as a string of letters.

There are 20 main amino acids (or residues), each with its own letter. The sequence describes the types of amino acids and the order they are connected in. When the protein is produced (by translation), the first residue to appear is the N-terminal residue, and the last is the C-terminal.

The primary sequence does not contain any specific terms describing structure except which residues are neighbouring along the backbone. However, the order and type of amino acids can give hints about local structures that may form [23]. Gathering multiple sequences from the similar proteins and aligning them gives better predictions of secondary structure [24] and gives a clearer picture of the overall characteristics of the proteins. The primary structure is useful as it gives a short definition of the protein, well suited for comparisons to other proteins. The primary sequence is often used to search for similar proteins in sequence databases and can be used to study the evolutionary relationship between proteins.

Secondary structure

A folded globular protein has a hydrophobic core and a hydrophilic exterior. The backbone of the protein chain is partially hydrophilic and passes through the core several times. For this to be possible without the protein unfolding, the hydrophilic groups of the backbone must form hydrogen bonds to minimise the effect of the partial charges in the core. The secondary structures, called α helices and β strands, describe local, regular three dimensional substructures in a protein that are the result of the backbone interactions.

An α helix is formed when backbone hydrogen bonds are made locally from a residue that is 4 residues further on in the sequence. All the hydrogen bonds are aligned in the same direction (along the axis of the helix) and the side chains of the residues point out from the helix, like stairs in a spiral stair case with the backbone forming the central pillar.

The other type of secondary structure, the β strand, is formed when two or more linear segments of backbone lie next to each other (either parallel or anti-parallel). Hydrogen bonds can then form between the adjacent strands perpendicular to the direction of the strand, and the result is a β sheet. The side chains of the residues will alternate between pointing above and below the sheet. A β sheet must have at least two strands, but can have more.

Most proteins have several secondary structure elements (SSEs). Some proteins have only helices or strands, while others have a mixture. The SSEs can be used to distinguish proteins on a different level from the sequence or function, although the sequence can give hints about which SSEs the chain will form. The different secondary structures have different characteristics, and can give some indication on the function of a protein.

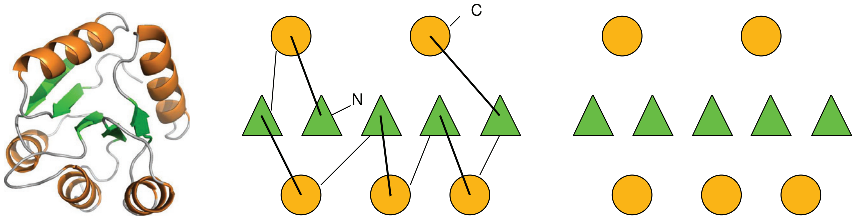
Tertiary structure

The tertiary structure describes, for a chain of protein, a snapshot of the positions of all atoms. Combining the atom positions and knowledge about bond formations, many bonds and interactions can be inferred. Atoms are positioned in a three dimensional space in relation to each other, so that it is possible to see which residues come close in space. This allows for a detailed inspection of the space a structure occupies. To find out if a protein can interact with certain molecules, the surfaces of both partners must fit, both with respect to the physical space they occupy, but also with respect to the charges and hydrophobic patches that are on the surface [25].

Structural genomics initiatives gather information about protein structure and function [26], but the structure itself is not always enough to understand how a protein works. Using the tertiary structure along with rules for how atoms and bonds move and can be changed, it is possible to investigate how a structure can change its shape and function.

A tertiary structure representation also allows for comparisons of different protein structures. A common measure of root mean square deviation (RMSD, see Methods section) quantifies the distances between equivalent points in two structures. This can be used to compare parts of, or whole tertiary structures.

Related to the tertiary structure are the concepts of protein domains, architecture and fold. Long protein chains can form one or several distinct domains. Domains can have function on their own or may need to be in contact with the rest of the structure in order to function. Protein architecture refers to the types of secondary structure a protein has along with their position and in some definitions direction relative to each other. The fold, or topology, of a protein is an extension of the architecture and defines the connectivity, as defined by the protein chain, between the SSEs in the architecture. Figure 1.2 shows a structure represented by its architecture, fold and how this relates to the actual structure. Comparing architecture or fold gives an impression of the overall similarity between two proteins without going into details about residue positioning.



(a) (b) (c)

Figure 1.2: A protein visualised by a cartoon representation of the actual coordinates (a), fold (b) and architecture (c). This protein has five β strands, coloured in green, and five α helices, coloured in orange.

Quartenary structure

When several chains interact, the quartenary structure of a protein describes this by specifying the chains in relation to each other. In some cases the individual chains can have a function on their own, or they may be dependent on the whole complex to work.

Studying how proteins in a complex interact can be done using the quartenary structure. Complexes can be compared in terms of which parts they have, how many chains interact and perhaps also the nature of the interactions between the parts.

1.1.3 Protein structures are not static

The Protein Data Bank contains structural information on thousands of proteins. The vast majority of atoms in the structures have a single coordinate set in three dimensions giving its location, but atoms are not stationary. The atoms themselves vibrate and the bonds between atoms stretch. Angles between different bonds also change depending on the surrounding atoms and groups. This allows the backbone of the protein to fold into different conformations, and it also allows side chains to position themselves differently. Flexibility in a protein is partially captured by the B-factor values supplied in the PDB file.

Relatively small adjustments of angles in the backbone in certain places in the structure can cause bigger movements, for example in loops or in subdomains. Loop movements cover several orders of magnitude on the time scale, as a small loop can fluctuate quickly and a large loop can travel a long way from one extreme to the other. Whole domain movements can take even more time and may require help from other molecules. The movements in a protein structure cover a time scale from approximately 10^{-15} to 10^{-3} seconds [27], and well-known methods, like Molecular Dynamics Simulations and Normal Mode Analysis, are used to analyse the dynamics of protein structures cover different time-scales.

1.2 How proteins relate to each other

Species that are closely related through evolution may look similar. Lions, tigers, lynx, and house cats, all in the cat family, are all related even though their sizes and shapes are somewhat diverse. Dogs and cats are also related, although more distantly, and so we can continue to trace related species. Continuing this we can group species into mammals, bacteria, crustaceans and so on. The study of genes and whole genomes of different species have helped to reveal how different organisms are related.

Related, or homologous genes have similar sequences, and the more closely related they are the more similar the sequences are. For close homologues, the DNA sequence can be used for comparison. Mutations in the DNA sequence does not necessarily generate a different protein sequence, as the same amino acid can be formed by different DNA triplets. Protein sequence thus evolves more slowly than DNA sequence, and comparisons between more diverse proteins can be carried out using the protein sequence. Structure is more conserved than the protein sequence again, and even when there is no discernible homology between the sequences, the overall structure, like architecture or fold, is robust to sequence variation and can retain similarities [17]. While there are millions of different proteins, the fact that many of them are similar can be of great help to understanding what they look like and how they function.

The majority of protein structures behave in this manner, but there are examples of proteins that have sequence similarity but have different three-dimensional structures [28]. To complicate this further, unrelated sequences can form similar structures. Some protein folds, called superfolds, seem to be easier than others to obtain and have been formed several times in the course of evolution [29].

2

Working with proteins

Blessed are the flexible, for they can tie themselves into knots.

Unknown

In order to understand the general characteristics of proteins we need to have a good vocabulary to describe proteins, methods to evaluate the quality of a model and methods that replicate and build protein-like models. This chapter provides a brief introduction to some methods that allow us to classify, compare, build and analyse models and structures.

2.1 Classification of protein structures

The sheer variety of protein structures makes it necessary to deal with proteins on different levels. While they all share some basic characteristics they are diverse, and it is possible to classify proteins according to their characteristics. Two well known classifications are SCOP [30] and CATH [31].

2.1.1 Classification databases

SCOP (Structural Classification of Proteins) is a manually curated database for domain classification directed towards function and evolution [30]. The topmost category is Class, distinguishing protein domains based on their overall secondary structure content. Domains are sorted into different folds, where the order and spatial relationship between SSEs are important. The folds are further divided into superfamilies, the criteria for separation being similar function and structure. Finally the proteins are sorted into families, where the similarity between sequences or structures should indicate an evolutionary relationship between members.

The CATH database has four levels of classification: Class, Architecture, Topology and Homologous superfamily [31]. Class refers to overall SSE content: all α, all β, mixed or low SSE content. Architecture refers to the positioning of the secondary structure elements and their direction. Topology brings in the connections between the SSEs, so that the order of SSEs along the protein chain is taken into account along with their orientation and relative position. Finally the structures are divided into Homologous Superfamilies, where structures have a detectable sequence similarity. The process of classifying structures in CATH is semiautomatic and the automatic part is based on how well the structures fit on each other using global structural alignments.

2.1.2 Topological classification

It is also possible to describe domains based on structural features only, even though this classification is not as detailed as the finer levels of the databases mentioned above. Domains can be assigned to Ideal Forms [32, 33], which classify domains based on the number of SSEs and the way these SSEs are positioned relative to each other. The protein in Figure 1.2 has three layers of secondary structure, an upper layer containing two helices, a middle layer with five strands and a lower layer with three helices. Figure 2.1 shows the forms for simple protein domains with both helices and strands. Ideal Forms can be compared to Architecture in CATH without directional information for each SSE, but the Forms, unlike CATH and SCOP, are not limited to what exists in determined structures in the PDB. Theoretical domains can be constructed by exploring all possible ways a set of SSEs can be arranged.

Ideal Forms that fit a domain can be found by representing the SSEs in a structure as line segments, or sticks [34], which are then matched to and superposed onto the Forms [33]. After finding the Ideal Form(s) for a domain, it is possible to extract the directional information and connectivity of the SSEs. This can be encoded in a topology string [35], reducing the whole topology of a domain into a single string. Using a three layer $\alpha\beta\alpha$ sandwich domain as an example, the letters A, B and C designate the layers of secondary structures. Each SSE is also given a direction (+ or -) and a number to indicate where in the layer it is with respect to the other elements in the same layer. The first SSE to enter a layer is designated 0, and elements to either side are given negative or positive numbers, see Figure 2.2 for an example.

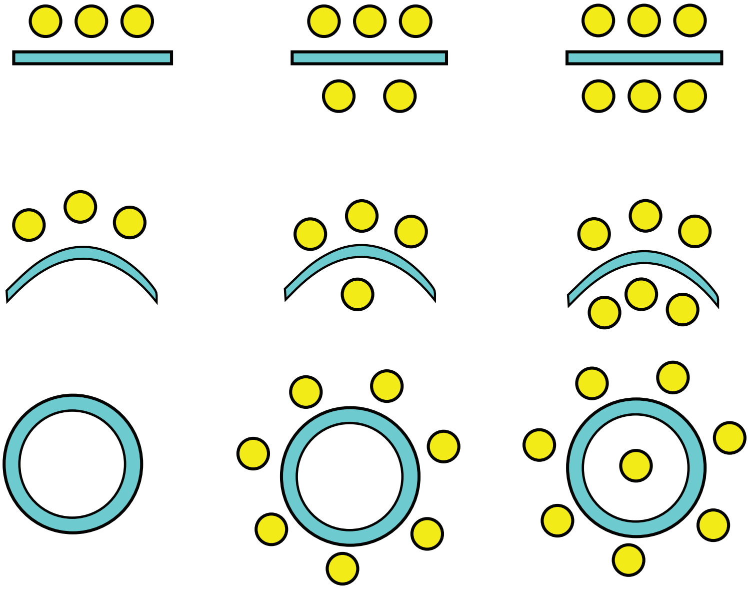


Figure 2.1: Examples of ideal forms. The blue bar indicates β strands, yellow circles α helices. The numbers of helices on either side of the sheet can vary as long as there is space under or over the sheet. The middle structures indicate a curved sheet, and the three bottommost variants shows a barrel formed of β strands.

TOPS (Topology of Protein Structure) diagrams can also be used to display the topology of a structure [36] and are drawn in a similar manner as shown in Figure 1.2 b. The types of secondary structure, their approximate orientation and relative positioning are used and diagrams are drawn based on this information.

2.2 Comparing protein structures

Classification of proteins is one way to describe the similarity between proteins. It is also possible to compare proteins directly and investigate in a more

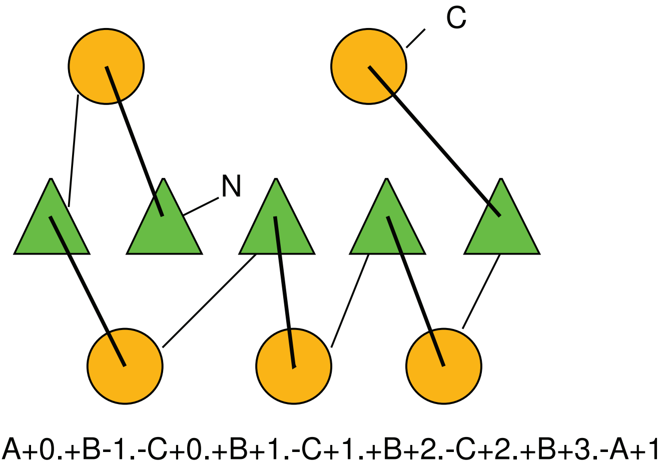


Figure 2.2: Rossman fold topology and fold string definition. The first β strand is numbered 0 as it is the first to enter a central layer (B), and is defined to stick out of the page (in direction). Period separates the different elements in the topology string. A designates the 'top' layer of helices, defined by the first element that is a helix, in this case the second SSE which goes into the page and is given direction - (coloured orange).

detailed manner how similar two proteins are. The methods described here compare structural traits and do not take function into account directly.

2.2.1 Geometric methods

Geometric comparison methods try to find the best way to superimpose two structures by using a combination of their tertiary structure and their sequence. Residues are represented by their coordinates in a three-dimensional space, usually one point per residue. The structures are usually treated as rigid bodies, and various measures can be applied to measure the quality of the superposition. Root Mean Square Deviation (RMSD) measures the average distance between two sets of points by the following formula:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\delta_i^2} \tag{2.1}$$

where δ is the Euclidean distance between two equivalent points and n is the number of points in each set.

In order to make a superposition of two structures, a set of equivalent points must be found. The set consists of pairs of residues, one from each structure, and ideally a large number of pairs should be found that give a low RMSD value. To do this, methods use a combination of sequence alignments and structural measures. Three commonly used programs that solve this task are DALI, SAP and TMAlign. The DALI program uses distance matrices as a starting point to compare structures [37]. A distance matrix represents a structure by the distances between all pairs of atoms. Similar submatrices indicate that two substructures match well, and these substructures are then superposed and the superposition expanded to cover larger parts of the two structures.

SAP uses two levels of dynamic programming to superpose structures [38, 39]. The first dynamic programming step searches for pairs of residues that may form part of the alignment. A good score in this step gives a bias in the position dependent scoring matrix that is used in the second, high-level dynamic programming step. The program alternates between finding residue pairs that are likely to be part of a good alignment and superposing the structures using the selected pairs. As good pairs are found, these are retained and more residue pairs are added.

TMAlign uses a combination of matching secondary structure elements and a structural alignment of the two structures to obtain an initial alignment, and uses TM-score to optimise the superposition in several steps [40]. The TM-score evaluates distances between pairs of residues, but unlike the RMSD measure, it takes into account the protein length and gives a preference to residue pairs with a short distance in 3D.

2.2.2 Topological measures

Rather than using the three-dimensional coordinates, structures can be compared in terms of their fold. This is a more coarse-grained measure and captures the similarity of the path the backbone takes through the protein. The TOPS diagrams can be used to compare protein structures and to search for similar topologies by graph matching methods and machine learning methods [41, 42].

Another way to describe similarities between structures is by counting how many changes one diagram (as seen in Figure 2.2) must undergo in order to be transformed into the other structure diagram. Each transformation step corresponds to removing or adding an SSE. If two structures are topologically identical, this distance will be zero. An SSE changing direction adds a step, and changing one SSE from one to the other type adds a step. Multiple operations on one SSE are only counted once. This measure of similarity is called edit distance and is presented in full in Paper III. It uses the topology strings described in Section 2.1.2.

2.3 Building models

By applying known principles about protein structure and the relationship between sequence and structure, models can be built from the information contained in the primary sequence. If the sequence is homologous to a determined structure, this can be used as a starting point for the model, but homology cannot always be detected. This section outlines a protein structure prediction pipeline which produces many potential models given a single protein sequence that can be applied also in the absence of any homology to a protein with known structure [43].

2.3.1 Fold generation

The first step is to determine the secondary structures that are likely to form. A multiple sequence alignment is automatically generated after searching for similar sequences and this alignment is then used to predict secondary structure using two programs, PsiPred [44] and YASPIN [45]. Secondary structure prediction methods have a limited accuracy, and by using different sequences as the main sequence in the prediction a large variety of secondary structures are generated. For any ambiguous secondary structures, all variants are retained, adding diversity at the cost of needing more models rather than miss out on good models because of a restrictive view of the number and position of SSEs.

The secondary structure prediction dictates the number of secondary structures the model will have, and matching Ideal Forms are found. By representing the SSEs as packed boxes, the forms are ranked according to a compactness score [46, 47]. A measure of conserved hydrophobicity is computed for the predicted SSEs [48], and this is matched against an estimate for the solvent exposure for each element in a form.

Different folds are then built on the Form scaffolds, only disallowing connections that are not seen in structures solved so far, e.g. left handed connections between β strands and crossing loops. A scoring function ensures that loops are not too long. The scores from this measure and the ones mentioned above are then used to score the initial models, along with a filter to discard models with poor secondary structure packing. After this step we are left with several thousand models with varying folds.

2.3.2 Model refinement

These models must then be refined, as only overall quality measures have been applied at this stage. To ensure that the structures fit well to the sequences, the sequences are threaded onto the C_α structures and the fit is optimised with respect to their residue burial and matching secondary structures [49]. This produces many variants around each template, and only the best models are retained.

Using a fragment based modelling approach, the remaining models are refined at the C_α coordinate level [50]. This method encodes the protein as a series of linear patterns, each describing the three-dimensional space around a residue. The exact spatial relationship is not retained, but relative distances along the backbone, amino acid type and secondary structure information is recorded for each element. Multiple sequence alignments can extend the amino acid types that can be allowed in any position, giving information about conservation as well. Patterns for a model are compared to a library of patterns derived from a non-redundant set of structures from the PDB. The structural fragments from the pattern matches are used to refine the model, and the process of encoding this new model as linear patterns is performed again. This whole process is repeated a number of times to ensure that the packing between SSEs and the SSEs themselves are regularised.

2.4 Characterising protein motion

Investigating the motions proteins can undergo can help us understand how proteins behave. Molecular dynamics simulations and normal mode analysis are computational analyses that can be used for this purpose. Both represent the protein in a system of particles, and interactions between the particles are represented by equations that are designed to approximate the actual energy and behaviour of the interactions.

In molecular dynamics (MD) the protein is simulated over time by solving Newton's equation of motion for the particles in the system, usually the atoms [51]. These calculations are designed to reproduce *in vitro* conditions and can also include molecules around a protein, like water and ions. MD simulations are time consuming in terms of CPU, limiting the time frame these simulations are used for.

Normal mode analysis is concerned with finding the inherent motions, or natural frequencies, of a protein represented as a mechanical object where residues, represented by balls, are connected with springs [52]. These concepts are also used in mechanics and acoustics. To give an example: A guitar string is connected at both ends, and when force is applied to the string it vibrates at a specific frequency. Applying force from different directions and with a different strength does not change the frequency with which the string oscillates, as it is the natural, harmonic frequency. When you strum a guitar string, it will create sound at the same frequency no matter where you apply the force, and the note will have a different sound level according to the amount of force applied. Only by changing the characteristics of the string, i.e. the length or the diameter, can the frequency be changed, and it will then oscillate with a different frequency. When the force is no longer applied, the string goes back to its original, native state.

The inherent motions can be characterised with normal modes and a set of frequencies. Computations are made without any direct link to time or distance. Interactions between particles are modelled as harmonically oscillating springs. The force of the spring is determined by the distance between the two points and by a force field developed to mimic actual dynamic behaviour in proteins (e.g. as predicted by MD and by B-factors in PDB structures).

A normal mode is constituted by a vector and a frequency and there are as many modes as there are points multiplied by dimensions (usually 3 multiplied by the protein length). The vector represents the direction of the motion, and the frequency indicates the speed and relative size of the motion. A low frequency corresponds to a slow and large structural change, while a high frequency indicates a rapid movement. The system will revert to its native state after oscillating along the normal modes. A wide range of motion can be detected through these computations, from large-scale, slow movements like domains moving relative to each other to small, rapid motions like bond-stretching and bending. A more mathematical introduction to normal mode analysis, examples of analyses and web servers can be found in a recent review [53].

3

Results and Discussion

Little by little, one travels far.

J. R. R. Tolkien

Four manuscripts contribute to the thesis. The first is concerned with refining models in a protein structure prediction pipeline, while the three others are concerned with exploring protein fold space and evaluating and characterising protein folds. In this chapter, we give a summary of the results of each manuscript and put it into context. A general discussion concludes this chapter.

3.1 Results

3.1.1 Model refinement

The goal of this work was to improve C_α models in the last step of the protein structure prediction pipeline summarised in Section 2.3. An existing refinement protocol based on structural patterns was the basis for this work [50].

Three-dimensional information about local substructures in models and PDB entries were encoded as linear patterns. The PDB entry strings were collected in a library and using a string matching algorithm, suitable library patterns were found for the model patterns. The structural fragments from the most similar patterns were then used to improve the model, before we generated patterns for the new model and the whole procedure was repeated.

The matching patterns were ranked according to how well they fitted onto the model by sequence fit, structural superposition as measured by RMSD (see Section 2.2.1), and a measure counting structural clashes between the library fragment and model structure. From this ranked list, several sets of patterns covering as many model residues as possible were tried out. Each pattern set was also ranked according to how well the library fragments fitted onto each other in structure and sequence. Both structural fit between overlapping fragment residues and the number of structural clashes between non-aligned residues was measured. The main result was that our method could improve the RMSD of rough C_α models for prediction purposes.

The performance of the method was investigated using two datasets. Firstly, we used a set of models generated by the prediction pipeline to test whether an overall improvement in RMSD could be seen using our new approach. Virtually all models saw a significant improvement. The average improvement was measured to around 0.5 Angstrom. Secondly, we chose four CASP 7 medium to hard targets with a suitable length and tested our method on the models that were ranked as number one by the different groups. As the models had undergone a more rigorous refinement in terms of the number of atoms and detailed modelling of interactions, we did not expect to improve these models. The improvement was not as obvious here, but we still saw an improvement in RMSD for many of these models.

3.1.2 Protein fold space

Leaving the details of model refinement, we now turn towards studying fold space at a theoretical level. Through the development and use of the structure prediction pipeline we inspected many models by hand, and some of the models we saw were not familiar in terms of fold. The goal of this study was to determine if any of our models had novel folds, i.e. that the folds could not be found in the present pool of determined structures. Any models with a novel fold were then characterised to determine whether the models were protein-like in terms of packing and fold characteristics.

The concept "fold" is not used consistently throughout the literature concerning proteins and any study of the properties of fold space is affected by the chosen definition. We used the approach summarised in Section 2.1.2 to define the fold of a structure in a non-redundant set of domains from the PDB, and compared them to a set of predicted models. All model folds were known by definition as they were created with a specific topology (see Section 2.3).

To determine whether any of the models had a novel fold, we checked if and of the folds from the PDB set could be mapped to the same Ideal Form and fold string as our models could. As we suspected, we found many new folds, and we investigated their characteristics to determine if they were protein-like or not in terms of their fold.

In our dataset, which included $\alpha\beta\alpha$ sandwich proteins of length between 100 and 150 residues, we found over 2000 models that had a unique fold definition that did not have a match in the PDB. If the native structures had an ambiguous fold definition, all possible fold matches were used. As our fold definitions had not been used to distinguish folds at this scale before, we compared all models to the non-redundant PDB set using the structural alignment program SAP (see Section 2.2.1). A subset of the comparisons with the best scores was checked manually to see if there was indeed a fold difference, even though the structural alignment had a combination of a low RMSD and long alignment length. In all but a few cases the fold of the model was unique and no PDB entries with identical fold could be found. Based on this we chose to trust our fold definition as an accurate measure of protein fold.

Our models were built using the rules and principles guided by what we see in real structures. While we knew that the inter-residue distances and packing were protein-like, we set up several tests to find other characteristics that might set the models apart from those found in the PDB. The sheet topology was tested, as was the packing and burial degree in the structure. Results from these tests showed that the novel folds were well within the range found in real structures, and we could find no reason why these protein folds could not exist.

3.1.3 Evaluation of comparison methods

The manual review of high scoring structural alignments between native structures and models we did for the fold space study showed that geometric comparison methods have problems distinguishing fold. Methods like TMAlign, DALI and SAP have been used to measure similarity between structures, both for detailed studies of two similar structures, and to determine the degree to which proteins are related structurally and evolutionary. Using geometric methods to measure similarity between structures, a loop becoming a sheet or a helix may not register as a significantly different score, as the average distance between pairs of residues may not change that much. Also, larger loops have an increased flexibility that can affect the score. While a loop turning into a small helix or a strand may not affect the characteristics of the protein, larger fold changes, like a strand swap, might.

In this study, we defined a method to describe the distance between two folds given the fold definition presented in Section 2.1.2. This measure, called edit distance (see Section 2.2.2), was used to determine to which extent the geometric methods could distinguish fold changes.

As all our models had unambiguous fold definitions given at the time of construction, we began by comparing different models to each other. While the models sharing the same fold definitions also had, on average, good structural alignments, models having different folds also received high scores and could not be distinguished from the real matches by the geometric methods. We then compared our full model set to a non-redundant set of domains from the PDB, where the results were comparable, although more noisy, as the structural diversity between the PDB structures was greater than between the models. The geometric methods yielded the same result as for the comparisons between models, and the comparisons with a fold match were even more spread out in terms of their structural alignment score. The edit distance showed that small and large fold changes could not be distinguished well by the geometric methods employed in this study.

Finally we compared native structures to each other. The results included pairs of structures with good structural alignments but different folds. This was verified manually, and as shown in Paper III we found a high scoring structural alignment where one structure had a strand swap.

Our main conclusion from this work was that geometric methods fail to distinguish even relatively large fold changes, as demonstrated by the edit distance measure. However, the fold of a protein can be ambiguous, making it hard to devise a single scoring function to distinguish protein folds. Poorly defined or small SSEs in loop regions, like one-turn helices and partial β strands at the edge of a sheet, can be hard to define and their effect on the structure is difficult to estimate.

3.1.4 Dynamics of protein folds

In this study we turned our focus from a static to a dynamic view of protein structure. The main goal of this work was to determine whether the architecture or the fold of a protein is the dominating factor for the slow dynamics characterised by normal modes. In addition we looked at how folds were differentiated in terms of their dynamics as their folds became more different, i.e.

that their edit distance increased.

We used the topology strings presented in Section 2.1.2 to describe the folds and the edit distance (Section 2.2.2) was used as a guide to describe fold differences. Rather than finding examples of determined protein structures with the fold differences we sought to investigate, we used our model set from which it was possible to choose models with ease. This had the added advantage that we were not limited to the folds that actually exist in the PDB. To be able to use the models to address these questions, we first needed to assess our models' behaviour in terms of dynamics. All models were refined with all backbone-atoms to ensure the quality of the models in terms of bond length, torsion angles and interactions between residues [54].

We calculated the dynamics using normal mode analysis on the C_α coordinates of the structures and the modes were characterised with flexibility profiles. In order to verify that our models had a protein-like behaviour, we compared our models to the native structures whose sequences were used to generate the models. The flexibility profiles of these were then compared to the profiles of all the models sharing the same fold definition. Our results showed that the model profiles were comparable to those of real proteins.

The architecture was found to be the dominating factor for the dynamics. This was shown by calculating normal modes of the model structures containing only the SSEs and not the loops connecting them. The flexibility profiles of these reduced structures had a very high rank correlation coefficient to the profiles of the original models. To determine the loop connections' influence on the dynamics, the equivalent parts of models with different folds were compared. For this analysis, we computed the normal modes on the complete models, but used only the flexibility values of the residues in the SSEs to measure the correlation. There was a significant contribution from the loop connections, but a larger edit distance did not necessarily mean that the difference between flexibility profiles was greater.

3.2 Discussion

3.2.1 Model quality

The refinement method we developed was included in the prediction pipeline (Section 2.3) for the 8th CASP experiment (Critical Assessment of Techniques for Protein Structure Prediction) [55]. Close inspection of the output models after refinement revealed that the protocol was not optimal, as most of the models had undesirable interactions and bond angles between C_αs. In the end we did not apply the refinement protocol to any of the models we submitted. Although the tests performed in Paper I were limited in terms of the range of protein classes, it was puzzling that the method performed poorly on proteins that should be comparable to those used in the paper. For this method to be applied in a general protein structure prediction setting further work is required, both to improve the model quality assessment and the method's behaviour according to a different quality measure.

If the input model does not get a good set of pattern matches from the library, or if the patterns do not cover the structure well, the results could be unsatisfactory. The core of the CASP models, however, were mostly covered and still did not look "good". Inspection of the models after refinement revealed a tendency to distort the secondary structure elements, strands being less extended and helices being more tightly wound (resembling 3-10 helices). In particular, the strand changes had an adverse effect on the structure as a whole, as the distances between secondary structures increased in the core. Why, then, did we not pick this up in the original work? We did not have any models with mainly β strands in our test set, as we could not find any suitable examples from CASP 7 with the desired length and difficulty category. The RMSD was improved in nearly all models we refined for Paper I and it is not trivial to find an automatic method that can assess the quality of the bond lengths and angles in the models.

The RMSD measure is global over the set of residues being measured, which could hide some of the problems we saw in the CASP 8 models. The same distance between two points will receive the same score although one distance set may be more protein-like than the other in terms of bond angles. Also, as several residues from different patterns are aligned to the same model residue, the residue's new position will be an average of the other positions. RMSD does not take into account bond angles and it is clear that this aspect of the refinement protocol is not sufficiently robust. To improve the protocol, another scoring function could be applied to pick model improvements that better satisfy ideal bond angles between residues in SSEs in addition to the scores we used. We have explored method variants including all backbone atoms, but the results were also measured by RMSD, and should be reviewed with a different quality measure before any conclusions about improvement can be made.

In terms of using RMSD to score models, the limitations of this measure should be addressed in tests. RMSD values are not independent with respect to protein length and is insensitive to local changes. Arguments and tests showing that the method performs significantly better than random should also be included.

In this setting, it is important to discuss what random means. Can a random background be constituted of numbers drawn from an interval, or should it constitute plausible positions of the points representing the residues? Should there be different intervals of plausible positions depending on the burial degree and packing of each residue? How should these values be treated, in a local or global view? Answers to these questions can be incorporated into benchmarking tests for refinement methods.

3.2.2 What is a fold?

At the level of model refinement, we take for granted that the fold is already determined. As discussed in Papers II and III, the term “fold” itself is not well defined. The automatic approach we used assumes that a fold can be clearly defined, an assumption that can be subject to debate but still allows our study to explore features of the protein structure universe.

Poorly defined secondary structures are prevalent throughout the determined structures in the PDB, for example in edge strands that only have a few hydrogen bonds to the core sheet or in loops that are likely to form partial bonds to the sheet, or in very short helices in loop regions. These are small structural differences, and the change from one to the other does not necessarily change the characteristics of the protein. This then begs the question whether or not such small differences are in fact differences at all, or just natural variation of a fold. If it is natural variation, how do we distinguish natural variation and distinct fold? If a central strand in a sheet loses its hydrogen bonding, the structural and dynamical effects are likely much larger than were it on the edge of the sheet. Similarly, a central strand cannot form a helix instead of a strand, as it would cause either a major rearrangement of interactions in the core or complete disruption of the fold. The interactions an SSE has, its relative position and size plays a role in how much variation it is natural to tolerate.

In our study of protein fold space in Paper II, we solved some of the challenges mentioned above by allowing a protein multiple fold definitions. This allowed us to compare the folds of determined structures to our models in a fully automated manner. Results from this showed that the vast majority of our models did not share the same fold as any known proteins, and our tests of the general characteristics did not reveal any reason why these folds should not exist. As our analysis of the models did not consider all possible aspects of protein structure, we cannot exclude the possibility that there exists reasons that these novel folds cannot exist. A more interesting possibility is that the folds are not present because they by chance have not been used throughout the course of evolution. That we have not seen these folds does not have to mean that they cannot exist or that we cannot, in the future, design proteins with these folds.

3.2.3 Evaluation of structure comparison methods

In our work to verify that the topology strings did indeed capture fold correctly in Paper II, we compared all of our models to a set of non-redundant PDB domains using the geometric comparison method SAP. It became clear to us that this method could not separate seemingly large fold differences like strand swaps in the core β sheet. The results from Paper III show that it is not only SAP, but also two other commonly employed geometric methods that suffer from this problem. Small fold changes, like those described in the previous section, are not likely to radically change the characteristics of a protein, but even larger differences such as strand swaps could not be reliably detected using geometric methods. While classification databases do not agree on how to classify proteins, strand swaps are separated in both SCOP and CATH. This implies that based on the characteristics used to form both databases, these folds should be considered distinct. The edit distance measure provides a supplement to evaluation of protein structures that geometric methods do not seem to be able to provide.

That the CATH and SCOP classifications do not agree on the classification of structures shows more clearly than any argument made here how difficult the classification problem is. Both these and our fully automatic classification schemas suffer from classifying structures into categories, which may introduce artificial divisions between folds. In addition, actual changes between folds could be set as equal (in distance) to these artificial changes. Allowing multiple classifications for a protein can avoid the problem, but the cost is that the boundaries between folds become blurred and it is harder to form a clear picture of how fold space spans out. A different possibility is to think of fold space as continuous rather than discrete, and move away from the fold concept. But how can we decide if fold is relevant for characterising proteins if we do not have an established idea of what a fold is, and how it behaves in relation to other protein characteristics?

3.2.4 Dynamic characteristics of folds

Studies of the dynamical properties of proteins should also look at the natural variation of dynamics in order to ensure that results are real differences rather than random variation. We simulated a wide range of proteins using models which, while sharing the same fold, exhibited structural differences. In this way we could analyse with larger confidence which motions were conserved and which were variable over the whole model set. A random background was provided using the reversed and permuted profile of the native structures. This gave us a background set with the same interval of flexibility values and the same secondary structure content. In this way, completely random numbers were avoided, as the backbone constrains the flexibility of the next residues in the chain. The distribution of correlation coefficients between our model profiles and the random profiles was wide, reflecting that folds are not random and that partial matches can yield high values of both correlation and anti-correlation. The use of the these distributions circumvents the issue with finding unrelated proteins, which in itself can be difficult. In addition, it circumvents the problems of aligning proteins that are very different in fold. These alignments will not cover the entire structure, and two different programs may give two very different alignments for the same structures.

The same arguments can be used for the model set. Using models of a uniform length made it easier to compare the proteins without dealing with alignments and their quality. Applying a procedure such as the one developed in this manuscript on real proteins does require careful attention to these issues, however.

Adopting the methods from Paper IV to a comprehensive set of folds covering all determined structures could give a database of representative flexibility profiles for all folds. Incorporating many structures for each fold would give a representation of the natural variability, e.g. the degree to which loops and secondary structures are expected to vary. Such a database would naturally face the classification problem as discussed earlier, but could perhaps give a different view of how protein structure evolves. Using models with folds not seen in the present pool of determined structures could also be of use, to help understand how folds in general behave in terms of dynamics.

Bibliography

[1] Harold Hartley. Origin of the word 'protein'. *Nature*, 168, 1951.

[2] James B. Sumner. The isolation and crystallization of the enzyme urease. *J. Biol. Chem.*, 69, 1926.

[3] Kendrew J.C., Bodo B., Dintzis H.M., Parrish R.G., Wyckoff H., and Phillips D.C. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181, 1958.

[4] Berman H. M., Henrick K., and Nakamura H. Announcing the worldwide protein data bank. *Nature, Structural Biology*, 2003.

[5] Combs M. D. and Yutzey K. E. Heart valve development: Regulatory networks in development and disease. *Circulation Research*, 105, 2009.

[6] Eldar A., Dorfman R., Weiss D., Ashe H., Shilo B.Z., and Barkai N. Robustness of the BMP morphogen gradient in drosophila embryonic patterning. *Nature*, 419, 2002.

[7] Xu M. and Lewis R. V. Structure of a protein superfiber: spider dragline silk. *Proc. Natl. Acad. Sci. USA*, 87, 1990.

[8] Prockop D.J. and Kivirikko K.I. Collagens: molecular biology, diseases, and potentials for therapy. *Ann. Rev. Biochem.*, 64, 1995.

[9] Popescu C. and Höcker H. Chapter 4 – cytomechanics of hair: Basics of the mechanical stability. *International Review of Cell and Molecular Biology*, 277, 2009.

[10] Ashe H. L. BMP signalling: Synergy and feedback create a step gradient. *Curr. Biol.*, 15, 2005.

[11] Alon U. Robustness of protein circuits: The example of bacterial chemotaxis. In *An introduction to systems biology – Design principles of biological circuits*, pages 135–157. Chapman & Hall/CRC, Taylor & Francis Group, 2007.

[12] Nelson D. L. and Cox M. M. Biosignaling. In *Lehninger Principles of biochemistry*, pages 437–481. Worth Publishers, 2000.

[13] Toyoshima C. How Ca2+-ATPase pumps ions across the sarcoplasmic reticulum membrane. *Biochim. Biophys. Acta*, 1793, 2009.

[14] Raunser S. and Walz T. Electron crystallography as a technique to study the structure on membrane proteins in a lipidic environment. *Annu. Rev. Biophys.*, 38, 2009.

[15] Chothia C. Proteins - 1000 families for the molecular biologist. *Nature*, 357:543–544, 1992.

[16] Bahar I. and Rader A. J. Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struc. Biol.*, 15, 2005.

[17] Chothia C. and Lesk A. M. The relation between divergence of sequence and structure in proteins. *EMBO J.*, 5:823–826, 1986.

[18] Jäckel C., Kast P., and Hilvert D. Protein design by directed evolution. *Annu. Rev. Biophys.*, 37, 2008.

[19] Flenniken M. L., Uchida M., Liepold L. O., Kang S., Young M. J., and Douglas T. A library of protein cage architectures as nanomaterials. In Manchester M. and Steinmetz N. F., editors, *Viruses and Nanotechnology*, pages 71–93. Springer Berlin Heidelberg, 2009.

[20] Gebeshuber I. C. Biotribology inspires new technologies. *Nano Today*, 2, 2007.

[21] Toksöz S. and Guler M. O. Self-assembled peptidic nanostructures. *Nano Today*, 4, 2009.

[22] Eidhammer I., Jonassen I., and Taylor W. R. Structure comparison and structure patterns. *J. Compu. Biol.*, 7:658–716, 2000.

[23] Parry D., Fraser B., and Squire J. M. Fifty years of coiled-coils and alpha-helical bundles: A close relationship between sequence and structure. *J. Struct. Biol.*, 163, 2008.

[24] Przybylski D. and Rost B. Alignments grow, secondary structure prediction improves. *Prot. Struct. Funct. Genet.*, 46, 2002.

[25] Pawson T. and Nash P. Protein–protein interactions define specificity in signal transduction. *Genes and Development*, 14, 2000.

[26] Edwards A. Large-scale structural biology of the human proteome. *Ann. Rev. Biochem.*, 78, 2009.

[27] Ding F. and Dokholyan V. Simple but predictive protein models. *TBIO*, 23, 2005.

[28] Grishin N. V. Fold change in evolution of protein structures. *Journal of Structural Biology*, 134, 2001.

[29] Salem G.M., Hutchinson E.G., and Orengo C.A. Correlation of observed fold frequency with the occurrence of local structural motifs. *J. Mol. Biol.*, 287:969–981, 1999.

[30] Murzin A. G., Brenner S. E., Hubbard T., and Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Molec. Biol.*, 247:536–540, 1995.

[31] Orengo C. A., Michie A. D., Jones S., Jones D. T., Swindells M. B., and Thornton J. M. CATH — a hierarchic classification of protein domain structures. *Structure*, 5:1093–1108, 1997.

[32] Taylor W. R. Searching for the ideal forms of proteins. *Biochem. Soc. Trans.*, 28:264–269, 2000.

[33] Taylor W. R. A periodic table for protein structure. *Nature*, 416:657–660, 2002.

[34] Taylor W. R. Defining linear segments in protein structure. *J. Molec. Biol.*, 310:1135–1150, 2001.

[35] Johannissen L. O. and Taylor W. R. Protein fold comparison by the alignment of topological strings. *Prot. Engng.*, 16:949–955, 2004.

[36] Flores T. P., Moss D. S., and Thornton J. M. An algorithm for automatically generating protein topology cartoons. *Prot. Engng*, 7:31–37, 1994.

[37] Holm L. and Park J. Dalilite workbench for protein structure comparison. *Bioinformatics*, 16:566–567, 2000.

[38] Taylor W. R. Protein structure comparison using iterated double dynamic programming. *Prot. Sci*, 8:654–665, 1999.

[39] Taylor W. R. Protein structure comparison using SAP. In Webster D.M., editor, *Protein structure prediction*, volume 143 of *Methods in Molecular Biology (ed. J. M. Walker)*, pages 19–32. Humana Press, Totowa, New Jersey, USA, 2000.

[40] Zhang Y. and Skolnick J. TM align: A protein structure alignment algorithm based on TM-score. *Nuc. Acids Res.*, 33:2302–2309, 2005.

[41] D. Gilbert, D. Westhead, N. Nagano, and J. Thornton. Motif-based searching in TOPS protein topology databases. *Bioinformatics*, in Press.

[42] Michalopoulos I., Torrance G. M., Gilbert D. R., and Westhead D. R. TOPS: an enhanced database of protein structural topology. *Nucleic Acids Res.*, 32, 2004.

[43] Taylor W. R., Bartlett G. J., Chelliah V., Klose D., Lin K., Sheldon T., and Jonassen I. Prediction of protein structure from ideal forms. *Proteins: struct., funct., bioinfo.*, 70:1610–1619, 2008.

[44] Jones D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Molec. Biol.*, 292:195–202, 1999.

[45] Lin K., Simossis V. A., Taylor W. R., and Heringa J. A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics*, 21:152–159, 2005.

[46] Chothia C. and Finkelstein A. V. The classification and origins of protein folding patterns. *Ann. Rev. Biochem.*, 59:1007–1039, 1990.

[47] Finkelstein A. V. and Ptitsyn O. B. Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys. Molec. Biol.*, 50:171–190, 1987.

[48] Taylor W. R., Lin K., Klose D., Fraternali F., and Jonassen I. Dynamic domain threading. *Proteins, Struct. Funct. Bioinfo.*, 64:601–614, 2006.

[49] Taylor W. R. Multiple sequence threading: an analysis of alignment quality and stability. *J. Molec. Biol.*, 269:902–943, 1997.

[50] Jonassen I., Klose D., and Taylor W. R. Protein model refinement using structural fragment tessellation. *Comp. Chem. Bioinformatics*, 30:360–366, 2006.

[51] Schlick T. *Molecular Modeling and Simulation.* Springer-Verlag New York, Inc., 2002.

[52] Hinsen K. Normal mode theory and harmonic potential approximations. In Cui Q. and Bahar I., editors, *Normal mode analysis – Theory and applications to biological and chemical systems*. Chapman and Hall/CRC, 2006.

[53] Skjaerven L., Hollup S. M., and Reuter N. Normal mode analysis for proteins. *Theochem*, 2008.

[54] MacDonald J. T., Maksimiak K., Sadowski M. I, and Taylor W. R. De novo backbone scaffolds for protein design. *Prot. Struct. Funct. Genet.*, 2009.

[55] Casp 8 proceedings. Prot. Struct. Funct. Genet., 2009. Volume 77, Issue S9, Pages 1-228.