Modeling and predicting super-secondary structures of transmembrane beta-barrel proteins

Thuong van Du Tran

THÈSE

présentée pour obtenir le grade de

DOCTEUR DE L'ÉCOLE POLYTECHNIQUE

Spécialité:

INFORMATIQUE

par

Thuong Van Du TRAN

Titre de la thèse:

Modeling and Predicting Super-secondary Structures of Transmembrane β-barrel Proteins

Soutenue le 7 décembre 2011 devant le jury composé de:

MM.	Laurent MOUCHARD Mikhail A. ROYTBERG	Rapporteurs
MM.	Gregory KUCHEROV Mireille REGNIER	Examinateurs
M.	Jean-Marc STEYAERT	Directeur

Laboratoire d'Informatique
UMR X-CNRS 7161
École Polytechnique, 91128 Plaiseau CEDEX, FRANCE

Composed with LaTeX

Contents

List of Figures

List of Tables

Acknowledgement

This thesis would not have been achieved without the aid of several people who, in one way or another, contributed their valuable assistance in the preparation and completion of my study; and it is my great pleasure to thank those who made it possible.

I owe my deepest gratitude to my supervisor, in fact more than a supervisor, Jean-Marc Steyaert, for his wise guidance throughout my research work as well as his warm support in my life. I have learned so much from his insight and personality.

I am heartily thankful to my co-advisor, Philippe Chassignet, who gave me many pieces of advice and helped me overcome various difficulties in all of the time of my research.

I would like to express my sincere gratitude to my Ph.D. committee. My heartiest thanks to Laurent Mouchard and Mikhail Roytberg for taking their time to thoroughly read my manuscript and giving me valuable assessments. My deepest appreciation to Gregory Kucherov and Mireille Régnier for the reviews and comments on my thesis.

I gratefully acknowledge Saad Sheikh for all his help, enthusiasm and valuable hints during the time we worked together.

Many thanks go in particular to my group of Bioinformatics, especially Peter Clote, Thomas Simonson, Julie Bernauer, Yann Ponty, Jérôme Waldispühl, Philippe Dessen, Mahsa Behzadi and Balaji Raman for their valuable suggestions and discussions, which helped me enrich my knowledge, broaden my view and develop my skills.

I would like to give special thanks to my colleagues, Morgan Barbier, Guillaume Quintin, Jérôme Milan, Pierre Saint-Geours, with whom I have shared the same office. Thanks for the helpful discussions, the humors and, especially, the collection of soft drink cans for which I still owe a work of art. I would also acknowledge Thomas Clausen who pushed me to finish the writing of this manuscript.

I am grateful to the secretaries, Evelyne Rayssac and Catherine Bensoussan, for all their help in administrative works. Thanks to James Regis and Mathieu Guionnet for all technical assistances.

Thanks to LIX with all of its staff members for the warm environment. Thanks to Ecole Polytechnique where I have been staying since my undergraduate, for the fundings and a lot of beautiful memories.

I would like to show my gratitude to all my friends in Polytechnique, especially the Vietnamese ones, who are always beside me.

And last, but not least, I am greatly indebted to my family. I wish to thank my parents and my sister for their love and spiritual support. Words fail me to express my appreciation to my wife, whose love and dedication has brought more energy to me. Many thanks not only for her correction of my manuscripts, but also for our future son. To them I dedicate this thesis.

Abstract

The transmembrane β-barrel proteins (TMBs) are found in the outer membrane of Gram-negative bacteria, mitochondria and chloroplasts. They entirely span the biological membrane and perform a wide range of important functions. As the number of TMB structures known today is very limited, due to difficulties in experimental methods, it is arguable whether the learning-based prediction methods could work well for recognizing and folding TMBs which are not homologous to those currently known. We present a novel graph-theoretic model for classification and prediction of permuted super-secondary structures of TMBs from their amino acid sequence, based on energy minimization. The model does not essentially depend on learning. The algorithms are fast, robust with comparable performance to the best currently known learning-based methods. This method can be thus a useful tool for the genome screening. Besides the performance on prediction and classification, this study gives an insight into TMB structures regarding the physicochemical constraints of biological membranes. The predicted permuted structures can also enhance the understanding on the folding mechanism of TMBs.

Keywords: transmembrane protein, β-barrel, super-secondary structure prediction, permuted structure, Greek key, *ab initio* modeling

Résumé

Les protéines transmembranaires canaux-β (TMBs) se trouvent dans les membranes externes des bactéries à Gram négatif, des mitochondries ainsi que des chloroplastes. Elles traversent entièrement la membrane cellulaire et exercent différentes fonctions importantes. Vu qu'il y a un petit nombre des structures des TMBs déterminées, en raison des difficultés avec les méthodes expérimentales, il est douteux que ces approches puissent bien trouver et prédire les TMBs qui ne sont pas homologues avec celles connues. Nous construisons un modèle de graphe pour la classification et la prédiction de structures super-secondaires permutées des TMBs à partir de leur séquence d'acides aminés, en se basant sur la minimisation d'énergie. Le modèle ne dépend essentiellement pas de l'apprentissage. Les algorithmes sont rapides, robustes avec des performances comparables à celles des meilleures méthodes actuelles qui utilisent l'apprentissage. Cette méthode peut être donc utile pour le screening des génomes. Outre la performance de prédiction et de classification, cette étude donne une vue plus profonde de la structure des TMBs en tenant compte des contraintes physicochimiques des membranes biologiques. Les structures permutées prédites peuvent aussi aider à mieux comprendre le mécanisme du repliement des TMBs.

Mots-clefs: protéine transmembranaire, canaux-β, prédiction de structure super-secondaire, structure permutée, clé grecque, modélisation *ab initio*

Introduction

Motivation

Proteins can be considered as major elements and tools of life at the molecular scale as they carry out various functions in living organisms. These functions are expressed through their three-dimensional conformations, i.e. the way that amino acids are arranged in the 3D space. Therefore, discovering the structures helps understand the functions associated to the proteins. Besides the experimental methods, the prediction of protein structure *in silico* from the amino acid sequence with high accuracy and reliability is one of the most important tasks, yet remains a challenge in bioinformatics and computational biology.

Transmembrane proteins play many important roles in the functioning of cells such as enzymes, receptors, transporters, and channels. They are also involved in many human diseases including heart disease, cancer, Alzheimer's, depression, migraine, retinitis pigmentosa, hereditary deafness, diabetes, cystis fibrosis, etc. [29, 42, 85]. As a result, they are the targets of a majority of current medicine and of an important research area. These proteins make up $20 - 30\%$ of identified proteins in most whole genomes. However, determining the structure of transmembrane proteins with experimental methods is difficult as they are totally destabilized by the change of environment after their removal from the membrane. Solved transmembrane protein structures constitute only about $1 - 2\%$ of the RCSB Protein Data Bank (PDB) [6, 13, 23, 40, 118]. Therefore, structure prediction by computational methods for this class of proteins is of particular importance for both biological and medical sciences.

Transmembrane proteins are divided into two main types according to their conformation: α-helical bundles and β-barrels, in which the transmembrane β-barrel (TMB) proteins are much less abundant than α-helical bundles in the PDB. These TMB proteins are found in the outer membrane of Gram-negative bacteria, mitochondria and chloroplasts. They entirely span the biological membrane and perform a wide range of functions, such as porins, passive or active transporters, enzymes, defense or structural support, multidrug resistance [54, 117]. Nevertheless, only a few non-homologous TMB structures have been experimentally determined due to difficulties in the experimental methods such as X-ray crystallography or nuclear magnetic resonance spectroscopy. Moreover, the folding mechanism of TMB proteins has not been well understood yet, though they are observed

in spontaneous folding process in certain experiments *in vitro* [17, 117, 131, 132].

We particularly concentrate, in this thesis, on the super-secondary structure of TMB proteins, which describes the arrangement and interaction of the β-strands in the 3D space.

State of the art

Contrarily to the great progress in structure prediction on α-helical bundles [40], due to a tiny number of determined TMB structures, the learning-based predictions for these proteins are still far from being reliable, although various techniques have been recently developed for discriminating TMB proteins from globular and transmembrane α-helical proteins [41, 50, 51, 130], and for predicting TMB secondary structures [7, 50, 51, 96, 103, 130].

Gromiha et al. [50, 51] used the amino acid compositions of both globular and outer membrane proteins (OMPs) to discriminate OMPs and developed a feed forward neural network-based method to predict the transmembrane segments. Bagos et al. [7] produced a consensus prediction from different methods based on hidden Markov models, neural networks and support vector machines [1, 9, 16, 51, 59, 86, 89, 94]. Waldispühl et al. [130] used a structural model and pairwise interstrand residue statistical potentials derived from globular proteins to predict the supersecondary structure of TMB proteins. Randall et al. [103] tried to predict the TMB secondary structure with 1D recursive neural network using alignment profiles. Ou et al. [96] proposed a method based on radial basis function networks to predict the number of β-strands and membrane spanning regions in β-barrel outer membrane proteins. Freeman et al. [41] introduced a statistical approach for recognition of TMB proteins based on known physicochemical properties. Most of these rely on the learning assumptions in the underlying models as well as the sampling of proteins in their training data set. As the number of TMB structures known today is very limited, it is arguable whether these approaches can work well for recognizing and folding TMB proteins which are not homologous to those currently known.

Moreover, the Greek key motifs are the topological signature of many β-barrel and β-sandwich structures [139]. This raises an open question whether the TMB structures are not merely a series of β-strands where each is bonded to the preceding and succeeding ones in the sequence order, but may contain Greek key or Jelly roll motifs as well: for instance, the C-terminal domain of the outer membrane usher protein PapC (PDB:3L48). This level of structure may be described as a permutation on the order of the bonded strands.

Contribution

We present a novel graph-theoretic model (see Chapter 2 and 3) for predicting the super-secondary structure of transmembrane β-barrel proteins from their amino acid sequence.

This structure is considered as a permuted arrangement or β-strands in a barrel, in which the β-strands are paired antiparallely or parallely. The problem consists in finding the thermodynamically most stable structure, i.e. the structure of minimum energy. This protein structure prediction problem can be modeled into finding the longest *cycle-attached path* in a graph with respect to a given permutation.

Each vertex in the graph represents an amino acid segment that satisfies the conformational constraints, for instance, the length of β-strands, the hydrophobicity of side chains, the propensity for each segment to be a β-strand... A probabilistic model is built from the determined structures to calculate these propensities. It is applied as a filter for potential β-strands. Each edge presents a pair of segments whose loop in between satisfies the constraints on length, flexibility, polarity, etc. The energies are assigned to the vertices, the edges, as well as to the interaction between each pair of pairing segments.

The amino acids are constructed in the three-dimensional space using the Dunbrack rotamer library. We then calculate the energies as the average on all rotamers. The hydrophobic interaction is computed on each pair of residue side chains using well-known hydrophobicity scales, while the electrostatic interactions between two amino acids are obtained thanks to the partial charges in the molecular mechanics force fields.

We prove the NP-completeness of the problem of finding the optimal permuted super-secondary structure. Then, a dynamic programming-based algorithm is proposed and implemented. This algorithm can find the optimum with a complexity in time of at most $\mathcal{O}(N^4)$ for the structures containing disjoint Greek key motifs (see Chapter 2). This complexity is improved to $\mathcal{O}(N^3)$ with another algorithm that uses the concept of tree decomposition (see Chapter 3).

To evaluate the performance of our method, we test the program on all TMB sequences with known structures in the PDBTM database (see Chapter 4). We show the accuracy of the approach with the F-score, sensitivity, specificity of more than 90% in the measure on β-strands and more than 74% in the measure on residues, which are comparable to the best learning-based methods. The ability of discrimination is also robust with 100% of α-helical transmembrane proteins and 97% β-barrel lipocalins being rejected. It also shows the ability to find the arrangement of β-strands with the "right permutation" locating in the zone of 0.7% - 1.5% of lowest-energy permutations. This method is thus potentially a useful tool for the genome screening. Beside the performance on prediction and classification, this study provides insight into TMB structures regarding the physicochemical constraints of biological membranes. The predicted permuted structures can also enhance the understanding on the folding mechanism of TMB proteins.

The program can be executed via the web-server BBP (Beta Barrel Predictor) (http://www.lix.polytechnique.fr/Labo/Van-Du.Tran/bbp/).

Organization

The manuscript is organized as follows:

Introduction

This chapter presents the motivation of the work, the state of the art in this research area, the summary of our contribution and an outline of the manuscript.

Chapter 1

Fundamental review of proteins. We remind the fundamental notions in biology concerning the proteins and the methods of protein structure prediction.

Chapter 2

Folding β-barrels. We introduce our model and algorithm for determining the protein structure of minimal energy, then provide an analysis on the computational complexity with regard to different types of structures.

Chapter 3

Tree-decomposition based algorithm. We present an algorithmic improvement based on the tree decomposition technique, followed by an analysis on its computation complexity.

Chapter 4

Evaluation of performance of BBP (Beta-Barrel Predictor). We assess the performance of our prediction model on the experimentally determined structures.

Conclusion and perspectives

The final chapter summarizes our work and suggests further research directions.

Chapter 1

Fundamental review of proteins

1.1 Introduction

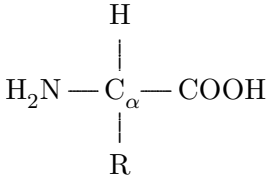
This chapter provides the reader with fundamental notions in biology that are mentioned throughout the manuscript and necessary for understanding the practical motivation of our work. The content is inspired from the Ecole Polytechnique text book of molecular and cellular biology by Yves Gaudin, Arnaud Echard and Sandrine Etienne-Manneville [44], the book on membrane structural biology by Mary Luckey [82], and Jérôme Waldispühl's PhD thesis [129].

We rapidly present the amino acids, constituent of proteins, before describing the properties and structures of the proteins themselves. Then, we focus on the class of transmembrane proteins, especially the β-barrels which are the subject of our whole work. We finally describe the problem of protein structure prediction and present the methods that have been developed to solve it.

1.2 Proteins

1.2.1 Amino acids

Amino acids have the general form:



They contain an amine group NH_2, a carboxylic group COOH and an organic substituent R. In aqueous solution at neutral pH, amino acids exist in the zwitterionic form where the amine functional group is protonated (NH_3^+) and the carboxylic functional group is deprotonated (COO^-). The substituent R, also called *side chain*, varies between 20 different standard amino acids. The four groups attached to the α-Carbon are distinguished (except for Glycine in which the side chain R consists of a hydrogen atom). Therefore, there exists two reflection-symmetric isomers L and D (see Figure 1.1), of which only L isomers are present in proteins.

Figure 1.1: Isomers L and D of amino acids

The 20 standard amino acids are shown in Figure 1.2. Each amino acid is associated with a 3-letter abbreviation and a 1-letter code which we will use throughout our work.

1.2.2 Properties of amino acids

The individual properties of constituent amino acids play a major role in determining the conformation and function of the protein. They are determined by the amino acid side chains. We make use of certain particular properties in this work, such as electric charge, polarity and hydrophobicity which are able to be quantified.

Among these, the hydrophobicity is the most important factor. It measures the capacity of the amino acid to interact with water molecules or more generally its behavior in the solvent. Several hydrophobic scales have been developed [31, 36, 37, 61, 72, 107, 108, 131, 133, 134] (see Table 1.1). They are clearly different due to the various methods that are used for measuring the hydrophobicity. Some methods examine proteins with known three-dimensional structures and define the hydrophobic character as the tendency for a residue to be found inside the protein rather than on its surface. Others result from the physiochemical properties of the amino acid side chains. The widely used Kyte-Doolittle scale [72] can help detect hydrophobic regions in proteins, in which regions with a positive value are considered hydrophobic. This scale can work for predicting surface-exposed regions as well as for finding transmembrane domains. The Engelman scale [37], or GES scale, is useful for prediction of transmembrane regions in proteins. Eisenberg *et al.* [36] proposed a normalized consensus scale which has many common features with

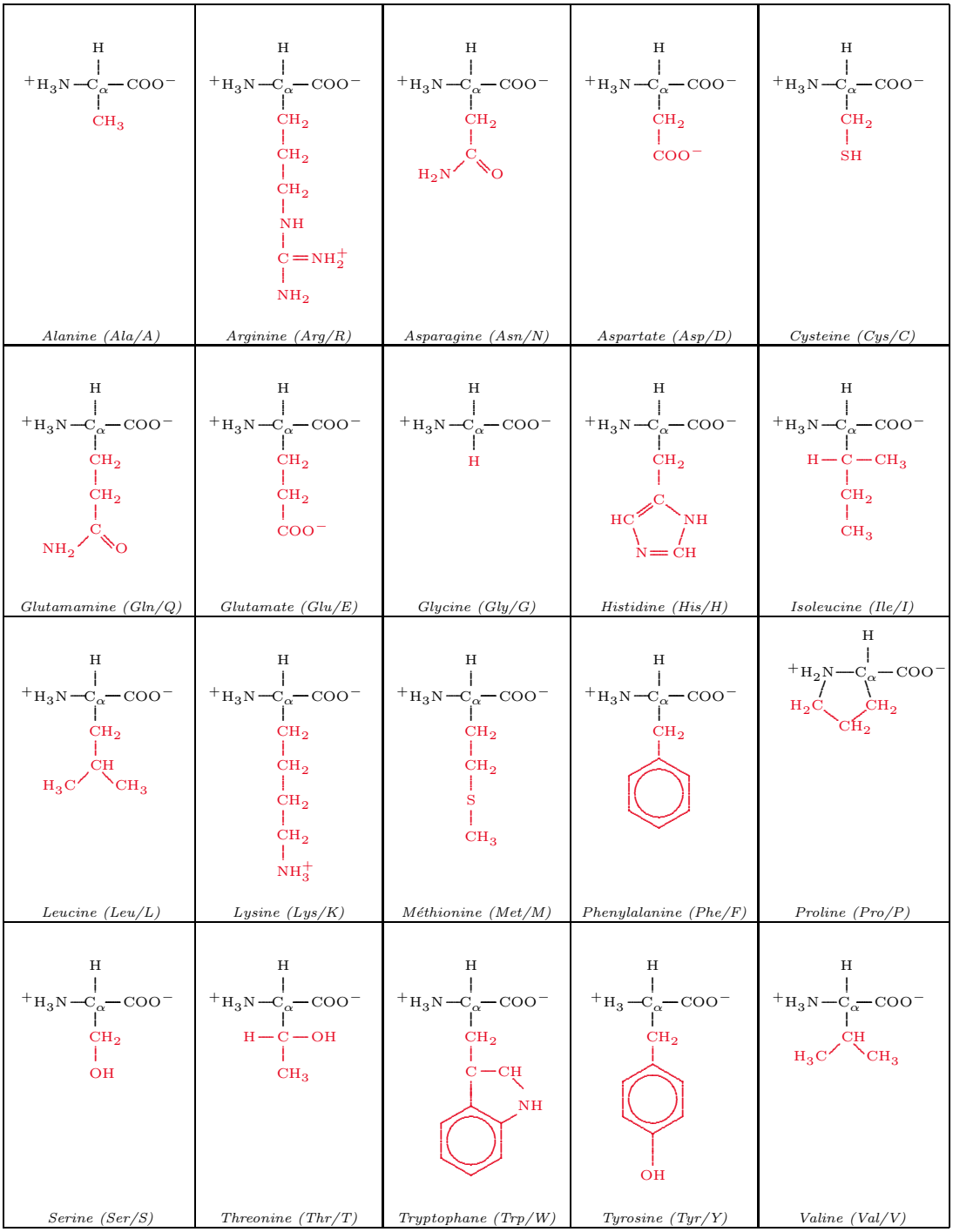


Figure 1.2: The 20 amino acids. The side chains are in red.

other hydrophobicity scales. Hopp-Woods scale [58] can be used for identification of putative antigenic sites in proteins. Cornette *et al.* [31] compared thirty-eight published hydrophobicity scales for their ability to identify the amphipathic α-helices and proposed an optimized scale using the eigenvector method. Janin scale [61] and Rose scale [107] evaluate the accessible and buried amino acid residues of globular proteins. Certain scales are calculated for specific classes of proteins: for instance, White & Wimley scale [131] evaluates the ability of amino acids to penetrate the hydrophobic membrane environment.

Amino acid	Kyte-Doolittle	Hopp-Woods	Cornette	Eisenberg	Rose	Janin	Engelman (GES)	Wimley-White
A	1.80	-0.50	0.20	0.62	0.74	0.30	1.60	0.50
R	-4.50	3.00	1.40	-2.53	0.64	-1.40	-12.3	1.81
N	-3.50	0.20	-0.50	-0.78	0.63	-0.50	-4.80	0.85
D	-3.50	3.00	-3.10	-0.90	0.62	-0.60	-9.20	0.43
C	2.50	-1.00	4.10	0.29	0.91	0.90	2.00	-0.02
Q	-3.50	0.20	-2.80	-0.85	0.62	-0.70	-4.10	0.77
E	-3.50	3.00	-1.80	-0.74	0.62	-0.70	-8.20	0.11
G	-0.40	0.00	0.00	0.48	0.72	0.30	1.00	1.15
H	-3.20	-0.50	0.50	-0.40	0.78	-0.10	-3.00	0.11
I	4.50	-1.80	4.80	1.38	0.88	0.70	3.10	-1.12
L	3.80	-1.80	5.70	1.06	0.85	0.50	2.80	-1.25
K	-3.90	3.00	-3.10	-1.50	0.52	-1.80	-8.80	2.80
M	1.90	-1.30	4.20	0.64	0.85	0.40	3.40	-0.67
F	2.80	-2.50	4.40	1.19	0.88	0.50	3.70	-1.71
P	-1.60	0.00	-2.20	0.12	0.64	-0.30	-0.20	0.14
S	-0.80	0.30	-0.50	-0.18	0.66	-0.10	0.60	0.46
T	-0.70	-0.40	-1.90	-0.05	0.70	-0.20	1.20	0.25
W	-0.90	-3.40	1.00	0.81	0.85	0.30	1.90	-2.09
Y	-1.30	-2.30	3.20	0.26	0.76	-0.40	-0.70	-0.71
V	4.20	-1.50	4.70	1.08	0.86	0.60	2.60	-0.46

Table 1.1: Hydrophobic scales

Table 1.2 shows other physicochemical properties, such as polarity [48], flexibility [15], volume [138] and surface area [27] associated to amino acids.

The 20 amino acids are classified into different categories regarding the properties of their side chain. The following is the most common classification.

- Glycine is the most simple amino acid with a hydrogen atom in the side chain.
- Alanine, valine, leucine and isoleucine possess an aliphatic side chain that makes them hydrophobic.

Amino acid	Polarity	Flexibility	Volume	Surface
A	8.1	0.36	88.6	115
R	10.5	0.53	173.4	225
N	11.6	0.46	114.1	160
D	13.0	0.51	111.1	150
C	5.5	0.35	108.5	135
Q	10.5	0.49	143.8	180
E	12.3	0.50	138.4	190
G	9.0	0.54	60.1	75
H	10.4	0.320	153.2	195
I	5.2	0.46	166.7	175
L	4.9	0.370	166.7	170
K	11.3	0.47	168.6	200
M	5.7	0.30	162.3	185
F	5.2	0.31	189.9	210
P	8.0	0.51	112.7	145
S	9.2	0.51	89.0	115
T	8.6	0.44	116.1	140
W	5.4	0.31	227.8	255
Y	6.2	0.42	193.6	230
V	5.9	0.39	140.0	155

Table 1.2: Polarity, flexibility and other physicochemical parameters of amino acids

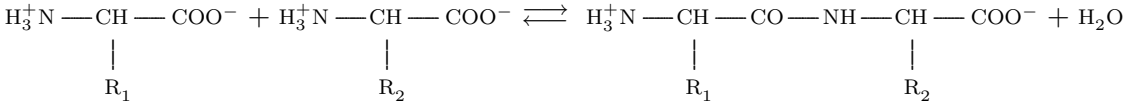
- Serine and threonine have an aliphatic side chain with a polar hydroxyl group.
- Phenylalanine, tyrosine and tryptophan contain an aromatic group. The hydroxyl function of tyrosine is a weak acid with $pK_a \sim 10$. Tyrosine is then ionizable but not ionized in physiological conditions.
- Lysine, arginine and histidine are basic. Lysine and arginine have a high pK_a in solution (10.5 and 12.5, respectively), and thus positively charged in physiological conditions. The low pK_a of histidine (~ 6) makes it neutral or protonated following the pH of the solution.
- Aspartate and glutamate are acid (with low pK_a of about 3.9 and 4.3, respectively) and negatively charged at neutral pH (named also aspartic acid and glutamic acid).
- Asparagine and glutamine are the amidated products of aspartate and glutamate, and thus not ionisable.
- Cysteine and methionine possess a sulphur atom in their side chain. The sulfhydryl group in cysteine is a highly potent nucleophile and also a weak acid. It can be

easily oxidized to form with another cysteine a disulfide bond which stabilizes the tridimensional conformation of proteins.

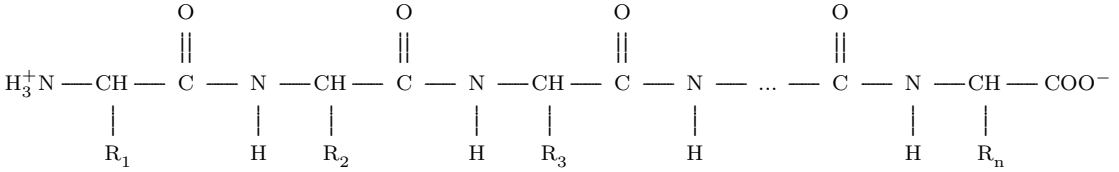
- Proline has a formula that is different from other amino acids. The cyclic secondary amino function gives it a specific role in the establishment of the tridimensional structure of proteins.

1.2.3 Peptide bond

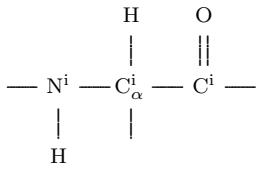
A peptide bond is a covalent bond formed between the α-carboxylic group of an amino acid and the α-amine group of the other one. This process combines two amino acids into an amide (dipeptide) and releases a molecule of water (H_2O). It is thus called a dehydration reaction or a condensation reaction, which is written as:



Amino acids in a protein are covalently linked together by peptide bonds to form a non-branching polypeptide chain. A unit of amino acid is called a residue. A polypeptide possesses an amino-terminal extremity (N-terminus) and an carboxy-terminal extremity (C-terminus). The synthesis of a polypeptide is carried out in a so-called "translation" process, where residues are consecutively added from its N-terminus. The N-terminus is then considered as the beginning of the chain.

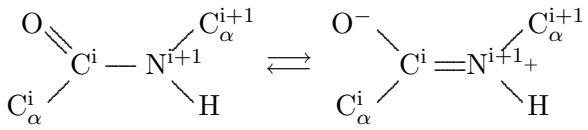


A polypeptide chain is composed of a series of repetitive bonded atoms, namely *backbone* or *main-chain*,

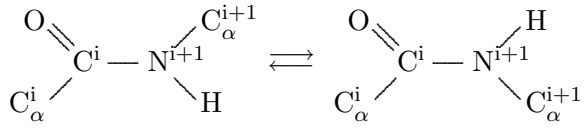


and a variable part of amino acid side chains R_i, where i denotes the residue position counting from the N-terminus. These side chains precisely determine the specific properties and functions of each protein. The sequence of amino acids of a polypeptide chain is known as its *primary structure*.

The peptide bond has characteristics of a double bond due to the mesomeric (resonance) effect, thus the six atoms above are coplanar, making a *peptide plan*.



Two configurations, called *trans* and *cis*, occur according to whether the two α-carbons are on the same or opposite side, respectively.



trans *cis*

The *trans* configuration is energetically favored as it causes less repulsion between non-bonded atoms. The crystallographic studies showed almost constant values of distances and angles of the peptide bond for every polypeptide chain (see Figure 1.3).

As the geometry of a peptide plane is fixed, the torsion angles ϕ and ψ are two degrees of freedom in determining the conformation of the polypeptide chain. ϕ is the dihedral angle around the N–C_α bond, determined by the two carbons CO. ψ is around C–C_α

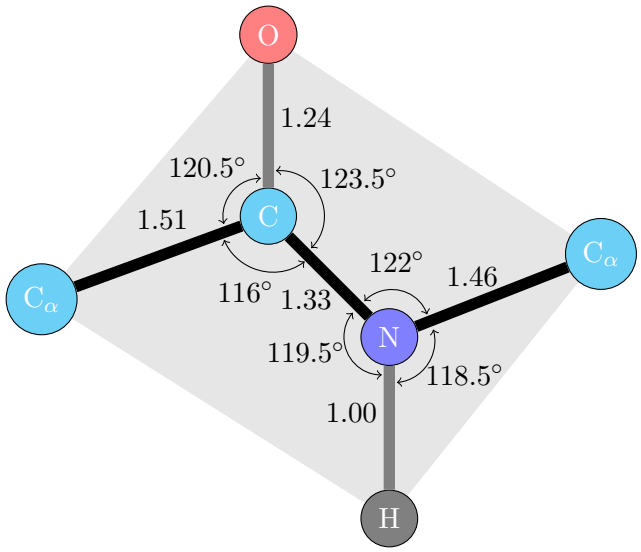


Figure 1.3: Peptide bond geometry in *trans* configuration

bond, determined by the two nitrogens N (see Figure 1.4. There are strong constraints on the angles ϕ and ψ. Certain combinations are clearly impossible, while some others are energetically unfavorable. Ramachandran *et al.* [100, 101] introduced Ramachandran diagram to visualize graphically the backbone dihedral angles ϕ and ψ in the polypeptide chain of proteins. Each amino acid in the protein is represented with the coordinate (ϕ, ψ) in the plot in the range of $[-180°, 180°]$ [81]. The Ramachandran diagram of the constituent amino acids of the outer membrane protein A (PDB:1BXW) is presented in Figure 1.5[1]. The limited regions of distribution of (ϕ, ψ) prove the restricted flexibility of the polypeptide chain.

1.2.4 Protein

Proteins are macromolecules constituted by a large number of amino acids, from a few dozens to several hundred. This is one of the four important organic macromolecules in living organisms, along with nucleic acids, carbohydrates and lipids. Many proteins are composed of only one polypeptide chain (namely *monomer*). Others can be formed of more than one chains, and thus are called *oligomers* (e.g., *dimer*, *trimer*, *tetramer*...). If these chains are identical, the protein is called *homo-oligomer*. Otherwise, it is a *hetero-oligomer*. Each constituent chain is a subunit, also known as a *protomer*.

Proteins are essential in organisms and take part in almost every process in the cells. They are usually classified into three major classes according to their overall three-

[1]Image generated by MolProbity web-server [26, 32]

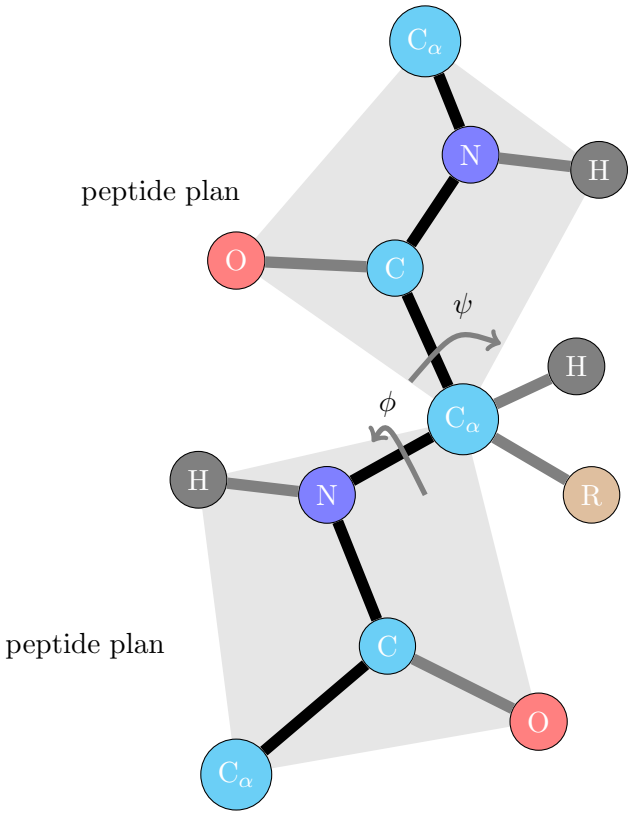


Figure 1.4: Torsion angles between two peptide plans

dimensional structures and their functional roles: fibrous, globular and membrane proteins.

- Fibrous proteins (or scleroproteins), which tend to be elongated fibers, are generally inert and insoluble. These proteins are usually constructed of repetitive amino acid sequences. These characteristics make them appropriate to play structural roles in organisms for supportive and protective function. For example, keratin constructs hair, nails, and skin...; collagen is abundantly found in connective tissues such as cartilage, tendons...; elastin is important in ligaments, blood vessels.... An example of collagen is given in Figure 1.6[2].

- Globular proteins, which comprise a large variety of proteins, are soluble and exist in an aqueous environment. Hence, these proteins generally have compact structures with polar residues on the surface and hydrophobic residues in the core. These

[2]Image generated by PyMOL [113]

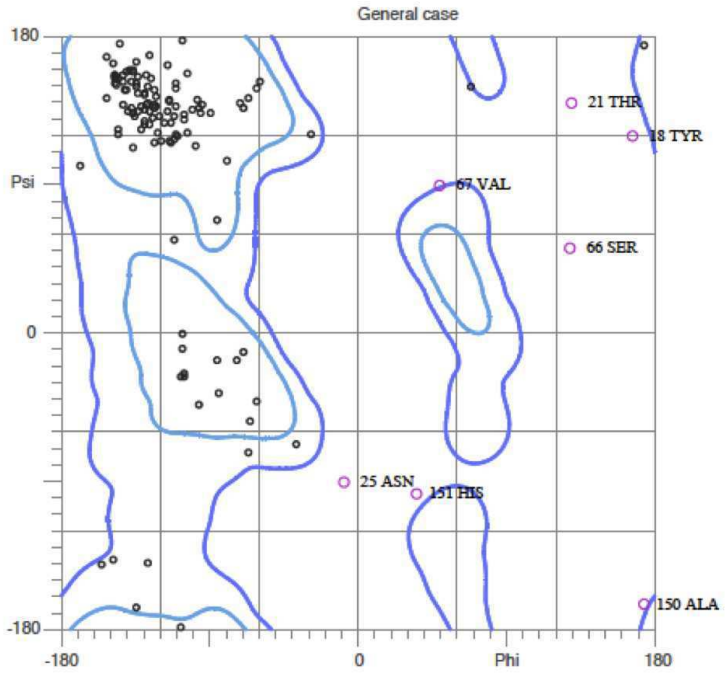


Figure 1.5: Ramachandran plot for the outer membrane protein A (PDB:1BXW)

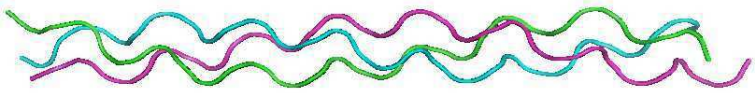


Figure 1.6: Structure of collagen (PDB:1BKV)

proteins are the most described in the Protein Data Bank (PDB) [13], since their structures are usually stable, and thus easy to determine experimentally. Two of the most known globular proteins, myoglobin and hemoglobin, are the first two experimentally determined structures by John Cowdery Kendrew [67] and Max Ferdinand Perutz [97], which led to them receiving a Nobel Prize in Chemistry in 1962. The structure of myoglobin is presented in Figure 1.7[3].

- Membrane proteins exist in the cell membranes – a phospholipid bilayer with hydrophobic core. They typically have hydrophobic exposed regions in order to be stable in such an environment. Some proteins slightly adhere to the membrane,

[3]Image generated by PyMOL [113]

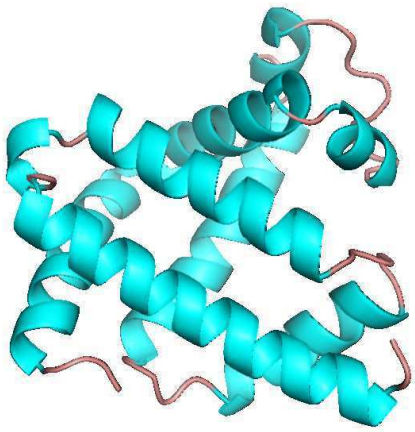


Figure 1.7: Structure of myoglobin (PDB:1A6M)

while others are embedded in the lipid bilayer. Among the latter, some proteins, namely transmembrane proteins, entirely span the biological membrane one or several times (*polytopic* proteins). Figure 1.8[4] illustrates the structure of insulin receptor, a well known transmembrane protein which helps induce glucose uptake, thus causes diabetes in case of its insensitivity.

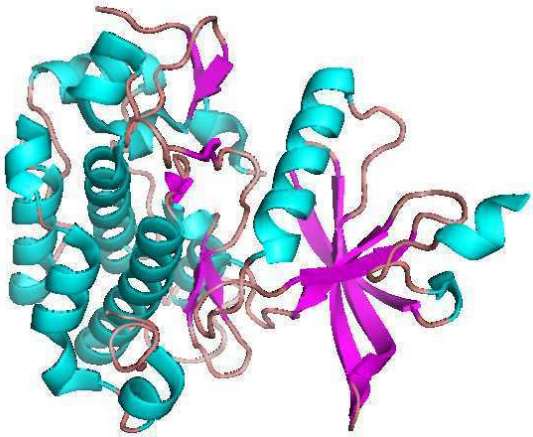


Figure 1.8: Structure of insuline receptor (PDB:1GAG)

[4]Image generated by PyMOL [113]

1.2.5 Protein structure

The structure of a protein can be decomposed into different structural elements which allow to describe it in some level of precision. The standard classification proposed by Linderstrom-Lang [78, 79] defined four structural levels: *primary*, *secondary*, *tertiary* and *quaternary*.

a. Primary structure

As mentioned in 1.2.3, the primary structure is the sequence of amino acids constituting the polypeptide chain: $R_1R_2 \dots R_n$.

b. Secondary structure

The secondary structure represents the local conformation of the polypeptide chain. Three main types of secondary structures are found: α-helices, β-sheets and loops.

α-helix

An α-helix is stabilized with hydrogen bonds between the C=O group in the main chain of residue i and the N−H group in the main chain of residue $i+4$. In such a regular structure, all residues are involved in hydrogen bonds. Generally, there are two other kinds of bonding though they are much less frequent. The 3.10-helices and π-helices are characterized by hydrogen bonds between residues i and $i+3$, and between residues i and $i+5$, respectively.

An α-helix is geometrically considered as a chain of periodic tours which correspond to a 5.4Å translation along the helix axis. Each tour contains, on average, 3.6 amino acids, thus the amino acids are translated 1.5Å along the axis. The structure of an α-helix is illustrated in Figure 1.9.

β-sheet

A β-sheet is composed of β-strand subunits. A β-strand can be considered as a degenerated helix with 2 amino acids per tour. Each strand interacts with its neighbors through hydrogen bonds between the C=O and N−H groups in the main chains. As in helices, all residues in a regular β-sheet are involved in hydrogen bonds. This bonding associates the β-strands to each other, making the β-sheet stable.

β-sheets are separated into two types regarding whether the constitutive β-strands are parallel or antiparallel, which is determined by the direction of the pairing β-strands (see Figure 1.10). The β-sheet structure generated by antiparallel pairing is found more frequently than the one with parallel pairing, as the former is naturally more stable thanks to a better arrangement of residues.

The torsion angles ϕ and ψ are respectively around $-119°$ and $+113°$ for parallel β-sheets, and around $-139°$ and $+135°$ for antiparallel ones. The distance between two consecutive residues in a strand is about 3.5Å. In addition, the large β-sheets are not

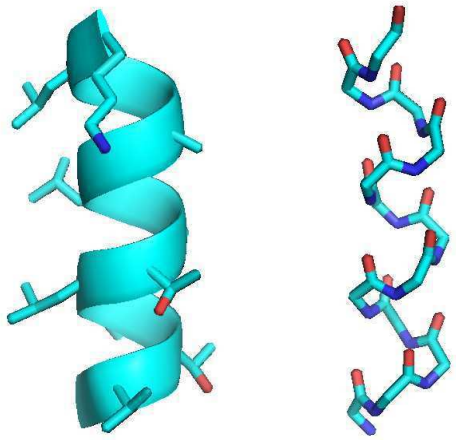


Figure 1.9: Structure of an α-helix

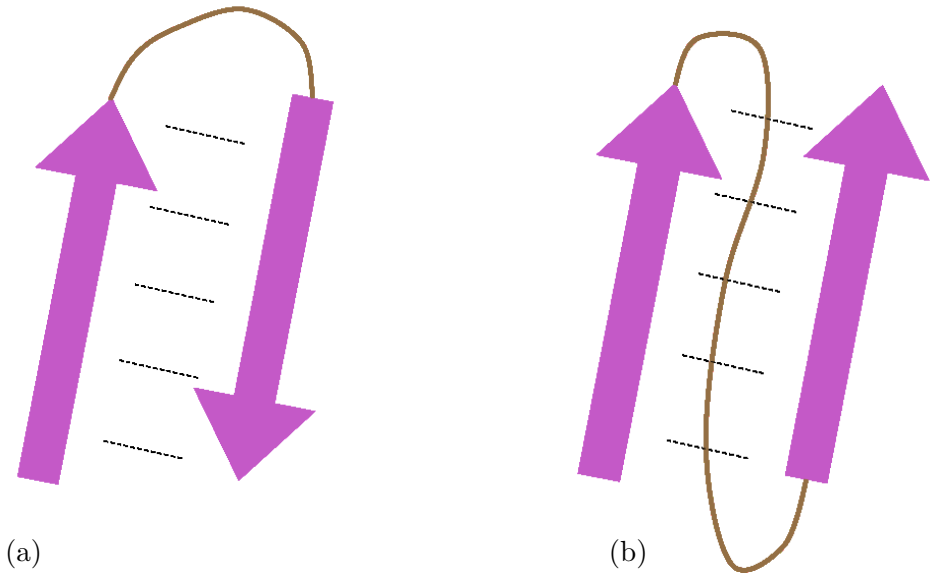


Figure 1.10: Antiparallel pairing (a) and parallel pairing (b) of β-strands

plane, but rather make the curved surfaces. The residue side chains are alternatively located on the two sides of the β-sheet. Frequently, the β-sheets possess a hydrophobic surface oriented towards the protein interior and a hydrophilic surface oriented towards the solvent. An illustration of β-sheet characteristics is presented in Figure 1.11[5].

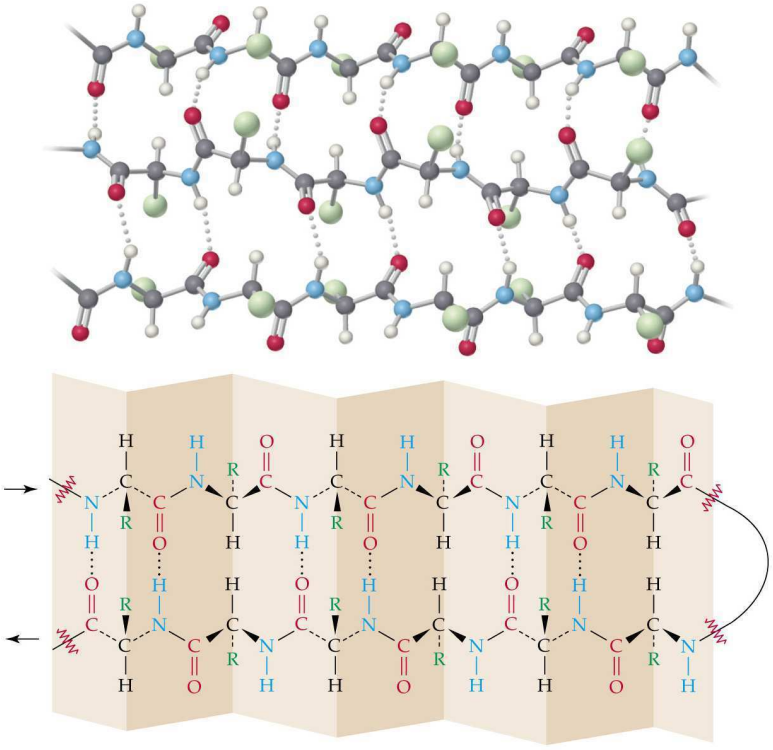


Figure 1.11: Characteristics of a β-sheet.

c. Tertiary structure

The tertiary structure is the tridimensional conformation of the polypeptide chain, i.e. the relative coordinates of all atoms constituting the protein. This level of structure is essentially stabilized by hydrophobic interaction. There is a considerable difference on the precision of description between secondary and tertiary structures. Hence, the super-secondary structure appears as an intermediary description level. This describes the secondary structure as well as its interactions. Figure 1.12[6] illustrates the tertiary and super-secondary structure of the cystic fibrosis transmembrane conductance regulator.

[5]Figure retrieved from http://wps.prenhall.com/wps/media/objects/602/616516/Chapter_24.html

[6]Image generated by PyMOL [113]

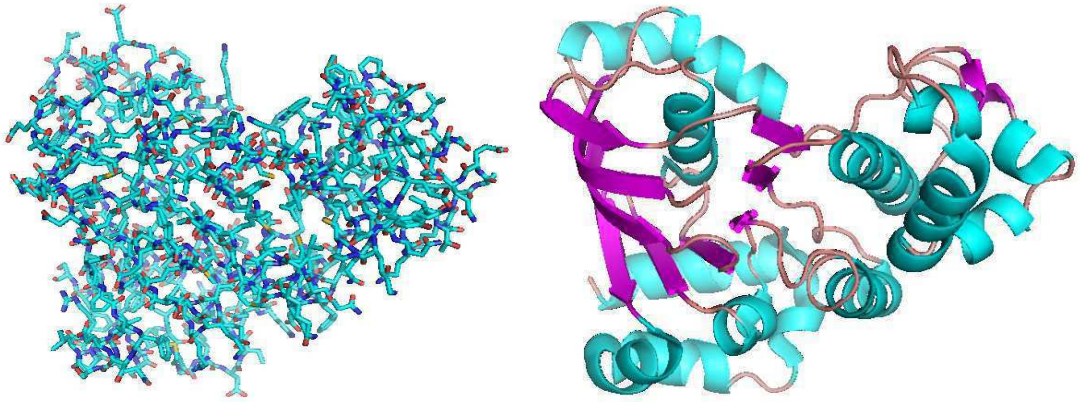


Figure 1.12: Tertiary structure (a) and super-secondary structure (b) of the cystic fibrosis transmembrane conductance regulator (PDB:1R0W)

d. Quaternary structure

When the protein is a multi-subunit complex, i.e. a composition of several polypeptide chains, the quaternary structure describes the arrangement of these chains (stoichiometry, interaction interface, symmetry,...). Figure 1.13[7] presents the quaternary structure of human hemoglobin, which is a heterotetramer ($\alpha_2\beta_2$) composed of two heterodimers ($\alpha\beta$).

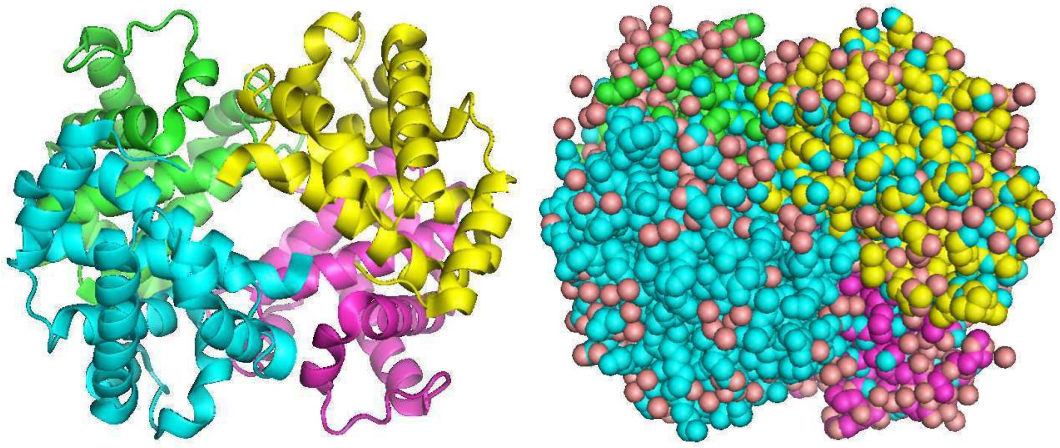


Figure 1.13: Quaternary structure of human hemoglobin (PDB:1MKO)

[7]Image generated by PyMOL [113]

1.3 Transmembrane proteins

1.3.1 Biological membrane

Before introducing the transmembrane proteins, it is appropriate to start with biological membranes, the environment where those proteins are located. The constitutive molecules of living organisms are contained in cells − compartments that allow the existence of a privileged environment in a restricted volume that differs from outside. This presents a thermodynamic advantage since it increases the probability of interaction of molecules, and thus the occurrence of chemical reactions. Such an enclosed space is defined by a *plasma membrane* (or cell membrane). This membrane separates the intracellular compartment, namely *cytoplasm*, and the extracellular environment. It not only determines the border of the cell, but it also helps maintain the difference of concentrations between the exterior and interior mediums, favor the entrance of nutrients into the cell, contribute to the elimination of waste of metabolism, and play an important role in intercellular communication.

All the biological membranes have a common structure. This is a two-layered sheet (also *bilayer*) composed of two layers of lipid molecules [2, 47, 82] with embedded proteins (see an illustration in Figure 1.14). The essential property of the membrane lipids, such as phospholipids, glycolipids and cholesterol, is their amphiphilic (or amphipathic) nature, i.e. they comprise both hydrophilic regions (dissolvable in water or "water-loving") and hydrophobic regions (insoluble in water or "water-fearing"). The lipid bilayer is spontaneously formed as an assemblage of lipid molecules, thanks to such a characteristic, with hydrophobic portions pointing toward the interior of the sheet, making this region free from water. The two hydrophilic surfaces of the sheet are then exposed to the aqueous mediums (intra- and extra-cellular environments). This gives the lipid bilayer two important properties. On the one hand, with a hydrophobic core, the membrane is impermeable to most biological molecules, such as nucleic acids, amino acids, proteins, sugars or ions. Thus, the membrane acts as barrier between intra- and extra-cellular mediums. On the other hand, the lipid bilayer forms a two-dimensional liquid in which the constituent molecules can be rapidly laterally rearranged.

Membrane proteins are embedded in the lipid bilayer and ensure most of membrane functions. They constitute about 50% of the membrane mass [115]. We distinguish membrane proteins according to their interaction with the membrane. These are illustrated in Figure 1.14[8].

- Transmembrane proteins are permanently attached to the membrane and span across the bilayer.
- Lipid-anchored proteins are attached to the lipid bilayer by a lipidated anchor.

[8]Figure retrieved from http://commons.wikimedia.org/wiki/File:Cell_membrane_detailed_diagram_en.svg

- Peripheral proteins are located at the membrane surface. They are essentially bound to lipid bilayer or transmembrane proteins by electrostatic interaction.

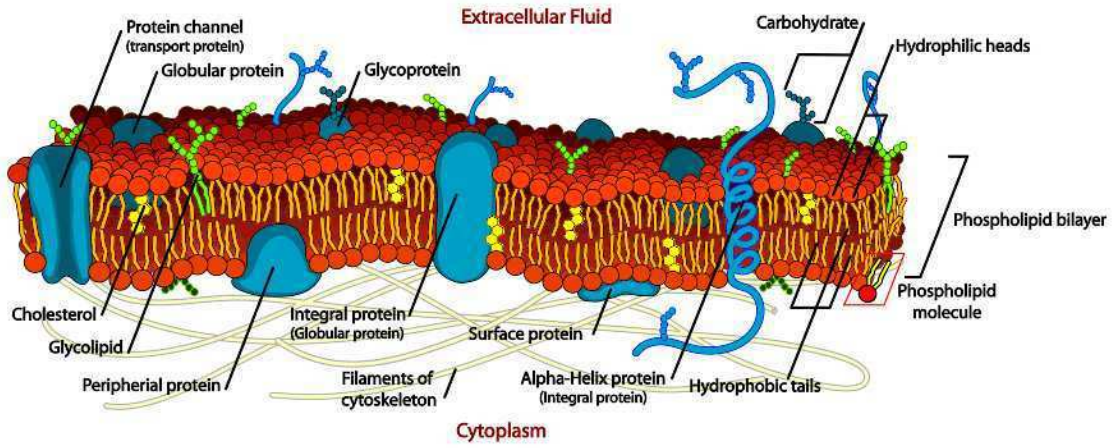


Figure 1.14: Illustration of a biological membrane and embedded membrane proteins.

1.3.2 Transmembrane proteins

Transmembrane proteins entirely span across the biological membranes. The hydrophobic domains included in the proteins allow them to interact with the hydrophobic center of the lipid bilayer (see Figure 1.15[9]). They can possess one or more successive hydrophobic domains, and thus, can traverse the membrane one or several times. Certain proteins can also partially penetrate the bilayer. The extraction of these proteins is difficult and requires detergents, nonpolar solvents or denaturing agents, causing a denaturation.

Transmembrane proteins play several key roles in the human body including inter-cell communication, transportation of nutrients, and ion transport, etc. They also play key roles in human diseases like heart disease, cancer, Alzheimer's, depression, migraine, retinitis pigmentosa, hereditary deafness, diabetes, cystis fibrosis, etc. [29, 42, 85], and thus are targeted by a majority of pharmaceuticals being manufactured today.

The transmembrane proteins are divided into two main types according to their conformation: α-helical bundles and β-barrels. These proteins make up $20-30\%$ of identified proteins in most whole genomes. However, due to difficulties in determination of their structures, solved TMB structures constitute only a meagre 2% of the RCSB Protein Data Bank (PDB) [6, 13, 23, 118].

[9]Figure retrieved from http://commons.wikimedia.org/wiki/File:Polytopic_membrane_protein.png

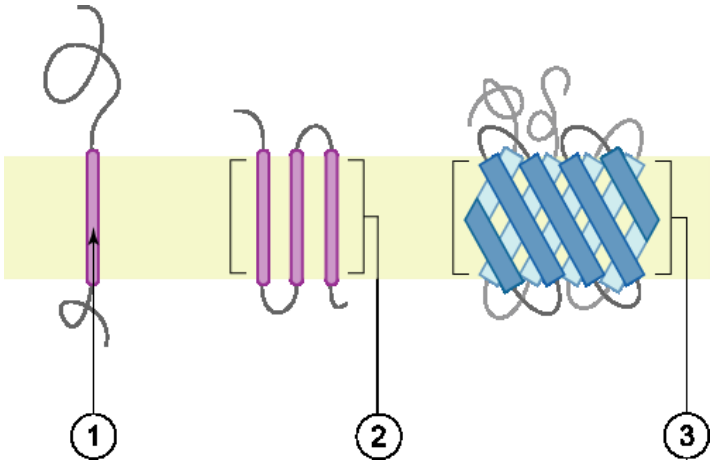


Figure 1.15: Transmembrane proteins: (1) a single transmembrane hydrophobic α-helix - bitopic membrane protein, (2) several transmembrane hydrophobic α-helices, (3) transmembrane β-barrel protein.

a. α-helical bundles

Transmembrane α-helices dominate the picture of transmembrane proteins with early structural information on bacteriorhodopsin in 1970s [57, 68] and with the first X-ray structure solved for membrane proteins, that of the photosynthetic reaction center [34] (which led to authors receiving a Nobel Prize in Chemistry in 1988). The majority of transmembrane proteins with solved structures fall in this class. These α-helical bundles are found in all types of biological membranes. A bundle is composed of a certain number of helices arranging in such a way as to create a channel through the membrane. These membrane spanning helices are generally constituted by a large majority of hydrophobic amino acids in order to adapt to the hydrophobic characteristics of the biological membrane.

The folding process of α-helical bundles is assumed to be decomposed into two stages [98]. In stage 1, the transmembrane α-helical segments are formed (stabilized by hydrogen bonds along the backbone) and insert independently into the bilayer (driven by the hydrophobic effect), and in stage 2, they assemble by packing together (driven by intrinsic forces such as packing, electrostatic interactions, hydrogen bonds between side chains, interactions between the loops between helices and components at the surface of the membrane, etc.).

Bacteriorhodopsin, which is shown in Figure 1.16[10], is the well-known representative of transmembrane α-helices.

[10]Image generated by PyMOL [113]

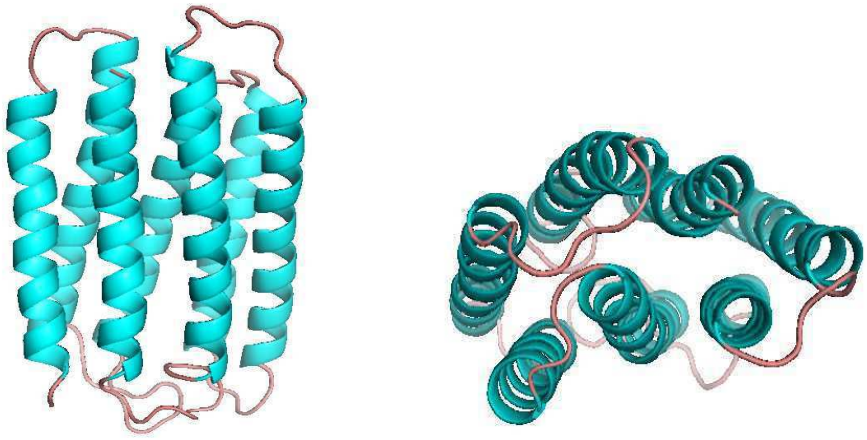


Figure 1.16: Bacteriorhodopsin in purple membrane (PDB:2BRD)

b. β-barrels

This class is central to our concern in this thesis. The transmembrane β-barrel (TMB) proteins whose solved structures are much less abundant than those of helical bundles are found in the outer membrane of Gram-negative bacteria, mitochondria and chloroplasts. Gram-negative bacteria characteristically possess two membranes: an inner cytoplasmic membrane and an outer membrane facing the extracellular environment. The latter is an asymmetric bilayer with an outer leaflet composed of lipopolysaccharide and an inner leaflet composed of phospholipids [117]. Beside the important roles in the interaction of symbiotic or pathogenic bacteria with the host organisms, the outer membrane usually acts as a permeability barrier to prevent the penetration of noxious substances and to allow the influx of nutrient molecules [95]. This is similar to mitochondria and chloroplasts. The TMB proteins located in those outer membranes perform diverse functions such as porins, passive or active transporters, enzymes, defense or structural support, multi-drug resistance [54, 117]. The structure of TMB proteins is thus very important for both biological and medical sciences.

As the number of determined TMB structures are very limited [125], the principles governing their formation are still not thoroughly clear. The folding mechanism of TMB proteins is unlike that of α-helical bundles, because each helix can be formed independently thanks to hydrogen bonds along the backbone while β-barrels necessitate hydrogen bonds between neighboring strands. Certain experiments *in vitro* result in observations that the outer membrane proteins spontaneously fold into lipid bilayers [17, 117, 131, 132]. TMB proteins are assumed to insert and fold into lipid bilayers in such a way that the transmembrane β-hairpins are concertedly translocated. The closure of β-barrels is synchronized to its formation, i.e. the hydrogen bonds between β-strands have to form along

with the translocation of the protein across the membrane [70, 117].

The TMB proteins are usually created by a succession of antiparallely paired β-strands forming a channel. A β-barrel can be considered as a self-closed β-sheet. The observed structures are formed by 8 to 22 β-strands which incline at an angle of 20°to 45°with respect to the barrel axis. Each of these β-strands comprises about 9 to 11 residues. While 8 appears to be the lower bound on the number of necessary β-strands to form a channel [112], the upper bound of 22 is only obtained by experimental observation [102]. The β-barrels are usually constituted by an even number of β-strands, which allows an antiparallel pairing at the barrel closure. An illustration of TMB protein is given in Figure 1.17[11].

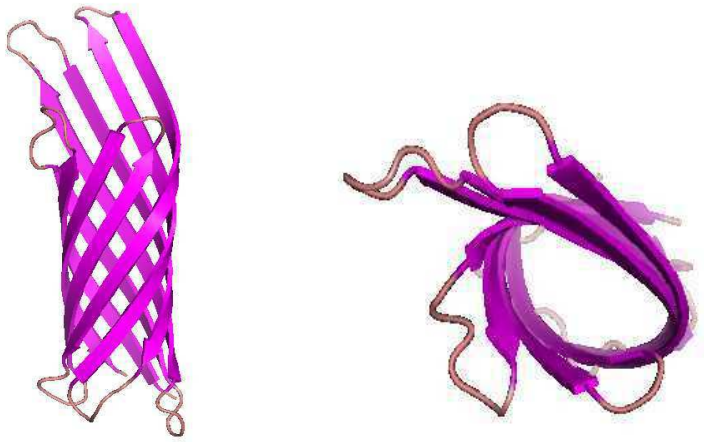


Figure 1.17: Outer membrane protein X (PDB:1QJ8)

1.4 Folding energy

The function of a protein is determined by the arrangement of its atoms in the 3D space. This conformation is stabilized by non-covalent interactions (except for disulfide bonds) between protein atoms as well as between protein atoms and water molecules in the medium. These interactions induce an energy, namely *folding energy*. It is widely assumed that the most stable structure is the one possessing the minimal folding energy, yet we will not discuss the pertinency of this assumption in this thesis. The folding energy involves various components that are briefly described below.

[11] Image generated by PyMOL [113]

1.4.1 Partial charges

A partial charge is a charge with a magnitude of less than one elementary charge unit (i.e. the charge of an electron). Partial charges of atoms are created due to the asymmetric distribution of electrons in chemical bonds. These charges are used to assess the energy of interactions. Their values are computed in various molecular mechanics force fields, such as AMBER [24], CHARMM [22], GROMOS [126], OPLS [63], etc. The values of partial charges from GROMOS force field (see Table 1.3) are used throughout our implementation.

1.4.2 Electrostatic interaction

Following Coulomb's law, two charged particles interact to each other with a potential energy:

$$\mathcal{V} = \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r_{ij}}$$

where q_i and q_j represent the charges of particles i and j, r_{ij} is the distance between them, $\epsilon_0 \approx 8.85\times10^{-12}$ F.m^{-1} is the vacuum permittivity and ϵ_r is the dielectric constant (or relative permittivity) of the medium (some examples are given in Table 1.4).

The amino acid side chains can carry a ionized group (such as the ammonium (NH_3^+) cation of lysine, the guanidinium ($[CH_5N_3]^+$) cation of arginine, carboxylate (COO^-) anion of aspartate and glutamate) or a polar group (such as the hydroxyl groups of serine, threonine and tyrosine). The polypeptide main chain also contains a positively charged amino-terminal extremity, a negatively charged carboxy-terminal extremity, as well as the polar groups C=O and N−H. These cause numerous electrostatic interactions between charged groups (potential $\sim \mathcal{O}(1/r)$), between a charge and a dipole (potential $\sim \mathcal{O}(1/r^2)$) or between two dipoles (potential $\sim \mathcal{O}(1/r^3)$), where r denotes their distance.

1.4.3 Hydrogen bond

The hydrogen bond is a particular type of electrostatic interaction, which can be considered as an intermediary between covalent and ionic bonds. It is, intermolecularly or intramolecularly, formed by a dipole-dipole attraction between a hydrogen covalently attached to an electronegative atom (*donor*) and another electronegative atom (*acceptor*). The hydrogen atom has a positive partial charge, while the electronegative atom, usually oxygen, nitrogen or fluorine, has a negative partial charge. The hydrogen bond is viewed as an in-between state in the proton transfer from the donor D to the acceptor A:

$$\mathrm{D-H+A} \;\rightleftarrows\; \mathrm{D^{\delta-}\cdots H^{\delta+}\cdots A} \;\rightleftarrows\; \mathrm{D^{-}\cdots H-A^{+}}$$

The energy of a hydrogen bond depends on its bonding geometry. The optimal energy is obtained when H is aligned with D and A. Figure 1.18 illustrates the two popular examples of hydrogen bond.

Amino acid	Atom type	PDB codes	Charge (e)
D, E	C	CG (CD)	0.270
	O	ODi(OEi), i = 1,2	-0.635
N, Q	N	ND2 (NE2)	-0.830
	H	HD2i (HE2i), i = 1,2	0.415
	C	CG (CD)	0.380
	O	OD1 (OE1)	-0.380
C	S	SG	-0.064
	H	HG	0.064
T	C	CB	0.150
	O	OG1	-0.548
	H	HG1	0.398
S	C	CB	0.150
	O	OG	-0.548
	H	HG	0.398
R	C	CD	0.090
	N	NE	-0.110
	C	CZ	0.340
	N	NHi, i = 1,2	-0.260
	H	HE, HH$ij, i, j$ = 1,2	0.240
K	C	CE	0.127
	N	NZ	0.129
	H	HZi, i =1,2,3	0.248
H (A/B)†	C	CD2/CG	0.130
	N	NE2/ND1	-0.580
	C	CE1	0.260
	H	HD1/HE2	0.190
F	C	CDi, CE$i, i$ = 1,2, CZ	-0.100
	H	HDi, HE$i, i$ = 1,2, HZ	0.100
Y	C	CDi, CE$i, i$ = 1,2	-0.100
	H	HDi, HE$i, i$ = 1,2	0.100
	C	CZ	0.150
	O	OH	-0.548
	H	HH	0.398
W	C	CG	-0.140
	C	CD1, CE3, CZ$i, i$ = 2,3, CH2	-0.100
	H	HD1, HE3, HZ$i, i$ = 2,3, HH2	0.100
	N	NE1	-0.050
	H	HE1	0.190

† The partial charges for Histidine represent two possible ionized states which carry neutral charge.

Table 1.3: Partial charges from the Gromos force field for standard amino acids. e is the absolute value of elementary charge unit.

Medium	ϵ_r
Vacuum	1.0 (by definition)
Paraffin	$2.0 - 2.5$
Methanol	33.6
Water $20°C$	80.3
Water $0°C$	87.7

Table 1.4: The dielectric constant of selected mediums

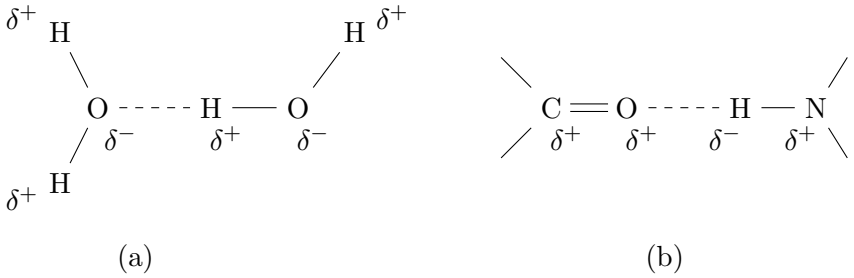


Figure 1.18: Hydrogen bonds represented in dash lines: (a) between water molecules and (b) between carboxylic and amino groups. δ^+ and δ^- are positive and negative partial charges, respectively.

Only oxygen, nitrogen and sulfur take part in hydrogen bonding in protein structures. The groups OH, NH, SH are donors, while oxygen and non-protonated nitrogen play the role of acceptors. Each residue, except for proline, in the polypeptide chain possesses a donor (N-H) and an acceptor (C=O) that can take part in hydrogen bonds in the main chain. Moreover, the side chains of more than half of residues are also capable to hydrogen bond with other residues or water molecules.

1.4.4 Van der Waals forces and steric repulsion

When two atoms approach each other, the modification of the electron distribution induces a polarization. There appears an attractive interaction by van der Waals forces. These forces include forces between polar molecules (Keesom force), between a polar molecule and a corresponding induced dipole (Debye force), and between two instantaneously induced dipoles (London dispersion force). Van der Waals forces have a potential of order $1/r^6$, where r is the distance between molecules.

Nevertheless, when two atoms are too close to each other, the steric repulsion becomes stronger and lead to a counterbalance to attractive forces. Its potential varies as $\mathcal{O}(1/r^{12})$.

These two forces cause a Lennard-Jones potential, in the case of interaction between

two atoms of the same type, given by:

$$\mathcal{V} = E_0 \left[\left(\frac{2r_0}{r} \right)^{12} - 2 \left(\frac{2r_0}{r} \right)^{6} \right]$$

where E_0 is the van der Waals well depth and r_0 is the van der Waals radius of the atom. The interaction is quite weaker than normal chemical bonds, yet these forces play an important role in folding stability thanks to their abundance. Table 1.5 shows typical values for these parameters of common atoms.

Atom	van der Waals well depth (kcal/mol)	van der Waals radius (Å)
H	0.02	1.00
C	0.12	1.85
N	0.16	1.75
O	0.20	1.60
S	0.20	2.00
P	0.20	2.10

Table 1.5: Typical values for van der Waals well depth and radius

1.4.5 Hydrophobic effect and interaction with the environment

The hydrophobic effect is the fact that a nonpolar molecule (or part of molecule) is incapable of hydrogen bonding with water molecules, thus agglomerate together in aqueous medium and exclude water molecules. It is not an attractive or repulsive force, but rather, it is entropically driven. Each water molecule is able to form four hydrogen bonds with its neighbors, thus in order that a nonpolar molecule dissolves into water, such hydrogen bonds have to be broken. The hydrogen bonding network of water disrupted by the nonpolar molecule will reform, by making a cage, around the molecule. This structure of cage is ordered, and thus is unfavored by the second law of thermodynamics which requires an increase in entropy. Hence, the corresponding free energy is unfavorable. The reorganization of water molecules is easier when the nonpolar surface exposed to the aqueous solution is reduced by aggregating the nonpolar molecules together. The hydrophobic effect plays the most important role in protein folding, compared to other non-covalent interactions. It helps polypeptide chains fold in a relatively compact form with a hydrophobic core.

Besides, due to the polarity of water molecules, amino acids with ionized or polar side chains have a tendency to interact with the aqueous medium through hydrogen bonds (see 1.4.3). This allows proteins to exist in water with a hydrophilic exterior.

1.4.6 Torsion energy around peptide bonds

The angles ϕ and ψ determining the polypeptide chain can differ from the theoretically optimal values which correspond to the equilibrium configuration. Such a deformation causes a energetic penalty, namely torsion energy.

1.4.7 Other interactions

Certain other interactions can also make an important contribution to the stability of a protein structure, such as salt bridge, cation-π interaction, $\pi - \pi$ stacking. Salt bridge which often occurs between the carboxylate anion of aspartic acid (D) or glutamic acid (E) and the ammonium cation of lysine (K) or guanidinium cation of arginine (R) can be considered as a combination of hydrogen bonding and electrostatic interactions. Cation-π interaction arises from the face of an electron-rich π system and a cation. $\pi-\pi$ stacking or aromatic-aromatic interaction consists of an attractive noncovalent interaction between aromatic rings. These interactions have an order of magnitude equivalent to hydrogen bonds.

The folding energy is finally defined as the sum of all the energies above.

1.5 Protein structure determination

The functions of proteins are performed through their conformations. Thus, it is crucial to determine the protein structures in order to understand the functions associated. Two different classes of methods have been used for protein structure determination: experimental methods which are based on physical measures and *in silico* prediction methods which used a wide range of computational tools.

1.5.1 Experimental methods

These methods are considered as providing the best *approximation* to real protein structures as they are based on observations and physical measures on *real* proteins. There currently exists a number of methods for protein structure determination, in which the most popular ones are X-ray crystallography and NMR spectroscopy.

X-ray crystallography

Most structures archived in the PDB were determined using X-ray crystallography [73]. Starting with the first two proteins crystallized (myoglobin and hemoglobin) at the end of the 1950s, the number of entries determined with X-ray crystallography reached over 55000 in 2010, following the annual report of the PDB [13]. For this method, a beam of X-rays strikes a purified and crystallized protein, and thus is diffracted by the protein crystal. Measuring the diffraction pattern allows to determine the distribution of electrons in the protein crystal. This distribution, or the map of electron density, is then

used to determine the location of each atom. The method of X-ray crystallography can give detailed atomic information, however, the crystallization process is difficult depending upon the type of proteins studied. It is well appropriate for rigid proteins forming well-ordered crystals, but not for flexible proteins with poor crystals.

NMR spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy [25] can also be used to determine protein structures. It is the use of NMR phenomenon to study the interaction of electromagnetic radiation with protein atoms. After being purified, the protein is placed in a strong magnetic field. The magnetic nuclei, with nonzero spin, absorb electromagnetic radiation at a resonance frequency which depends on the magnetic field strength and the magnetic properties of the isotope of the atoms. The resultant NMR spectra reflect the transitions between energy levels when the nuclei, which are close to one another, change their spin from up to down or inversely. This allows to characterize the local conformation of atoms that are bonded together, and then lead to determining the location of each atom. The method of NMR spectroscopy provides detailed information not only about the structure, but also about the molecular dynamics of the protein. This technique does not necessitate to crystallize the protein, thus can be studied in a medium similar to the one *in vivo*. As opposed to X-ray crystallography, it is useful for studying the atomic structure of flexible structures. However, the technique is still limited to small proteins (about a hundred residues) due to difficulties with overlapping peaks in the NMR spectra.

Some other techniques have also been used, such as electron microscopy [56, 57], X-ray microscopy [69], etc. Each of them has advantages and disadvantages. An atomic model cannot be entirely constructed with only the experimental information obtained in each method. Some additional knowledge about the molecular structure is required to build a model which is consistent with both the experimental data and the expected composition and geometry of the molecule.

1.5.2 *In silico* prediction

The computational methods of protein structure prediction are much more various. Although several approaches show a reasonable prediction performance, none of them appear to dominate the others. However, the CASP (Critical Assessment of Techniques for Protein Structure Prediction) [90, 128] competition allows to assess the efficiency of published methods.

Since it is difficult, expensive and time-consuming to obtain the protein structures from experimental methods, the *in silico* can propose useful structural models for generating hypotheses about protein's functions and pointing to further experimental work. The reliability of a prediction is determined by the prediction concept and the refinement of the used model.

Three standard approaches are widely used for predicting protein structures. The

first one includes the methods of molecular dynamics which are based on thermodynamic models. They allow to simulate the folding of a protein and then to determine its three-dimensional structure [11, 52, 55, 64]. These methods are implemented in the software packages like CHARMM [22, 83] or GROMACS [12, 77]. However, they are not quite helpful in practice as only polypeptide chains of very limited size can be studied due to an enormous complexity of computing. Other than that, the projects of distributed computing also aim to predicting the tertiary structure of proteins based on *ab initio* modeling, such as Folding@Home (http://folding.stanford.edu), POEM@Home (http://boinc.fzk.de/poem/), Predictor@Home (http://predictor.chem.lsa.umich.edu), Rosetta@Home (http://boinc.bakerlab.org), etc. They make use of the help of several active volunteered computers around the world to deal with the problem on huge complexity.

The second approach, namely *comparative* or *homology modeling*, tries to approximate the tertiary structure by aligning the sequences or the structural subunits [10, 49, 62, 60, 87, 109, 111]. Several softwares like COMPOSER [116], MODELLER [38, 110], PRISM [137], SEGMOD [74] or SWISS-MODEL [114] are developed based on this concept. Their predicting quality depends on the homology of the analyzed sequence to some ones in the database. Hence, the resulting prediction is far from correct with the proteins whose structural topology does not exist in the database. Moreover, these techniques are not suitable for discovering new protein structures. The most reliable strategy for finding the tertiary structure of a protein is proposed as a combination of comparative modeling and refinement by an optimization of force fields.

The third approach is known as *protein threading* or *fold recognition*. It is used to predict the structures of the proteins which have the same fold as proteins of known structures, but do not have homologous sequences with the latter. This is distinguished from homology modeling, even though they are both template-based methods. When no significant homology between sequences is found (for instance, the sequence identity is less than 30%), homology modeling is not helpful and protein threading can be used for prediction using the structural information of the target protein. This method has been applied in several applications, such as 3D-PSSM [66], PHYRE/PHYRE2 [65], RAPTOR [136], etc.

The prediction of secondary structures appears simpler than that of tertiary structures but it is still a difficult problem. It consists in assigning regions of an amino acid sequence to secondary motifs (α-helix, β-strand or turn). Due to a limited number of characteristics for determining the formation of those motifs, it is more appropriate to specialize a particular class of proteins for each predictor. The existing predicting methods can be classified into two categories: those aiming to globular proteins and those aiming to transmembrane proteins. With a large number of globular proteins with known structures in the PDB at present, the machine learning based techniques seem to be an efficient approach for this class of proteins. This is still reasonable for transmembrane α-helical proteins, although their known structures are much less abundant. However, with less than 200 available structures of TMB proteins which are reduced to

about 40 non-redundant ones [125], the structure prediction problem becomes intractable while the reliability of learning based methods is far from being approved.

Chapter 2

Folding β-barrels

2.1 Introduction

We present in this chapter the model that we developed for classification and structure prediction of TMB proteins [119, 121, 123]. TMB proteins are hard to identify, however, it is relatively easy to identify a majority of other proteins which are not TMB. We use physicochemical properties and a simple probabilistic model based on a sliding window for filtering amino acid segments that are obviously not involved in any β-barrel structures as a membrane spanning β-strand. Proteins that are considered to be putative TMB proteins by this initial phase are then further analyzed. Next, we try to fold the given protein, treating it as a TMB protein, using the pseudo-energy minimization model. If the protein cannot be folded into β-barrels according to the energy minimization framework, the protein is rejected and classified as a non-TMB protein.

Before presenting the simple model that we used for filtering the transmembrane β-strands in Section 2.4, we discuss some geometric constraints (Section 2.2) and physicochemical constraints (Section 2.3) that a protein must obey to be a TMB protein. We enforce these constraints in both the filtering and folding steps of our algorithm. We give our concrete folding problem definition in the next section before describing a dynamic programming approach to solve the problem [120, 124].

2.2 Geometric framework for β-barrels

The backbone geometry of a regular β-barrel [84, 91, 92] is entirely determined by n, the number of strands composing the barrel, and by S, the *shear number*, which is defined below.

Definition 2.1. Shear number of a β-barrel

In a regular β-barrel, the shear number S is unambiguously defined as the ordinal distance between an amino acid A and an amino acid B that is located on the same

strand as A and linked to A through a path of hydrogen bonds. B is the projection of the "copy" of A after one turn on the first strand of the barrel.

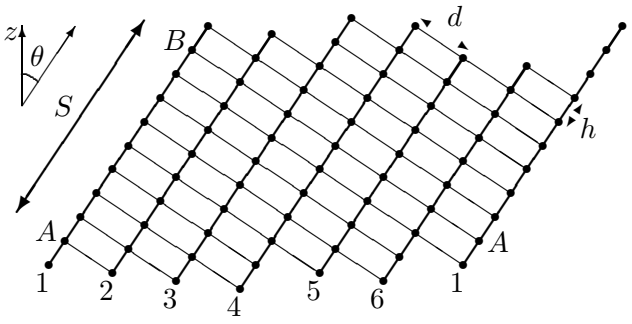


Figure 2.1: The simplified geometry of a β-barrel, a schematic planar view for 6 strands (strand 1 is duplicated for clarity). Thick lines denote the peptide bonds that link consecutive amino acids along their strand. Thin lines denote the hydrogen bonds that link the amino acids of two adjacent strands. In this example, the *shear number* is $S = 8$, which is the ordinal distance between amino acids A and B. We note that all known β-barrels have a positive *shear number* [80] and are slanted "to the right", as illustrated here.

Structural constants are $h(\approx 3.3\text{Å})$, the jump per amino acid along a strand, and $d(\approx 4.4\text{Å})$, the mean distance between adjacent strands, given respectively by the peptide bond and hydrogen bond geometries. The other geometric characteristics, such as θ, the slant angle of the strands relative to the barrel z-axis, are given from n, S, h and d [28]:

$$\tan\theta = \frac{hS}{dn}$$

Angle θ, in association with a given membrane thickness, is involved in the energetic rules and restricts the membrane spanning β-strand length. Then, n and S have to be fixed as parameters.

Definition 2.2. Relative shear number

Given a shear number S, the relative shears *between adjacent strands remain as $n-1$ degrees of freedom. As a convention, we consider the* relative shears *on the extracellular side of the barrel. So, $\forall\, i > 1$, s_i, the* relative shear *of strand $i+1$ with respect to strand i (strand $n+1$ being identified with 1), is measured on strand i as the ordinal distance between the undermost amino acid of strand i and the one that is directly bound to the undermost amino acid of strand $i+1$.*

On the example of Figure 2.1, the sequence of *relative shears* (s_i) is $(1\,1\,1\,2\,1\,2)$. The sum of consecutive *relative shears* naturally defines the *shear* between two extreme strands, thus we have the constraint for the β-barrel, where the two extreme strands are strand 1, for instance, and itself after a round on the barrel:

$$\sum_{1\leq i\leq n} s_i = S$$

We define the shear number, by extension, for the case of a β-sheet (i.e. an open β-barrel) to make our algorithms capable of dealing with the structure of β-sheets.

Definition 2.3. Shear number of a β-sheet

The shear number of a n-strand β-sheet is defined as the sum of relative shears on consecutive pairs of adjacent strands:

$$S = \sum_{1 \leq i \leq n-1} s_i$$

where s_i is the relative shear of strand $i+1$ with regard to strand i.

Each β-strand is directed with respect to the sequence order from N-terminal to C-terminal. A strand is said to be *upward* if it is oriented from the extracellular environment to the periplasmic space, i.e. the N-terminal of the strand is located on the extracellular side and its C-terminal is on the periplasmic side. Inversely, the strand is said to be *downward*. The *upward/downward* orientation of the strand, relatively to the barrel axis, defines another degree of freedom.

Finally, considering a β-strand as a ribbon where the amino acids direct their side-chains alternatively on both sides, toward the barrel interior (channel) or toward the surrounding lipid (membrane), we will distinguish two ways of facing, neglecting small swivel adjustments. A strand is said to be *odd inward* if the odd indexed amino acids face to the channel and *odd outward* if those face to the membrane (see Section 2.3 for more details). We have one more degree of freedom.

These notions of orientation are illustrated in Figure 2.2.

2.3 Physicochemical constraints

On the amphipathic β-strand of TMB proteins, the side-chains of amino acids are directed towards the membrane and the channel alternatively. Hydrophilic and polar side-chains orient towards the aqueous interior while hydrophobic ones contact the hydrophobic bilayer [117]. We use the Kyte-Doolittle scale [72] (see Section 1.2.2) to measure the hydrophobicity $H(r)$ of each amino acid r. In this scale, a higher value represents higher hydrophobicity, and vice versa. The necessary condition for a segment $r_i....r_j$ to be a potential membrane spanning β-strand is that one side is hydrophobic and the other side is hydrophilic. Formally, we define

$$\begin{aligned} H^e_{i,j} &= \big\langle H(r_{2k}) \big\rangle, i \leq 2k \leq j \\ H^o_{i,j} &= \big\langle H(r_{2k+1}) \big\rangle, i \leq 2k+1 \leq j, k \in \mathbb{N} \end{aligned}$$

as the average hydrophobicity on the respective even and odd numbered sides. Hence, the constraints

$$\max\{H^e_{i,j}, H^o_{i,j}\} > \zeta^- \quad \text{and} \quad \min\{H^e_{i,j}, H^o_{i,j}\} < \zeta^+$$

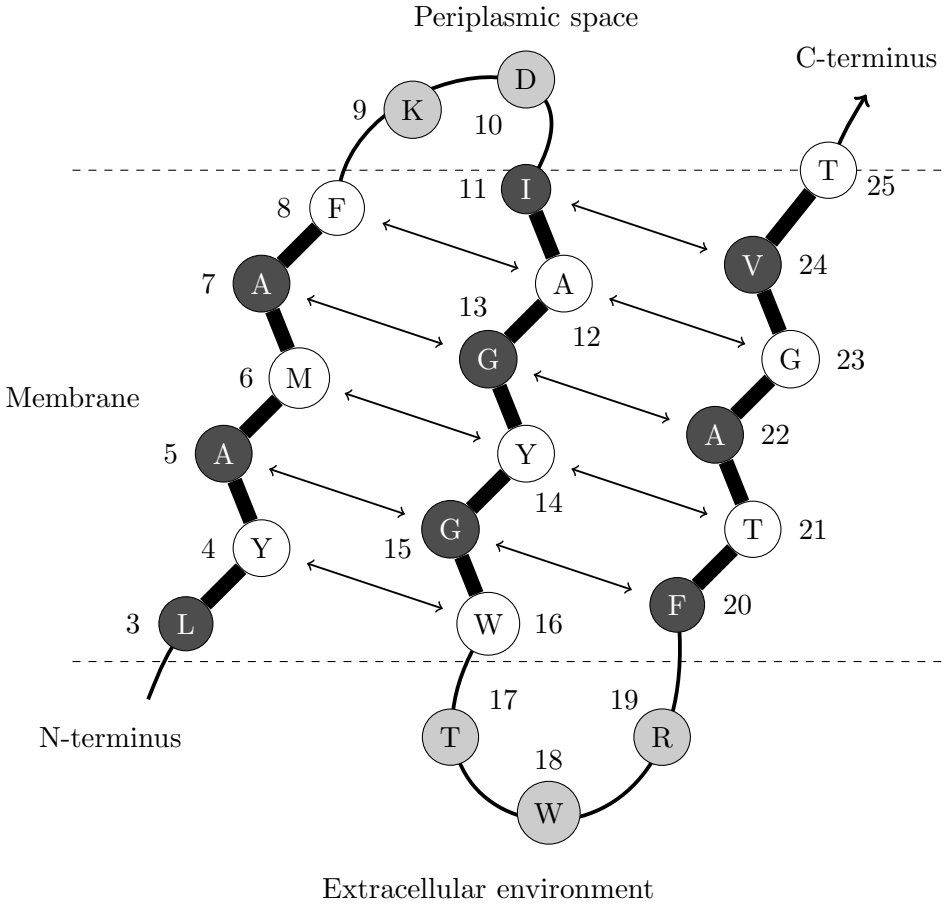


Figure 2.2: A schematic planar representation of 3 β-strands in a transmembrane β-barrel. The black residues direct their side chains toward the membrane and white ones toward the channel. The first and third strands are *upward* and the second one is *downward*. The first and second strands are *odd outward* and the third one is *odd inward*.

are necessary for a segment of $j-i+1$ consecutive amino acids $r_i....r_j$ to be a potential membrane spanning β-strand, where ζ^- is a lower bound for the hydrophobic side and ζ^+ is an upper bound for the hydrophilic side. We use the values $\zeta^- = -1$ and $\zeta^+ = 1$, which were obtained through an statistical data analysis on known TMB structures (see Figures 2.3, 2.4). Then, with respect to the TMB structure, the segment $r_i....r_j$ is defined as *odd inward* oriented if $H^o_{i,j} < H^e_{i,j}$ and *odd outward* oriented if $H^e_{i,j} < H^o_{i,j}$.

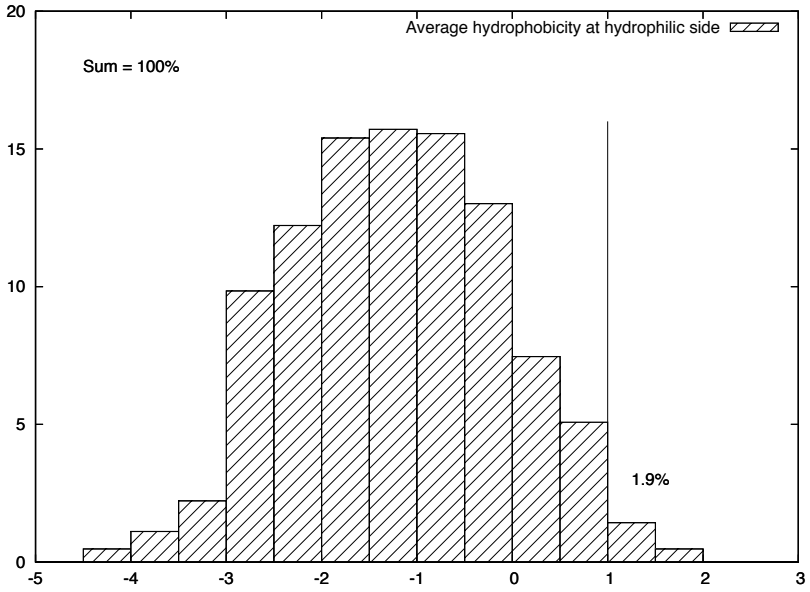


Figure 2.3: The distribution of average hydrophobicity index of the hydrophilic side of the membrane spanning β-strands from PDBTM40 (see Section 4.2)

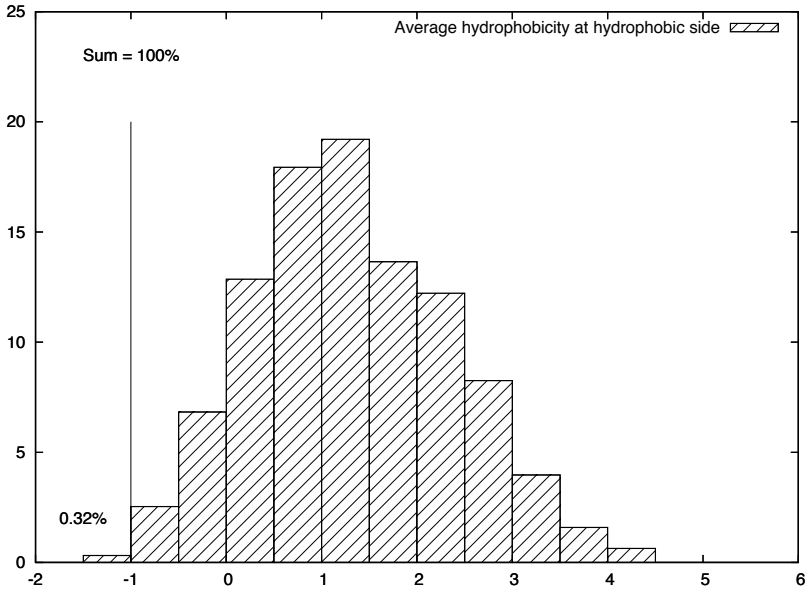


Figure 2.4: The distribution of average hydrophobicity index of the hydrophobic side of the membrane spanning β-strands from PDBTM40 (see Section 4.2)

2.4 Classification filtering

In order to identify substrings as potential membrane spanning β-strands (the vertices) or turns/loops (the edges), we introduce a simple probabilistic model that acts as a primary filter. We use a sliding window (segment) as a sequence of consecutive l-residue subsegments (or blocks) ($l = 3$ in our implementation). Let r denote the occurrence of a given block ($r = r_1 r_2 ... r_l$) and let τ be the event that a block is found in a given conformation (β-strand or turn/loop). The information that τ gets from r is defined as:

$$I(\tau\,;r) = \log \frac{P(\tau|r)}{P(\tau)} = \log \frac{f_{\tau,r}/f_{\cdot,r}}{f_{\tau,\cdot}/f_{\cdot,\cdot}},$$

where $f_{\tau,r}$ represents the frequency observed in the training dataset for a block r to be found in conformation τ and we denote for short [39]:

$$\begin{aligned} f_{\cdot,r} &= \sum_{\tau} f_{\tau,r} \\ f_{\tau,\cdot} &= \sum_{r} f_{\tau,r} \\ f_{\cdot,\cdot} &= \sum_{\tau}\sum_{r} f_{\tau,r} \end{aligned}$$

Thus, $I(\tau\,;r)$ measures the influence of r on the occurrence of τ. If $I(\tau\,;r) = 0$, there is no influence; whereas $I(\tau\,;r) > 0$ indicates that r is favorable to the occurrence of τ and vice versa. Formally, the preference of r in favor of τ as opposed to $\overline{\tau}$, any conformation different from τ [45], is:

$$I(\tau : \overline{\tau}\,;r) = I(\tau\,;r) - I(\overline{\tau}\,;r) = \log \frac{f_{\tau,r}/f_{\tau,\cdot}}{f_{\overline{\tau},r}/f_{\overline{\tau},\cdot}}$$

A simple measure is associated to each segment $r_1 r_2 ... r_p$ that helps determine if it is likely a β-strand or a coil. It is defined as the sum of informations on all the l-residue blocks:

$$\tilde{I}(\tau : \overline{\tau}\,; r_1 r_2 ... r_p) = \sum_{i=1}^{p-l+1} \frac{I(\tau : \overline{\tau}\,; r_i r_{i+1} ... r_{i+l-1}) - \log \rho}{p - l + 1}$$

The segment is then considered as a candidate for conformation τ if $\tilde{I}(\tau : \overline{\tau}\,; r_1 r_2 ... r_p) > 0$.

The non-redundant training set of TMB proteins described in Section 4.2 is used to learn this probabilistic model. Due to the small size of the training set, we apply the filter with a relatively low threshold at $\rho = \frac{2}{3}$ to avoid overfitting. This ensures that on average, each block r is accepted in conformation τ if the propensity for r to be in τ (i.e. $f_{\tau,r}/f_{\tau,\cdot}$) is at most 1.5 times less than the propensity to be in $\overline{\tau}$ (i.e. $f_{\overline{\tau},r}/f_{\overline{\tau},\cdot}$). Only substrings that pass these very stringent criteria are considered to be putative strands.

2.5 Folding problem definition

Let $\mathcal{S}$ be the sequence of the N amino acids constituting the primary structure of a given protein. We will consider $\mathbf{G}(\mathbf{V}, \mathbf{E}, \mathcal{E}_{\text{intr}}, \mathcal{E}_{\text{adj}}, \mathcal{E}_{\text{loop}})$, the weighted directed acyclic graph (DAG) [30] built from $\mathcal{S}$ as follows:

2.5.1 Vertices

Let $\mathbf{V} = \mathbf{V}^* \cup \{\top, \bot\}$ be the set of vertices. Each vertex of $\mathbf{V}^*$ represents a candidate secondary structure item as a β-strand associated with a given set of parameters. It corresponds to a contiguous part (a substring, defined by its starting and ending indices $1 \leq \nu < \kappa \leq N$) of $\mathcal{S}$ that satisfies given conformational constraints (such as length, propensity to be a β-strand, ...). The associated parameters provide information about the discretized spatial laying of this part relatively to the whole structure. So, combining the *upward/downward* and *inward/outward* degrees of freedom introduced in 2.2, we consider 4 different orientations for each given candidate β-strand. We could also consider the different instances of *relative shear* to multiply the number of vertices, but we do not for reasons to be clarified later.

A canonical order is defined on $\mathbf{V}^*$ as the lexicographic order on tuples formed by the respective starting/ending indices in $\mathcal{S}$ and the associated parameters. The length constraint implies that the number of candidate substrings and thus $|\mathbf{V}|$, the number of vertices, are bounded above by kN for a small value k. To simplify further definitions, a dummy vertex $\top$ will be used to represent an empty substring at the start of $\mathcal{S}$ and, similarly, $\bot$ will represent an empty substring at the end of the sequence. To extend the order on all of the vertices, we set $\top < v < \bot, \forall\, v \in \mathbf{V}^*$ (see Figure 2.5).

2.5.2 Edges

Let $\mathbf{E} \subset \mathbf{V} \times \mathbf{V}$ be the set of directed edges. Intuitively, an edge corresponds to a turn or a loop that connects two consecutive β-strands. To be more precise, $\forall\, v, w \in \mathbf{V}^*$, with $\nu_v, \kappa_v, \nu_w, \kappa_w$ denoting their respective starting and ending indices, (v, w) is an edge, if $\kappa_v < \nu_w - 2$ and the substring of amino acids from $\kappa_v + 1$ to $\nu_w - 1$ satisfies the constraints that allow to form a turn or a loop (such as conditions on length, flexibility, propensity, ...) also depending on the relative laying of the two substructures. We have the elementary property:

$$\forall\, v, w \in \mathbf{V}^*, (v, w) \in E \implies v < w$$

for the lexicographic order, and this ensures the DAG structure.

The set $\mathbf{E}$ also contains edges of the form $(\top, v)$ that define the subset of starting vertices - the leading substrings satisfying specific constraints. Similarly, $\mathbf{E}$ contains edges of the form $(v, \bot)$ that define the subset of ending vertices, with a satisfactory trailing

substring. Again, the length constraints applied to the substrings associated to edges imply that $|\mathbf{E}|$, the number of edges, is $\mathcal{O}(|\mathbf{V}|)$ or $\mathcal{O}(N)$.

Figure 2.5 gives a small example of such a graph (to simplify, only one orientation has been considered). An edge like (v_1, v_2) is forbidden, since the two corresponding substrings overlap. Edges like (v_2, v_3) or (v_2, v_6) are also forbidden, since the inserted substrings are respectively too short for a turn or too long for a loop.

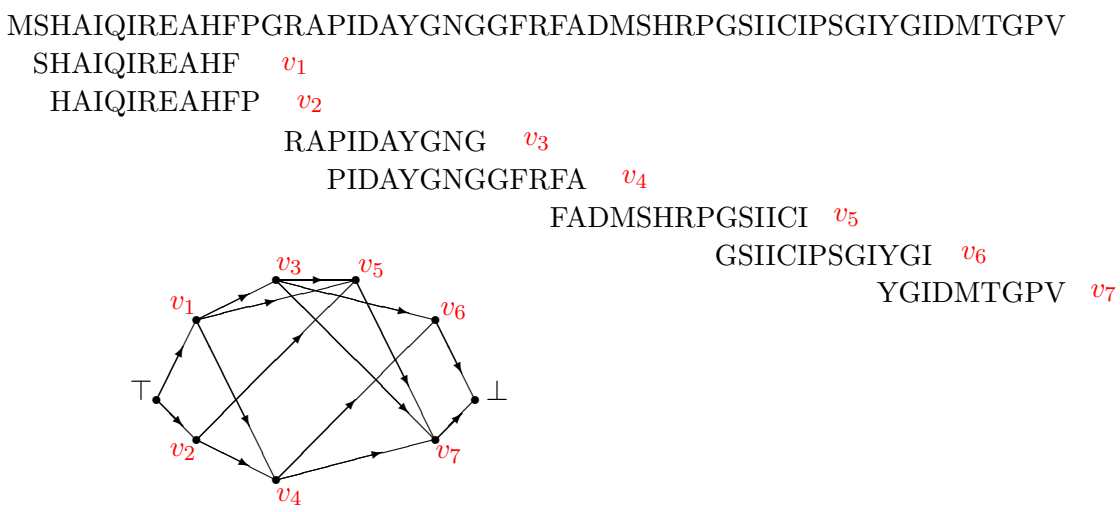


Figure 2.5: A short example of the graph structure. Edge (v_1, v_2) is not allowed, since the two corresponding substrings overlap. Edges (v_2, v_3) or (v_2, v_6) are not allowed, since the substrings in between are respectively too short for a turn or too long for a loop, etc.

2.5.3 Energy attributes:

The attributes that complete the definition of the graph $\mathbf{G}$ are pseudo-energy functions defined as follows:

- $\forall\, v \in \mathbf{V}^*$, $\mathcal{E}_{\text{intr}}(v)$ represents the intrinsic energy of the given strand in the given orientation. This term is the sum of both the internal energy of the substructure, i.e. the interactions between its own amino acids, and the interaction energy with the environment (e.g. membrane and channel) apart from the rest of the considered protein.
  Note that $\mathcal{E}_{\text{intr}}(\top) = \mathcal{E}_{\text{intr}}(\bot) = 0$.

- $\forall\, (v, w) \in \mathbf{V}^* \times \mathbf{V}^*$, $\mathcal{E}_{\text{adj}}(v, w, s)$ represents the interaction energy of the pair (v, w) when the two corresponding strands are placed side by side along the barrel, with

respect to the respective orientation parameters associated to the vertices and accordingly to the *relative shear* s. The energy will take into account the number of contacts and different side-chain interactions such as the packing of hydrophobic cores and bonding abilities.

Then, $\forall\,(v,w) \in \mathbf{V}^* \times \mathbf{V}^*$, $\mathcal{E}_{\text{adj}}(v,w) = \min_s \mathcal{E}_{\text{adj}}(v,w,s)$ is the interaction energy of the pair (v,w) for an optimal relative shear. It is further assumed that $\mathcal{E}_{\text{adj}}$ is defined over a superset of $\mathbf{E}$, since we will consider the case where two adjacent strands are not consecutive along the sequence.

We also introduce the particular values:

$$\mathcal{E}_{\text{adj}}(\top, v) = \mathcal{E}_{\text{adj}}(v, \bot) = 0, \forall\, v \in \mathbf{V}.$$

- An associated function s_{adj} is defined such that:
  $\forall\,(v,w) \in \mathbf{V}^* \times \mathbf{V}^*$, $\mathcal{E}_{\text{adj}}(v,w,s_{\text{adj}}(v,w)) = \mathcal{E}_{\text{adj}}(v,w)$, which is a *relative shear* that leads to the optimal interaction energy.

  An arising question is why the orientation degrees of freedom are described as a multiplicity of nodes but the *relative shear* degrees of freedom are considered when calculating the $\mathcal{E}_{\text{adj}}$ terms. A first answer comes from the fact that wrong orientations are rather absolute and will result in pruning the sets $\mathbf{E}$ and $\mathbf{V}$ while the *shear* parameters are not so discriminative. The main reason is that we will consider "floating" parts in which adjacencies are already set, while a *relative shear* between any two parts is not yet known. In such a situation, attaching the *relative shears* to node pairs allows a significant factorization.

- $\forall\,(v,w) \in \mathbf{E}$, $\forall\, t \in \{1,2,\ldots,n-1\}$ and $\forall\, s-$a *relative shear*, $\mathcal{E}_{\text{loop}}(v,w,t,s)$ is related to the intrinsic energy of the turn/loop between the strands v and w (consecutive along the sequence) when they are placed at a distance t along the barrel with a *relative shear* s. The distance $t = 1$ corresponds to the case where the strands are placed consecutively on the barrel, while an integer value $t > 1$ will correspond to the case where $t-1$ other strands are interleaf.

  To simplify, we will also use $\mathcal{E}_{\text{loop}}(\top, v)$ or $\mathcal{E}_{\text{loop}}(v, \bot)$ for denoting the intrinsic energy of the outer fragment attached respectively to a starting or an ending vertex v. As such a fragment has a free side, the position parameters may be dropped.

Then, in the usual case of two β-strands that fold as a hairpin, the related energy is considered to be $\mathcal{E}_{\text{adj}}(v,w) + \mathcal{E}_{\text{loop}}(v,w,1,s_{\text{adj}}(v,w))$. It is supposed a relative flexibility for turns and loops, so, when a fold is feasible, $\mathcal{E}_{\text{loop}}$ is weak compared to $\mathcal{E}_{\text{adj}}$ and the relative placement of the two β-strands is enforced to be close to s_{adj}. Nevertheless, $\mathcal{E}_{\text{loop}}$ will result in a strong penalty in the case of an unfeasible turn or loop, for example a loop with a majority of hydrophobic residues.

2.5.4 Protein folding problem

Given a graph $\mathbf{G}(\mathbf{V}, \mathbf{E}, \mathcal{E}_{\text{intr}}, \mathcal{E}_{\text{adj}}, \mathcal{E}_{\text{loop}})$ defined as above, two integers n, S, and a permutation[1] σ as 3 parameters, we look for the path $\mathcal{P}$ in $\mathbf{G}$ that maximizes the following objective function:

$$\mathcal{E} = \sum_{v \in \mathcal{P}} \mathcal{E}_{\text{intr}}(v) + \sum_{(v,w) \in \mathcal{P}} \mathcal{E}_{\text{loop}}(v, w) + \sum_{(v,w) \in \sigma(\mathcal{P})} \mathcal{E}_{\text{adj}}(v, w)$$

such that $\sum_{(v,w) \in \mathcal{P}} s_{\text{adj}}(v, w) = S$.

Such a path $\mathcal{P}$ whose vertices are arranged onto a circle is called a *circle-attached path*. The adjacent vertices in the path are not necessarily successive on the circle. This order of succession is determined by the given permutation σ (see Figure 2.6).

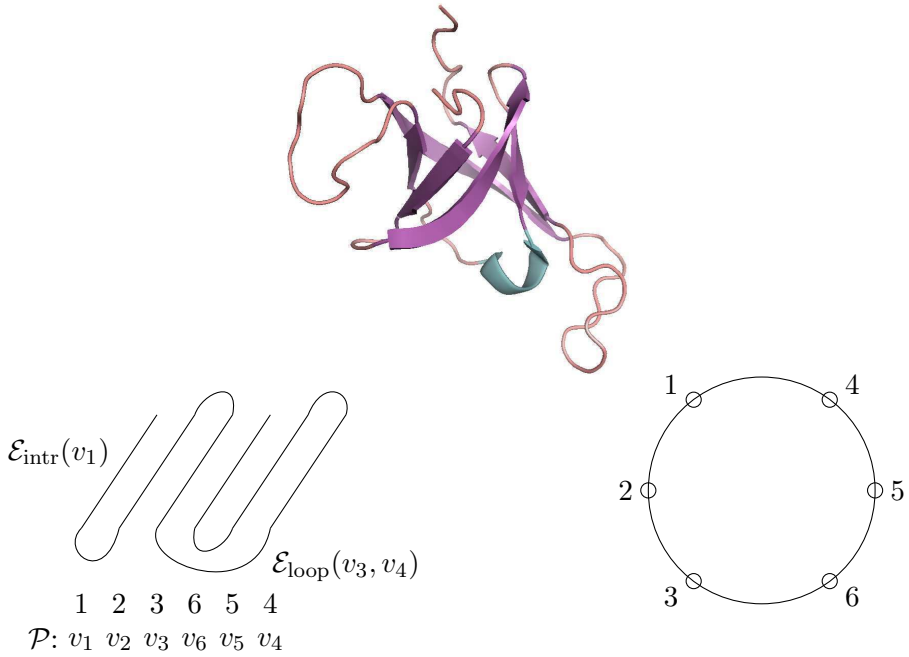


Figure 2.6: Different views of a β-barrel with a Greek key motif 3654, $\sigma = 1\,2\,3\,6\,5\,4$

[1] The notion of permutation is described in detail in 2.7.1

2.6 Dynamic programming approach

2.6.1 Solving as the longest path problem

We will first consider an open structure, as a β-sheet, where the adjacency of strands follows their natural order along the amino acid sequence, i.e. σ is an identity permutation. We involve here the constraint $\sum_{1<i\leq n} s_i = S$. Hence, solving such a structure will result in finding a path $\mathcal{P}$ in $\mathbf{G}$ whose overall "energy" is given by the sum:

$$\mathcal{E} = \Sigma_{v\in\mathcal{P}}\mathcal{E}_{\text{intr}}(v) \ + \Sigma_{(v,w)\in\mathcal{P}}[\ \mathcal{E}_{\text{adj}}(v,w) + \mathcal{E}_{\text{loop}}(v,w,1,s_{\text{adj}}(v,w))\ ]$$

Aiming at minimizing $\mathcal{E}$, the protein folding problem will turn into finding the path from $\top$ to $\bot$ that maximizes the criterion $\mathbf{C} = -\mathcal{E}$. Let $\mathbf{C}_v^h$ be the maximum value for $\mathbf{C}$ over all the paths from $\top$ to v, with a shear number of h of the corresponding β-sheet, then $\mathbf{C}_\top^0 = 0$ and, $\forall\, v \in \mathbf{V}\backslash\{\top\}, \forall h, \mathbf{C}_v^h$ is defined as:

$$\mathbf{C}_v^h = \max_{u\in\mathbf{V},(u,v)\in\mathbf{E}} \left[\ \mathbf{C}_u^{h-s_{\text{adj}}(u,v)} - \mathcal{E}_{\text{intr}}(v) - \mathcal{E}_{\text{adj}}(u,v) - \ \mathcal{E}_{\text{loop}}(u,v,1,s_{\text{adj}}(u,v))\ \right]$$

Since the graph is a DAG, the longest path problem is solved with a well known dynamic programming scheme [30] of complexity $\mathcal{O}(|\mathbf{V}|)$ in space and $\mathcal{O}(|\mathbf{V}| + |\mathbf{E}|)$ in time, that is also $\mathcal{O}(N)$ for both, from the structural constraints that relate $|\mathbf{V}|$, $|\mathbf{E}|$ and N. The objective is the computation of $\mathbf{C}_\bot^S$ and the optimal structure is then reconstructed by a usual traceback post-processing. Note that, for each path, we only have to consider its last vertex, so, we have to track single index states.

2.6.2 Solving as the longest closed path problem

For a barrel secondary structure, we have to consider a closing spatial adjacency between the last and the first strands. σ is still an identity permutation. The constraint on the shear number becomes $\sum_{1<i\leq n+1} s_i = S$. The dynamic programming scheme is almost the same as previously, except that we also have to keep track of the first vertex of any path. So, $\forall\, v \in \mathbf{V}^*$, such that $(\top, v) \in \mathbf{E}$, let $\mathbf{C}_{(v,v)}^0 = -\mathcal{E}_{\text{intr}}(v) - \mathcal{E}_{\text{loop}}(\top, v)$, then the general recurrence is: $\forall\, v, w \in \mathbf{V}^*, \forall h$, such that $(\top, v) \in \mathbf{E}$,

$$\mathbf{C}_{(v,w)}^h = \max_{u\in\mathbf{V},(u,w)\in\mathbf{E}} \left[\ \mathbf{C}_{(v,u)}^{h-s_{\text{adj}}(u,w)} - \mathcal{E}_{\text{intr}}(w) - \mathcal{E}_{\text{adj}}(u,w) - \ \mathcal{E}_{\text{loop}}(u,w,1,s_{\text{adj}}(u,w))\ \right]$$

and a special closing step is needed: $\forall\, v \in \mathbf{V}^*, \forall h$, such that $(\top, v) \in \mathbf{E}$,

$$\mathbf{C}_{(v,\bot)}^h = \max_{u\in\mathbf{V},(u,\bot)\in\mathbf{E}} \left[\ \mathbf{C}_{(v,u)}^{h-s_{\text{adj}}(u,v)} - \mathcal{E}_{\text{adj}}(u,v) - \mathcal{E}_{\text{loop}}(u,\bot)\ \right]$$

The goal is to calculate $\max_{v,(\top,v)\in\mathbf{E}} \mathbf{C}_{(v,\bot)}^S$. Thus the scheme is of complexity $\mathcal{O}(|\mathbf{V}|^2)$ in space and $\mathcal{O}(|\mathbf{V}|.|\mathbf{E}|)$ in time, that is also $\mathcal{O}(N^2)$ for both, from the structural constraints. This may produce paths of any length and the constraint of n strands is applied as a cut in the recurrence.

2.6.3 Generalization

In a more general case, we consider permutations to deal with the fact that the arrangements of the strands along the barrel do not necessarily follow their order along the sequence. This usually occurs with Greek key motifs or more rarely with Jelly roll motifs. Hence, the protein folding problem becomes finding the longest path $\mathcal{P}$ in a graph with respect to a given permutation σ, i.e. the vertices of $\mathcal{P}$, seen on a circle as in Figure 2.6 are permuted according to σ.

Let σ be a circular permutation of $\{1, 2, \ldots, n\}$. When $1, 2, \ldots, n$ are numbering the positions along the barrel, values $\sigma(1), \sigma(2), \ldots \sigma(n)$ will give the respective ranks of the strands in the sequence order. A position of reference along the barrel is fixed by setting $\sigma(1) = 1$. The Greek key example of Figure 2.6 is described by the permutation $\sigma = (1, 2, 3, 6, 5, 4)$. Hereafter, we will consider $\sigma = (1, 2, 5, 4, 3, 6)$ which is a bit trickier situation (see Figure 2.7).

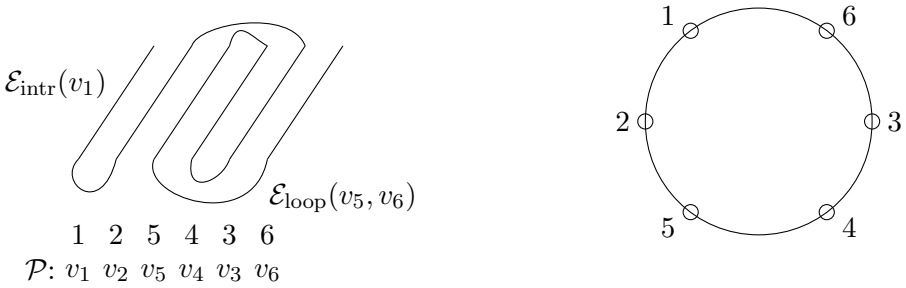


Figure 2.7: A permuted β-barrel with a Greek key motif 5436, $\sigma = 1\,2\,5\,4\,3\,6$

The dynamic programming scheme now consists in building a barrel, by adding a next candidate strand, taken in the sequence with respect to the graph edges, but that is inserted at the position defined by the given permutation. Useful values are the ranks (in the sequence order) of the two strands between which a given one will be inserted. For instance, with the given example, the 5^{th} strand will be inserted between the 2^{nd} and the 4^{th} strands.

Let now k denote the level of construction ($1 \leq k \leq n$), that is the number of strands already placed.

Proposition 2.1. *The k^{th} strand (in the sequence order) is inserted between the two strands whose ranks (in the sequence order) are* $\mathbf{left}_k$ *and* $\mathbf{right}_k$*, defined as:*

$$\mathbf{left}_k = \begin{cases} \sigma(\sigma^{-1}(k) - 1) & \text{if } \sigma^{-1}(k) > 1 \\ \sigma(n) & \text{otherwise} \end{cases}$$

$$\mathbf{right}_k = \begin{cases} \sigma(\sigma^{-1}(k) + 1) & \text{if } \sigma^{-1}(k) < n \\ 1 & \text{otherwise} \end{cases}$$

With the current example, we get:

$$\begin{array}{llllllllll} \mathbf{left}_1 & = & 6 & \mathbf{left}_2 & = & 1 & \mathbf{left}_3 & = & 4 \\ \mathbf{left}_4 & = & 5 & \mathbf{left}_5 & = & 2 & \mathbf{left}_6 & = & 3 \\ \mathbf{right}_1 & = & 2 & \mathbf{right}_2 & = & 5 & \mathbf{right}_3 & = & 6 \\ \mathbf{right}_4 & = & 3 & \mathbf{right}_5 & = & 4 & \mathbf{right}_6 & = & 1 \end{array}$$

An important piece of information to be stored for the dynamic programming scheme is the set of *active* indices, i.e. ranks of the strands (in the sequence order) that are either not definitively bonded on both sides along the barrel or not linked along the sequence, and thus have to be kept as degrees of freedom. So, in the given example, we have to keep in mind every valid instance as 2^{nd} and 4^{th} strands until an optimal choice is recorded for each instance as a 5^{th} strand. At that time, any instance as a 5^{th} strand is kept as a candidate for a link with a 6^{th}, by a turn or loop, while the different instances as the 3^{rd} and 1^{st} are kept for proceeding to an insertion in between.

Definition 2.4. *Two ranks i and j, which refer to the sequence order, are said* adjacent *if:*

$$|\sigma^{-1}(i) - \sigma^{-1}(j)| \in \{1, n-1\},$$

where the case $n-1$ is intended for the adjacency that will close the barrel.

Proposition 2.2. *The set of* active *indices (in the sequence order) at level k is defined by:*

$$\mathbf{conf}_k = \{k\} \cup \{\, i \,|\, (1 \leq i < k) \ \wedge\ (\exists j : k < j \leq n \,|\, i, j \text{ are adjacent}) \,\} \tag{2.1}$$

With the current example, we get:

$$\begin{array}{llllllllll} \mathbf{conf}_1 & = & \{1\} & \mathbf{conf}_2 & = & \{1,2\} & \mathbf{conf}_3 & = & \{1,2,3\} \\ \mathbf{conf}_4 & = & \{1,2,3,4\} & \mathbf{conf}_5 & = & \{1,3,5\} & \mathbf{conf}_6 & = & \{6\} \end{array}$$

Thus, in this example, the maximal complexity in space, $\mathcal{O}(N^4)$, is reached for the set of subsolutions with 4 strands. Then looping over this set, for computing the set of subsolutions with 5 strands, will also cost $\mathcal{O}(N^4)$ in time, since the choice for the 5^{th} strand is bounded by the structural constraints embedded as edges in the graph.

Proposition 2.3. $\forall i < j$,

$$\mathbf{conf}_i \cap \mathbf{conf}_j \subset \mathbf{conf}_k, \forall k \in [i+1, j-1]$$

Proof. For any i, let k_{max} be the maximum index such that $i \in \mathbf{conf}_{k_{max}}$. We have $k_{min} = i$ is the minimum index such that $i \in \mathbf{conf}_{k_{min}}$. Following 2.1, there exists $j > k_{max} \geq k, \forall k \in [k_{min}, k_{max}]$, so that i and j are *adjacent*. Hence, $i \in \mathbf{conf}_k, \forall k \in [k_{min}, k_{max}]$. □

This property proves the necessity to keep definitely an *active* index since it is "activated" until it is "deactivated", i.e. the rank of a strand must be stored since it is involved in a substructure in the dynamic programming process until it is totally absorbed in another substructure (see Figure 2.8).

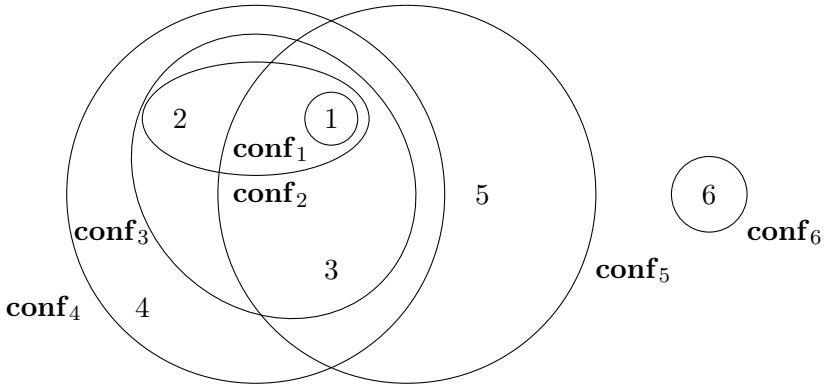


Figure 2.8: Schema of sets $\mathbf{conf}_k$ corresponding to $\sigma = \{1, 2, 5, 4, 3, 6\}$

Now we have to decide at which minimal level k the quantities $\mathcal{E}_{\text{adj}}$ and $\mathcal{E}_{\text{loop}}$ are determined and can be integrated in the dynamic programming scheme. For the $\mathcal{E}_{\text{adj}}$ terms, it is easily checked that the previous or the next strand along the barrel is already placed when $\mathbf{left}_k < k$ or $\mathbf{right}_k < k$, respectively.

Proposition 2.4. *For all k, we have:*

$$\begin{aligned} \mathbf{left}_k < k \quad &\Longleftrightarrow \quad \mathbf{left}_k \in \mathbf{conf}_{k-1}, \\ \mathbf{right}_k < k \quad &\Longleftrightarrow \quad \mathbf{right}_k \in \mathbf{conf}_{k-1} \end{aligned}$$

Proof. This results from the definition of the *active* indices of $\mathbf{conf}_{k-1}$ (2.1).
($\Rightarrow$)

- If $\mathbf{left}_k = k - 1$, then $\mathbf{left}_k \in \mathbf{conf}_{k-1}$.
- If $\mathbf{left}_k < k - 1$, as $\mathbf{left}_k$ and k are *adjacent*, we have also $\mathbf{left}_k \in \mathbf{conf}_{k-1}$.

($\Leftarrow$)
$\mathbf{left}_k \in \mathbf{conf}_{k-1}$ implies $\mathbf{left}_k \leq k - 1 < k$. □

To simplify the energy expression, we use the following notation for an *ifelse* function:

$$\mathbf{if}_k(i, \mathcal{E}) = \begin{cases} \mathcal{E} & \text{if } i < k \\ 0 & \text{otherwise} \end{cases}$$

For the $\mathcal{E}_{\text{loop}}$ terms, the problem is to wait until the *relative shear* between the two ends of a turn or loop is solved by the interleaf adjacencies. So, in the given example, the energy of the loop between the 2^{nd} and 3^{rd} strands can only be evaluated when the 5^{th} strand has been laid and the optimal *relative shear* $s^*_{\text{adj}}(v_2, v_3) = s_{\text{adj}}(v_2, v_5)+s_{\text{adj}}(v_5, v_4)+s_{\text{adj}}(v_4, v_3)$ is known.

Definition 2.5. *Let Δ_k be the relation on positive integers, defined as: $\forall\, i, j$,*

$$i\,\Delta_k\,j \iff \left[\begin{array}{l} i = j \\ (i \le k)\ \wedge\ (j \le k)\ \wedge\ (i, j \text{ are adjacent}) \end{array}\right.$$

then let Δ^*_k denote the equivalence relation defined by the transitive closure of Δ_k and let $\mathbf{A}_k = \{\, i < k \mid i\,\Delta^*_k\,(i+1)\,\}$.

Thus, $i \in \mathbf{A}_k$ means that the i^{th} and $(i+1)^{\text{th}}$ strands are geometrically linked by adjacency when the k^{th} substructure is laid (see Figure 2.9) and we can compute by composition an optimal *relative shear* $s^*_{\text{adj}}(v_i, v_{i+1})$. We temporarily forget here the closure of the substructures.

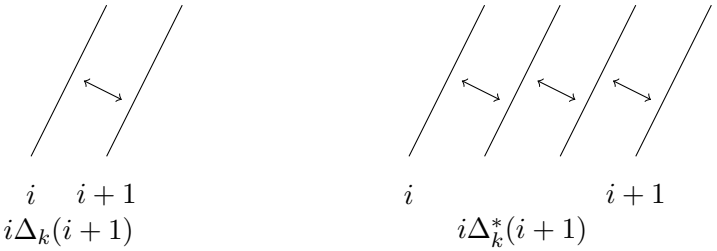


Figure 2.9: Relation Δ_k and its transitive closure Δ^*_k on the k^{th} substructure

Corollary 2.5. *The sequence $\{\mathbf{A}_k\}_{k=2,3\ldots}$ is increasing: $\forall k \ge 2, \mathbf{A}_k \subset \mathbf{A}_{k+1}$.*

This explains the fact that if the i^{th} and $(i+1)^{\text{th}}$ strands are linked to each other by adjacency in the $(k-1)^{\text{th}}$ substructure, then they are also linked in the k^{th} substructure. We will now focus on the set $\delta\mathbf{A}_k = \mathbf{A}_k \backslash \mathbf{A}_{k-1}$, $\forall\, k > 1$.

Proposition 2.6. *For all k, we have:*

$$(k-1) \in \delta\mathbf{A}_k \iff \mathbf{left}_k\,\Delta^*_{k-1}\,(k-1) \quad \vee \quad \mathbf{right}_k\,\Delta^*_{k-1}\,(k-1)$$

Proof. We have straightforwardly:

$$\begin{array}{ll} & (k-1) \in \delta\mathbf{A} = \mathbf{A_k} \backslash \mathbf{A_{k-1}} \\ \iff & (k-1) \in \mathbf{A_k}, (\text{as } (k-1) \notin \mathbf{A_{k-1}}) \\ \iff & (k-1)\,\Delta^*_k\,k \\ \iff & (k-1)\,\Delta^*_{k-1}\,\mathbf{left}_k \quad \vee \quad \mathbf{right}_k\,\Delta^*_{k-1}\,(k-1), \\ & (\text{ since } k \text{ is } \textit{adjacent} \text{ to } \mathbf{left}_k \text{ and } \mathbf{right}_k) \end{array}$$

□

Proposition 2.7. *For all* $i < k-1$,

$$i \in \delta\mathbf{A}_k \iff \begin{cases} i \notin \mathbf{A}_{k-1} \\ \left[\begin{array}{lcl} \mathbf{left}_k\, \Delta^*_{k-1}\, i & \wedge & \mathbf{right}_k\, \Delta^*_{k-1}\, (i+1) \\ \mathbf{right}_k\, \Delta^*_{k-1}\, i & \wedge & \mathbf{left}_k\, \Delta^*_{k-1}\, (i+1) \end{array}\right. \end{cases}$$

Proof.

($\Rightarrow$) If $i \in \delta\mathbf{A}_k$ then $i \notin \mathbf{A}_{k-1}$ and $i \in \mathbf{A}_k$, which means that the i^{th} and $(i+1)^{\text{th}}$ strands are linked in the k^{th} substructure, but not in $(k-1)^{\text{th}}$ substructure. This implies that the k^{th} strand is located between the i^{th} and $(i+1)^{\text{th}}$ strands in the k^{th} substructure. We then deduce the links in the $(k-1)^{\text{th}}$ substructure, that is:

$$\left[\begin{array}{lcl} \mathbf{left}_k\, \Delta^*_{k-1}\, i & \wedge & \mathbf{right}_k\, \Delta^*_{k-1}\, (i+1) \\ \mathbf{right}_k\, \Delta^*_{k-1}\, i & \wedge & \mathbf{left}_k\, \Delta^*_{k-1}\, (i+1) \end{array}\right.$$

($\Leftarrow$) Reversely, the links determined in the $(k-1)^{\text{th}}$ substructure by

$$\left[\begin{array}{lcl} \mathbf{left}_k\, \Delta^*_{k-1}\, i & \wedge & \mathbf{right}_k\, \Delta^*_{k-1}\, (i+1) \\ \mathbf{right}_k\, \Delta^*_{k-1}\, i & \wedge & \mathbf{left}_k\, \Delta^*_{k-1}\, (i+1) \end{array}\right.$$

lead to the link of the i^{th} and $(i+1)^{\text{th}}$ strands in the k^{th} substructure, i.e. $i \in \mathbf{A}_k$. Thus, $i \in \delta\mathbf{A}_k$. □

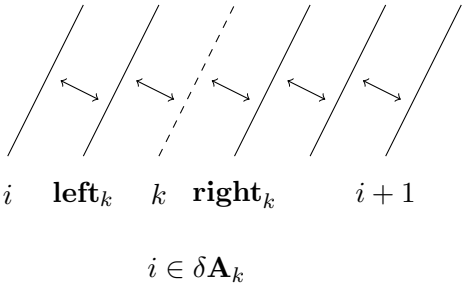


Figure 2.10: Illustration for property 2.7

Definition 2.6. *Let* $\mathbf{T}_k \subset \mathbf{V}^{*|\mathbf{conf}_k|}$ *denote the set of all tuples of* $|\mathbf{conf}_k|$ *vertices such that there is at least one path (of* k *edges) starting from* $\top$ *and passing through these vertices in order.*

For any instance $\mathbf{z} \in \mathbf{T}_k$ *of such a tuple and,* $\forall\, i \in \mathbf{conf}_k$, *let* $\mathbf{z}[i]$ *denote the* i^{th} *vertex of the corresponding path.*

This notation (not to be confused with $\mathbf{z}_i$, the i^{th} component of tuple $\mathbf{z}$) is not ambiguous since, from definition, the vertex $\mathbf{z}[i]$ is in common to any path associated to $\mathbf{z}$. Particularly, $\mathbf{z}[k]$ is the last vertex of any path associated to $\mathbf{z}$.

Proposition 2.8. *For all $\mathbf{z} \in \mathbf{T}_k$, the set of tuples corresponding to the paths of length $k-1$ that can be extended to a path corresponding to $\mathbf{z}$ is defined as:*

$$\mathbf{pre}(\mathbf{z}) = \{\, \mathbf{y} \in \mathbf{T}_{k-1} \mid ((\mathbf{y}[k-1], \mathbf{z}[k]) \in \mathbf{E}) \ \wedge \ (\forall\, i \in \mathbf{conf}_k \cap \mathbf{conf}_{k-1},\ \mathbf{y}[i] = \mathbf{z}[i]) \,\}$$

Let $\mathbf{C}^h_{k,\mathbf{z}}$ be the maximum value for $\mathbf{C}$ over all paths starting from $\top$ and leading in order through the vertices of a given tuple $\mathbf{z} \in \mathbf{T}_k$ with a shear number of h of the corresponding β-barrel. The general recurrence relation is: $\forall \mathbf{z} \in \mathbf{T}_k$,

$$\begin{aligned}\mathbf{C}^h_{k,\mathbf{z}} \quad = \quad \max_{\mathbf{y} \in \mathbf{pre}(\mathbf{z})} \Big(& \mathbf{C}^{h - s_{\text{adj}}(y[\mathbf{left}_k], \mathbf{z}[k]) - s_{\text{adj}}(\mathbf{z}[k], y[\mathbf{right}_k]) + s_{\text{adj}}(y[\mathbf{left}_k], y[\mathbf{right}_k])}_{k-1,\mathbf{y}} \ - \ \mathcal{E}_{\text{intr}}(\mathbf{z}[k]) \\ & - \mathbf{if}_k(\mathbf{left}_k, \mathcal{E}_{\text{adj}}(\mathbf{y}[\mathbf{left}_k], \mathbf{z}[k]) - \ \mathbf{if}_k(\mathbf{right}_k, \mathcal{E}_{\text{adj}}(\mathbf{z}[k], \mathbf{y}[\mathbf{right}_k]) \\ & - \sum_{i \in \delta \mathbf{A}_k} \mathcal{E}_{\text{loop}}\big(\mathbf{y}[i], \mathbf{y}[i+1], \sigma^{-1}(i+1) - \sigma^{-1}(i), s^*_{\text{adj}}(\mathbf{y}[i], \mathbf{y}[i+1])\big) \ \Big)\end{aligned}$$

Note that, from proposition 2.4, $\forall \mathbf{y} \in \mathbf{T}_{k-1}$, if $\mathbf{left}_k < k$ then the vertex $\mathbf{y}[\mathbf{left}_k]$ is defined (and the same is worth for $\mathbf{right}_k$). We can check that each $\mathcal{E}_{\text{adj}}$ term is finally counted exactly once in the sum, at the level corresponding to the position of its further vertex in the sequence order. The optimum is found at $k = n$ and $h = S$.

2.7 Complexity on permuted structures

Corollary 2.9. *The complexities both in time and space are $\mathcal{O}(\sum_{k=2}^{n} (\frac{|V|}{n})^{|\mathbf{conf}_k|})$, that is $\mathcal{O}(nN^{\max_k |\mathbf{conf}_k|})$.*

For any permutation, we have

$$|\mathbf{conf}_{n-k}| \leq \min\{1 + 2k, n - k\} \ , \forall k = 0, ..., n-1$$

Hence $\max_k |\mathbf{conf}_k| \leq 1 + (2n-2)/3$.

We study below the complexity of our dynamic programming scheme for certain classes of permuted structures. We first remind some notions of *permutation* and *group theory*.

2.7.1 Preliminaries

Definition 2.7. Permutation

A permutation on a set of objects is a sequential arrangement of these objects into certain order. In other words, it is a bijection from the set of objects to itself.

A permutation σ is noted as:

$$\begin{pmatrix} 1 & 2 & \cdots & n-1 & n \\ \sigma(1) & \sigma(2) & \cdots & \sigma(n-1) & \sigma(n) \end{pmatrix}$$

where the first row is the list of objects, and the image of each object under permutation σ is put below itself in the second row.

It can also be briefly written as $\sigma(1)\sigma(2)\ldots\sigma(n-1)\sigma(n)$. Obviously, the number of permutations on a set of n distinct objects is $n! = n.(n-1)...2.1$.

Example 2.1.

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 2 & 5 & 6 & 1 \end{pmatrix} = 342561$$

represent the sequential arrangement where the object with label 3 is first, the item with label 4 is second, etc. ◁

The *composition*, or *product*, of two permutations σ and π, denoted $\sigma \bullet \pi$ is defined as a bijection from the set of objects to itself that maps any object i to $\sigma(\pi(i))$ (the permutations are applied from right to left). This is again a permutation on this set of objects.

As the composition of functions is always associative, so is the composition of permutations:

$$\sigma \bullet (\pi \bullet \rho) = (\sigma \bullet \pi) \bullet \rho = \sigma \bullet \pi \bullet \rho, \text{ for all permutations } \sigma, \pi, \rho$$

It should be also noted that the composition is not *commutative*.

Example 2.2.

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 2 & 5 & 6 & 1 \end{pmatrix} \qquad \pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 4 & 5 & 1 & 3 & 6 \end{pmatrix}$$

The computation of $\sigma \bullet \pi$ can be represented in three rows. The second row is the image of the objects in the first row under π. The third row is the image of the second one under σ.

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 4 & 5 & 1 & 3 & 6 \\ 4 & 5 & 6 & 3 & 2 & 1 \end{pmatrix}$$

By eliminating the intermediary rows, we finally have:

$$\sigma \bullet \pi = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 3 & 2 & 1 \end{pmatrix}$$

◁

The *identity* permutation $\mathrm{Id}_n = 12\ldots(n-1)n$ which maps each object to itself is the neutral element for the composition.

$$\sigma \bullet \mathrm{Id}_n = \mathrm{Id}_n \bullet \sigma = \sigma, \text{ for all permutations } \sigma$$

Any permutation, as a bijection, σ has its *inverse* σ^{-1} that is also a permutation:

$$\sigma(i) = j \iff \sigma^{-1}(j) = i$$

The inverse σ^{-1} can be obtained by interchanging the two rows of σ, then sorting the first row accordingly.

Example 2.3.

$$\sigma^{-1} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 4 & 2 & 5 & 6 & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 3 & 4 & 2 & 5 & 6 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 6 & 3 & 1 & 2 & 4 & 5 \end{pmatrix}$$

◁

Definition 2.8. Cycle

A permutation $\sigma = \sigma_1\sigma_2\ldots\sigma_t$ is said to be a cycle if and only if

$$\begin{cases} \sigma(k) = \sigma_{k+1}, \ \forall k = 1, \ldots, t-1 \\ \sigma(t) = \sigma_1 \end{cases}$$

It is written as $(\sigma_1 \ \sigma_2 \ \ldots \ \sigma_t)$.

A permutation can be represented in cycle form by a decomposition into disjoint cycles. Thus, an element in a permutation of size n belongs to a unique cycle of length from 1 to n, and the permutation is comprised of a set of from 1 to n cycles. We can decompose a permutation σ as follows: choose some element i from σ, the cycle containing i is constructed by taking successively images under σ until the image would be i:

$$(i \ \ \sigma(i) \ \ \sigma(\sigma(i)) \ \ldots).$$

We repeat this process by choosing an element of σ that is not taken into account until all elements have been considered.

Example 2.4.

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 5 & 1 & 3 & 2 & 6 \end{pmatrix} = (1\ 4\ 3)(2\ 5)(6) = (2\ 5)(1\ 4\ 3)(6) = (2\ 5)(6)(4\ 3\ 1)$$

◁

This might be read ad "1 goes to 4 goes to 3 goes to 1", and so on. The elements i such that $i = \sigma(i)$ are called *fixed points* of σ, for instance 6 in the permutation above. Without confusion, we can ignore the cycles of length 1, or fixed points, in the notation. For example,

- $(1\ 4\ 3)(2\ 5)(6) = (1\ 4\ 3)(2\ 5)$
- $\mathrm{Id}_4 = (1)$

In this cycle notation, the reverse of a permutation can be obtained by reversing the order of the elements in each of its cycles.

Example 2.5.

$$\sigma = (1\ 4\ 3)(2\ 5)(6),\ \text{then}\ \sigma^{-1} = (3\ 4\ 1)(5\ 2)(6) = (1\ 3\ 4)(2\ 5)(6)$$

◁

We use in this thesis a notion of *circular permutation* that can be defined in a different way in the literature. A *circular permutation* is a sequential arrangement of the objects along a fixed circle. We distinguish here clock-wise and anti-clock-wise orders to take into consideration the slant angle of β-barrels in our application. Since the circle can be rotated, the number of circular permutations on a set of n distinct objects is $(n-1)!$. For example, 123 is the same as 231, but different from 132.

Definition 2.9. Group

Let G be a finite or infinite set of elements and $\bullet$ be a binary operation. We note, for simplicity, $a \bullet b$ as ab. A group is the pair $(G, \bullet)$ that satisfies:

i. Closure: $\forall a, b \in G, ab \in G$.

ii. Associativity: $\forall a, b, c \in G, (ab)c = a(bc)$.

iii. Identity: $\exists e \in G, \forall a \in G, ae = ea = a$. *The identity element e is also denoted* $\mathbf{1}_G$.

iv. Inverse: $\forall a \in G, \exists a^{-1} \in G, aa^{-1} = e$.

• *is also called the* group operation. *G is said to be a group* under *this operation. The* order *of a group is its cardinality, i.e. the number of its elements.*

If the binary function is commutative, i.e. $\forall a, b \in G, ab = ba$, *then the group is called an* abelian *group, or* commutative *group.*

Definition 2.10. Subgroup

A subset H of G is a subgroup *of group G under the operation $\bullet$ if H also forms a group under $\bullet$. In other words, H is nonempty and closed under operations $\bullet$ and inverse:* $\forall a, b \in H, ab \in H$ *and* $a^{-1} \in H$. *It is written as* $H \leq G$ *and read as "H is a subgroup of G".*

The order of any subgroup of a group of order n must be a divisor of n.

Let S be a subset of G. There exists a minimum subgroup of G containing S. It is said to be the subgroup generated by S and is denoted $\langle S \rangle$.

Definition 2.11. Symmetric group
The symmetric group S_n *is the group whose elements are permutations on* n *symbols, and whose group operation is the composition of such permutations.*

The order of S_n is the number of possible permutations, i.e. $n!$. In case of circular permutations, the order of S_n is $(n-1)!$.

Definition 2.12. Permutation group
A permutation group *is a subgroup of the* symmetric group S_n*. The order of a permutation group is then a divisor of* $n!$*, or* $(n-1)!$ *with circular permutations.*

Example 2.6. S_4 is the symmetric group on the set $M = \{1, 2, 3, 4\}$. We consider the set G of permutations that contains:

- $e = (1)(2)(3)(4) = \mathrm{Id}_4 = (1)$
- $a = (1\ 2)(3)(4) = (1\ 2)$
- $b = (1)(2)(3\ 4) = (3\ 4)$
- $ab = (1\ 2)(3\ 4)$

G forms a permutation group, since

- $aa = bb = e, ba = ab, aab = aba = b, abb = bab = b, abab = e$
- $a^{-1} = a, b^{-1} = b, (ab)^{-1} = ab$

◁

Theorem 2.10. *The set AP of circular permutations corresponding to permuted barrel structures of size* n *that ensure the antiparallel pairing is a permutation group or a subgroup of the symmetric group* S_n *under composition, where* n *is even.*

Proof. A permuted barrel structure of size n ensures the antiparallel pairing if and only if the corresponding circular permutation has the cycle-decomposition form of:

$$(e_1\ \dots\ e_{i_1})(e_{i_1+1}\ \dots\ e_{i_2})\dots(e_{i_r+1}\ \dots\ e_k)(o_1\ \dots\ o_{j_1})(o_{j_1+1}\ \dots\ o_{j_2})\dots(o_{j_s+1}\ \dots\ o_k)$$

where $n = 2k$, e_i is even, and o_i is odd for all i. We can choose 1 as a fixed point since the permutation is circular.

- AP contains Id_n
- The composition of two such permutations gives a circular permutation in which 1 is still a fixed point and the parity of cycles is kept unchanged. AP is then closed under composition.

- The inverse is obtained by reversing the order of the elements in each cycle, thus the fixed point 1 and the parity of cycles are unchanged. Therefore, AP is also closed under inversion.

So, AP is a subgroup of S_n. □

Lemma 2.11. *Any permutation containing a Greek key motif can be written as* $(k\ k+2)$ *for some* k.

Proof. This can be straightforwardly deduced from Greek key motifs that have the form $k(k+3)(k+2)(k+1)$ (denoted g_+) or $(k+2)(k+1)k(k+3)$ (denoted g_-) [139]. The permutation has all cycles of length 1, except for $((k+1)\ (k+3))$ or $(k\ (k+2))$. □

Theorem 2.12. *The subgroups* $H_1 = \left\langle \{((4k+1)\ (4k+3))\}_{k=0,1\ldots} \right\rangle$, $H_2 = \left\langle \{((4k+2)\ (4k+4))\}_{k=0,1\ldots} \right\rangle$, $H_3 = \left\langle \{((4k+3)\ (4k+5))\}_{k=0,1\ldots} \right\rangle$, $H_4 = \left\langle \{((4k+4)\ (4k+6))\}_{k=0,1\ldots} \right\rangle$ *represent the barrel structures with disjoint Greek key motifs. These subgroups are abelian.*

Proof. We prove the theorem for the subgroup $H_1 = \left\langle \{((4k+1)\ (4k+3))\}_{k=0,1\ldots} \right\rangle$. The proof is the same for the others.

Straightforwardly, the cycles $((4k+1)\ (4k+3))$'s are either disjoint or identical. The composition applied on them is then commutative, and thus the subgroup H_1 is abelian.

For every permutation σ in H_1, the Greek key motifs in σ are of form: $(4k)(4k+3)(4k+2)(4k+1)$. There does not exist two different values k_1 and k_2 such that $(4k_1)(4k_1+3)(4k_1+2)(4k_1+1)$ and $(4k_2)(4k_2+3)(4k_2+2)(4k_2+1)$ are overlapped. Hence, the Greek key motifs in σ are disjoint. □

We also note that $\left\langle \{((4k+1)\ (4k+3))\}_{k=0,1\ldots}, \{((4k+2)\ (4k+4))\}_{k=0,1\ldots}, \{((4k+3)\ (4k+5))\}_{k=0,1\ldots}, \{((4k+4)\ (4k+6))\}_{k=0,1\ldots} \right\rangle = \left\langle \{((2k-1)\ (2k+1))\}_{k=1,2\ldots}, \{((2k)\ (2k+2))\}_{k=1,2\ldots} \right\rangle$ is the subgroup AP.

We study different possible configurations for disjoint Greek key motifs in permutations. The regular expression is used to describe the permutation. We consider the alphabet $\Sigma = \{\mathrm{Id}, g_+, g_-\}$, where Id represents the identity motifs, g_+ represents Greek key motifs of form $k(k+3)(k+2)(k+1)$ and g_- represents $(k+2)(k+1)k(k+3)$. A permutation with disjoint Greek key motifs can be written as a word of Σ^*. For example, $14325678 = g_+\mathrm{Id} = \mathrm{Id}g_-\mathrm{Id}$, $14327658 = g_+g_-$.

- For $\sigma \in H_1$:
  - $\sigma = \mathrm{Id}$: $\max_k |\mathbf{conf}_k| = 2$. The complexity of the prediction algorithm is $\mathcal{O}(nN^2)$.

  - $\sigma = (1\ 3) = g_-\mathrm{Id} : \max_k |\mathbf{conf}_k| = 2$. The complexity is $\mathcal{O}(nN^2)$.
  - $\sigma = \mathrm{Id}g_-(\mathrm{Id} + g_-)^* + g_-(\mathrm{Id} + g_-)^*g_-(\mathrm{Id} + g_-)^* : \max_k |\mathbf{conf}_k| = 4$. The complexity is $\mathcal{O}(nN^4)$.

- For $\sigma \in H_2$:
  - $\sigma = \mathrm{Id} : \mathcal{O}(nN^2)$.
  - $\sigma = (2\ 4) = g_+\mathrm{Id} : \max_k |\mathbf{conf}_k| = 3$. The complexity is then $\mathcal{O}(nN^3)$.
  - $\sigma = \mathrm{Id}g_+ : \max_k |\mathbf{conf}_k| = 2$. The complexity is then $\mathcal{O}(nN^2)$.
  - $\sigma = g_+\mathrm{Id}g_+ + g_+g_+ : \max_k |\mathbf{conf}_k| = 3$. The complexity is then $\mathcal{O}(nN^3)$.
  - $\sigma = (\mathrm{Id} + g_+)^+g_+(\mathrm{Id} + g_+)^+ : \max_k |\mathbf{conf}_k| = 4$. The complexity is $\mathcal{O}(nN^4)$.

- For $\sigma \in H_3$:
  - $\sigma = \mathrm{Id} : \mathcal{O}(nN^2)$.
  - $\sigma = \mathrm{Id}g_+(\mathrm{Id} + g_+)^+ : \max_k |\mathbf{conf}_k| = 4$. The complexity is $\mathcal{O}(nN^4)$.

- For $\sigma \in H_4$:
  - $\sigma = \mathrm{Id} : \mathcal{O}(nN^2)$.
  - $\sigma = \mathrm{Id}g_+ : \max_k |\mathbf{conf}_k| = 2$. The complexity is then $\mathcal{O}(nN^2)$.
  - $\sigma = \mathrm{Id}g_+(\mathrm{Id} + g_+)^+ : \max_k |\mathbf{conf}_k| = 4$. The complexity is $\mathcal{O}(nN^4)$.

Thus, the complexity of the prediction algorithm for the subgroups H_1, H_2, H_3, H_4 is from $\mathcal{O}(nN^2)$ to $\mathcal{O}(nN^4)$, according to the given permutation. For a β-barrel structure with identity permutation, which we observed the most in nature, it is possible to compute the optimal structure in $\mathcal{O}(nN^2)$ running time.

More generally, for a permutation σ that differs from the identity permutation by disjoint Greek key motifs, i.e. $\sigma = (\mathrm{Id} + g_+ + g_-)^+$, we also have a complexity in time and space from $\mathcal{O}(nN^2)$ to $\mathcal{O}(nN^4)$.

Chapter 3

Tree-decomposition based algorithm

3.1 Introduction

Our previous dynamic programming scheme can be seen as a way to extract from a graph optimal paths following a given pattern. It can find the optimal permuted β-barrel structures with disjoint Greek key motifs in time from $\mathcal{O}(N^2)$ to $\mathcal{O}(N^4)$. We describe in this chapter yet another algorithm based on tree decomposition that predicts more efficiently these structures. Our tree decomposition based algorithm is able to deal with such β-barrel structures in time at most $\mathcal{O}(N^3)$ [122], a non trivial improvement.

In Section 3.2, we introduce the essential graph-theoretic background on tree decomposition and modular decomposition. The NP-completeness of the problem of finding the arbitrarily permuted structure of minimum energy is discussed in Section 3.3. We describe the algorithm in Section 3.4, followed by a complexity analysis regarding the Greek key motifs.

3.2 Graph-theory background

We recall some standard notions from graph theory. Let $G = (V, E)$ be an undirected graph with vertex set V and edge set E that has no edge connecting a vertex to itself (no loop) and no more than one edge between any two different vertices. A *subgraph* H of a graph G is a graph whose vertex set is a subset of V, and whose edge set is a subset of E restricted to its vertex set. A subgraph H is said to be *induced* if the edges of H are the ones appearing in G over the same vertex set, i.e.

$$\forall x, y \in V(H), (x, y) \in E(H) \iff (x, y) \in E(G).$$

H can be constructed from G by removing all vertices in $V(G) \backslash V(H)$ and their incident edges. For a subset X of $V(G)$, $G[X]$ denotes the induced subgraph of G, and is said to

be *induced by* X.

A set of vertices X is called a *separator* of a connected graph G if $G[V \backslash X]$ is disconnected.

An *outerplanar graph* is a graph that can be drawn in the plane in such a way that no edges cross each other and all the vertices belong to the unbounded face.

3.2.1 Tree decomposition

Roberton and Seymour introduced the concept of tree decomposition, treewidth, path decomposition and pathwidth in their studies on graph minors in 1980's [106, 105]. This concept has been widely studied and applied to solve several combinatorial problems that are NP-hard for general graphs. Such problems can be efficiently solved in polynomial time by using dynamic programming on a tree decomposition (or path decomposition) of graphs of bounded treewidth (or pathwidth) [3, 5, 14, 18]. The protein structure prediction problems such as protein threading for backbone prediction and protein side-chain prediction can also be solved using this technique [135].

Definition 3.1. Tree decomposition - Treewidth

A **tree decomposition**, *denoted* $D_T(G)$, *of a graph* $G(V, E)$ *is a pair* $(\mathcal{X}, \mathcal{T})$, *where* $\mathcal{X} = \{X_i | i \in I\}$ *is a family of subsets of* V, *and* $\mathcal{T}$ *a tree whose nodes are the subsets* X_i *satisfying:*

- $\bigcup_{i \in I} X_i = V$
- $\forall (u, v) \in E, \exists i \in I : u, v \in X_i$
- $\forall i, j, k \in I$: *if* X_j *is in the path from* X_i *to* X_k, *then* $X_i \cap X_k \subseteq X_j$

The **width** *of a tree decomposition* $D_T(G)$ *is* $\max_i |X_i| - 1$. *The* **treewidth** *of a graph* G, *denoted* $tw(G)$, *is the minimum width among all tree decompositions of* G.

Definition 3.2. Path decomposition - Pathwidth

A **path decomposition**, $D_P(G)$, *is a tree decomposition where the tree* $\mathcal{T}$ *is reduced to a path. The* **pathwidth** *of a graph* G, *denoted* $pw(G)$, *is the minimum width among all path decompositions of* G.

The treewidth and pathwidth of a graph G measure the distance from G to a tree and a chain, respectively. The smaller the treewidth (pathwidth), the more "tree-like" ("chain-like") the graph is. For any graph, its treewidth is always less than its pathwidth, as every path decomposition is also a tree decomposition. The simplest tree decomposition or path decomposition of a graph G is a single set containing all vertices of G that gives the width of $|V| - 1$. For example,

- A graph G has treewidth 1 if and only if G is a forest;

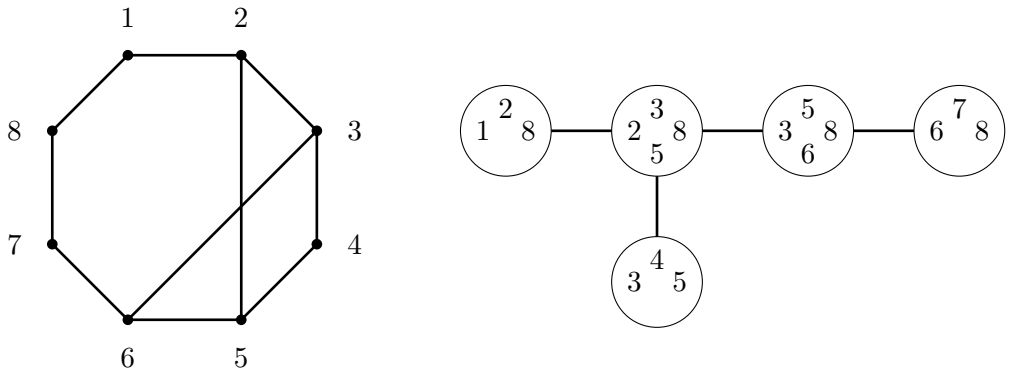


Figure 3.1: A graph and a tree decomposition of width 3

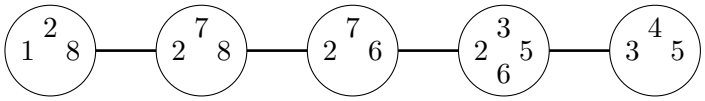


Figure 3.2: A path decomposition of width 3 of the graph in 3.1

- If G is a cycle then $tw(G) = pw(G) = 2$.
- If G is an outerplanar graph then $tw(G) = 2$.
- If G is a k-clique then $tw(G) = pw(G) = k - 1$.

It is worth reminding the fundamental properties observed and proved in [21, 19, 46, 106] which are usually used for the analysis of tree decomposition based dynamic programming algorithms.

i. *If H is a subgraph of G then $tw(H) \leq tw(G)$.*

ii. *Let $(\mathcal{X} = \{X_i | i \in I\}, \mathcal{T})$ be a tree decomposition of G. For any clique $G[X]$, $X \subseteq V$, there exists $i \in I$ such that $X \subseteq X_i$.*

iii. *If graph G has treewidth at most k then G has a vertex of degree at most k.*

iv. *If graph $G = (V, E)$ has treewidth at most k then G has at most $k|V| - \binom{k+1}{2}$ edges.*

v. *Let $(\mathcal{X}, \mathcal{T})$ be some tree decomposition of G, ij an edge of $\mathcal{T}$, and $\mathcal{T}_1, \mathcal{T}_2$ the two connected components of $\mathcal{T} - ij$, then $X_i \cap X_j$ is a separator between $\cup_{i \in \mathcal{T}_1}$ and $\cup_{i \in \mathcal{T}_2}$.*

vi. Graph G has treewidth at most k if and only if G can be decomposed using only separators of size at most k.

In an arbitrary graph G, finding $tw(G)$ or $pw(G)$ are NP-complete problems [4]. However, the problem with a fixed parameter is tractable: testing if a graph G has treewidth (or pathwidth) at most k and construct a tree decomposition (or path decomposition) accordingly [19]. This can be solved in a time that is linear in the size of the graph but exponential in parameter k. It is also NP-hard to absolutely approximate treewidth and pathwidth of arbitrary graphs [20]. It is still an open question whether there is a polynomial-time approximation scheme (PTAS) for treewidth and pathwidth.

3.2.2 Modular decomposition

The technique of modular decomposition has been introduced by Gallai [43]. This concept arises in various algorithmic topics. It is an important preprocessing step of several combinatorial algorithms [53, 93].

Definition 3.3. Module

A **module** *of a graph $G(V, E)$ is a subset of vertices $M \subseteq V$ such that, for every vertex $v \notin M$, either v is a neighbor of every element of M or v is not a neighbor of any element of M. In other words, M is a module if and only if all elements of M have the same neighbors that are not in M.*

$\emptyset$, singletons, V are trivial modules. A graph is **prime** *if it admits only trivial modules.*

A **strong module** *of a graph G is a module M that does not strictly overlap any other module M': for any module $M'$of G, either $M \cup M' = \emptyset$ or $M \subseteq M'$ or $M' \subseteq M$*

Definition 3.4. Modular decomposition

A **modular partition** *of a graph $G(V, E)$ is a partition $\mathcal{P}$ of the vertex set V where each part is a module of G.*

The **quotient graph** *$G_{/\mathcal{P}}$ is the induced subgraph obtained by assigning each part of $\mathcal{P}$ to a vertex.*

3.3 NP-Completeness

We first prove the NP-completeness of the traveling salesman problem where we look for the *longest* tour in which a salesman can visit each city exactly once. Note that this is similar, but not the same, to the problem of finding the *shortest* tour which is mentioned more frequently in the literature [30].

TRAVELING SALESMAN:

Input Given n cities $c_1, c_2, \ldots, c_n$, a distance $d_{ij} > 0$ between each pair (c_i, c_j) and a positive m.

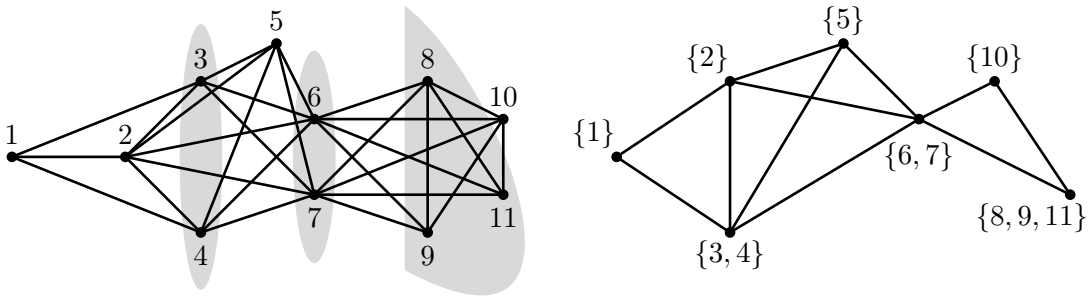


Figure 3.3: A graph and its modular decomposition are on the left. The quotient graph is on the right.

Question Is there a circular tour that visits each city exactly once of distance at least m?

Corollary 3.1. TRAVELING SALESMAN *is* NP*-complete.*

Proof. TRAVELING SALESMAN is in NP.

The same to the traveling salesman problem where we look for a shortest tour, to prove that TRAVELING SALESMAN is NP-complete, we describe a reduction from Hamiltonian Cycle Problem.

Let G an instance of Hamiltonian Cycle Problem, with n vertices, we create an instance of TRAVELING SALESMAN. For each vertex v, create a city c_v. If there is an edge (u, v), then the distance between c_u and c_v is 1; otherwise, the distance is 1/2. Let $m = n$.

We now prove that G has a Hamiltonian cycle if and only if there is a tour of distance at least n.

($\Rightarrow$) If G contains a Hamiltonian cycle, then this cycle forms a tour of distance n through all the cities.

($\Leftarrow$) If there is a tour of distance at least n through the n cities, where each city is visited exactly once, then the distance between each pair of cities along the tour must be 1. Thus each corresponding pair of vertices are adjacent in G. G has therefore a Hamiltonian cycle. □

We recall *circle-attached path* (see Section 2.5.4) a path in which the vertices are arranged onto a circle. The weight is defined on the adjacency of vertices in the path and the succession of vertices on the circle. Note that adjacent vertices in the path are not necessarily successive on the circle. We are interested in finding the order of vertices in such a circle-attached path. The problem is defined as followed:

PERMUTED BARREL:

Input Given a directed acyclic graph $G(V, E)$, a weight w defined on every vertex, a positive weight c on every edge, a positive weight e on every pair of vertices and a positive m. A circle-attached path has a weight of $\sum w(v_i) + \sum c(v_i, v_j) + \sum e(v_h, v_k)$, where v_i, v_j are adjacent in the path and v_h, v_k are successive on the circle.

Question Is there a circle-attached path of weight at least m?

Corollary 3.2. PERMUTED BARREL *is* NP*-complete.*

Proof. The weight of a circle-attached path is easily computed. PERMUTED BARREL is in NP.

We describe a reduction from TRAVELING SALESMAN.

Let (C, d, m) be an instance of TRAVELING SALESMAN, where C is the set of cities, d is the distance function between cities. For each city, we create a vertex in G. These vertices have weight $w = 0$. We add randomly directed edges of weight $c = 0$ to form a unique path through all vertices of G. Weight e between every pair of vertices is set to d.

It is clear that there is a tour of distance at least m if and only if there is a circle-attached path of weight at least m.

Then, finding the right permuted β-barrel structures is an NP-complete problem. □

We are interested in finding the permuted super-secondary structure of transmembrane β-barrel proteins corresponding to a given permutation σ. That is to find the maximum weighted circle-attached path, with the succession on the circle defined by the permutation σ of vertices in the path.

CONSTRAINT PERMUTED BARREL:

Input Given a directed acyclic graph $G(V, E)$, a weight w defined on every vertex, a positive weight c on every edge, a positive weight e and a shift s on every pair of vertices, a permutation σ of size n, an integer S, and a positive m. A circle-attached path has a weight of $\sum w(v_i) + \sum c(v_i, v_j) + \sum e(v_h, v_k)$, where v_i, v_j are adjacent in the path and v_h, v_k are successive on the circle.

Question Is there a circle-attached path corresponding to σ, which satisfies the constraint $\sum s_{v_h, v_k} = S$, of weight at least m?

Conjecture 3.1. CONSTRAINT PERMUTED BARREL *is* NP*-complete?*

We propose a dynamic programming approach that is described in the next section to solve the problem.

3.4 Algorithm for finding barrel structures of minimum energy

We call *n-strand barrel graph* corresponding to a permutation σ the contact graph $G_c = (V_c, E_c)$ of n vertices named by the ranks of the β-strands along the amino acid sequence, with edges representing the contact of strands in the barrel (see Figure 3.4). Thus, G_c is the superposition of the open path $(1, 2, .., n)$ and the σ-permuted closed path of $\{1, 2, ...n\}$, i.e. the closed path $(\sigma_1, \sigma_2, ..., \sigma_n)$.

We claim in Propositions 3.3 and 3.4 some fundamental properties of the *n-strand barrel graph* G_c.

Proposition 3.3. *Every vertex v in G_c has degree at least 2 and at most* 4.

Proof. Every vertex v has 1 or 2 neighbors in the path and 2 neighbors in the σ-permuted cycle. We have then $2 \leq \deg(v) \leq 4$ □

Proposition 3.4. $n \leq |E_c| \leq 2n - 1$

Proof. The σ-permuted cycle has n edges while the path $(1, 2, .., n)$ has $n - 1$ edges. $|E_c|$ gets the minimum value of n when σ is the identity permutation and gets the maximum value of $2n - 1$ when there is no common edge between the path and the cycle. □

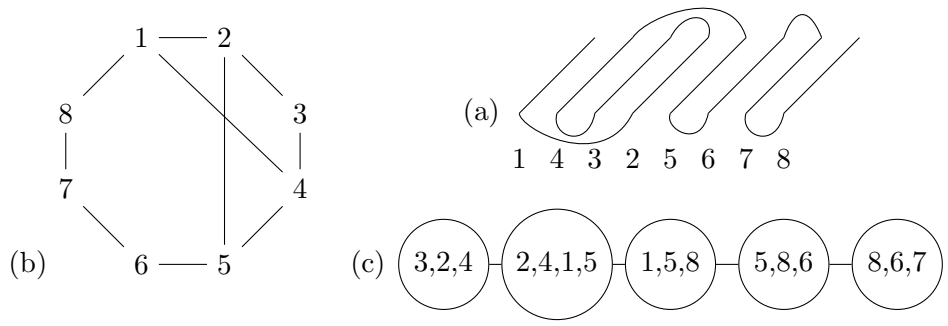


Figure 3.4: The β-barrel(a), G_c(b) and the tree/path decomposition(c) of $\sigma = 1\ 4\ 3\ 2\ 5\ 6\ 7\ 8$

We consider a path decomposition constructed by an *elimination process* as described below (see Procedure 1). G_c is modified at each step by removing some vertices and edges. Let $\mathcal{N}_c(r)$ be the set of neighbors of r in G_c, $d_c(r)$ the degree of r in G_c or $d_c(r) = |\mathcal{N}_c(r)|$, $\deg(r)$ the degree of r in the initial G_c, $md_c(A)$ be the vertex of A that has the lowest degree in G_c, and $G_c[A] = (A, E_c[A])$ be the graph induced by the set of vertices A.

Procedure 1 Elimination process

Input: *n-strand barrel graph* G_c

1: $X_0 \leftarrow \{md_c(V_c), \sigma(\sigma^{-1}(md_c(V_c)) + 1)\}$, where $\sigma_{n+1} = \sigma_1$
2: Remove $E_c[X_0]$ from G_c.
3: $k \leftarrow 1, r_1 \leftarrow md_c(X_0), X_1 \leftarrow X_0 \cup \mathcal{N}_c(r_1)$,
4: **repeat**
5: Remove $E_c[X_k]$, then all unconnected vertices, from G_c.
6: $r_{k+1} \leftarrow md_c(X_k \setminus \{r_k\})$
7: $X_{k+1} \leftarrow (V_c \cap (X_k \setminus \{r_k\})) \cup \mathcal{N}_c(r_{k+1})$
8: $k \leftarrow k + 1$.
9: **until** V_c is empty.

Output: $\{X_k\}_{k \in I = \{0,1,\ldots,K\}}$

The construction of r_k and X_k implies that r_k has at least one neighbor in $\bigcup_{i=0}^{k-1} X_i$. So, $\forall k \geq 1, d_c(r_k) \leq deg(r_k) - 1$, and thus, $1 \leq d_c(r_k) \leq 3$. We also have $\sum_{i=1}^{K} d_c(r_i) = |\bigcup_{i=1}^{K} \mathcal{N}_c(r_i)| = n - 2$, hence, $K \leq n - 2 \leq 3K$, where $K = |I| - 1$,

We derive then the bounds on the number K of subsets X_k:

Corollary 3.5. $\dfrac{n-2}{3} \leq K \leq n - 2$

The cardinal of X_k's is bounded above as:

Lemma 3.6. $\forall k \geq 1, |X_k| \leq 3 + \sum_{i=1}^{k} (d_c(r_i) - 1) \leq 3 + \sum_{i=1}^{k} (\deg(r_i) - 2)$

Proof. We have: $|X_1| = 1 + \deg(r_1) = 3 + (d_c(r_1) - 1)$.

By induction,

$$\begin{aligned} \forall k \geq 1, |X_{k+1}| &\leq |X_k| - 1 + |\mathcal{N}(r_{k+1})| \\ &\leq 3 + \sum_{i=1}^{k} (d_c(r_i) - 1) + (d_c(r_{k+1}) - 1) \\ &= 3 + \sum_{i=1}^{k+1} (d_c(r_i) - 1) \end{aligned}$$

Moreover, $d_c(r_k) \leq deg(r_k) - 1, \forall k \geq 1$.

Thus, $\forall k \geq 1, |X_k| \leq 3 + \sum_{i=1}^{k}(d_c(r_i) - 1) \leq 3 + \sum_{i=1}^{k}(\deg(r_i) - 2)$ □

As $X_k \cap X_{k+1} \subsetneq X_k$, then $|X_k \cap X_{k+1}| \leq |X_k| - 1, \forall k \geq 1$, we deduce:

Lemma 3.7. $\forall k \geq 1, |X_k \cap X_{k+1}| \leq 2 + \sum_{i=1}^{k}(d_c(r_i) - 1) \leq 2 + \sum_{i=1}^{k}(\deg(r_i) - 2)$

We firstly prove the following lemma in order to establish an upper bound on the cardinal of the intersections of X_k's.

Lemma 3.8.

- *If there exists k such that $|X_k \cap X_{k+1}| \geq \left\lceil \frac{n}{2} \right\rceil + 1$, then:*

$$\left[\begin{array}{l} k = K - 1 \\ |X_{k+1} \cap X_{k+2}| \leq |X_k \cap X_{k+1}| \end{array} \right.$$

- *If there exists k such that $|X_k \cap X_{k+1}| \geq \left\lceil \frac{n}{2} \right\rceil$, then:*

$$\left[\begin{array}{l} k = K - 1 \\ |X_{k+1} \cap X_{k+2}| \leq |X_k \cap X_{k+1}| + 1 \end{array} \right.$$

Proof.

- If there exists k such that $|X_k \cap X_{k+1}| \geq \left\lceil \frac{n}{2} \right\rceil + 1$, then:

$$\begin{aligned} 2 + \sum_{i=1}^{k}(d_c(r_i) - 1) &\geq \left\lceil \frac{n}{2} \right\rceil + 1 \quad \text{(Lemma 3.7)} \\ \Rightarrow \quad 2k &\geq \left\lceil \frac{n}{2} \right\rceil \quad \text{(since } d_c(r_1) \leq 2, d_c(r_i) \leq 3, \forall i > 1\text{)} \\ \Rightarrow \quad k &\geq \frac{1}{2}\left\lceil \frac{n}{2} \right\rceil \end{aligned}$$

Let u_k be the number of non-visited vertices ($u_k = n - |\bigcup_{i=1}^{k} X_i|$) and t_k the number of edges of G_c after removing $E_c[X_k]$ and all unconnected vertices (i.e. step 5 in

the elimination process). We have:

$$\begin{aligned} u_k = n - |\bigcup_{i=1}^{k} X_i| = n - (\sum_{i=1}^{k-1} |X_i \backslash X_{i+1}| + |X_k|) &\leq n - (k - 1 + |X_k|) \\ &\leq n - (k + |X_k \cap X_{k+1}|) \\ &\leq n - \left(\frac{1}{2}\left\lceil\frac{n}{2}\right\rceil + \left\lceil\frac{n}{2}\right\rceil + 1\right) \\ &\leq \frac{n}{4} - 1 \end{aligned}$$

Each non-visited vertex has degree at most 4. Therefore,

$$t_k \leq 4u_k \leq n - 4 \tag{3.1}$$

Following the construction of X_{k+1} from X_k, r_{k+1} is the vertex of minimum degree in $X_k \cap X_{k+1}$. Then,

$$t_k \geq d_c(r_{k+1})|X_k \cap X_{k+1}| \geq \left(\left\lceil\frac{n}{2}\right\rceil + 1\right) d_c(r_{k+1}) \tag{3.2}$$

(3.1) and (3.2) infer that, if there exists k such that $|X_k \cap X_{k+1}| \geq \left\lceil\frac{n}{2}\right\rceil + 1$, then $d_c(r_{k+1}) \leq 1$. Hence, $k + 1 = K$ or $|X_{k+1}| = |X_k \cap X_{k+1}| + 1$. We have then $k + 1 = K$ or $|X_{k+1} \cap X_{k+2}| \leq |X_{k+1}| - 1 = |X_k \cap X_{k+1}|$.

- Similarly, if there exists k such that $|X_k \cap X_{k+1}| \geq \left\lceil\frac{n}{2}\right\rceil$, then $k \geq \frac{1}{2}\left\lceil\frac{n}{2}\right\rceil - \frac{1}{2}$. So, $u_k \leq n-(k+|X_k \cap X_{k+1}|) \leq n - \left(\frac{1}{2}\left\lceil\frac{n}{2}\right\rceil + \left\lceil\frac{n}{2}\right\rceil - \frac{1}{2}\right) \leq \frac{n}{4} + \frac{1}{2}$. Hence, $t_k \leq n+2$. We have also $t_k \geq d_c(r_{k+1})|X_k \cap X_{k+1}| \geq \left\lceil\frac{n}{2}\right\rceil d_c(r_{k+1})$. Then, $n + 2 \geq \frac{n}{2} d_c(r_{k+1})$. This implies that for $n \geq 4$, $d_c(r_{k+1}) \leq 2$. So, $k + 1 = K$ or $|X_{k+1} \cap X_{k+2}| \leq |X_{k+1}| - 1 \leq |X_k \cap X_{k+1}| + 1$.

□

So, the cardinal of $(X_k \cap X_{k+1})$ is bounded by:

Theorem 3.9. $\forall k \geq 1, 2 \leq |X_k \cap X_{k+1}| \leq \left\lceil\frac{n}{2}\right\rceil + 1$

Proof. We first prove by contradiction that $|X_k \cap X_{k+1}| \geq 2$. The path decomposition requires that $X_k \cap X_{k+1}$ is the separator of two non-empty sets $(\bigcup_{i=1}^{k} X_i)\backslash(X_k \cap X_{k+1})$ and $(\bigcup_{i=k+1}^{K} X_i)\backslash(X_k \cap X_{k+1})$. If $\exists k, |X_k \cap X_{k+1}| = 1$, or $X_k \cap X_{k+1} = \{r_{k+1}\}$, then there would be no Hamiltonian cycle in the *n-strand barrel graph*, as in every complete tour, r_{k+1} is visited at least twice.

We now prove $|X_k \cap X_{k+1}| \leq \left\lceil\frac{n}{2}\right\rceil + 1$.

- For $n \leq 3$, we have $K = 1$. There is only X_1.
- For $3 < n \leq 5$, if $K \geq 2$, then $4 \geq |X_k| > |X_k \cap X_{k+1}|$, so $|X_k \cap X_{k+1}| \leq 3 \leq \left\lceil \frac{n}{2} \right\rceil + 1, \forall k$.
- For $n \geq 6$, $\forall k \geq 1$, we have: $(X_{k+1} \cap X_{k+2}) \backslash (X_k \cap X_{k+1}) \subseteq X_{k+1} \backslash (X_k \cap X_{k+1}) = X_{k+1} \backslash X_k = \mathcal{N}_c(r_{k+1})$ and $(X_k \cap X_{k+1}) \backslash (X_{k+1} \cap X_{k+2}) \supseteq \{r_{k+1}\}$

  Hence,

$$\begin{aligned} & |X_{k+1} \cap X_{k+2}| - |X_k \cap X_{k+1}| \\ = \; & |(X_{k+1} \cap X_{k+2}) \backslash (X_k \cap X_{k+1})| - |(X_k \cap X_{k+1}) \backslash (X_{k+1} \cap X_{k+2})| \\ \leq \; & |\mathcal{N}_c(r_{k+1})| - |\{r_{k+1}\}| = \; d_c(r_{k+1}) - 1 \leq \; 2 \end{aligned}$$

  So,

$$|X_{k+1} \cap X_{k+2}| \leq |X_k \cap X_{k+1}| + 2, \forall k \tag{3.3}$$

  Following the elimination process,

$$|X_1 \cap X_2| \leq |X_1| - 1 = \deg(r_1) \leq 3 \leq \left\lceil \frac{n}{2} \right\rceil \tag{3.4}$$

  Let k_0 be the minimum index such that $|X_{k_0} \cap X_{k_0+1}| \geq \left\lceil \frac{n}{2} \right\rceil$ (if k_0 is not determined, the theorem is proved). (3.3) and (3.4) imply that $|X_{k_0} \cap X_{k_0+1}| = \left\lceil \frac{n}{2} \right\rceil$ or $|X_{k_0} \cap X_{k_0+1}| = \left\lceil \frac{n}{2} \right\rceil + 1$. This, with regard to Lemma 3.8, ensures that $|X_k \cap X_{k+1}| \leq \left\lceil \frac{n}{2} \right\rceil + 1, \forall k = k_0, ..., K$.

□

Such an elimination process allows to construct a path decomposition with a bounded width in a linear time with regard to the number of edges in the *n-strand barrel graph*. A question arises as to whether there is a polynomial time algorithm to find out an optimum tree decomposition of such a graph.

Conjecture 3.2. *Finding treewidth of an n-strand barrel graph is NP-hard?*

We describe here the algorithm based on dynamic programming with a constraint on the *shear number* of the barrel. Let $\mathbf{W}_i$ denote the set of potential vertices in $\mathbf{V}$ for the i^{th} β-strand in sequence order, $\mathbf{U}_i \subset \prod_{k \in X_i} \mathbf{W}_k$ and $\mathbf{T}_i \subset \prod_{k \in X_i \cap X_{i+1}} \mathbf{W}_k$, the set of tuples of $|X_i|$ and $|X_i \cap X_{i+1}|$ vertices, respectively, such that there is at least a substructure of the barrel through these vertices, $\mathcal{E}(G_c[A](\mathbf{z}))$ the weight of the contact graph $G_c[A]$ where the tuple $\mathbf{z}$ of size $|A|$ is assigned to A.

Definition 3.5. *For all* $\mathbf{x} \in \mathbf{T}_i$*, the set of tuples which determine the substructures corresponding to* $\mathbf{x}$ *is defined as:*

$$\mathbf{ext}(\mathbf{x}) = \{\mathbf{z} \in \mathbf{U}_i \,|\, \forall k \in X_i \cap X_{i+1}, \mathbf{z}[k] = \mathbf{x}[k]\}$$

Definition 3.6. *For all* $\mathbf{z} \in \mathbf{U}_i$*, the reduced tuple of* $\mathbf{z}$ *is defined as:*

$$\mathbf{red}(\mathbf{z}) = \mathbf{x} \in \mathbf{T}_{i-1}, \text{ such that } \forall k \in X_{i-1} \cap X_i, \mathbf{z}[k] = \mathbf{x}[k]$$

We have the recurrence: $C^k_{\mathbf{x}} = -\mathcal{E}(G_c[X_0](\mathbf{x})), \forall \mathbf{x} \in \mathbf{T}_0$, where k is the *relative shear* defined by pairing vertices in $G_c[X_0](\mathbf{x})$, and

$$\forall \mathbf{x} \in \mathbf{T}_i,\ C^k_{\mathbf{x}} = \max_{\mathbf{z} \in \mathbf{ext}(\mathbf{x})} (C^{k'}_{\mathbf{red}(\mathbf{z})} - \mathcal{E}(G_c[X_i] \setminus G_c[X_{i-1}](\mathbf{z})))$$

where k is defined by k' and the *relative shears* of pairing vertices in $G_c[X_i] \setminus G_c[X_{i-1}](\mathbf{x})$. $G_1 \setminus G_2$ is determined by removing from G_1 all the edges of G_2 and then the unconnected vertices.

The solution is obtained when we reach the optimum at the end of the path decomposition with the *shear number* $k = S$. The sum of the *relative shears* gives a constant factor $\tau \sim 2n$. Length constraints on turns or loops between two consecutive strands and on themselves imply that the number of assignments to a strand with regard to the other one is bounded by a constant $\lambda \sim \mathcal{O}(1)$. Hence, the complexity can be reduced by 1 in the exponent of N, the constant being then multiplied by a factor of λ. The dynamic programming runs in time and space $\mathcal{O}(nN^{\max_i |X_i \cap X_{i+1}|})$.

Theorem 3.9 gives an upper bound of $\left\lceil \frac{n}{2} \right\rceil + 1$ on the exponent of N, which is strictly smaller than the previous upper bound of $1 + (2n-2)/3$ (see Section 2.7). In standard β-barrels, where σ is the identity permutation Id, we have $|X_k \cap X_{k+1}| = pw(G_c) = 2, \forall k$. The complexity is then $\mathcal{O}(nN^2)$ in time and space.

3.5 About Greek key motifs in β-barrels

Following the standard structure corresponding to the identity permutation, the β-barrels are found more commonly in such a way that the β-strands are paired in an antiparallel manner to each other. Among this, the most popular structures are those containing disjoint Greek key motifs (see Figure 2.6), for which, our approach can efficiently solve the optimization problem.

We study different possible configurations for disjoint Greek key motifs in permutations. For such structures, we can apply the elimination process to the quotient graph of the *n-strand barrel graph* G_c to construct its tree decomposition. The notations mentioned in this section are those of Section 2.7. The regular expression is used to describe the permutation. We consider the alphabet $\Sigma = \{\mathrm{Id}, g_+, g_-\}$, where Id represents the identity motifs, g_+ represents Greek key motifs of form $k(k+3)(k+2)(k+1)$ and g_- represents $(k+2)(k+1)k(k+3)$. A permutation with disjoint Greek key motifs can be written as a word of Σ^*. For example, $14325678 = g_+\mathrm{Id} = \mathrm{Id}g_-\mathrm{Id}$, $14327658 = g_+g_-$.

- For $\sigma \in H_1$:
  - $\sigma = \mathrm{Id}$: G_c is a cycle, thus has treewidth 2. The complexity is then $\mathcal{O}(nN^2)$.
  - $\sigma = (1\ 3) = g_-\mathrm{Id}$: The *n-strand barrel graph* is an outerplanar graph, thus has treewidth 2 (see Figure 3.5) The complexity is then $\mathcal{O}(nN^2)$.

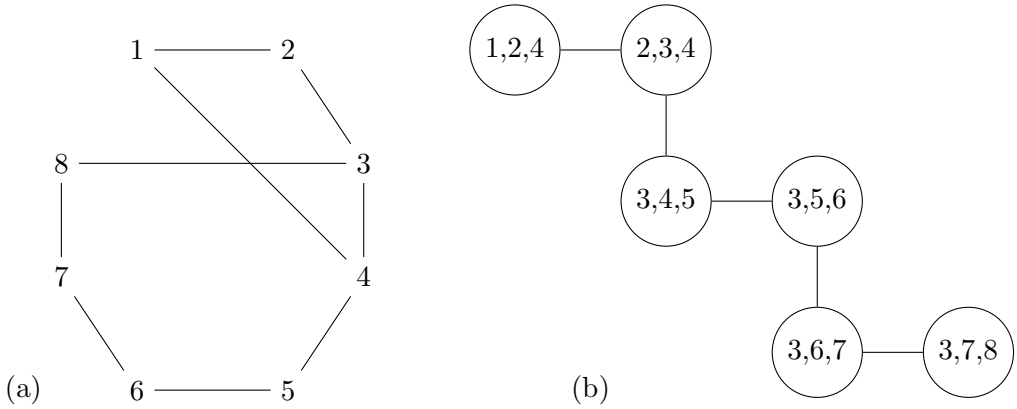


Figure 3.5: G_c(a) and its tree decomposition(b) of $\sigma = 3\ 2\ 1\ 4\ 5\ 6\ 7\ 8$

  - $\sigma = \mathrm{Id}g_-(\mathrm{Id}+g_-)^*+g_-(\mathrm{Id}+g_-)^*g_-(\mathrm{Id}+g_-)^*$: The quotient graph of G_c is an outerplanar graph, in which each module is of size at most 2 and the modules of size 2 are not adjacent. Hence, we can easily construct a tree decomposition of G_c that has width 3 (see Figure 3.6). The complexity is $\mathcal{O}(nN^3)$.

- For $\sigma \in H_2$:
  - $\sigma = \mathrm{Id} : \mathcal{O}(nN^2)$.
  - $\sigma = (2\ 4) = g_+\mathrm{Id}$: G_c has treewidth 3, thus the complexity is $\mathcal{O}(nN^3)$ (see Figure 3.7).
  - $\sigma = \mathrm{Id}g_+ : G_c$ is outerplanar, thus has treewidth 2. The complexity is then $\mathcal{O}(nN^2)$ (see Figure 3.8).
  - $\sigma = g_+\mathrm{Id}g_+ + g_+g_+ + (\mathrm{Id}+g_+)^+g_+(\mathrm{Id}+g_+)^+$: The quotient graph of G_c is also an outerplanar graph, in which each module is of size at most 2 and the modules of size 2 are not adjacent. Hence, the complexity is $\mathcal{O}(nN^3)$ (see Figure 3.9).

- For $\sigma \in H_3$:
  - $\sigma = \mathrm{Id} : \mathcal{O}(nN^2)$.

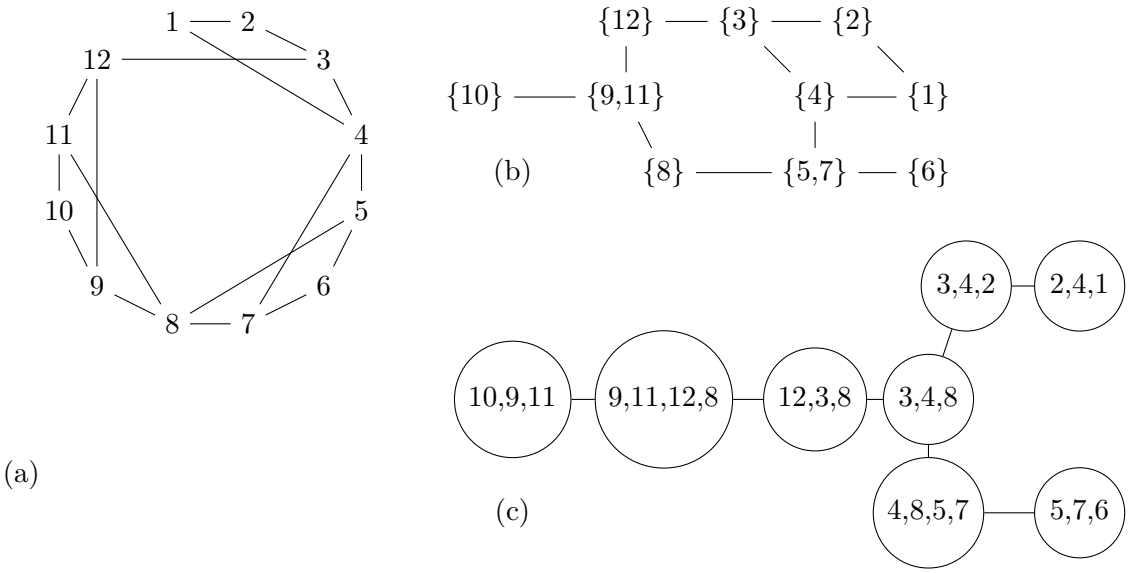


Figure 3.6: G_c(a), its quotient graph(b) and its tree decomposition(c) of $\sigma = 3\ 2\ 1\ 4\ 7\ 6\ 5\ 8\ 11\ 10\ 9\ 12$

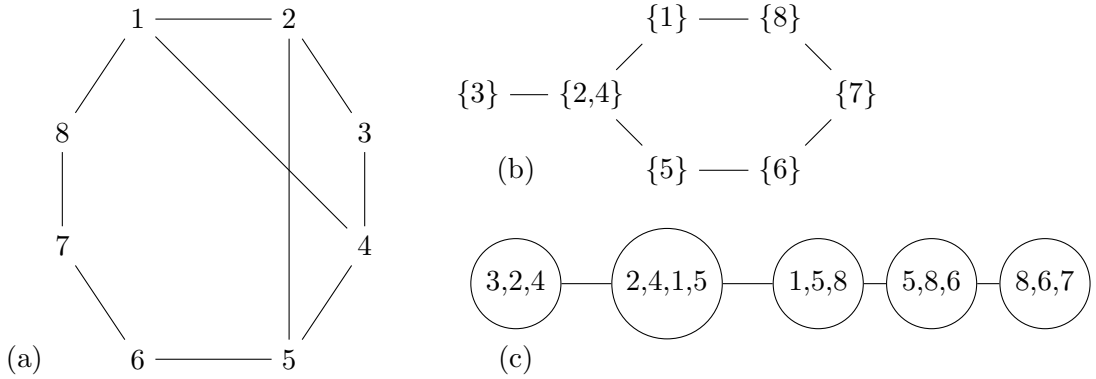


Figure 3.7: G_c(a), its quotient graph(b) and its tree decomposition(c) of $\sigma = 1\ 4\ 3\ 2\ 5\ 6\ 7\ 8$

  - $\sigma = \mathrm{Id}g_+(\mathrm{Id} + g_+)^+$: The quotient graph of G_c is also an outerplanar graph, in which each module is of size at most 2 and the modules of size 2 are not adjacent. Hence, the complexity is $\mathcal{O}(nN^3)$ (see Figure 3.10).

- For $\sigma \in H_4$:
  - $\sigma = \mathrm{Id} : \mathcal{O}(nN^2)$.
  - $\sigma = \mathrm{Id}g_+$: G_c is outerplanar, thus has treewidth 2. The complexity is then

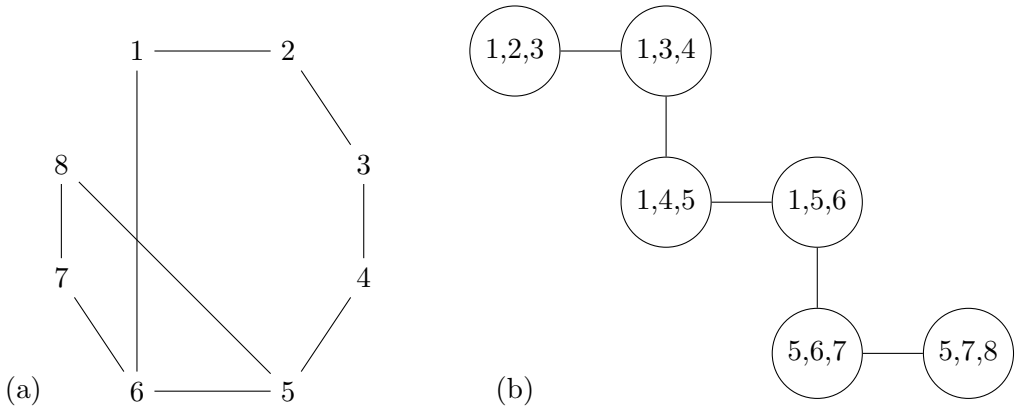


Figure 3.8: G_c(a) and its tree decomposition(b) of $\sigma = 1\ 2\ 3\ 4\ 5\ 8\ 7\ 6$

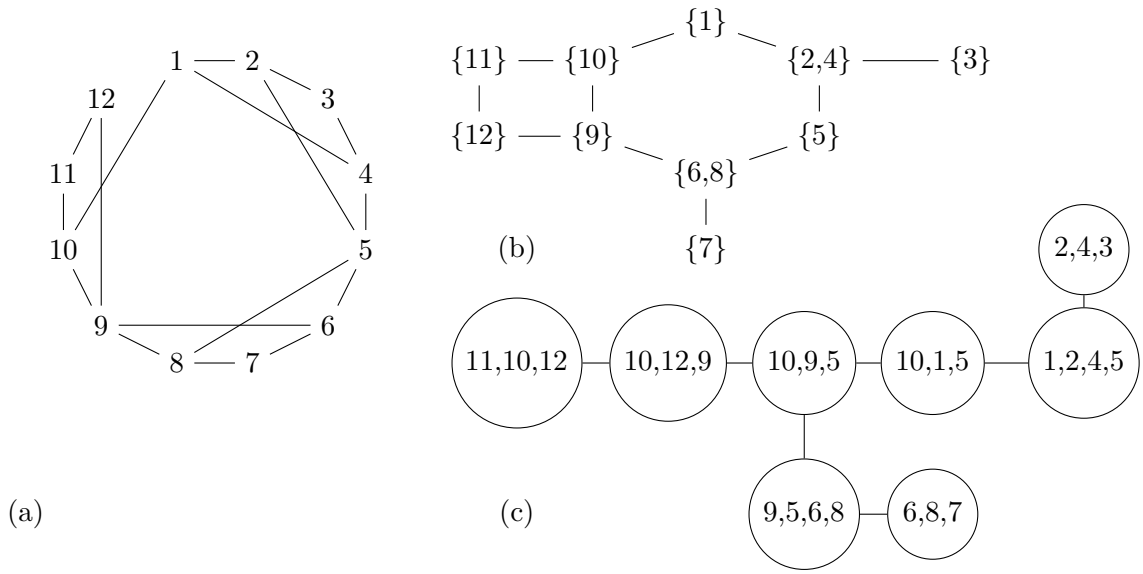


Figure 3.9: G_c(a), its quotient graph(b) and its tree decomposition(c) of $\sigma = 1\ 4\ 3\ 2\ 5\ 8\ 7\ 6\ 9\ 12\ 11\ 10$

$\mathcal{O}(nN^2)$ (see Figure 3.11).

– $\sigma = \mathrm{Id}g_+(\mathrm{Id} + g_+)^+$: Similarly to the case $\sigma \in H_3$, the complexity is $\mathcal{O}(nN^3)$ (see Figure 3.12).

More generally, for a permutation σ that differs from the identity permutation by disjoint Greek key motifs, i.e. $\sigma = (\mathrm{Id} + g_+ + g_-)^+$, the width of the tree decomposition is determined at the motifs g_-g_+ or g_+g_-. The motifs g_-g_- or g_+g_+ can be reduced to

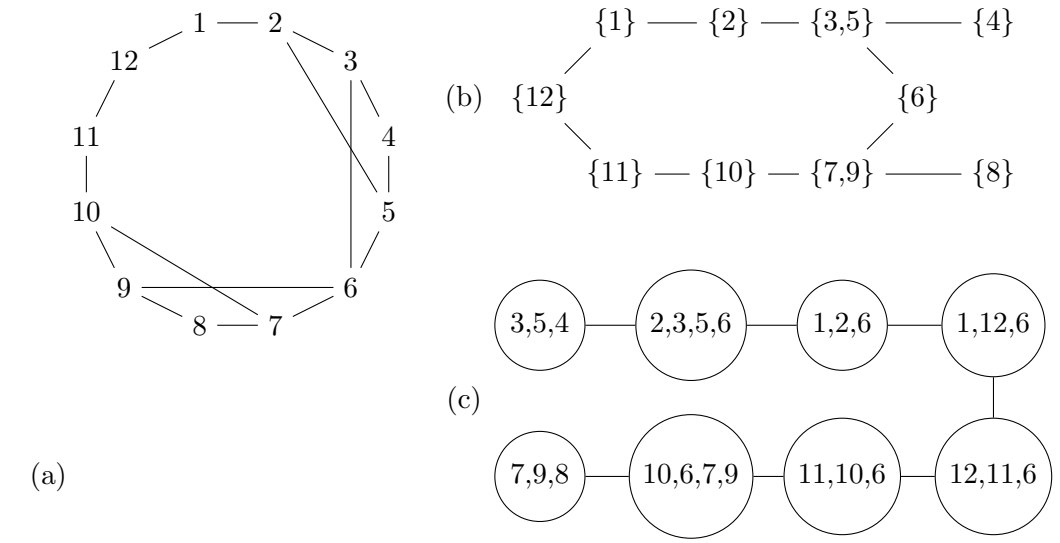


Figure 3.10: G_c(a), its quotient graph(b) and its tree decomposition(c) of σ = 1 2 5 4 3 6 9 8 7 10 11 12

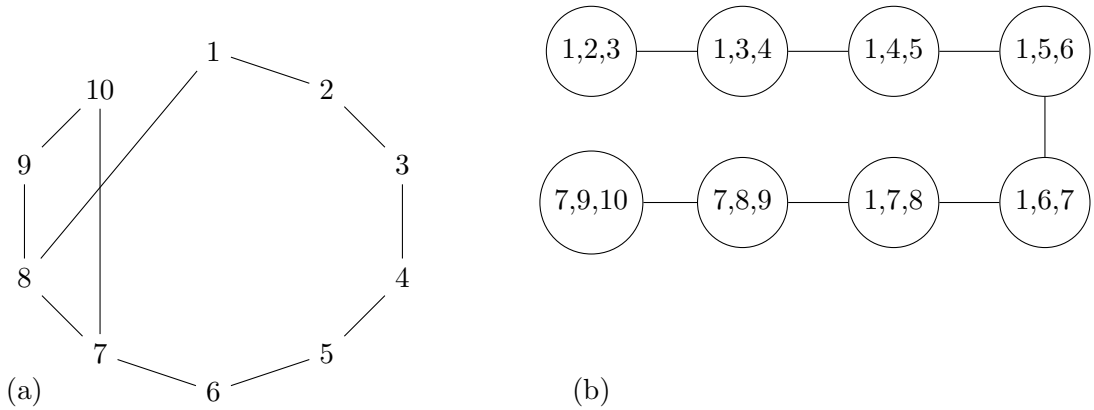


Figure 3.11: G_c(a) and its tree decomposition(b) of σ = 1 2 3 4 5 6 7 10 9 8

the previous cases of H_1, H_2, H_3 and H_4, thus give the width at most 3.

- For motifs g_-g_+, either g_- = 3214 or g_- takes part in a standard complete Greek key motif 14325 [99], the quotient graph of G_c is always an outerplanar graph, in which each module is of size at most 2 and the modules of size 2 are not adjacent. Hence, the complexity is $\mathcal{O}(nN^3)$.
- For motifs g_+g_-, the width of the tree decomposition is $\max\{3, tw(G_{+-})\}$, where $tw(G_{+-})$ is the treewidth of the graph G_{+-} in Figure 3.13. Since all vertices of

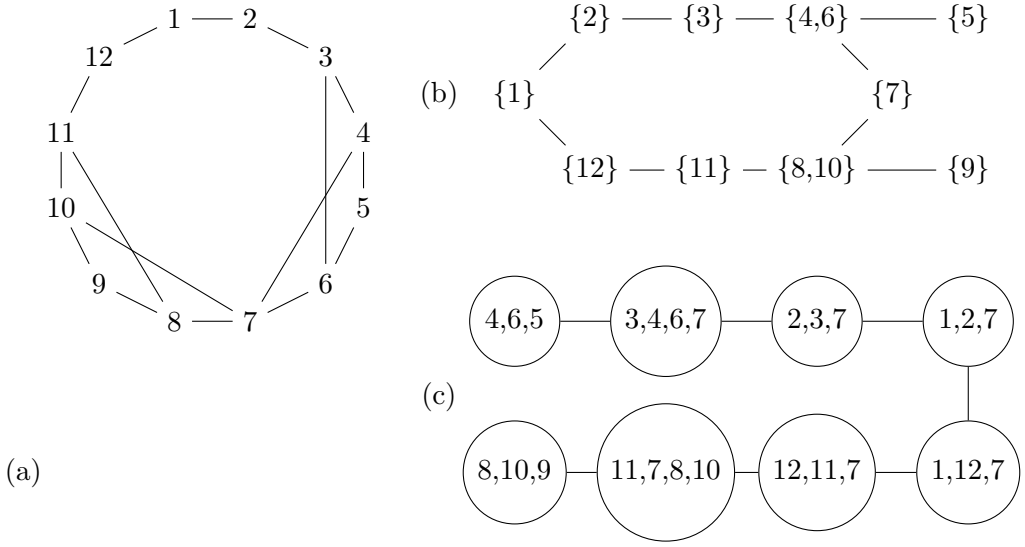


Figure 3.12: G_c(a), its quotient graph(b) and its tree decomposition(c) of $\sigma = 1\ 2\ 3\ 6\ 5\ 4\ 7\ 10\ 9\ 8\ 11\ 12$

G_{+-} have degree 3, $tw(G_{+-}) \geq 3$. We can easily construct a tree decomposition of width 3, thus $tw(G_{+-}) = 3$. Therefore, the complexity is also $\mathcal{O}(nN^3)$.

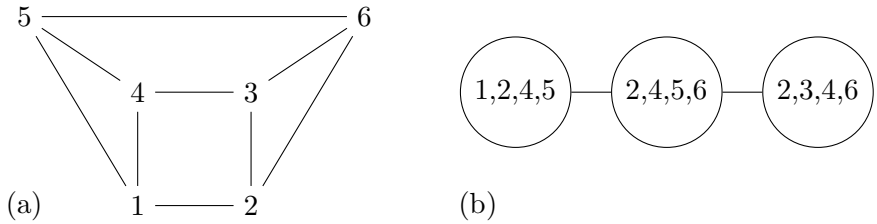


Figure 3.13: The reduced graph G_{+-} for g_+g_-(a) and its tree decomposition of width 3(b)

So, we also have a complexity in time and space $\mathcal{O}(nN^2)$ to $\mathcal{O}(nN^3)$ for this tree decomposition based algorithm for this popular class of structures. The algorithm favorably compares to our previous algorithm in Section 2.7 regarding the complexity in time and space ($\mathcal{O}(nN^2)$ to $\mathcal{O}(nN^4)$).

Chapter 4

Evaluation of performance of BBP (Beta-Barrel Predictor)

4.1 Introduction

Beside the theoretical study, our major focus in this work is to deal with the super-secondary prediction of transmembrane β-barrel proteins. We describe in this chapter the evaluation of the performance of our predictor, namely Beta-Barrel Predictor (BBP), in comparison with other existing approaches [119, 120, 123].

We describe at first several TMB datasets used for the comparison with different softwares. The details of our implementation are discussed in the next section, followed by the concepts and measures that will be used for the assessment. Finally, we present and discuss the evaluation results in the last section.

4.2 Experimental setup

4.2.1 Software

We compare our folding prediction accuracy to TMBpro [103] and TMBETAPRED-RBF [96]. We compare our classification results to Freeman et al. [41], TMBETAPRED-RBF [96], PRED-TMBB [9] and transFold [130]. These are currently state-of-the-art softwares for prediction and discrimination of TMB proteins which perform better than other approaches in literature. The results of these approaches are executed from their web-server.

4.2.2 Datasets

We used TMB proteins from the PDBTM database [125] to train and test our approaches.

- Folding: We used CD-HIT [76] to constrain the redundancy in proteins. A threshold of 40% similarity was applied to reduce the dataset, resulting in 49 sequences (**setPDBTMB40**). We retain only the monomeric barrels, i.e. the sequences that form a unique complete barrel. Thus, **setPDBTMB40** contains 41 sequences 1OH2_Q, 3A2R_X, 3AEH_A, 3BRZ_A, 3CSL_A, 2R4P_A, 3DWO_X, 2FGQ_X, 3EFM_A, 3EMN_X, 2ERV_A, 2IWW_A, 2F1T_A, 1FEP_A, 3FHH_A, 3FID_A, 1ILZ_A, 1BY3_A, 2GSK_A, 1BH3_A, 2HDF_A, 2J1N_A, 2IAH_A, 3JTY_A, 1BXW_A, 2VDF_A, 1PNZ_A, 3GP6_A, 1AF6_A, 3NJT_A, 2O4V_A, 2ODJ_A, 1QJ8_A, 1P4T_A, 2POR_A, 1TLW_A, 1UXF_A, 1UYN_X, 2WJQ_A, 2X4M_A, 1XKW_A. It is important to note that while other learning based methods use the available entire dataset of TMB structures for training, we use these known structures to build a statistical model which only plays the role of a filter to discard the obviously non-putative β-strands and does not take part in our folding algorithm. While this may result in overfitting for a learning-based approach, the effect on our approach should be very small.

  In order to evaluate the performance of BBP, with regard to mutation, geometrical details, we use a subset of **setPDBTMB40**, namely **setECOLI40**, which contain the TMB proteins from Escherichia coli. This choice of TMB structures from a specific species is to make our prediction under the physicochemical properties of the membrane, given that these are not quite varied in the same species. **setECOLI40** contains then 17 sequences: 1AF6_A, 1BXW_A, 1BY3_A, 1FEP_A, 1ILZ_A, 1PNZ_A, 1QJ8_A, 1TLW_A, 2F1T_A, 2GSK_A, 2HDF_A, 2IWW_A, 2J1N_A, 2R4P_A, 2WJQ_A, 3AEH_A, 3GP6_A.

  We also used the two sets of TMB proteins reported in [103]. The first dataset which is described in [130] contains 14 non-redundant TMB proteins with PDB codes of 1A0S, 1E54, 1I78, 1K24, 1PRN, 1QJ8, 1QJP, 2OMF, 2POR, 1QD6, 1P4T, 1AF6, 1THQ and 1TLY. This set will be referred as **setTransFold**. The second dataset described in [8] also contains 14 non-redundant TMB proteins: 1A0S, 1E54, 1I78, 1K24, 1PRN, 1QJ8, 1QJP, 2OMF, 2POR, 1QD5, 1FEP, 2MPR, 1KMO and 2FCP, where the first nine are in common with **setTransFold**. We refer to this second set as **setPREDTMBB**.

- Classification: We used a set of 177 α-helical transmembrane proteins of length from 140 to 800 residues, at 40% redundancy reduction, from PDBTM, that is named **setPDBTMH40** and 32 non-redundant lipocalins taken from PDB (**setLIPOC**). **setPDBTMH40** contains 1AIG_H, 1AIG_L, 1AR1_B, 1BCC_C, 1C17_M, 1C51_B, 1DOP_D, 1ET2_S, 1EYS_M, 1F6G_A, 1FFT_A, 1FFT_B, 1FFT_C, 1FX8_A, 1IZL_A, 1J4N_A, 1JB0_F, 1JB0_L, 1KAD_A, 1KPW_A, 1KQF_B, 1KQF_C, 1L7V_A, 1LBN_A, 1LNQ_A, 1LVI_A, 1M0K_A, 1O5W_A, 1OED_B, 1OED_C, 1ORQ_C, 1OZ5_A, 1P49_A, 1P7B_A, 1PB2_A, 1PB4_C, 1PB4_D, 1PRC_H, 1PW4_A, 1Q90_A, 1QLE_C, 1RH5_A, 1S6E_A, 1SR1_A, 1SUK_A, 1UPE_A, 1XIO_A, 1Y36_A, 1Y8S_A, 1Y9C_A, 1YEW_C, 1YG7_A, 1YO9_L, 1Z8E_A, 1ZAS_A, 1ZC7_A, 1ZCD_A, 1ZTI_A, 2A06_E, 2A0D_A, 2A65_A, 2AC6_A, 2AKH_Y, 2AMK_A, 2AUI_A, 2AXT_B, 2AXT_C, 2B0X_A, 2B2F_A,

PDB	Species	Protein	Chain	Length
1AF6	Escherichia coli	Maltoporin	A	421
1BH3	Rhodobacter blasticus	Porin	A	289
1BXW	Escherichia coli	Outer membrane protein A	A	172
1BY3	Escherichia coli	Ferrichrome-Iron receptor precursor	A	714
1FEP	Escherichia coli	Ferric enterobactin receptor	A	724
1ILZ	Escherichia coli	Outer membrane phospholipase A	A	275
1OH2	Salmonella typhimurium	Sucrose specific porin	Q	413
1P4T	Neisseria meningitidis	Outer membrane protein NspA	A	155
1PNZ	Escherichia coli	Ferric citrate transporter	A	751
1QJ8	Escherichia coli	Outer membrane protein X	A	148
1TLW	Escherichia coli	Nucleoside-specific channel-forming protein Tsx	A	278
1UXF	A. peuropneumoniae	Hemoglobin binding protein HgbA	A	550
1UYN	Neisseria meningitidis	Autotransporter Nalp	X	308
1XKW	Pseudomonas aeruginosa	Fe(III)-pyochelin receptor	A	665
2ERV	Pseudomonas aeruginosa	Outer membrane enzyme PagL	A	150
2F1T	Escherichia coli	Outer membrane protein W	A	197
2FGQ	Delftia acidovorans	Porin outer membrane protein 32	X	332
2GSK	Escherichia coli	Vitamin B12 transporter BtuB	A	590
2HDF	Escherichia coli	Colicin I receptor	A	639
2IAH	Pseudomonas aeruginosa	Ferripyoverdine receptor	A	772
2IWW	Escherichia coli	Outer membrane protein G	A	281
2J1N	Escherichia coli	Outer membrane protein C	A	346
2O4V	Pseudomonas aeruginosa	Porin P	A	411
2ODJ	Pseudomonas aeruginosa	Porin D	A	428
2POR	Rhodobacter capsulatus	Porin	A	301
2R4P	Escherichia coli	Long-chain fatty acid transport protein	A	427
2VDF	Neisseria meningitidis	Outer membrane protein	A	253
2WJQ	Escherichia coli	Outer membrane protein NanC	A	215
2X4M	Yersinia pestis	Coagulase/Fibrinolysin	A	298
3A2R	Neisseria meningitidis	Outer membrane protein II	X	355
3AEH	Escherichia coli	Hemoglobin-binding protease autotransporter	A	308
3BRZ	Pseudomonas putida	Toluene transporter TodX	A	439
3CSL	Serratia marcescens	HasR protein	A	753
3DWO	Pseudomonas aeruginosa	Outer membrane protein	X	451
3EFM	Bordetella pertussis	Ferric alcaligin siderophore receptor	A	707
3EMN	Mus musculus	Voltage-dependent anion-selective channel 1	X	295
3FHH	Shigella dysenteriae	Outer membrane heme receptor ShuA	A	640
3FID	Salmonella typhimurium	Outer membrane protein LpxR	A	296
3GP6	Escherichia coli	Protein pagP	A	163
3JTY	Pseudomonas fluorescens	BenF-like porin	A	402
3NJT	Bordetella pertussis	Filamentous hemagglutinin transporter fhaC	A	566

Table 4.1: Transmembrane β-barrel proteins in **setPDBTMB40**

2B6S_A, 2BG9_A, 2BG9_E, 2BL2_A, 2BMN_A, 2BS2_C, 2C3E_A, 2CFP_A, 2D2C_A, 2D2C_B, 2D2C_D, 2D57_A, 2EVU_A, 2F75_A, 2F93_A, 2F95_B, 2FBW_C, 2FYN_A, 2FYN_B, 2FYN_C, 2G1X_A, 2G2A_A, 2GFP_A, 2GFZ_A, 2H8A_A, 2HE6__, 2HYD_A, 2IC8_A, 2IIL__, 2IK3__, 2IQP_A, 2IUB_A, 2J58_A, 2J7A_C, 2JIZ_G, 2JLN_A, 2K73_A, 2KSE_A, 2NQ2_A, 2NR9_A, 2OAR_A, 2OAU_A, 2Q7M_A, 2QFI_A, 2R6G_F, 2RH1_A, 2VL0_A, 2VPW_C, 2W1P_A, 2WCD_A, 2WIT_A, 2WLH_A, 2WSC_1, 2WSC_2, 2WSC_3, 2WSC_A, 2WSC_F, 2WSC_G, 2WSC_H, 2YVX_A, 2Z73_A, 2ZJS_Y, 2ZW3_A, 3A3Y_B, 3A7K_A, 3ABK_B, 3ABK_D, 3B4R_A, 3B5W_A, 3B9W_A, 3BEH_A, 3BVD_B, 3C02_A, 3C1G_A, 3C9L_A, 3CHX_A, 3CHX_B, 3CN5_A, 3CX5_D, 3D31_C, 3DDL_A, 3DET_A, 3DH4_A, 3DHW_A, 3DTU_A, 3DWW_A, 3EAM_A, 3EGW_C, 3EH3_A, 3FH6_G, 3FWL_A, 3G67_A, 3GI8_C, 3H9V_A, 3HD6_A, 3HFX_A, 3HGC_A, 3HKK_A, 3HQK_A, 3IXZ_B, 3JYC_A, 3K3F_A, 3KBC_A, 3KCU_A, 3KP9_A, 3LLQ_A, 3LNM_B, 3M71_A. **setLIPOC** contains 1AVG_I, 1BEB_A, 1BJ7_A, 1DZK _A, 1E5P_A, 1GT1_A, 1I4U_A, 1JYD_A, 1JZU_A, 1KXO_A, 1LF7_A, 1MUP_A, 1OEJ_A, 1PM1_X, 1QFT_A, 1QWD_A, 1VPR_A, 1X8Q_A, 1XKI _A, 1Y0G_A, 2CM4_A, 2HZQ _A, 2RA6_A, 2WEW_A, 2WWP_A, 3BRN_A, 3BS2_A, 3CQN_A, 3DSZ_A, 3KQ0_A, 3L4R_A.

4.3 Implementation details

The number of strands n and the shear number S determine the geometry of the barrel, particularly the membrane spanning part of the segments, and are thus involved in the computation of the energy terms. If they are known, the algorithm can enforce these values and fold the protein accordingly. The values for n, which are usually even, are governed by the consideration on the length of the sequence, the thickness of membrane and the length of turns or loops and vary between 8 and 22 [117]. The values for S are usually even and comprised between n and $2n$ [82, 91, 92]. The problem is then solved by the constraint dynamic programming with the constraints of given n and S. A small number of couples (n, S) have to be explored and our algorithm is fast enough for that.

Side-chain interactions between contiguous residues along a segment on the same side and interactions with the environment of channel or bilayer define the intrinsic energy of the corresponding vertex. The pairing energy of two adjacent segments in the barrel is computed by optimizing the relative positions between the constitutive amino acids. These energies involve hydrogen bonds in main chains, electrostatic interactions between side-chains, hydrophobic effect as well as environmental effect. More specifically, the extracellular and intracellular environments with distinct hydrophobicity indices can have significantly different hydrophobic effects. In addition, the membrane thickness gives constraints on the segment size and helps identify the interactions inside or outside the membrane region. We use here by default a parameter of 3nm for the membrane thickness, thus making it about 8 residues thick [75, 104]. The features on size, polarity [48], and flexibility [15] of turns and loops are also taken into consideration, i.e. turns and loops satisfy threshold constraints on their polarity and flexibility indices and their length.

PDB	Species	Protein	Chain	Length
1AVG	Triatoma pallidipennis	Triabin	I	142
1BEB	Bos taurus	Beta-lactoglobulin	A	162
1BJ7	Bos taurus	Allergen Bos d 2	A	156
1DZK	Sus scrofa	Odorant-binding protein	A	157
1E5P	Mesocricetus auratus	Aphrodisin	A	151
1GT1	Bos taurus	Odorant-binding protein	A	159
1I4U	Synthetic construct	alpha-crustacyanin	A	181
1JYD	Homo sapiens	Plasma retinol-binding protein	A	183
1JZU	Coturnix coturnix	Q83	A	157
1KXO	Pieris brassicae	DigA16	A	184
1LF7	Homo sapiens	Complement Protein C8gamma	A	182
1MUP	Mus musculus	Major urinary protein	A	166
1OEJ	Escherichia coli	Yoda	A	193
1PM1	Rhodnius prolixus	Nitrophorin 2	X	180
1QFT	Rhipicephalus appendiculatus	Female-specific histamine binding protein 2	A	175
1QWD	Escherichia coli	Lipoprotein blc	A	177
1VPR	Lingulodinium polyedrum	Luciferase	A	374
1X8Q	Rhodnius prolixus	Nitrophorin 4	A	184
1XKI	Homo sapiens	Von Ebner's gland protein	A	162
1Y0G	Escherichia coli	YceI	A	191
2CM4	Ornithodoros moubata	Complement inhibitor OmCI	A	150
2HZQ	Homo sapiens	Apolipoprotein D	A	174
2RA6	Trichosurus vulpecula	Trichosurin	A	166
2WEW	Homo sapiens	Apolipoprotein M	A	172
2WWP	Homo sapiens	Prostaglandin-H2 D-Isomerase	A	176
3BRN	Argas monolakensis	Amine-binding protein	A	157
3BS2	Argas monolakensis	Monomine	A	148
3CQN	Arabidopsis thaliana	Violaxanthin de-epoxidase	A	185
3DSZ	Homo sapiens	Engineered human lipocalin 2	A	186
3FIQ	Rattus norvegicus	Odorant-binding protein 1F	A	157
3KQ0	Homo sapiens	Alpha-1-acid glycoprotein 1	A	192
3L4R	Canis familiaris	Minor allergen Can f 2	A	170

Table 4.2: β-barrel proteins in **setLIPOC**

Their energies are approximated by hydrophobicity, using Kyte-Doolittle scale [72].

We use the Dunbrack backbone-dependent rotamer library [35] and the partial charges from GROMOS force field [126] to compute pairwise interaction energies. The hydrophobic interaction between two side-chains u, v is assessed by the amount of contacts between non-polar groups, calculated by taking the average on all rotamer pairs of the two side-chains $e_{uv} =< e_{uv|rotamers} >$. Each side-chain plays the role of a group of partial charges in the electrostatic interaction. The main-chain hydrogen bond is measured by the electrostatic potential energy between peptide C=O and N−H groups.

The probabilistic model and the constraints on hydrophobicity help discard the unlikely membrane spanning β-strands (see Chapter 2). A threshold on overall energy can also be involved to enhance the discrimination. We studied the per-strand energy value for a variety of TMB proteins including the training dataset and other TMB proteins. Even though this value is always higher than 0.9 for these proteins, we chose 0.85 as a threshold to avoid overfitting. Note that this does not affect the prediction results, and is only used for discrimination.

4.4 Method of evaluation

4.4.1 Concepts on predicted secondary structures

We first introduce the notions of *secondary structure assignment* and *elementary secondary structure*. These are followed by the concepts of *overlap of secondary structures*, on which we define a *correctly predicted elementary secondary structure* and a *correctly predicted structure*. The concepts are inspired from Waldispühl's PhD thesis [129] with modifications according to our context.

Notion 4.1. Secondary structure assignment

A secondary structure assignment *of an amino acid sequence $\mathcal{S}$ is a sequence of designations of a secondary structure type (α, β or turn/loop) to residues of $\mathcal{S}$. Particularly, given an alphabet $\Sigma = \{\mathrm{S}, -\}$, as β-barrel structures is the main target of our work, the secondary structure assignment can be described as a word of Σ^* with the same length to $\mathcal{S}$.* S *corresponds to a residue belonging to a membrane spanning β-structure, $-$ to other structures in non-membrane regions.*

Notion 4.2. Elementary secondary structure

Let Γ be a secondary structure assignment. We call elementary secondary structure *a maximal segment of consecutive residues that belongs to the same kind of secondary structure (*S *or $-$).*

Example 4.1.

An N-terminal subsequence of protein OmpX (1QJ8) aligned with its secondary structure assignment:

```
         10        20        30        40        50        60        70
         |         |         |         |         |         |         |
ATSTVTGGYAQSDAQGQMNKMGGFNLKYRYEEDNSPLGVIGSFTYTEKSRTASSGDYNKNQYYGITAGPAYR
---SSSSSSSSS------SSSSSSSSSS--------SSSSSSSS-------------SSSSSSSSS--
```

It comprises 4 elementary secondary structures. The first one is a membrane-spanning strand stretching from residue 4 to residue 13, the second strand contains residues from 20 to 30, the third one from 39 to 47 and the fourth one from 61 to 70. ◁

Definition 4.1. Overlap of secondary structures

Given an alignment of two secondary structure assignments Γ_1 *and* Γ_2 *of an amino acid sequence. Let* E_1^i *and* E_2^j *be two elementary secondary structures in* Γ_1 *and* Γ_2*, respectively. We say that* E_1^i *and* E_2^j overlap *each other if and only if the two corresponding* β*-strands have at least 4 common residues.*

Definition 4.2. Correctly predicted elementary secondary structure

Given an alignment of two secondary structure assignments Γ_{obs} *and* Γ_{pred} *of an amino acid sequence that correspond to the experimentally observed structure and the predicted structure, respectively. Let* E_{obs}^i *and* E_{pred}^j *be two elementary secondary structures in* Γ_{obs} *and* Γ_{pred}*, respectively. We say that the elementary secondary structure* E_{obs}^i *is* correctly predicted *by* E_{pred}^j *if and only if* E_{obs}^i *overlaps* E_{pred}^j *and only* E_{pred}^j*, and reversely,* E_{pred}^j *overlaps* E_{obs}^i *and only* E_{obs}^i*.*

Example 4.2.

An alignment of two secondary structure assignments corresponding to an observed structure (the first line) and a predicted structure (the second line):

```
         10        20        30        40        50        60        70
         |         |         |         |         |         |         |
---SSSSSSSSSS---------SSSSSSSSSS----SSSSSSSSSS-----SSSSSSSSSS---------
----SSSSSSSSSSS-----------SSSSSSSSSSSSS----SSSSSSSS------SSSSSSSSS---
```

The first observed elementary secondary structure (the membrane-spanning strand from residue 4 to 13) only overlaps the first predicted strand (from residue 5 to 16) and reversely. The second predicted strand overlaps both the second and third observed strands while the third predicted strand does not overlap any observed strand. The fourth predicted and observed strands overlap each other. Hence, only the first and the fourth elementary secondary structures are predicted. ◁

Definition 4.3. Correctly predicted structure

Given an alignment of two secondary structure assignments Γ_{obs} *and* Γ_{pred} *of an amino acid sequence* $\mathcal{S}$ *that correspond to the experimentally observed structure and the predicted structure, respectively. The structure of protein* $\mathcal{S}$ *is said to be* correctly predicted *if and only if every observed elementary secondary structure overlaps one and only one predicted elementary secondary structure and, reversely, if every predicted elementary secondary structure overlaps one and only one observed elementary secondary structure.*

Example 4.3. Two alignments of secondary structure assignment, in which the first line corresponds to the observed structure and the second line is the predicted structure.

• Correctly predicted structure:

```
         10        20        30        40        50        60        70
         |         |         |         |         |         |         |
---SSSSSSSSSS---------SSSSSSSSSS----SSSSSSSSSS-----SSSSSSSSSS---------
----SSSSSSSSSSS----SSSSSSSSSSSS-----SSSSSSS-------------SSSSSSSSS---
```

• Non-correctly predicted structure:

```
         10        20        30        40        50        60        70
         |         |         |         |         |         |         |
---SSSSSSSSSS---------SSSSSSSSSS----SSSSSSSSSS-----SSSSSSSSSS---------
----SSSSSSSSSSS-----------SSSSSSSSSSSSS----SSSSSSSS------SSSSSSSSS---
```

◁

4.4.2 Measures of performance

We have just presented the notions that allow to evaluate the quality of super-secondary structure prediction. The prediction can be considered as a binary classification, which is to classify a set of objects into two different classes. We now describe the measures for the performance of the our prediction and other approaches, starting with the fundamental measures of *true positive* (also known as *hit*), *true negative* (or *correct rejection*), *false positive* (or *false alarm*), *false negative* (or *miss*), which are the four different possible outcomes of a binary classification. For two classes, let's say "**positive**" and "**negative**", or "**yes**" and "**no**", *true positive* (TP) is a correct classification of an object into "positive" class, *true negative* (TN) is a correct classification into "negative" class. *False positive* (FP) is when an object is incorrectly classified as "positive", and *false negative* (FN) is when it is incorrectly classified as "negative". In this work, we consider two classes of membrane spanning β-strands (S) and non-membrane region ($-$). Without confusion, we also use these two notations to mention the two classes. The measures are defined *on residues* as well as *on segments*, in order to evaluate not only the capacity to assign some sort of secondary structure to residues, but also to recognize membrane-spanning segments.

On residues:

- TP = number of residues S which are predicted in S.
- TN = number of residues $-$ which are predicted in $-$.
- FP = number of residues $-$ which are predicted in S.
- FN = number of residues S which are predicted in $-$.

On segments:

- TP = number of elementary secondary structures S which are correctly predicted.
- TN = number of elementary secondary structures − which are correctly predicted.
- FP = number of elementary secondary structures − which are not correctly predicted.
- FN = number of elementary secondary structures S which are not correctly predicted.

These four outcomes can represented in a *contingency table* (also known as *confusion matrix*), as follows:

		prediction outcome	
		S	−
actual value	S	*True Positive*	*False Negative*
	−	*False Positive*	*True Negative*

Based on these basic quantities, the principal measures of the performance of a binary classifier are defined:

- *Sensitivity* (or *true positive rate*, *recall*) is the proportion of actual positive objects which are correctly identified, i.e. the percentage of residues S (or elementary secondary structures S) which are correctly predicted.

$$Sensitivity = \frac{TP}{TP + FN}$$

i.e.

$$Sensitivity = \frac{\text{number of residues S correctly predicted}}{\text{number of residues observed in S}}$$

or

$$Sensitivity = \frac{\text{number of elementary secondary structures S correctly predicted}}{\text{number of elementary secondary structures observed in S}}$$

- *Specificity* (or *true negative rate*) is the proportion of actual negative objects which are correctly identified, i.e. the percentage of residues − (or elementary secondary structures −) which are correctly predicted.

$$Specificity = \frac{TN}{TN + FP}$$

- *Positive predictive value* (PPV or *precision*) measures the proportion of objects with positive prediction results which are correctly predicted, i.e. the percentage of residues S (or elementary secondary structures S) among all residues (or elementary secondary structures) that are predicted in S.

$$PPV = \frac{TP}{TP + FP}$$

i.e.

$$PPV = \frac{\text{number of residues S correctly predicted}}{\text{number of residues predicted in S}}$$

or

$$PPV = \frac{\text{number of elementary secondary structures S correctly predicted}}{\text{number of elementary secondary structures predicted in S}}$$

- *F-score* [127] is a measure of accuracy of the prediction. It is the harmonic mean of the *recall* and the *precision*. The *F-score* has a value between 0 and 1. The prediction is ideal when the *F-score* reaches 1.

$$\textit{F-score} = 2 \times \frac{recall \times precision}{recall + precision}$$

- Matthews correlation coefficient (MCC) [88] is also a measure of quality of the binary classification. This measure takes into account all the four outcomes of true positive, true negative, false positive and false negative. It can be considered as a correlation coefficient between the observed and predicted secondary structures. Its value is included between -1 and $+1$. An MCC value of $+1$ ensures a perfect prediction, while -1 represents an inverse prediction. When MCC is 0, the prediction shows an average random. The MCC is calculated using the formula:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

We also use the score Q_2 which evaluates the proportion of correctly predicted residues over the whole sequence [129] for the measure of prediction performance:

$$Q_2 = 100\% \times \frac{\text{number of correctly predicted residues}}{\text{number of residues}}$$

4.5 Experimental results

BBP can execute the prediction rapidly. The results reported here were obtained through an Intel Pentium IV 3.2-GHz processor with 4 GB of memory.

4.5.1 Folding

On setTransFold with *transFold* and *TMBpro*

The evaluation results in comparison with *transFold* and *TMBpro-SS* in Tables 4.3 show that our method outperforms *transFold*, which is based on pseudo-energy minimization, and is equivalent to *TMBpro-SS* which is based on 1D recursive neural network using alignment profiles.

Method	Q_2	MCC	$Sensitivity$	PPV
transFold	69.9	0.38	94.9	85.2
TMBpro-SS	77.8	0.54	97.2	88.2
BBP	79.1	0.56	96.5	92.6

Table 4.3: Comparison of prediction accuracy on **setTransFold**. Q_2 and MCC are measures on residues. $Sensitivity$ and PPV are measures on β-strands.

On setPREDTMBB with *PRED-TMBB* and *TMBpro*

Method	Q_2	MCC	TP	FP	FN	TOP
PRED-TMBB	84.2	0.72	203	13	11	8
TMBpro-SS	88.3	0.75	204	6	10	11
BBP	79.0	0.57	199	21	15	11

Table 4.4: Comparison of prediction accuracy on **setPREDTMBB**. Q_2 and MCC are measures on residues. TP, FP, TN are measures on β-strands. TOP is the number of proteins with correctly predicted topology, i.e. the proteins with correctly predicted number of β-strands.

In **setPREDTMBB** with the bigger barrels, our method performs worse considering the residues, but gives as good results as the others with regard to the topology. We point out the fact that, in our work, the probabilistic model only plays the role of a filter for potential β-strands, but does not take part in the pseudo-energy function. Furthermore, our method is fairly independent of the learning set. The refinements we are carrying out on structural constraints, hydrophobicity may help to improve the prediction accuracy. Our scores Q_2 and MCC are equivalent for the two datasets while there are deviations in TMBpro-SS's score which might come from their two different training sets.

On *setPDBTMB40* with TMBpro

The folding prediction results are presented in Table 4.5 and Figure 4.1. Figure 4.1 plots the Matthews Correlation Coefficient for our approach BBP and TMBpro for different

proteins along the x-axis. The results of our approach are comparable to those of TMBpro but more consistent as we do not rely on training for folding. We note that in the cases the program predicts an optimal structure with a wrong number of strands, the optimal energy is really close to the energy of the topologically right structure.

(a) Residues

Method	Q_2	$Specificity$	$Sensitivity$	$F-score$	MCC
TMBpro	$81.2 \pm 6.1^*$	79.3 ± 7.9	84.2 ± 11.2	0.76 ± 0.1	0.61 ± 0.14
BBP	79.2 ± 5.4	78.4 ± 6.3	80.4 ± 9.9	0.74 ± 0.1	0.57 ± 0.12

(b) Segments

Method	$Specificity$	$Sensitivity$	$F-score$	MCC
TMBpro	$90.1 \pm 15.0^*$	94.2 ± 12.5	0.93 ± 0.12	0.85 ± 0.26
BBP	91.4 ± 12.0	91.4 ± 11.3	0.92 ± 0.11	0.83 ± 0.22

* Standard Deviation

Table 4.5: Comparison of prediction accuracy on **setPDBTMB40**

On *setPDBTMB40* with TMBETAPRED-RBF

The TMBETAPRED-RBF web-server predicted non-TMB for 24 over 41 proteins of **setPDBTMB40**, or 58.5%. The structures for correctly identified proteins were completely accurate. This might be because they were included in the training set.

4.5.2 Evaluation of the shear numbers

We studied the energy distribution of 17 TMB structures in Escherichia coli taken from **setPDBTMB40** (**setECOLI40**: 1AF6_A, 1BXW_A, 1BY3_A, 1FEP_A, 1ILZ_A, 1PNZ_A, 1QJ8_A, 1TLW_A, 2F1T_A, 2GSK_A, 2HDF_A, 2IWW_A, 2J1N_A, 2R4P_A, 2WJQ_A, 3AEH_A, 3GP6_A) with regard to the slant angle, hence the shear number (see Figure 4.2). Most optimal structures incline with an angle of $41° - 49°$, as observed in databases. This suggests that our model takes well into account the physicochemical properties of TMB structures. It should be also noted that there is no natural way to define the shear number *a priori*.

4.5.3 Influence of the filtering threshold

We applied the filtering thresholds $\rho = \frac{1}{3}, \frac{1}{2}$ and $\frac{2}{3}$ on **setECOLI40**. These thresholds ensure that on average, considering 3-residue blocks as subunits, each segment is accepted as a β-strand if its propensity to be β-strand is at most $3, 2, 1.5$ times, respectively, less than its propensity to be other structure (α-helices or turns/loops). The observed minor

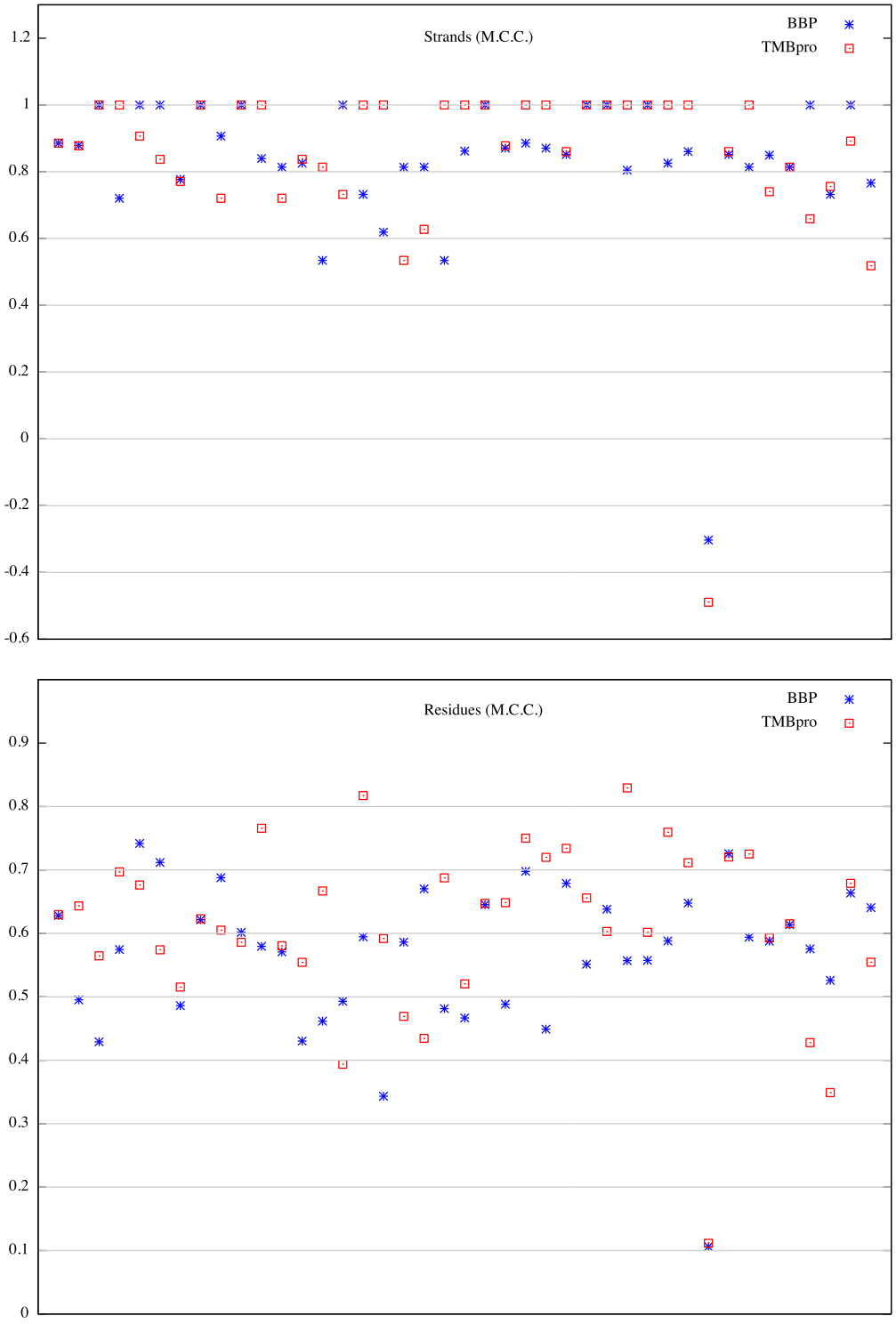


Figure 4.1: Comparison of BBP and TMBpro on structure prediction results.

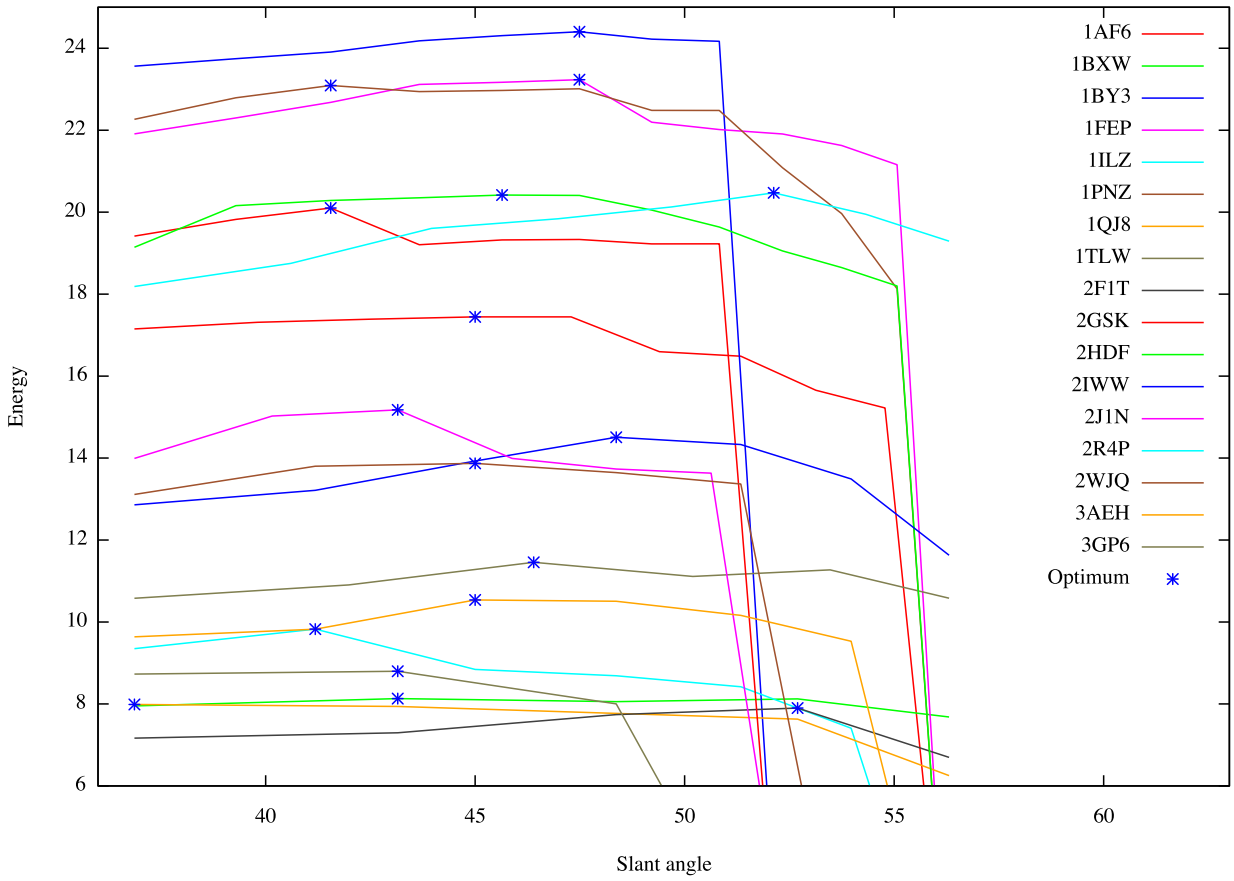


Figure 4.2: Energy distribution of **setECOLI40**, $\theta = \arctan \frac{hS}{dn}$

PDB	Protein	n	S	Angle	Length
1AF6	Maltoporin	18	24	45.0	421
1BXW	Outer membrane protein	8	10	43.2	172
1BY3	Ferrichrome-Iron receptor precursor	22	32	47.5	714
1FEP	Ferric enterobactin receptor	22	32	47.5	724
1ILZ	Outer membrane phospholipase	12	14	41.2	275
1PNZ	Ferric citrate transporter	22	26	41.6	751
1QJ8	Outer membrane protein X	8	8	36.9	148
1TLW	Nucleoside-specific channel-forming protein Tsx	10	14	46.4	278
2F1T	Outer membrane protein W	8	14	52.7	197
2GSK	Vitamin B12 transporter BtuB	22	26	41.6	590
2HDF	Colicin I receptor	22	30	45.6	639
2IWW	Outer membrane protein G	12	18	48.4	281
2J1N	Outer membrane protein C	16	20	43.2	346
2R4P	Long-chain fatty acid transport protein	14	24	52.1	427
2WJQ	Outer membrane protein NanC	12	16	45.0	215
3AEH	Hemoglobin-binding protease autotransporter	12	16	45.0	308
3GP6	Protein pagP	8	10	43.1	163

Table 4.6: Predicted optimal structures of transmembrane β-barrel proteins in **setECOLI40**. n is the number of β-strands, S is the shear number, the slant angles are expressed in degrees.

difference in accuracy with such considerably distinguished thresholds reinforces the fair independence of our approach from the training data. The results in Table 4.7 show the strong predicting ability of BBP from a poor known database. The lower the parameter ρ, the more independent to the training the predictor. This reduced the prediction performance of the model on the known structures, however, it may be useful to discover new TMB proteins.

4.5.4 Evaluation on mutated sequences

We generate the mutated sequences from **setECOLI40** by substituting the amino acids at turns or loops using the PAM250 substitution matrix [33]. Each sequence in **setECOLI40** is mutated up to 5% of amino acids into 10 new sequences. Figures 4.3 and 4.4 show the Matthews Correlation Coefficient and F-score for residues and β-strands. We observe from these results the stability of our predictions. It also suggests that the TMB proteins are stable against these mutations at their turns and loops. The difference in structures of those mutated proteins may merely come from the shift of membrane spanning β-strands when their two extremities are mutated.

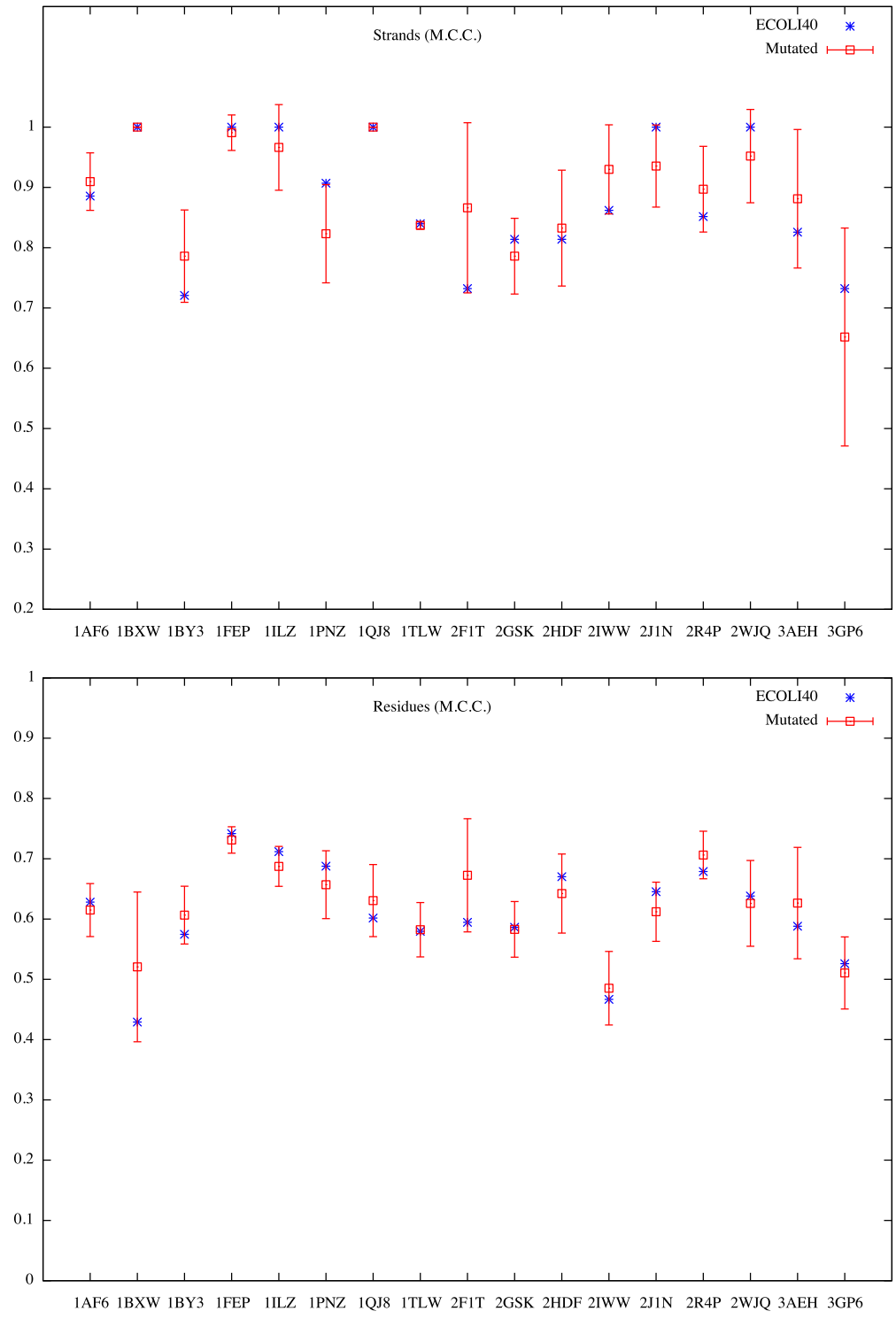


Figure 4.3: MCC of mutated **setECOLI40**

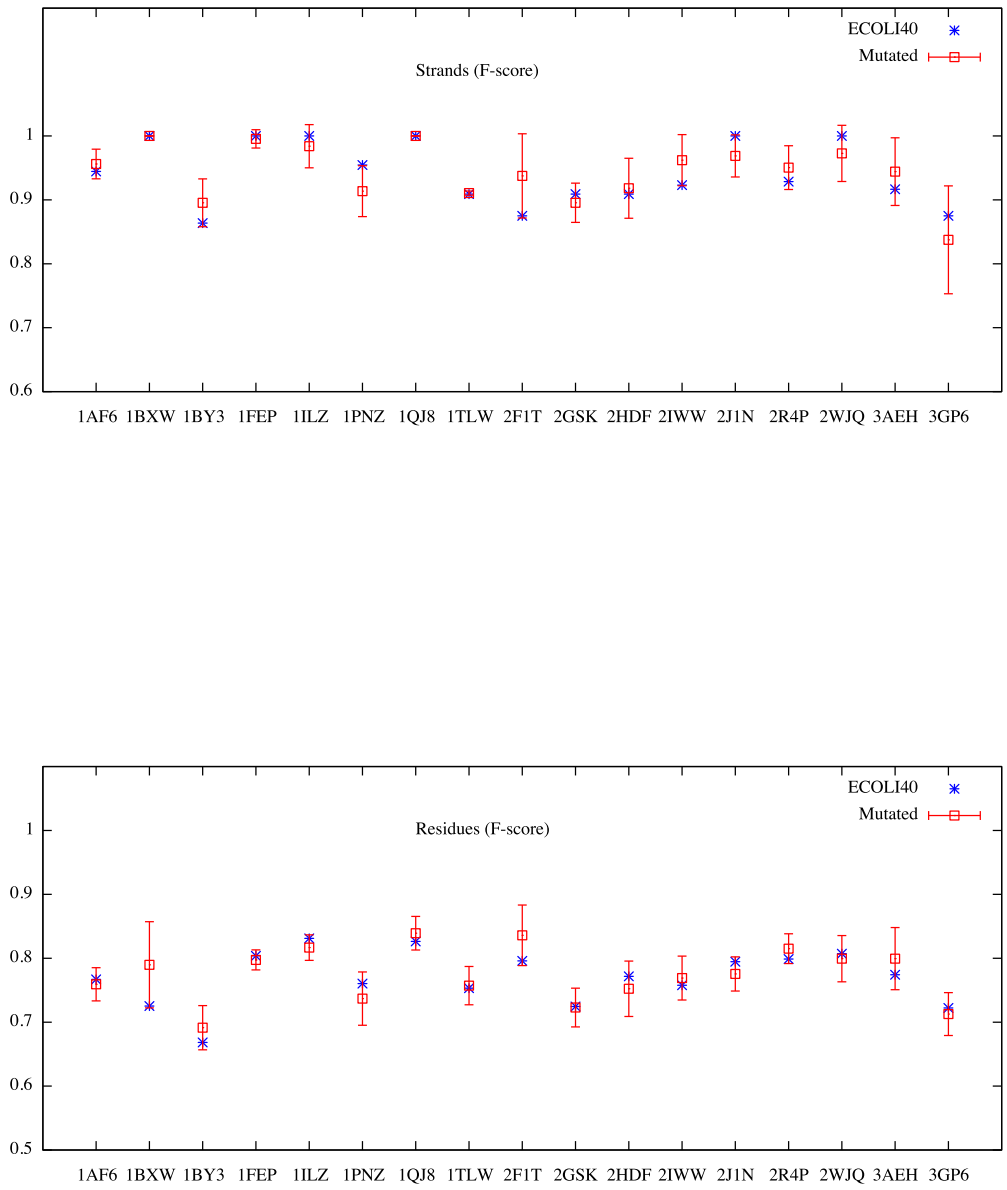


Figure 4.4: F-score of mutated **setECOLI40**

(a) Residues

ρ	Q_2	Specificity	Sensitivity	F-score	MCC
2/3	$80.9 \pm 4.8^*$	80.4 ± 5.2	82.7 ± 8.4	0.77 ± 0.04	0.61 ± 0.08
1/2	79.7 ± 6.0	78.5 ± 5.1	82.4 ± 8.6	0.76 ± 0.05	0.58 ± 0.11
1/3	77.7 ± 5.6	75.6 ± 6.5	81.1 ± 8.6	0.74 ± 0.05	0.55 ± 0.11

(b) Strands

ρ	Specificity	Sensitivity	F-score	MCC
2/3	94.8 ± 5.7	93.3 ± 5.9	0.94 ± 0.05	0.88 ± 0.1
1/2	96.1 ± 4.8	95.4 ± 5.3	0.96 ± 0.05	0.91 ± 0.09
1/3	91.7 ± 9.2	94.9 ± 6.5	0.94 ± 0.07	0.87 ± 0.07

* Standard Deviation

Table 4.7: Comparison of prediction accuracy on **setECOLI40** with different thresholds

4.5.5 Permuted structures

For 3L48, the C-terminal domain of the PapC usher in E. coli, the observed structure topology containing a Greek key motif corresponds to the permutation $\sigma = (1, 4, 3, 2, 5, 6, 7)$ and is predicted with an accuracy (Q_2) of 70.2% at $\rho = 0.2$.

Following the experimental observations that were published previously on the efficiency of the *in vivo* membrane assembly of OmpA variants [71], we tested our algorithm with different given permutations. OmpA (1BXW) consists of eight β-strands, thus without feasibility being taken into account, there are (8-1)! = 5040 circular permutations to check (see Figure 4.5). The pseudo-energy 10.21 of the observed permutation is found in the lowest energy zone. 41 permuted structures, or 0.81%, reach an energy of (10.21 ± 0.3). A ratio of about 1.31% is found in the case of OmpX 1QJ8 (see Figure 4.6). These results are not surprising since a protein may be folded into more than one spatial conformation. In both cases, a Poisson-like distribution is found. This observation may help to discriminate most of infeasible conformations with the use of a threshold on the global energy. Hence, the method is expected to rapidly find a small set containing the right structure within a threshold of, for instance, 2% from the lowest energy and with structural feasibility conditions on permutations. This set might be much smaller be refining the biologically plausible permutations. Other proposed solutions in this set may be the candidates for *in vivo* and *in vitro* studies.

4.5.6 Classification

100% of the non-redundant set of 177 α-helical transmembrane proteins of length from 140 to 800 residues in **setPDBTMH40** are rejected, whereas 31 out of 32 non-redundant lipocalins taken from PDB are predicted as non-TMB. Though lipocalins are also β-

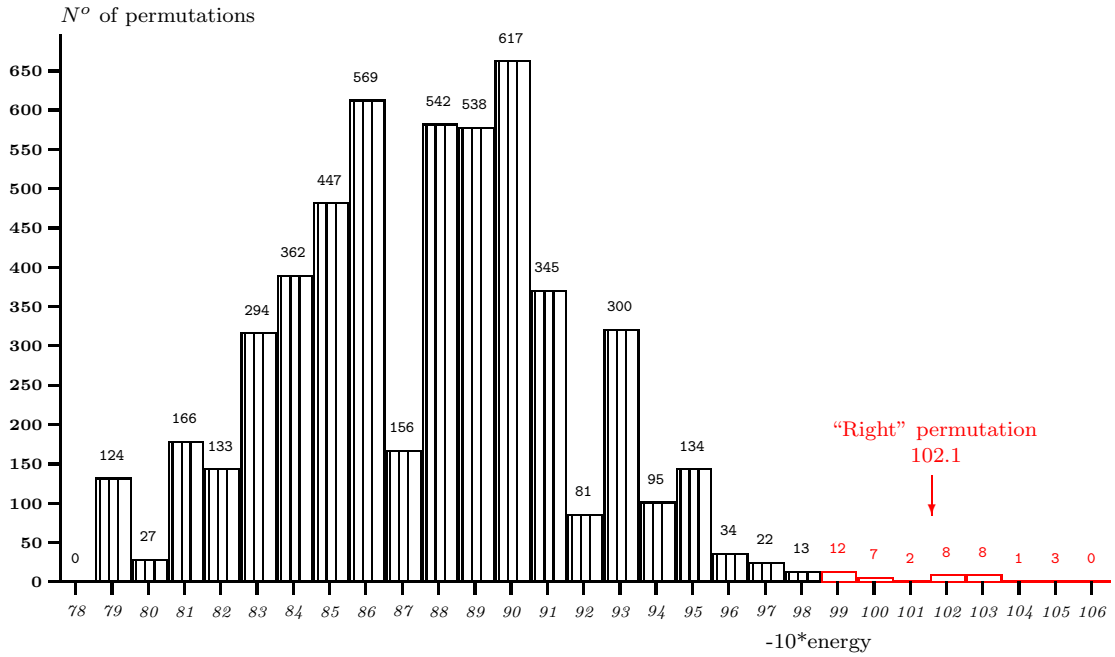


Figure 4.5: Distribution of 7! permutations on E. Coli OmpA 1BXW 8-strand barrel

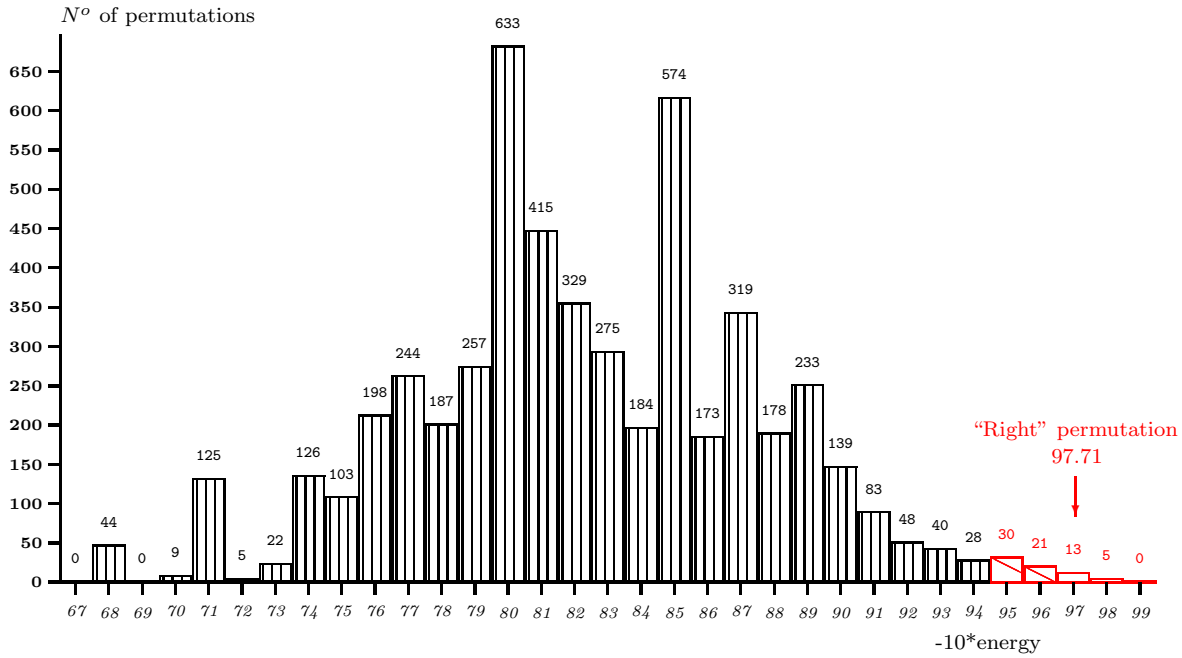


Figure 4.6: Distribution of 7! permutations on E. Coli OmpX 1QJ8 8-strand barrel

barrels which reverse the TMB pattern with a hydrophobic core, the environmental effects on both sides of the barrel are still different. Our pseudo-energy model yields unfavorably on such structures and discriminates considerably better than the learning-based methods like Freeman-Wimley [41], TMBpro [103], PRED-TMBB [9] and TMBETAPRED-RBF [96], but also of transFold [130].

Conclusion and perspectives

Conclusion

We have presented, in this thesis, a new pseudo-energy minimization method for the classification and prediction of transmembrane protein super-secondary structure based on a variety of potential structures. Our approach takes into account many physicochemical constraints and minimizes the global free energy of the structure. It also accounts for permuted structures, thus giving more complete information on the folded structure. Our method is quite accurate with more than 90% sensitivity and F-score, over 80% M.C.C. score on strands; and over 74% accuracy and F-score on residues. The results are comparable to those given by the best currently known approaches, which are based on learning. Moreover, our results are more consistent and have a significantly less variation across different TMB proteins. This is especially interesting given that our algorithm is based mainly on pseudo-energy minimizations, and the probabilistic model only plays a small role. While the model presented here is only for TMB proteins, it can be easily extended to accommodate α-helical bundles. We did not use a more sophisticated statistical model for classifying β-barrel strands because that would risk overfitting and reliance on the training dataset. It is also interesting to note that our approach performs very well for identification of TMB proteins, rejecting all the α-helical bundles and most of the globular β-barrels. Our approach provides the best overall classification results amongst the methods that try to predict structures. Our model learns the probabilistic model from the training dataset, but it is mainly to screen out obvious non-TMB strands. Therefore, there are no concerns about the size of the training data or overfitting.

Even though the results presented in our evaluation are comparable to other methods, the methodology presented here is novel and gives insight into the actual physicochemical constraints and energy. Moreover, our approach should be able to predict TMB proteins which are significantly different from known proteins. Finally, our approach provides more information than the current approaches by providing the permutations of the strands. This can give an insight into the understanding of the folding mechanism of TMB proteins.

We show that it is possible to design models for classification and structure prediction for transmembrane β-barrel proteins which do not essentially depend on training sets but on combinatorial properties of the structures to be proved. These models are fairly

accurate, robust and can be run very efficiently on PC-like computers. Such models are useful for the genome screening.

The BBP program allows users to set up freely the physicochemical parameter values according to their own choice. This is helpful for discovering the structure as well as the folding process of specific proteins. BBP is also available for use as a web server.

Future work

The model can be applied to the the prediction of TM α-helical bundles and a mixed helice-strand structures, as well as globular β-barrels like lipocalins or membrane targeting proteins (C2 domain) where permuted structures are usually found. Appropriate energetic functions should be developed to embed into the current model.

Similar to the other methods, at present, we only propose single-domain protein structures. A pretreatment to determine protein domains will be necessary for long sequences with several domains.

The refinements in structural constraints and hydrophobicity, which may help to improve the accuracy of our predicted structure, can always be improved. A context-dependent physiochemical model, depending on specific biological membranes, can provide insight into predicted structures. Finally, it will be interesting to investigate more sophisticated statistical models for the initial screening, both to improve the results and understand how effective a mixed approach can be.

Bibliography

[1] C. Ahn, S. Yoo, and H. Park. Prediction for beta-barrel transmembrane protein region using HMM. *Journal of Korea Information Science Society*, 30(2):802–804, 2003.

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular biology of the cell*. Garland Science Taylor & Francis Group, 4 edition, 2002.

[3] S. Arnborg. Efficient algorithms for combinatorial problems on graphs with bounded decomposability a survey. *BIT Numerical Mathematics*, 25:1–23, 1985.

[4] S. Arnborg, D. G. Corneil, and A. Proskurowski. Complexity of finding embeddings in a k-tree. *SIAM J. Alg. Disc. Meth.*, 8:277–284, 1987.

[5] S. Arnborg and A. Proskurowski. Linear time algorithms for np-hard problems restricted to partial k-trees. *Discrete Applied Mathematics*, 23(1):11–24, 1989.

[6] A. Arora and L. K. Tamm. Biophysical approaches to membrane protein structure determination. *Current Opinion in Structural Biology*, 11:540–547, 2001.

[7] P. Bagos, T. Liakopoulos, and S. Hamodrakas. Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method. *BMC Bioinformatics*, 6:7, 2005.

[8] P. Bagos, T. Liakopoulos, I. Spyropoulos, and S. Hamodrakas. A Hidden Markov Model method, capable of predicting and discriminating β-barrel outer membrane proteins. *BMC Bioinformatics*, 5:29, 2004.

[9] P. G. Bagos, T. D. Liakopoulos, I. C. Spyropoulos, and S. J. Hamodrakas. PRED-TMBB: a web server for predicting the topology of β-barrel outer membrane proteins. *Nucleic Acids Res.*, 32:W400–W404, 2004.

[10] J. Bajorath, R. Stenkamp, and A. Aruffo. Knowledge-based model building of proteins: concepts and examples. *Protein Sci.*, 2:1798–1810, 1993.

[11] O. M. Becker, A. D. MacKerell Jr, B. Roux, and M. Watanabe. *Computational biochemistry and biophysics*. Marcel Dekker, Inc., New York, 2001.

[12] H. Berendsen. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1-3):43–56, 1995.

[13] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28:235–242, 2000.

[14] M. W. Bern, E. L. Lawler, and A. L. Wong. Linear-time computation of optimal subgraphs of decomposable graphs. *Journal of Algorithms*, 8(2):216–235, 1987.

[15] R. Bhaskaran and P. Ponnuswamy. Amino acid scale: average flexibility index. *Int. J. Pept. Protein Res.*, 32:242–255, 1988.

[16] H. R. Bigelow, D. S. Petrey, J. Liu, D. Przybylski, and B. Rost. Predicting transmembrane beta-barrels in proteomes. *Nucleic Acids Res.*, 32:2566–2577, 2004.

[17] C. M. Bishop, W. F. Walkenhorst, and W. C. Wimley. Folding of beta-sheets in membranes: specificity and promiscuity in peptide model systems. *J. Mol. Biol.*, 309:975–988, 2001.

[18] H. Bodlaender. Dynamic programming on graphs with bounded treewidth. In T. Lepist and A. Salomaa, editors, *Automata, Languages and Programming*, volume 317 of *Lecture Notes in Computer Science*, pages 105–118. Springer Berlin / Heidelberg, 1988.

[19] H. L. Bodlaender. A linear-time algorithm for finding tree-decompositions of small treewidth. *SIAM J. Comput.*, 25:1305–1317, 1996.

[20] H. L. Bodlaender, J. R. Gilbert, H. Hafsteinsson, and T. Kloks. Approximating treewidth, pathwidth, frontsize, and shortest elimination tree. *Journal of Algorithms*, 18(2):238–255, 1995.

[21] H. L. Bodlaender and R. H. Möhring. The pathwidth and treewidth of cographs. *SIAM J. Disc. Math.*, 6:181–188, 1993.

[22] B. R. R. Brooks, C. L. L. Brooks, A. D. D. Mackerell, L. Nilsson, R. J. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. W. Pastor, C. B. B. Post, J. Z. Z. Pu, M. Schaefer, B. Tidor, R. M. M. Venable, H. L. L. Woodcock, X. Wu, W. Yang, D. M. M. York, and M. Karplus. CHARMM: The biomolecular simulation program. *J. Comput. Chem.*, 30(10):1545–1614, 2009.

[23] R. Casadio, P. Fariselli, and P. L. Martelli. In silico prediction of the structure of membrane proteins: Is it feasible? *Briefings in Bioinformatics*, 4(4):341–348, 2003.

[24] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods. The amber biomolecular simulation programs. *Journal of computational chemistry*, 26(16):1668–1688, 2005.

[25] J. Cavanagh, W. J. Fairbrother, A. G. Palmer III, M. Rance, and N. J. Skelton. *Protein NMR spectroscopy: principles and practice*. Academic Press, 2nd edition, 2007.

[26] V. B. Chen, W. B. Arendall, III, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson. *MolProbity*: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D*, 66(1):12–21, Jan 2010.

[27] C. Chothia. The nature of the accessible and buried surfaces in proteins. *Journal of Molecular Biology*, 105(1):1–12, 1976.

[28] K. C. Chou, L. Carlacci, and G. M. Maggiora. Conformational and geometrical properties of idealized beta-barrels in proteins. *J. Mol. Biol.*, 213:315–326, 1990.

[29] C. Cobbold, A. P. Monaco, A. Sivaprasadarao, and S. Ponnambalam. Aberrant trafficking of transmembrane proteins in human disease. *Trends Cell Biol*, 13(12):639–647, 2003.

[30] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 3rd edition, 2009.

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

[32] I. W. Davis, A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, W. B. Arendall, J. Snoeyink, J. S. Richardson, and D. C. Richardson. Molprobity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Research*, 35(suppl 2):W375–W383, 2007.

[33] M. O. Dayhoff, R. M. Schwartz, and B. C Orcutt. A model of evolutionary change in proteins. *Atlas of protein sequence and structure*, 5(Suppl 3):345–352, 1978.

[34] J. Deisenhofer, O. Epp, K. Miki, R. Huber, and H. Michel. Structure of the protein subunits in the photosynthetic reaction centre of rhodopseudomonas viridis at 3åresolution. *Nature*, 318:618–624, 1985.

[35] R. L. Dunbrack and F. E. Cohen. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.*, 6(8):1661–1681, 1997.

[36] D. Eisenberg. Three-dimensional structure of membrane and surface proteins. *Annual Review of Biochemistry*, 53(1):595–623, 1984.

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

[38] N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper, and A. Sali. Comparative protein structure modeling using MODELLER. *Curr Protoc Protein Sci*, Chapter 2:Unit 2.9, 2007.

[39] R. Fano. *Transmission of Information*. Wiley, New York, 1961.

[40] S. J. Fleishman and N. Ben-Tal. Progress in structure prediction of [alpha]-helical membrane proteins. *Current Opinion in Structural Biology*, 16(4):496–504, 2006. Membranes / Engineering and design.

[41] T. C. J. Freeman and W. C. Wimley. A highly accurate statistical approach for the prediction of transmembrane beta-barrels. *Bioinformatics*, 26(16):1965–74, 2010.

[42] D. C. Gadsby, P. Vergani, and L. Csanady. The ABC protein turned chloride channel whose failure causes cystic fibrosis. *Nature*, 440:477–483, 2006.

[43] T. Gallai. Transitiv orientierbare graphen. *Acta Mathematica Hungarica*, 18:25–66, 1967.

[44] Y. Gaudin, A. Echard, and S. Etienne-Manneville. *Biologie moléculaire et cellulaire*. Ecole Polytechnique, 2010. Biology course book.

[45] J.-F. Gibrat, J. Garnier, and B. Robson. Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J. Mol. Biol.*, 198:425–443, 1987.

[46] M. C. Golumbic. *Algorithmic Graph Theory and Perfect Graphs*. Academy Press, New York, 1980.

[47] E. Gorter and F. Grendel. On bimolecular layers of lipoids on the chromocytes of the blood. *J. Exp. Med.*, 41(4):439–443, 1925.

[48] R. Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185:862–864, 1974.

[49] J. Greer. Comparative modeling methods: application to the family of the mammalian serine proteases. *Proteins*, 7:317–334, 1990.

[50] M. Gromiha, S. Ahmad, and M. Suwa. Neural network-based prediction of transmembrane β-strand segments in outer membrane proteins. *J. Comp. Chem.*, 25:762–767, 2004.

[51] M. M. Gromiha, S. Ahmad, and M. Suwa. TMBETA-NET: discrimination and prediction of membrane spanning β-strands in outer membrane proteins. *Nucleic Acids Res.*, 33:W164–W167, 2005.

[52] W. Gu, S. J. Rahi, and V. Helms. Solvation free energies and transfer free energies for amino acids from hydrophobic solution to water solution from a very simple residue model. *Journal of Physical Chemistry B*, 108(18):5806–5814, 2004.

[53] M. Habib and C. Paul. A survey of the algorithmic aspects of modular decomposition. *Computer Science Review*, 4(1):41–59, 2010.

[54] E. M. Hearn, D. R. Patel, B. W. Lepore, M. Indic, and B. van den Berg. Transmembrane passage of hydrophobic compounds through a protein channel wall. *Nature*, 458:367–370, 2009.

[55] V. Helms and J. A. McCammon. Conformational transitions of proteins from atomistic simulations. In P. Deuflhard, J. Hermans, B. Leimkuhler, A. Mark, S. Reich, and R. D. Skeel, editors, *Computational Molecular Dynamics: Challenges, Methods, Ideas*, volume 4 of *Lecture Notes in Computational Science and Engineering*, pages 66–77. Springer-Verlag, 1998.

[56] R. Henderson, J. M. Baldwin, T. A. Ceska, F. Zemlin, E. Beckmann, and K. H. Downing. Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *Journal of Molecular Biology*, 213(4):899–929, 1990.

[57] R. Henderson and P. N. Unwin. Three-dimensional model of purple membrane obtained by electron microscopy. *Nature*, 257(5521):28–32, 1975.

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

[59] I. Jacoboni, P. L. Martelli, P. Fariselli, V. D. Pinto, and R. Casadio. Prediction of the transmembrane regions of β-barrel membrane proteins with a neural network-based predictor. *Protein Sci.*, 10:779–787, 2001.

[60] M. Jacobson and A. Sali. Comparative protein structure modeling and its applications to drug discovery. In *Annual Reports in Medicinal Chemistry*, volume 39, pages 259–276. Academic Press, 2004.

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

[62] M. S. Johnson, N. Srinivasan, R. Sowdhamini, and T. L. Blundell. Knowledge-based protein modeling. *Crit. Rev. Biochem. Mol. Biol.*, 29:1–68, 1994.

[63] W. L. Jorgensen and T. J. Rives. The opls force field for proteins. energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, 110:1657–1666, 1988.

[64] H. Kamberaj and V. Helms. Monte carlo simulation of biomolecular systems with biomcsim. *Computer Physics Communications*, 141(3):375 – 402, 2001.

[65] L. Kelley and M. Sternberg. Protein structure prediction on the Web: a case study using the Phyre server. *Nature protocols*, 4(3):363–371, 2009.

[66] L. A. Kelley, R. M. MacCallum, and M. J. Sternberg. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, 299:499–520, 2000.

[67] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. G. Parrish, H. Wyckoff, and D. C. Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666, 1958.

[68] H. G. Khorana, G. E. Gerber, W. C. Herlihy, C. P. Gray, R. J. Anderegg, K. Nihei, and K. Biemann. Amino acid sequence of bacteriorhodopsin. *Proc. Natl. Acad. Sci. USA*, 76(10):5046–5050, 1979.

[69] J. Kirz, C. Jacobsen, and M. Howells. Soft x-ray microscopes and their biological applications. *Quarterly Reviews of Biophysics*, 28(1):33–130, 1995.

[70] J. H. Kleinschmidt and L. K. Tamm. Secondary and tertiary structure formation of the beta-barrel membrane protein ompa is synchronized and depends on membrane thickness. *J. Mol. Biol.*, 324(2):319–330, 2002.

[71] R. Koebnik and L. Krämer. Membrane assembly of circularly permuted variants of the E. coli outer membrane protein OmpA. *J. Mol. Biol.*, 250:617–626, 1995.

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

[73] M. F. C. Ladd and R. A. Palmer. *Structure Determination by X-Ray Crystallography*. Springer, 4th edition, 2003.

[74] M. Levitt. Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.*, 226:507–533, 1992.

[75] B. A. Lewis and D. M. Engelman. Lipid bilayer thickness varies linearly with acyl chain length in fluid phosphatidylcholine vesicles. *Journal of Molecular Biology*, 166(2):211–217, 1983.

[76] W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

[77] E. Lindahl, B. Hess, and D. van der Spoel. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Journal of Molecular Modeling*, 7(8):306–317, 2001.

[78] K. U. Linderstrom-Lang. *The Lane Medical Lectures*. Stanford University Press, 1952.

[79] K. U. Linderstrom-Lang and J. A. Shellman. Protein structure and enzyme activity. *The Enzymes*, pages 443–510, 1959.

[80] W.-M. Liu. Shear numbers of protein β-barrels: definition, refinements and statistics. *J. Mol. Biol.*, 275:541–545, 1998.

[81] S. C. Lovell, I. W. Davis, W. B. Arendall, P. I. W. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson. Structure validation by c_α geometry: ϕ, ψ and c_β deviation. *Proteins: Structure, Function, and Bioinformatics*, 50(3):437–450, 2003.

[82] M. Luckey. *Membrane structural biology: with biochemical and biophysical foundations*. Cambridge University Press, 2008.

[83] A. MacKerell Jr., C. Brooks III, L. Nilsson, B. Roux, Y. Won, and M. Karplus. *CHARMM: The Energy Function and Its Parameterization with an Overview of the Program*, volume 1 of *The Encyclopedia of Computational Chemistry*, pages 271–277. John Wiley & Sons: Chichester, 1998.

[84] D. Marsh. Infrared dichroism of twisted beta-sheet barrels. The structure of E. coli outer membrane proteins. *J. Mol. Biol.*, 297:803–808, 2000.

[85] A. Marsico, D. Labudde, T. Sapra, D. J. Muller, and M. Schroeder. A novel pattern recognition algorithm to classify membrane protein unfolding pathways with high-throughput single-molecule force spectroscopy. *Bioinformatics*, 23(2):e231–e236, 2007.

[86] P. Martelli, P. Fariselli, A. Krogh, and R. Casadio. A sequence-profile-based HMM for predicting and discriminating β-barrel membrane proteins. *Bioinformatics*, 18 Suppl 1:S46–S53, 2002.

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

[88] B. W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta*, 405(2):442–451, 1975.

[89] L. J. McGuffin, K. Bryson, and D. T. Jones. The PSIPRED protein structure prediction server. *Bioinformatics*, 16:404–405, 2000.

[90] J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, and A. Tramontano. Critical assessment of methods of protein structure predictionround viii. *Proteins*, 77(S9):1–4, 2009.

[91] A. G. Murzin, A. M. Lesk, and C. Chothia. Principles determining the structure of β-sheet barrels in proteins I. A theoretical analysis. *J. Mol. Biol.*, 236:1369–1381, 1994.

[92] A. G. Murzin, A. M. Lesk, and C. Chothia. Principles determining the structure of β-sheet barrels in proteins II. The observed structures. *J. Mol. Biol.*, 236:1382–1400, 1994.

[93] R. H. Mhring. Algorithmic aspects of the substitution decomposition in optimization over relations, set systems and boolean functions. *Annals of Operations Research*, 4:195–225, 1985.

[94] N. K. Natt, H. Kaur, and G. Raghava. Prediction of transmembrane regions of β-barrel proteins using ANN- and SVM-based methods. *Proteins: Struct. Funct. Bioinf.*, 56:11–18, 2004.

[95] H. Nikaido. Molecular basis of bacterial outer membrane permeability revisited. *Microbiol Mol Biol Rev*, 67(4):593–656, 2003.

[96] Y.-Y. Ou, S.-A. Chen, and M. M. Gromiha. Prediction of membrane spanning segments and topology in β-barrel membrane proteins at better accuracy. *J. Comp. Chem.*, 13:217–223, 2010.

[97] M. F. Perutz, M. G. Rossmann, A. F. Cullis, H. Muirhead, G. Will, and A. C. T. North. Structure of haemoglobin: A three-dimensional fourier synthesis at 5.5 Angstroms resolution, obtained by X-ray analysis. *Nature*, 185(4711):416–422, 1960.

[98] J. L. Popot and D. M. Engelman. Membrane protein folding and oligomerization: the two-stage model. *Biochemistry*, 29(17):4031–4037, 1990.

[99] O. B. Ptitsyn and A. V. Finkelstein. Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding? *Q. Rev. Biophys*, 13:339–386, 1980.

[100] G. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1):95–99, 1963.

[101] G. Ramachandran and V. Sasisekharan. Conformation of polypeptides and proteins. In J. T. E. C.B. Anfinsen, M.L. Anson and F. M. Richards, editors, *Advances in Protein Chemistry*, volume 23 of *Advances in Protein Chemistry*, pages 283–437. Academic Press, 1968.

[102] R. Ramachandran, A. P. Heuck, R. K. Tweten, and A. E. Johnson. Structural insights into the membrane-anchoring mechanism of a cholesterol-dependent cytolysin. *Nature Structural & Molecular Biology*, 9(11):823–827, 2002.

[103] A. Randall, J. Cheng, M. Sweredoski, and P. Baldi. TMBpro: secondary structure, β-contact and tertiary structure prediction of transmembrane β-barrel proteins. *Bioinformatics*, 24:513–520, 2008.

[104] W. Rawicz, K. Olbrich, T. McIntosh, D. Needham, and E. Evans. Effect of chain length and unsaturation on elasticity of lipid bilayers. *Biophysical Journal*, 79(1):328–339, 2000.

[105] N. Robertson and P. Seymour. Graph minors. I. Excluding a forest. *Journal of Combinatorial Theory, Series B*, 35(1):39–61, 1983.

[106] N. Robertson and P. D. Seymour. Graph minors. III. Planar tree-width. *Journal of Combinatorial Theory, Series B*, 36(1):49–64, 1984.

[107] G. Rose, A. Geselowitz, G. Lesser, R. Lee, and M. Zehfus. Hydrophobicity of amino acid residues in globular proteins. *Science*, 229(4716):834–838, 1985.

[108] G. D. Rose and R. Wolfenden. Hydrogen bonding, the hydrophobic effect, packing, and protein folding. *Ann. Rev. Biophysics and Biological Structure*, 22:381–415, 1993.

[109] A. Sali. Modeling mutations and homologous proteins. *Curr. Opin. Biotechnol.*, 6:437–451, 1995.

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

[111] R. Sanchez and A. Sali. Advances in comparative protein-structure modelling. *Curr. Opin. Struct. Biol.*, 7:206–214, Apr 1997.

[112] M. S. Sansom and I. D. Kerr. Transbilayer pores formed by beta-barrels: molecular modeling of pore structures and properties. *Biophys. J.*, 69:1334–1343, 1995.

[113] L. Schrödinger. The PyMOL molecular graphics system, version 1.3r1. The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC., August 2010.

[114] T. Schwede, J. Kopp, N. Guex, and M. C. Peitsch. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.*, 31:3381–3385, 2003.

[115] T. L. Steck. The organization of proteins in the human red blood cell membrane: A review. *The Journal of Cell Biology*, 62(1):1–19, 1974.

[116] M. J. Sutcliffe, I. Haneef, D. Carney, and T. L. Blundell. Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng.*, 1:377–384, 1987.

[117] L. K. Tamm, H. Hong, and B. Liang. Folding and assembly of β-barrel membrane proteins. *Biochim. et Biophys. Acta*, 1666:250–263, 2004.

[118] P. D. Taylor, C. P. Toseland, T. K. Attwood, and D. R. Flower. Beta-barrel transmembrane proteins: Enhanced prediction using a Bayesian approach. *Bioinformation*, 1(6):231–233, 2006.

[119] V. D. Tran, P. Chassignet, S. Sheikh, and J.-M. Steyaert. Energy-based classification and structure prediction of transmembrane beta-barrel proteins. In *Proceedings of the 2011 IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), IEEE Xplore*, Orlando, FL, USA, February 2011.

[120] V. D. Tran, P. Chassignet, S. Sheikh, and J.-M. Steyaert. A graph-theoretic approach for classification and structure prediction of transmembrane beta-barrel proteins. *BMC Genomics*, 13(Suppl 13):S5, 2012.

[121] V. D. Tran, P. Chassignet, and J.-M. Steyaert. Prediction of super-secondary structure in alpha-helical and beta-barrel transmembrane proteins. *BMC Bioinformatics*, 10(Suppl 13):O3, 2009.

[122] V. D. Tran, P. Chassignet, and J.-M. Steyaert. On permuted super-secondary structures of beta-barrel proteins. Technical report, INRIA, 2011.

[123] V. D. Tran, P. Chassignet, and J.-M. Steyaert. Prediction of permuted super-secondary structures in beta-barrel proteins. In *Proceedings of the 2011 ACM Symposium on Applied Computing SAC'11, ACM Digital Library*, Taichung, Taiwan, March 2011.

[124] V. D. Tran, P. Chassignet, and J.-M. Steyaert. Super-secondary structure prediction of transmembrane beta-barrel proteins. In A. Kister, editor, *Methods in Molecular Biology: Protein Super-secondary Structure*. Humana Press, NY, 2011. In press.

[125] G. E. Tusnády, Z. Dosztányi, and I. Simon. PDB_TM: selection and membrane localization of transmembrane proteins in the Protein Data Bank. *Nucleic Acids Res.*, 33:D275–D278, 2005.

[126] W. F. van Gunsteren, S. R. Billeter, A. A. Eising, P. H. Hünenberger, P. Krüger, A. E. Mark, W. R. P. Scott, and I. G. Tironi. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*. vdf Hochschulverlag AG an der ETH Zürich and BIOMOS b.v.: Zürich, Groningen, 1996.

[127] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 2 edition, 1979.

[128] C. Venclovas, A. Zemla, K. Fidelis, and J. Moult. Comparison of performance in successive casp experiments. *Proteins*, Suppl 5:163–170, 2001.

[129] J. Waldispühl. *Modélisation et prédiction de la structure des protéines transmembranaires*. PhD thesis, Ecole Polytechnique, 2004.

[130] J. Waldispühl, B. Berger, P. Clote, and J.-M. Steyaert. Predicting transmembrane β-barrels and interstrand residue interactions from sequence. *Proteins: Struct. Funct. Bioinf.*, 65:61–74, 2006.

[131] S. H. White and W. C. Wimley. Membrane protein folding and stability: physical principles. *Annual Review of Biophysics and Biomolecular Structure*, 28(1):319–365, 1999.

[132] W. C. Wimley, K. Hristova, A. S. Ladokhin, L. Silvestro, P. H. Axelsen, and S. H. White. Folding of beta-sheet membrane proteins: a hydrophobic hexapeptide model. *J. Mol. Biol.*, 277:1091–1110, 1998.

[133] C. R. Woese, D. H. Dugre, S. A. Dugre, M. Kondo, and W. C. Saxinger. On the fundamental nature and evolution of the genetic code. *Cold Spring Harbor Symposia on Quantitative Biology*, 31:723–736, 1966.

[134] R. Wolfenden, L. Andersson, P. M. Cullis, and C. C. B. Southgate. Affinities of amino acid side chains for solvent water. *Biochemistry*, 20(4):849–855, 1981.

[135] J. Xu, F. Jiao, and B. Berger. A tree-decomposition approach to protein structure prediction. In *Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference*, Washington, DC, USA, 2005.

[136] J. Xu, M. Li, D. Kim, and Y. Xu. RAPTOR: optimal protein threading by linear programming. *J Bioinform Comput Biol*, 1:95–117, 2003.

[137] A. S. Yang and B. Honig. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J. Mol. Biol.*, 301:665–678, 2000.

[138] A. A. Zamyatnin. Protein volume in solution. *Prog. Biophys. Mol. Biol.*, 24:107–123, 1972.

[139] C. Zhang and S.-H. Kim. A comprehensive analysis of the Greek key motifs in protein β-barrels and β-sandwiches. *Proteins: Struct. Funct. Genet.*, 40:409–419, 2000.