

Title      Development of a suite of bioinformatics tools  
for the analysis and prediction of membrane  
protein structure

Name      Roberto Coiti Togawa

This is a digitised version of a dissertation submitted to the University of Bedfordshire.

It is available to view only.

This item is subject to copyright.

DEVELOPMENT OF A SUITE OF BIOINFORMATICS  
TOOLS FOR THE ANALYSIS AND PREDICTION OF  
MEMBRANE PROTEIN STRUCTURE

by

Roberto Coiti Togawa

A thesis submitted for the degree of Doctor of Philosophy  
of the University of Luton

UNIVERSITY OF BEDFORDSHIRE
B/COD: 3403180508
CLASS 574.875 Tog
SEQUENCE Reference only

October 2006

kept at Enquiry  
Desk



Development of a suite of bioinformatics tools for the analysis and prediction  
of membrane protein structure

by Roberto Togawa

**Abstract**

This thesis describes the development of a novel approach for prediction of the three-dimensional structure of transmembrane regions of membrane proteins directly from amino acid sequence and basic transmembrane region topology.

The development rationale employed involved a knowledge-based approach. Based on determined membrane protein structures, 20x20 association matrices were generated to summarise the distance associations between amino acid side chains on different alpha helical transmembrane regions of membrane proteins. Using these association matrices, combined with a knowledge-based scale for propensity for residue orientation in transmembrane segments (kPROT) (Pilpel *et al.*, 1999), the software predicts the optimal orientations and associations of transmembrane regions and generates a 3D structural model of a given membrane protein, based on the amino acid sequence composition of its transmembrane regions. During the development, several structural and biostatistical analyses of determined membrane protein structures were undertaken with the aim of ensuring a consistent and reliable association matrix upon which to base the predictions.

Evaluation of the model structures obtained for the protein sequences of a dataset of 17 membrane proteins of determined structure based on cross-validated leave-one-out testing revealed generally high accuracy of prediction, with over 80% of associations between transmembrane regions being correctly predicted. These results provide a promising basis for future development and refinement of the algorithm, and to this end, work is underway using evolutionary computing approaches. As it stands, the approach gives scope for significant immediate benefit to researchers as a valuable starting point in the prediction of structure for membrane proteins of hitherto unknown structure.

Deal with difficult tasks while they are easy.

Act on large issues while they are small.

- Lao Tzu

## **Dedication**

I dedicate the thesis to Leila,  
André and Filipe.

**Table of contents**

**Abstract.....ii**

**Dedication ..... v**

**Table of contents.....vi**

**List of tables..... x**

**List of figures .....xii**

**Preface ..... xv**

**Acknowledgements.....xvii**

**Declaration..... xxi**

**Abbreviations.....xxii**

**Chapter 1 - Membrane proteins..... 24**

**1.1. Introduction ..... 24**

**1.2. Different classes of membrane proteins..... 26**

**1.3. Helix bundle and beta barrel integral membrane proteins ..... 29**

**1.4. Helix packing..... 33**

**1.5. Helix location and topology predictive tools..... 36**

**Chapter 2 - Bioinformatics ..... 39**

**2.1. Introduction ..... 39**

**2.2. DNA sequences..... 40**

---

<b>2.3. Protein sequence .....</b>	<b>41</b>
<b>2.4. The computational era .....</b>	<b>43</b>
<b>2.5. Protein structure.....</b>	<b>43</b>
<b>2.6. Secondary structure prediction .....</b>	<b>45</b>
<b>2.7. Prediction of 3D structure.....</b>	<b>46</b>
2.7.1. Homology modelling .....	47
2.7.2. Threading .....	48
2.7.3. <i>Ab initio</i> prediction.....	49
<b>2.8. Availability of tools and databases.....</b>	<b>49</b>
<b>2.9. Experimental and computational data .....</b>	<b>50</b>
<b>Chapter 3 – Structural Bioinformatics of Membrane Proteins .....</b>	<b>52</b>
<b>Chapter 4 – Objectives of the Project.....</b>	<b>57</b>
<b>Chapter 5 - Material and Methods .....</b>	<b>59</b>
<b>5.1. Hardware.....</b>	<b>60</b>
<b>5.2. Software .....</b>	<b>60</b>
<b>5.3. Databases.....</b>	<b>61</b>
<b>5.4. Amino acid colour code .....</b>	<b>63</b>
<b>5.5. Knowledge based approaches.....</b>	<b>64</b>
5.5.1. Association matrix .....	64
5.5.2. Variability between datasets used in the generation of the association matrices..	66
5.5.2.1. Bacteriorhodopsin structures .....	66
5.5.2.2. All 7 TM protein structures .....	69
5.5.2.3. Eukaryotic vs. prokaryotic.....	70
5.5.2.4. Photosynthetic reaction centres .....	70
5.5.2.5. High resolution vs. medium resolution.....	71
5.5.3. kPROT scale .....	74
<b>5.6. Distances and angles used in the project .....</b>	<b>75</b>
<b>5.7. Cartesian mathematics.....</b>	<b>76</b>
5.7.1. The use of Cartesian mathematics.....	77
<b>5.8. Permutation process .....</b>	<b>78</b>

---

<b>Chapter 6 - <i>TMCompare</i></b>	<b>80</b>
6.1. Introduction	80
6.2. Description	81
6.3. The algorithm	83
6.4. Software development	83
6.5. Discussion	85
6.5.1. <i>TMLimits</i> algorithm	87
<b>Chapter 7 – <i>TMDistance</i></b>	<b>88</b>
7.1. Introduction	88
7.2. Description	89
7.3. The algorithm	91
7.4. Software development	91
7.5. Results	92
7.6. Brief discussion	93
<b>Chapter 8 - <i>TMRelate</i></b>	<b>95</b>
8.1. Introduction	95
8.2. Description	96
8.3. The algorithm	99
8.4. The software development	99
8.4.1. <i>TMRelate_K</i>	101
8.4.2. The algorithm	102
8.5. Results	102
8.6. Brief discussion	104
<b>Chapter 09 – Evaluation</b>	<b>109</b>
9.1. Associations between TM regions – end on view	109
9.2. End on view evaluation	110
9.3. Relationship between the number of proteins used to build the matrix and the percentage of correctly predicted TM region adjacencies	116

---

9.4. 3D model evaluation .....	119
9.5. The 3D evaluation pipeline using <i>TMEvaluation_3D</i> .....	120
9.6. Structural evaluation.....	123
9.7. Test for sensory rhodopsin against a matrix of bacterial rhodopsin. ....	123
<b>Chapter 10 – Discussion.....</b>	<b>125</b>
<b>10.1. The developed software .....</b>	<b>125</b>
10.1.1. <i>TMCompare</i> .....	126
10.1.2. <i>TMDistance</i> .....	127
10.1.3. <i>TMRelate</i> .....	129
<b>10.2. Advances to the field.....</b>	<b>132</b>
<b>10.3. Future work.....</b>	<b>135</b>
10.3.1. Improvements to the developed software .....	136
10.3.2. Software availability .....	137
<b>Chapter 11 - Conclusion .....</b>	<b>138</b>
<b>Glossary .....</b>	<b>140</b>
<b>Appendix I.....</b>	<b>146</b>
<b>Appendix II .....</b>	<b>149</b>
<b>Appendix III.....</b>	<b>173</b>
<b>Appendix IV .....</b>	<b>193</b>
<b>Bibliography .....</b>	<b>203</b>



List of tables

Table 3.1 - Topology predictive tools ..... 53

Table 5.1 - Colours used to display physical and chemical characteristics of amino acids ..... 63

Table 5.2 - Bacteriorhodopsin protein divided by resolution ..... 67

Table 5.3 - Proteins with Seven TM regions ..... 69

Table 5.4 - Photosynthetic reactions centres ..... 71

Table 5.5 - The group of 17 proteins divided into high and medium resolution ..... 72

Table 5.6 - significantly lower resolution structures ..... 73

Table 5.7 - The used kPROT scale ..... 74

Table 9.1 - Evaluation of *TMRelate* by assessment of the percentage of correctly predicted associations between TM regions ..... 114

Table 9.2 - Evaluation of *TMRelate\_K* by assessment of the percentage of correctly predicted associations between TM regions ..... 115

Table 9.3 - Test for sensory rhodopsin against a matrix of bacterial rhodopsin ..... 124

Table 10.1 - The evaluation of the prediction of TM region associations by *TMRelate* ..... 131

Table 10.2 - The number of membrane proteins classified by TM regions in the Swiss-Prot database ..... 134

Table A.1 - Rasmol script for the 3 defined buttons ..... 151

Table A.2 - The obtained score for each pair of TM regions ..... 161

Table A.3 - An example of the permutation file ..... 163

Table A.4 - The permutation file size statistics ..... 163

Table A.5 - The neighbour association table ..... 164

Table A.6 - *TMRelate\_K* algorithm: helix packing definition ..... 169

Table A.7 - The buried angle ..... 170

## List of figures

Figure 1.1 - The lipid bilayer .....	26
Figure 1.2 - Integral membrane protein .....	29
Figure 1.3 - Helix bundle membrane proteins .....	32
Figure 1.4 - $\beta$ -strands integral membrane protein.....	33
Figure 1.5 - $\alpha$ -helix characteristic.....	34
Figure 1.6 - Hydropathy plot.....	37
Figure 5.1 - Sample of PDB and Swiss-Prot files.....	62
Figure 5.2 - Bacteriorhodopsin variability: High resolution dataset.....	67
Figure 5.3 - Bacteriorhodopsin variability: Medium resolution dataset .....	68
Figure 5.4 - Bacteriorhodopsin variability: Inter-helical associations .....	68
Figure 5.5 - seven TM membrane protein variability .....	69
Figure 5.6 - Variability – Eukaryotic vs. prokaryotic .....	70
Figure 5.7 - Variability – Photosynthetic reaction centres.....	71
Figure 5.8 - Variability – high vs. medium.....	73
Figure 5.9 - Variability – high vs. low .....	73
Figure 6.1 - <i>TMCompare</i> : Structure frame .....	82
Figure 6.2 - <i>TMCompare</i> : Sequence frame details .....	83
Figure 6.3 - Stand-alone version of <i>TMCompare</i> .....	84
Figure 6.4 - Web version of <i>TMCompare</i> .....	84
Figure 6.5 - <i>TMCompare</i> running with different versions of Swiss-Prot file.....	86
Figure 7.1 - 20x20 association matrix with the number of associations within given distance range .....	89

Figure 7.2 - Number of association histogram.....	90
Figure 7.3 - Atom list with distances between interacting amino acids from different TM region.....	91
Figure 7.4 - A screen shot showing the 3D confirmation of distance associations. ..	92
Figure 8.1 - <i>TMRelate</i> : Helix wheel representation.....	97
Figure 8.2 - <i>TMRelate</i> : Predicted 3D model .....	98
Figure 8.3 - The original 5x5 grid.....	99
Figure 8.4 - Transforming the grid into linear form.....	100
Figure 8.5 - The end on configuration for up to 10 TM regions.....	100
Figure 8.6 - <i>TMEvaluation</i> user interface .....	104
Figure 8.7 - Differences in results obtained with <i>TMRelate_K</i> due to changes between Swiss-Prot versions. ....	106
Figure 9.1 - <i>TMEvaluation</i> user interface .....	110
Figure 9.2 - Recording the TM region position using <i>TMCompare</i> .....	111
Figure 9.3 - Relationship between the number of proteins used to build the association matrix and the average percentage of correctly predicted TM region adjacencies.....	117
Figure 9.4 - Relationship between the number of proteins used to build the association matrix and the average percentage of correctly predicted TM region adjacencies using up to 26 proteins to build the matrix .....	117
Figure 9.5 - Relationship between the number of proteins used to build the association matrix and the average percentage of correctly predicted TM region adjacencies using only 7 helix bundle proteins. ....	118
Figure 9.6 - Comparison of the 3 bootstrap analyses.....	119
Figure 9.7 - <i>TMDistance</i> results with distances between residues in different TM regions .....	121
Figure 9.8 - 3D evaluation pipeline using <i>TMEvaluation_3D</i> program .....	122
Figure A. 1 - The “DBREF” tag from the PDB file and its’ use in <i>TMCompare</i> ....	150
Figure A. 2 - Option menu in <i>TMCompare</i> .....	152
Figure A. 3 - Scheme of loops and alpha helix distribution.....	153
Figure A. 4 - Output of the <i>TMLimits</i> program.....	155
Figure A. 5 - An example of the designated angle values for each TM region .....	159

Figure A. 6 - Residues of similar membrane depth ( $\pm 1.5$ Å) on different TM regions by alignment using colour coding .....	162
Figure A.7 - 10-digit string and the corresponding end on view configuration.....	162
Figure A. 8 - The development of the configuration buttons.....	164
Figure A. 9 - <i>TMRelate</i> : The user interface .....	165
Figure A.10 - Buried angle.....	170
Figure A.11 - Example how the algorithm consider the buried angle .....	170
Figure A.12 - The <i>TMRelate_K</i> user interface.....	171

## **Preface**

Working in the fascinating area of bioinformatics since 1994 with Dr. Goran Neshich, I have learnt about and practiced many different disciplines such as the system management of computers and peripherals installed in our Bioinformatics laboratory at Embrapa – Genetic Resources and Biotechnology located in Brasilia, Brazil; management and support of our local network and its connection to this amazing network we know as the Internet; management, manipulation and support of databases such as the PDB, Swiss-Prot and NCBI databases; development and management of our laboratory's web page; development of many different scripts written in the 'perl' programming language to help my colleagues to extract variety of information from the protein files in the Brookhaven format, among other administrative activities. At the end of 1996, I began to be interested in the research of protein structures. I spent the summer of 1997 at Columbia University in Dr. Barry Honig's laboratory, located in New York City, learning about the integration and manipulation of protein structure files (PDB format) into web browsers. There, we started the SMS project (STING Millennium Suite – <http://trantor.bioc.columbia.edu/SMS/>) (Neshich *et al.*, 2003). In 1998, our laboratory launched the first version of STING, and I became more and more interested in this fascinating area.

I'd like to express my gratitude to my wife Leila Barros who visited Rothamsted Research to carry out part of her PhD, and encouraged me to specialise in the protein structure area, that later resulted in this PhD Thesis. Since she had decided to go to Rothamsted Research in Harpenden – UK, I looked for a training program in this institute, and I contacted Dr. Paul Verrier, and he accepted me as a visiting researcher to work with Dr. John Antoniw, where I started the project described in this thesis.

In September 1999, I was introduced by Dr. John Antoniw to Dr. Jonathan Mullins who was working in the membrane protein field. At that time, Dr. Mullins had just started his research of the prediction of membrane protein structures from primary sequence and we became mutually interested in developing collaboration. Since I had a computer science background, I was very interested in understanding and learning more about protein structures and so decided to specialise in this area. So, when Dr. Mullins accepted becoming my supervisor, I started my PhD, developing a suite of programs to predict the overall associations between transmembrane regions and subsequently 3D models starting from the primary sequence.

During the project, a paper describing *TMCompare* was submitted and published in the December 2001 issue of *Bioinformatics* journal.

I am especially grateful to Embrapa, my home institute and CAPES. The former one provided me with financial support and air tickets, the latter one provided a scholarship that paid my expenses and fees at the University of Luton.

## **Acknowledgements**

I would like to thank my supervisor Dr. Jonathan Mullins, firstly for accepting me as a PhD. student and always supporting me when I needed it, for his encouragement of my work, for his valuable discussions about the project, for passing to me his enthusiasm about the membrane protein area and also for his friendship.

I would like to thank my second supervisor Dr. John Antoniwi, who introduced me to Dr. Jonathan Mullins, and also for having the patience to teach me a new programming language (Delphi) and a new concept in Oriented Object Programming. I appreciate his valuable discussion about the programming methodology, programming techniques and for helping me to speed up my programs. I would like to thank him for his encouragement in the development of the project and for his friendship.

I would like to thank Dr. Paul Verrier, from Rothamsted Research who accepted me in his department as a visiting researcher and for his valuable discussions about the project.

I would like to thank Dr. Rosane Curtis, from Rothamsted Research who helped my family when we arrived in Harpenden – UK.



I would like to thank Dr. Goran Neshich, from Embrapa – Information Technology, who introduced me to this area and always encourages me to learn about bioinformatics and the protein structure area in the best way: pushing me to work. I would like to thank him for teaching me how to be organised; this organisation helped me a lot during the development of my project. I would like to thank him for believing in my capacity and giving me the opportunity to work in his lab. And also I would like to thank him for his companionship and friendship.

I would like to thank Peter Erdélyi and Nina Federley both from the Department of Linguistics (English for academic purposes Postgraduate Course) University of Luton, for the opportunity they gave me to learn about how to develop the writing skills that eventually helped me to write my thesis.

Thanks to Kevin Crowley for his help, friendship and companionship during the period I was working at Rothamsted Research. I would also like to thank him for many English corrections in my thesis.

Thanks to my friend Fernando Ribeiro who encouraged me to apply as a candidate for visiting researcher at Rothamsted, for his friendship and companionship. And his belief in my capability to work and to be a researcher.

Thanks to Konstadina Sarakinou, another PhD student, for her initial help in the membrane protein area and her friendship.

Special thanks to my other friends from the membrane protein group at University of Luton, also doing PhDs and M.Phils: Gareth Hannaford, Gorka Lasso Cabrera, Athina Kalaitzoglou and Noushin Minaji-Moghaddam.

Thanks to my friends from my home institute laboratory who have given me help during my PhD.: Dr. Marcos Costa, Dr. Natalia Martins, Dr. Luciane Mello-

Rigden, Dr. Daniel Rigden and Dr. Felipe Silva. Especially to Natalia and Marcos for their comments reading the Thesis. I also would like to thank Dr. Georgios Pappas and Dr. Wellington Martins from Universidade Catolica de Brasilia. I would like to thank Dr. Marcelo Brigido, Dr. Ildinete Pereira, Dr. Fernando Torres, Dr. Fernando Fortes and Dr. Sonia Freitas for their lectures at the Universidade de Brasilia. A special thank to Dr. Alexandre Coelho from Universidade Federal de Goiás for the statistical discussions.

I would like also to thank Antonio Américo for his English corrections in the thesis and for his friendship.

Special thanks to my friends in Harpenden and Luton: Roisin Mullins, Renato, Josi, Shantal, Ms. Gray and Natalie.

Thanks to my landladies Cath and Joanna for their company and shelter.

I would like also to thank the people from “Living well”, especially my yoga and spinning teachers An See and Joel. Thanks to Teresa and Navarro, also Brazilian yoga and spinning instructors, respectively. These people kept my mind and body working well during my PhD.

Thanks to my nephew Eduardo Togawa for his help in solving some mathematics problems.

Special thanks to my brother Kazuro for his advice, help and friendship during the course of my entire life.

Many thanks to all my colleagues from UK and Brazil that I did not mention.

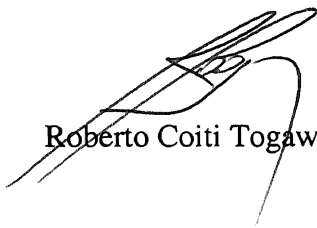
I would like also to thank all my family for the support given during my whole life.

Special thanks to my sons André and Filipe for their understanding about my absence for the achievement of my PhD. Thank you for being great company and friends for me and my wife.

Special thanks to Leila, my wife and companion for all the years we have been together, for her support in all my decisions and for her love.

## **Declaration**

I declare that this thesis is my own unaided work. It is being submitted for the degree of Doctor of Philosophy at the University of Luton. It has not been submitted before for any degree or examination in any other University.



Roberto Coiti Togawa

October, 2006

## **Abbreviations**

ASP	-	Microsoft Active Server Page
HTML	-	Hyper Text Mark-up Language
IIS	-	Internet Information Service
NMR	-	Nuclear Magnetic Resonance
OOP	-	Object Oriented Programming
PDB	-	Protein Data Bank
PWS	-	Personal Web Service
RAD	-	Rapid Application Development
RSMD	-	Root mean square deviation
TM	-	Transmembrane

Amino acid one and three letter code:

A	ALA - Alanine
C	CYS - Cysteine
D	ASP - Aspartic acid
E	GLU - Glutamic acid
F	PHE - Phenylalanine
G	GLY - Glycine
H	HIS - Histidine
I	ILE - Isoleucine
K	LYS - Lysine

L	LEU - Leucine
M	MET - Methionine
N	ASN - Asparagine
P	PRO - Proline
Q	GLN - Glutamine
R	ARG - Arginine
S	SER - Serine
T	THR - Threonine
V	VAL - Valine
Y	TYR - Tyrosine
W	TRP - Tryptophan

## **Chapter 1 - Membrane proteins**

### **1.1. Introduction**

Proteins play a variety of roles in life processes and many different classes of proteins are known. There are structural proteins like viral coat proteins; molecules of the cytoskeleton, epidermal keratin; catalytic proteins known as enzymes; transport and storage proteins like haemoglobin, myoglobin and ferritin; regulatory proteins including hormones and many proteins that control genetic transcription; proteins of the immune system and the immunoglobulin superfamily, including proteins involved in cell-cell recognition and signaling (Lesk, 2001). Several studies suggest that around 25% of all protein types in a cell are membrane proteins (Boyd *et al.*, 1998; Wallin and von Heijne, 1998; Chen and Rost, 2002). Their importance is also highlighted by their likely representation in a high proportion of preferred pharmaceutical targets. Some estimates show that 60% of drug targets in the pharmaceutical industry are membrane proteins (Yeagle and Lee, 2002).

Membranes are vital for living cells; they separate the cell from the outer world, they also separate compartments inside the cell (organelles) to protect important processes and events. These membranes are extremely thin (4.5nm) films of lipids and embedded proteins (Branden and Tooze, 1999). The lipid molecules in

the cell membranes are amphipathic, i.e., one end is hydrophilic and the other end is hydrophobic. The main category of lipid molecules used to build biological membrane is the phospholipids. They have a hydrophilic head group and two hydrophobic hydrocarbon tails. The tails are usually fatty acids and they can be different in length. One tail may have one or more cis-double bonds (unsaturated) creating a kink in the tail and the other not (saturated). These differences in the length and saturation of the fatty acid tails are important because they influence the ability of phospholipid molecules to pack against one another, affecting the fluidity of the membrane. The shape and the amphipathic nature of the lipid molecules cause the aggregation and hiding of the hydrophobic tails in the interior and exposure of the hydrophilic heads to water. They can aggregate in two ways: forming spherical vesicles, or they can form bimolecular sheets or bilayers, with the hydrophobic tails 'sandwiched' between the hydrophilic head groups. One of the most important characteristics of the lipid bilayer is the fluidity, which is fundamental to many membrane functions (Alberts *et al.*, 2002).

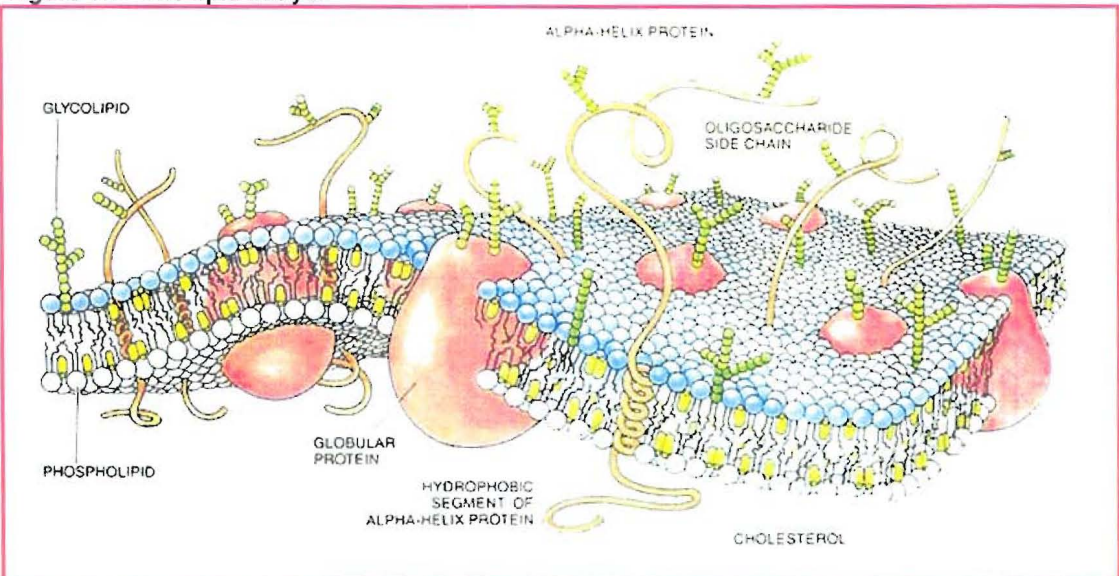
The first membrane model to be generally accepted was proposed by Danielli and Davson in 1935, their model was basically a 'sandwich' of lipids covered on both sides with proteins (Danielli and Davson, 1935). This was the basic model for membrane structure accepted by biologists for many years until the early 1970s. This model was eventually replaced in 1972 by the current model of the membrane, known as the 'fluid mosaic model' and was proposed by the biochemists Singer and Nicolson (1972). This model retains the basic lipid bilayer structure, but the proteins, are thought to be globular and to float within the lipid bilayer rather than form the layers of the sandwich-type model. Floating within this bilayer are the proteins, some



of which span the entire bilayer and may contain channels or pores to allow passage of molecules through the membrane.

Some protein molecules, known as transmembrane proteins, are embedded in this bilayer, crossing it entirely, and they are usually arranged within three distinct regions: One or more hydrophobic transmembrane (TM) segments ( $\alpha$ -helix or  $\beta$ -strands in the interior of the membrane) and two hydrophilic loop regions, one at each side of the membrane. They serve as highly active mediators between the cell and its environment or the interior of an organelle and the cytosol. Membrane proteins have many functions such as acting as receptors for hormones, pumps for transporting materials across the membrane, ion channels, adhesion molecules for holding cells to the extracellular matrix, and cell recognition antigens among others (Chen and Rost, 2002). The following chapter will provide a brief background to membrane proteins.

**Figure 1.1 –The lipid bilayer**



This image was taken from Wikipedia and it is classified as a public domain.

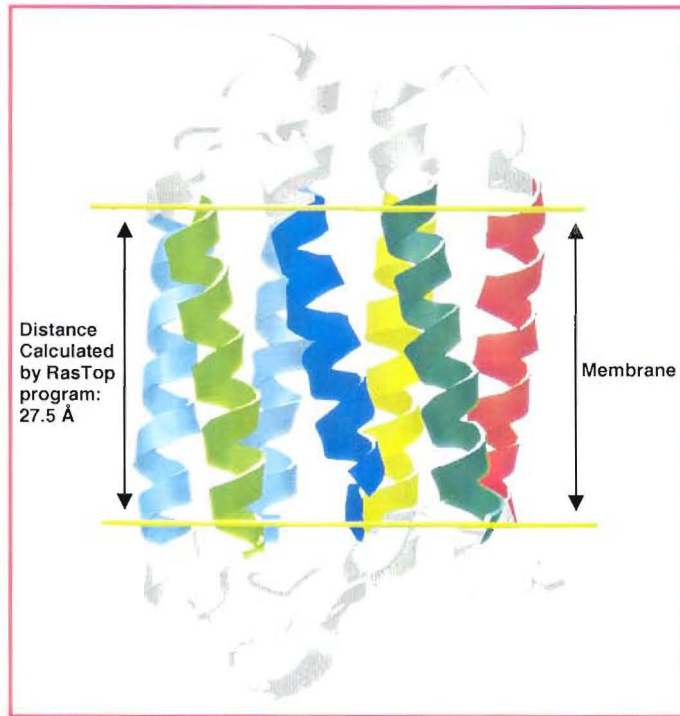
## 1.2. Different classes of membrane proteins

Membrane proteins can be classified into two main categories: integral and peripheral. Integral membrane proteins are those whose polypeptide chain traverses the membrane at least once. They are difficult to extract and can only be removed from the membrane with the use of detergents. This process however can disrupt the structure of the protein, making the study of this category of membrane proteins more difficult. These proteins possess segments immersed in the non-polar interior of the membrane, which mainly have hydrophobic surface residues, while the portions that extend into the aqueous environment are by and large sheathed with polar residues. The structural conformation of integral membrane proteins enables them to completely span the membrane and mediate the flow of nutrients and waste. Consequently they are the focus of a continually increasing number of studies (Branden and Tooze, 1999). Peripheral membrane proteins are anchored to the membrane by non-covalent bonds, and may be attached to integral proteins. They can be easily extracted from the membrane for further studies using high salt conditions or alkaline pH. This class of membrane proteins will not be discussed further in the thesis.

The integral membrane proteins can be divided in three main types according to their transmembrane (TM) regions: (1) Porin class of proteins displaying the characteristic  $\beta$ -barrel structure; (2) those that span the lipid bilayer with one single  $\alpha$ -helix known as single-span or bitopic; (3) and those that cross the lipid bilayer with two or more  $\alpha$ -helices known as multi-span or polytopic membrane proteins. These last ones will be focused on this thesis. The  $\alpha$ -helices run generally perpendicular to the membrane plane and connections are formed between neighbouring helices, while the  $\beta$ -barrels contain meandering antiparallel sheets, with

topologies merely dependent on the strand number (Schulz, 2000). The membrane lipid bilayer reduces the degrees of freedom for  $\alpha$ -helices facilitating computational methods for the prediction of its secondary and tertiary structure from the primary sequence (Chen and Rost, 2002). However, this constraint does not apply to the porin-like proteins that form pores by  $\beta$ -strands barrels. This can be explained by the greater stability of the  $\alpha$ -helix compared by the  $\beta$ -strand structure. Due the lack of experimental information available on different porin-like membrane protein, it is difficult to develop prediction methods and estimate the prediction accuracy for this class (Chen and Rost, 2002).

The most frequently observed secondary structure in integral TM segments is the helix bundle conformation, making up about 90% of membrane proteins sequences (Jones *et al.*, 1994), highlighting the importance of the development of tools for predicting the associations between TM regions for this class of proteins. Figure 1.2 shows multi-span integral membrane protein.

**Figure 1.2 – Integral membrane protein**

**Multi-span membrane protein (Bacteriorhodopsin with seven transmembrane regions - PDB code 1at9); The TM regions were coloured using *TMCompare* program (Togawa *et al.*, 2001). The picture was created by RasTop molecular visualization v.2.0.2 (Valadon, 2002).**

### 1.3. Helix bundle and beta barrel integral membrane proteins

The  $\alpha$ -helix was first described in 1951 by Linus Pauling (Branden and Tooze, 1999); he made his remarkable prediction on the basis of accurate geometrical parameters that he had derived for the peptide unit from the results of crystallographic analyses of the structures of a range of small molecules.

In the literature are found different families of helix bundles, classified by the number of TM regions. In archaea, eubacteria and plants, membrane proteins with 4, 10 and 12 TM regions are dominant, while 4 and 7 TM regions protein appear to be more common in yeast and in higher eukaryotes (Ubarretxena-Belandia and Engelman, 2001). Studies of functional processes have also revealed that the

numbers of TM regions are correlated with their functions. Proteins with more than seven TM regions are related to transport systems (Paulsen *et al.*, 2000; Kihara and Keneshisa, 2000) and multicellular organisms have a greater proportion of seven TM region proteins belonging presumably to the GPCR family (Jones, 1998; Remm and Sonnhammer, 2000).

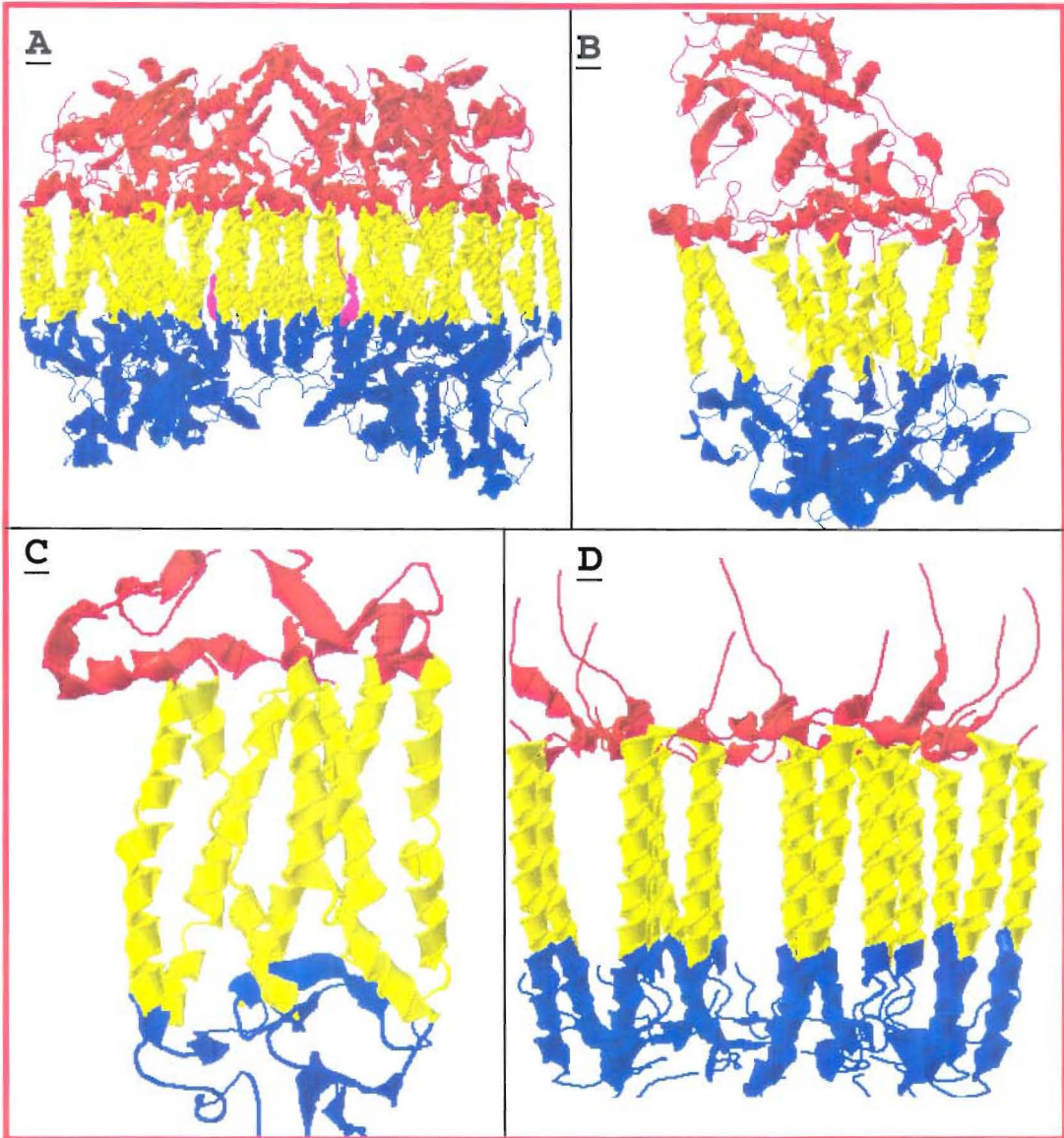
However, the most studied and best-characterised helix bundle membrane protein is bacteriorhodopsin from *Halobacterium halobium* (Swiss-Prot accession code P02945), with 7 TM regions. It consists of 262 amino acid residues, and contains one molecule of retinal which is bound deep inside the protein and connected to the polypeptide by a lysine residue. This retinal molecule changes its conformation when absorbing a photon, resulting in a conformational change of the surrounding protein and the proton pumping action. The first structural model of bacteriorhodopsin was obtained in 1975 (Henderson and Unwin, 1975) by electron microscopy; it gave the first insight as to how membrane proteins are constructed, showing that they have a number of TM  $\alpha$ -helices. This work had a great impact on subsequent theories and experiments on membrane proteins (Branden and Tooze, 1999). It is also the most studied in terms of 3D structures, with 62 different entries found in the PDB. “Bacteriorhodopsin is not only one of the best structurally and functionally characterised integral membrane protein, but has also served as the test-bed for the development of both hardware and software for electron crystallography” (Von Heijne, 1997).

Another helix bundle family of great interest to the pharmaceutical companies is the G protein coupled receptors (GPCRs), also of 7 helix bundles. They present novel targets for drugs (Stadel *et al.*, 1997; Wong, 2003). GPCRs, include receptors

for hormones, neurotransmitters, growth factors, light and odour-related ligands (Dewji and Singler, 1997; Hildebrand and Shepherd, 1997; Pierce *et al.*, 2002). The importance of the GPCRs is illustrated by the observation that 30~50% of drugs act on GPCRs (Wise *et al.*, 2004; Dahl and Sylte, 2005; Sarramegna *et al.*, 2006). For this family there is only one solved 3D structure with resolution at 2.8 Å (Palczewski *et al.*, 2000). In contrast, over 1000 GPCRs amino acid sequences are known due to the Human Genome Project and other genome projects (Karchin *et al.*, 2002), showing again the importance of tools for predicting the interactions between TM regions and the creation of 3D structures from the primary sequence. The following figure shows the structures of different helix bundle integral membrane proteins.



Figure 1.3 – Helix bundle membrane proteins

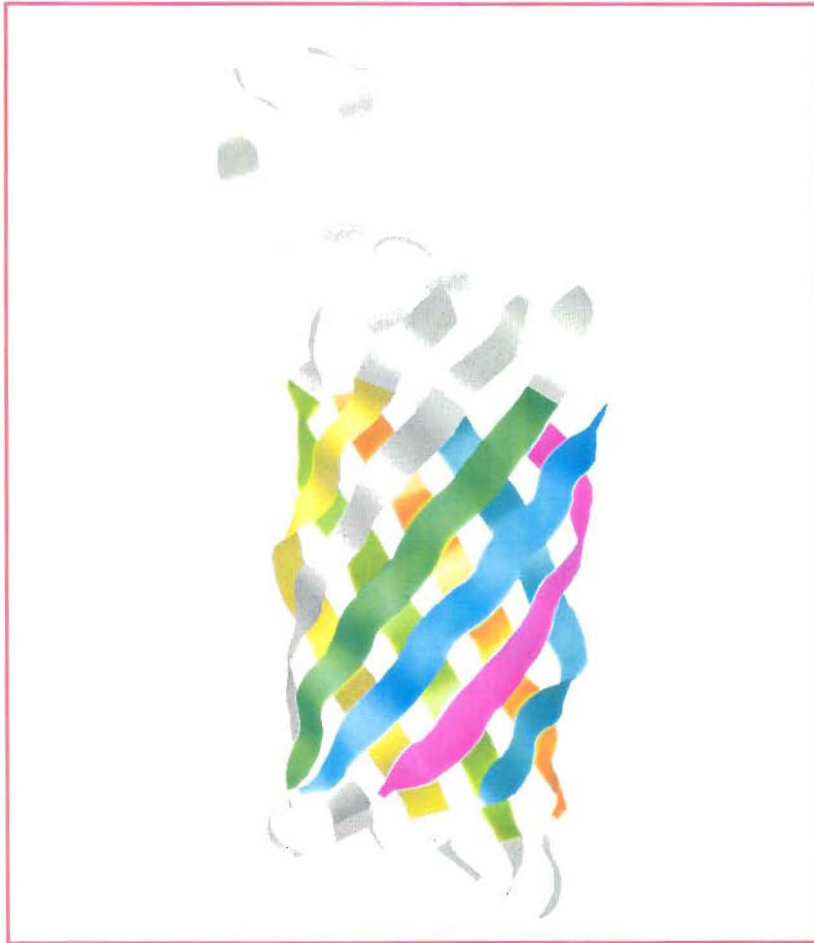


Crystal structures showing the helix bundle membrane proteins. (A) Bovine heart *Cytochrome C oxidase* (PDB code: 1OCC - Tsukihara *et al.*, 1996). (B) Photosynthetic reaction center from *T. bacterium*, *T. tepidum* (PDB code: 1EYS – Nogi *et al.*, 2000). (C) Bovine *Rhodopsin* (PDB code: 1U19 – Okada *et al.*, 2004). (D) Light-harvesting protein from *Rhodospseudomonas acidophila* (PDB code: 1NKZ – Papiz *et al.*, 2003 ). The TM regions were coloured using PDBTM tool (Tusnády *et al.*, 2004).

There are also TM structures based on the  $\beta$ -strand conformation, mainly for the porin family of proteins (figure 1.4). The porin family is a group of proteins consisting of an anti-parallel  $\beta$ -barrel that creates the basic pore while an 'eyelet' loop of polypeptide lining the inner barrel wall defines the characteristics of the pore, and allows the passage of small molecules across the bilayer (Garavito, 1998). Porins are

the most abundant membrane proteins in bacteria. To illustrate this abundance, each *Escherichia coli* cell contains about 100,000 copies of porin molecules in its outer membrane. Each porin forms an open water-filled channel that allows passive diffusion of nutrients and waste elements across the outer membrane (Branden and Tooze, 1999).

**Figure 1.4 -  $\beta$ -strands integral membrane protein**



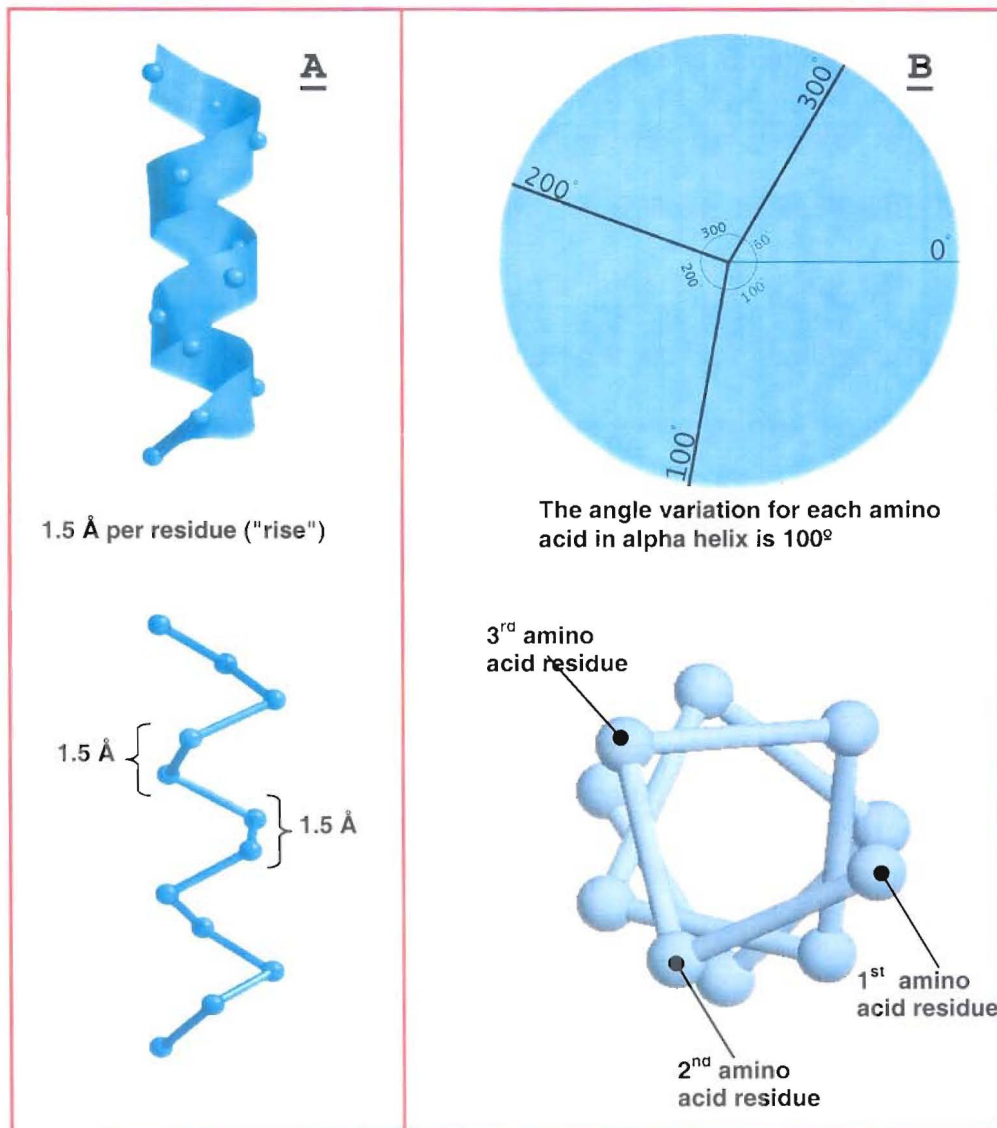
The  $\beta$  conformation of the integral membrane protein ompx porin from *Escherichia coli* (PDB code 1QJ8 – Vogt and Schulz 1999). The coloured  $\beta$ -strands are the TM regions and are obtained using *TMCompare* program (Togawa *et al.*, 2001). The image was created using RasTop 2.0.2 (Valadon, 2002).

#### 1.4. Helix packing



An alpha helix is a common element of secondary structure in proteins. The amino acids are arranged in a helical structure, and the stability is maintained by the hydrogen bonds between the C=O group of amino acid  $n$  and the N-H group of amino acid  $n+4$ . There is a  $100^\circ$  rotation about the axis from one residue to the next, making 3.6 residues per turn and the distance along the axis from one residue to the next is  $1.5\text{\AA}$ . This is called the rise of the helix. Figure 1.5 shows the  $\alpha$ -helix structural nature.

**Figure 1.5 -  $\alpha$ -helix characteristic**



A) The rise of the helix by  $1.5\text{\AA}$  per residue. B) The angle between each amino acid residue. The protein  $\alpha$ -helix image was created using RasTop 2.2 (Valadon, 2002) and the 'angle variation' image was created using CorelDraw software (<http://www.corel.com>).

The study of helix packing is important to the stability, folding, and associations of membrane proteins. The helix associations occur through a combination of hydrogen bonding, electrostatic and van der Waals interactions (Eilers *et al.*, 2000).

The analysis of helix packing has been critical for evaluating structural models, designing novel proteins and for the general understanding of how the final tertiary structure of proteins is encoded in its primary sequence (Eilers *et al.*, 2000; Russ and Engelman, 2000). The packing arrangements between amino acids on adjacent  $\alpha$ -helices has revealed very significant patterns of fitting together residues, the foremost of which is called the ridge-groove arrangement (Chothia *et al.*, 1981). The arrangement of the side chains in a helical row along the surface of the helix results in the formation of ridges separated by shallow grooves on the surface. The ridges and the grooves are formed by amino acids that are usually three or four residues apart and occur as a packing arrangement between particular amino acids, so that their detailed geometry is dependent not only on the geometry of the helix but also on the actual amino acid sequence. Russ and Engelman (2000) described that an amino acid that has been found to be of great significance in packing TM regions is the glycine residue as a single residue and as a part of the GxxxG motif, both of which have been highlighted in a number of different studies (Senes *et al.*, 2000). Another study about internal packing, by Eilers and colleagues, using the method of occluded surfaces that provides a direct measure of molecular packing and allows the fractionation of the atomic or molecular surface given by the packing value, revealed that the highest packing values in integral membrane proteins originate from small hydrophobic (glycine and alanine) and small hydroxyl-containing (serine and

threonine) amino acids (Eilers *et al.*, 2000). This packing involving ‘big’ and ‘small’ residues was also confirmed by Adamian and Liang: they describe that large residues like phenylalanine, tryptophan, and histidine have the highest propensity to be in a TM void or a pocket, whereas small residues such as serine, glycine, alanine and threonine are least likely to be found in a void or a pocket (Adamian and Liang, 2001).

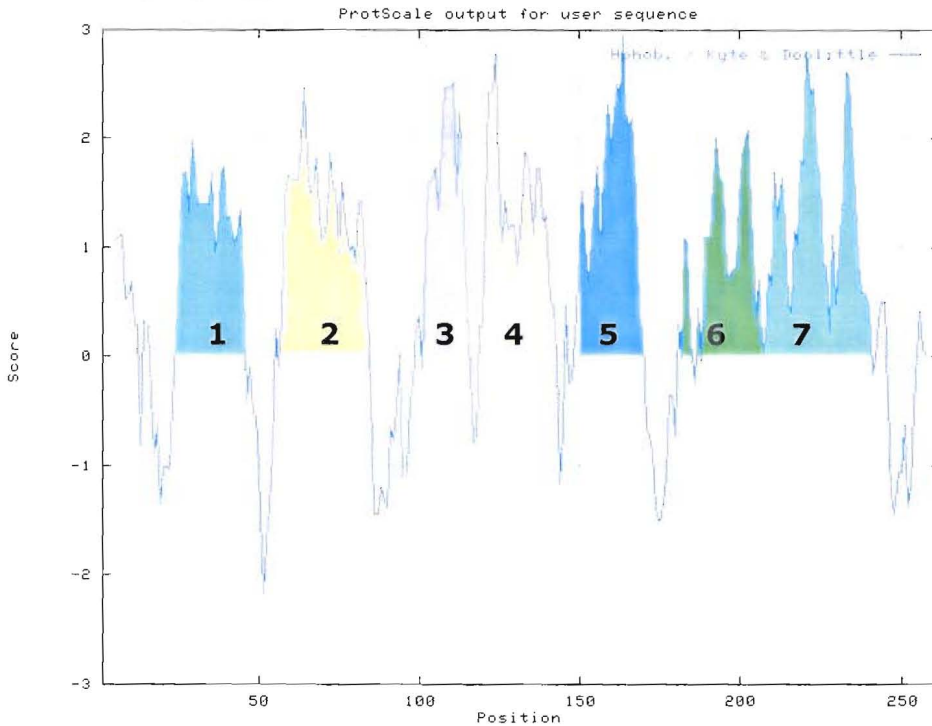
### 1.5. Helix location and topology predictive tools

TM  $\alpha$ -helix location can be reliably predicted from amino acid sequences due to the fact that membrane-spanning domains structures are generally encoded by unusually long hydrophobic stretches of 20-30 residues (Sonnhammer *et al.*, 1998). They can be predicted based on the patterns of hydrophobic and polar regions within the primary sequence. Each amino acid side chain forming the transmembrane helix has a different hydrophobicity. However, to consider all side chains according to hydrophobicity and to assign actual numbers that represent their degree of hydrophobicity is not trivial. Many different hydrophobicity scales have been developed on the basis of solubility measurements of the amino acids in different solvents, vapour pressures of side-chain analogues, analysis of side-chain distributions within soluble proteins, and theoretical energy calculations (Branden and Tooze, 1999).

A high hydrophobicity value indicates a preference to be in a non-polar environment like the interior of the membrane. Kyte and Doolittle (1982) introduced the first and most often used hydrophobicity scale to predict TM regions. The

following figure shows the hydropathy plot of bacteriorhodopsin using the Kyte and Doolittle hydrophobicity scale.

**Figure 1.6 – Hydropathy plot**



The hydropathy index is plotted against the residue number for bacteriorhodopsin, showing the 7 peaks corresponding to the 7 TM regions using the Kyte and Doolittle hydrophobicity scale (Kyte and Doolittle, 1982) with window size 9. The plot was created using the ProtScale tool from the ExPASy web server (Gasteiger *et al.*, 2005). The coloured TM regions were adapted from *Lehninger Principles of Biochemistry* – Chapter 11 (Nelson and Cox, 2004).

Most TM proteins have a specific distribution of positively charged amino acids; this rule is known as the ‘positive-inside-rule’ and it describes the observation that the inter-helix connecting loop regions on the inside of the membrane have more positive charges than the loop regions on the outside (von Heijine, 1992). These observations have been the main basis of a variety of the topology prediction methods developed over the last two decades (Chen and Rost, 2002).

Many different methods to predict TM regions and their topology have been developed in the past two decades, and most of them are available on the Internet. Some of those methods, use hydropathy analysis (TMPred, Hofmann and Stoffel,

1993; SOSUI, Hirokawa *et al.*, 1998), and others use different approaches such as dynamic programming (MEMSAT, Jones *et al.*, 1994), evolutionary information from protein families (PHDhtm, Rost *et al.*, 1994; DAS, Cserző *et al.*, 1997) and rules reflecting global aspects of membrane region using Hidden Markov Model (HMM) (TMHMM, Krogh *et al.*, 2001; HMMTOP, Tusnády and Simon, 1998). Table 3.1 in chapter 3 shows a list of the available programs with their respective web addresses.

Significant advances in the area of membrane protein structure have been achieved in the last 10 years. The prediction methods are more accurate in terms of the correct localisation of the TM segments and topology. These programs are more user-friendly showing the results with graphical interfaces, making it easier to interpret the output, like SOSUI (Hirokawa *et al.*, 1998), HMMTOP (Tusnády and Simon, 1998), TMHMM (Krogh *et al.*, 2001) among others.

In terms of 3D structure determination, new techniques like atomic force microscopy and electron microscopy are making an increasing contribution. The first structure of aquaporin (AQP1 at 3.8 Å resolution) was elucidated by electron crystallography in 2000 by Murata and colleagues (Murata *et al.*, 2000). One year later the high-resolution structure determined by X-ray analyses at 2.2 Å resolution was made available (Sui *et al.*, 2001). The two structures have been assessed and found to agree in significant details (de Groot *et al.*, 2003). This comparison has stimulated further efforts into electron crystallography, with the goal of improving both data processing technology as well as 2D crystallogenesis (Werten *et al.*, 2002).

## **Chapter 2 - Bioinformatics**

### **2.1. Introduction**

Over the course of the past decade, the number of sequenced genes has increased exponentially, due to the development of new laboratory sequencing techniques and driven by the various genome projects. However, the development of laboratory techniques for protein sequencing began slowly. The first protein to be sequenced was the hormone insulin, in 1955 (Ryle *et al.*, 1955). Five years later, the first enzyme was sequenced – ribonuclease (Hirs *et al.*, 1960). By 1965, around 20 proteins with more than 100 residues were sequenced (Attwood and Parry-Smith, 1999). The development of new protein sequencing techniques in the following years increased this number to about 1,500 sequences in 1980. Today there are more than 204,000 protein sequences available in the Swiss-Prot repository (Release 48.7 of 20-Dec-2005) (Bairoch and Apweiler, 2000).

The number of DNA sequences, due in large part to the many genome projects currently being undertaken worldwide, is even bigger. As a direct consequence of the necessity to store, search, compare and analyse the huge amount of data created, an increase in the importance of the existing area known as computational biology has occurred as well as the creation of a sub-discipline called

bioinformatics (Gibas and Jambeck, 2001). This chapter will introduce some important aspects of bioinformatics, focusing mainly on protein structure prediction.

## 2.2. DNA sequences

Protein sequences are mainly obtained from the DNA sequences of the genes coding for particular proteins. The gene contains genetic information in the form of a linear molecule composed of four types of nucleotide bases (adenine, thymine, cytosine, and guanine). The DNA sequences generated by the sequencing projects are validated and deposited in the three primary DNA repositories:

NCBI-GenBank      (<http://www.ncbi.nlm.nih.gov:80/Database/index.html>)

EMBL                (<http://www.ebi.ac.uk/embl/index.html>)

DDBJ                (<http://www.ddbj.nig.ac.jp/>)

These institutes form an international collaboration, providing reliable and up-to-date DNA sequence databank information to researchers around the world. Each of the three institutes collects a portion of the total sequence data reported worldwide and all new and updated database entries are exchanged between the groups on a daily basis (Stoesser *et al.*, 2002). These databases are publicly available and it is possible to search and retrieve these resources for a particular sequence using programs such as BLAST (Altschul *et al.*, 1990; Altschul *et al.*, 1997) or fastA (Pearson and Lipman, 1988), which perform comparisons between pairs of sequences, searching for regions of local similarity. Since these databases are publicly available, it is possible to download them to a local computer to be used for further investigation.

The growth rates of these databases are impressive. The number of sequences in the NCBI-GenBank more than doubled in 2 years, from 14,976,310 different DNA entries (December, 2001) to 30,968,418 (December, 2003). The current release and numbers of entries for each DNA database are:

**NCBI-GenBank<sup>1</sup>:**

Release 151 / December, 2005

52,016,762 entries

56,037,734,462 nucleotides

**EMBL<sup>2</sup>:**

Release 85 / December, 2005

52,651,500 entries

56,476,719,034 nucleotides

**DDBJ<sup>3</sup>:**

Release 64 / December, 2005

52,727,669 entries

56,098,558,378 nucleotides

### 2.3. Protein sequence

On the other hand, the numbers of protein sequences are relatively low when compared with the number of DNA sequences; in the release 48.7 of 20-Dec-2005: there are 204,086 proteins sequences in the Swiss-Prot database. However, there are 2,506,886 protein sequences in the TrEMBL database (31.7 of 20-Dec-2005). There are two primary protein sequence repositories:

Swiss-Prot      (<http://www.expasy.ch>)

TrEMBL        (<http://www.expasy.ch/sprot>)

---

<sup>1</sup> <ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt>

<sup>2</sup> [http://www.ebi.ac.uk/embl/Documentation/Release\\_notes/current/relnotes.html](http://www.ebi.ac.uk/embl/Documentation/Release_notes/current/relnotes.html)

<sup>3</sup> [http://www.ddbj.nig.ac.jp/breakdown\\_stats/dbgrowth-e.html](http://www.ddbj.nig.ac.jp/breakdown_stats/dbgrowth-e.html)



and one repository for the processing and distribution of 3-D biological macromolecular structure data:

PDB (<http://www.pdb.org>)

Swiss-Prot is a curated protein sequence database. The aim of this database is to provide a high level of annotation (such as the description of the protein function, its domain structure, transmembrane regions, variants, etc), a minimal level of redundancy and a high level of integration with other databases (Bairoch and Apweiler, 2000). TrEMBL is a computer-annotated protein sequence database supplementing the Swiss-Prot protein sequence data bank. It contains the protein translations of all coding sequences present in the EMBL (The European Molecular Biology Laboratory) nucleotide sequence database not yet integrated in Swiss-Prot (Bairoch and Apweiler, 2000). The Protein data bank (PDB) is a weekly updated archive of experimentally determined three-dimensional structures of biological macromolecules, serving a global community of researchers, educators, and students. The archives contain atomic co-ordinates, bibliographic citations, primary and secondary structure information, as well as crystallographic structure factors and NMR experimental data (Berman *et al.*, 2000).

The current release and numbers of entries for each main protein database are:

**Swiss-Prot<sup>4</sup>:**

48.7 of 20-Dec-2005:  
204,086 entries.

**TrEMBL<sup>4</sup>:**

31.7 of 20-Dec-2005:  
2,506,886 entries.

---

<sup>4</sup> <http://ca.expasy.org/sprot/>

**PDB<sup>5</sup>:**

Release 27-Dec-2005:

34,376 entries.

**2.4. The computational era**

To store and manipulate the huge number of DNA and protein sequences, electronic databases were created, providing valid and reliable information for the many groups interested in the field. In order to extract information from raw data, biologists require new programs and algorithms for investigating sequence homology (Altschul *et al.*, 1990; Sánchez *et al.*, 1997), sequence alignment analysis (Thompson *et al.*, 1994), protein structural classification (Murzin *et al.*, 1995), protein structure modelling (Sali and Blundell, 1993), among others, resulting in the development of a new computational biology discipline called Bioinformatics.

A reasonable definition of bioinformatics is “information technology applied to the management and analysis of biological data” (Attwood and Parry-Smith, 1999). Researchers working in bioinformatics laboratories are mainly either biologists learning computer science or programmers learning biology, though there are also many physicists, chemists, mathematicians, statisticians and even designers. It is indeed a very multidisciplinary area with researchers from many backgrounds working toward one goal: to increase understanding about the relationship between DNA/protein sequences, structures and functions.

**2.5. Protein structure**

---

<sup>5</sup> <http://www.pdb.org>

Protein structure is commonly classified at four levels: primary structure consists of a sequence of amino acids linked together by peptide bonds and includes any disulphide bonds; secondary structure is the resulting polypeptide coiled into regularly occurring structure such as  $\alpha$ -helices or  $\beta$ -strands; tertiary structure describes the packing of the secondary structure units into one or several compact globular units called domains; quaternary structure is used to describe proteins composed of multiple subunits (chains).

The primary structure of a protein can readily be deduced from the nucleotide sequence of the corresponding messenger RNA. Based on primary structure, many features of secondary structure can be predicted with the aid of computer programs. The next goal in the post-genomic era is to understand more about protein structure and function, translating all of available sequence data into structural knowledge (Maggio and Ramnarayan, 2001). However, information about structure is more complex to extract, store and manipulate than the sequence information. Two main approaches can be used to determine the three-dimensional structure of macromolecules: (1) Nuclear magnetic resonance spectroscopy yields information on the structure of proteins in solution, with a size limitation of approximately 30 kD. This technique is used for small proteins. (2) X-ray crystallography apparently has no size limit for generation of structural data, but it requires purification and crystallization of the protein under study. Owing to recent technical advances, X-ray crystallography is now the preferred method for precise structural determination of proteins (Montelione and Anderson, 1999).

## 2.6. Secondary structure prediction

One reason for studying and attempting to predict secondary structure is to understand better the effects of amino acid substitution in the catalytic and regulatory regions of a protein. The first method to predict secondary structure was developed by Chou and Fasman (Chou and Fasman, 1974). This is a one-dimensional prediction of the secondary structure of each residue that may be  $\alpha$ -helix,  $\beta$ -sheet, turn or coil. Today, many programs that predict secondary structure using different methods are available through the Internet and give an average accuracy of 75% (the percentage of residues correctly predicted) (Petersen *et al.*, 2000). Using a web browser interface, these predictive tools take as input the primary linear sequence, execute the algorithm on their servers, returning the result usually by e-mail. This is because some methods use intensive computer processing (CPU) and tend to be run in a batch queue. The following list shows some of the secondary structure prediction programs available on the web and their addresses:

Program Name: **PHD**  
Description: PHD uses a neural network system (a sequence-to-structure level and a structure-structure level) to predict secondary structure. PHD focuses on hydrogen bond prediction. The use of evolutionary information contained within a multiple sequence alignment increases the prediction accuracy. The inputs to the neural network are multiple alignments.  
Web Address: [http://cubic.bioc.columbia.edu/predictprotein/submit\\_def.html](http://cubic.bioc.columbia.edu/predictprotein/submit_def.html)  
Reference: Rost and Sander, 1993

Program Name: **PSIPRED**  
Description: This program combines neural network predictions with a multiple sequence alignment derived from a PSI-BLAST<sup>6</sup> database search.  
Web Address: <http://bioinf.cs.ucl.ac.uk/psipred/>

---

<sup>6</sup> See glossary

Reference: Jones, 1999a

Program Name: **Jpred**

Description: The program takes a protein sequence or multiple alignment of protein sequences and predicts secondary structure using a neural network called Jnet. The prediction is the definition of each residue into either alpha helix, beta sheet or random coil secondary structures.

Web Address: <http://www.compbio.dundee.ac.uk/~www-jpred/>

Reference: Cuff *et al.*, 1998.

Program Name: **PREDATOR**

Description: PREDATOR combines multiple sequence alignment information with the hydrogen bonding characteristics of the amino acids to predict the secondary structure.

Web Address: [http://npsa-pbil.ibcp.fr/cgi-bin/npsa\\_automat.pl?page=/NPSA/npsa\\_preda.html](http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_preda.html)

Reference: Frishman and Argos, (1995,1996)

Program Name: **PSA**

Description: PSA is based in the Hidden-Markov Model approach to secondary structure prediction. It has a detailed graphical output, which represents predicted probabilities of helix, sheet and coil states for each position in the protein sequence.

Web Address: <http://bmerc-www.bu.edu/psa/>

Reference : Stultz *et al.*, 1993

The secondary structure predictions provide important information for 3D structure prediction. These predictions are widely applicable to the analysis of proteins and are a starting point for fold recognition methods for tertiary structure prediction.

## 2.7. Prediction of 3D structure

In order to increase the number of characterised protein structures, many bioinformatics laboratories are attempting to predict 3D protein structures *in silico*. The two main approaches are (1) comparative model building methods (homology and threading modelling) and (2) knowledge-based prediction methods to deduce 3D structure directly from the linear sequence using tables for possible interactions between amino acids and specific mathematical models.

### 2.7.1. Homology modelling

Historically, the most successful techniques of protein structure prediction have been those based on inference from evolution (homology). “If a sequence can be shown to be sufficiently similar to another sequence of known structure, then the implied evolutionary relationship will guarantee structural similarity” (Westhead and Thornton, 1998). Homology modelling techniques consist of the use of a structural template derived from a known structure to build a new 3D model of a protein. After finding the homologous structure file(s) (in PDB format) using alignment programs like BLAST (Altschul *et al.*, 1990; Altschul *et al.*, 1997) or FastA (Pearson and Lipman, 1988), one can start model-building by using some of the free available tools for protein modelling such as Modeller (Sali and Blundell, 1993), DeepView Swiss PdbViewer (Guex and Peitsch, 1997). With these, it is possible to create, manipulate, modify, test and evaluate the new predicted 3D structure. It is also possible to use some of the available services at the Internet for automatic homology modelling like the Swiss-Model server - An Automated Comparative Protein Modelling Server - [http://www.expasy.org/swissmod/SM\\_TOPPAGE.html](http://www.expasy.org/swissmod/SM_TOPPAGE.html) (Guex *et*

*al.*, 1999), where the user can submit a primary sequence or the Swiss-Prot accession number and receive a predicted 3D model in PDB format.

### 2.7.2. Threading

The threading technique is most profitably used for fold recognition, rather than for model building. The threading approach is designed to assess sequences as likely candidates to fit into particular folds, not to build usable models, but it can be used as a basis for homology modelling (Gibas and Jambeck, 2001). In threading, a new sequence is mounted on a series of known folds with the goal of finding a fold that provides the best score (lowest energy). Four key components of a threading approach are necessary: a) construction of a structural template library; b) development of a scoring function for the threading alignment; c) design of a search algorithm for the best threading alignment and d) evaluation of a best-scoring threading alignment (Xu and Xu, 2000). Some web servers for fold recognition are available, where the user provides the primary sequence and the results are sent by e-mail. Some available threading web based services are: 3D-PSSM (Web-based Method for Protein Fold Recognition - <http://www.sbg.bio.ic.ac.uk/~3dpssm/>) (Kelley *et al.*, 2000); PSIRED- GenTHREADER (attempts to make inferences about possible evolutionary relationships - <http://bioinf.cs.ucl.ac.uk/psipred/>) (Jones, 1999b); and UCLA/DOE Fold Server (<http://fold.doe-mbi.ucla.edu/>) (Fischer and Eisenberg, 1996).

### 2.7.3. Ab initio prediction

The *ab initio* prediction methods consist of modelling all the energetics involved in the process of protein folding and then finding the structure with lowest free energy. This approach is based on the ‘thermodynamic hypothesis’, which states that the native structure of a protein is the one for which the global free energy achieves the minimum (Bonneau and Baker, 2001). The *ab initio* prediction is clearly the most difficult one, requiring more computer processing time and more complex algorithms. However, it is the ideal method for proteins without similar structures available for homology modelling.

Every two years, structure prediction research groups compete in the community wide experiment in the Critical Assessment of Techniques for Protein Structure Prediction (CASP - <http://predictioncenter.org/>). According to CASP3 and CASP4 (Moult *et al.*, 2001), Rosetta (Simons *et al.*, 1999) is one of the best current methods for structure prediction in the absence of similarity to a known structure (Bonneau *et al.*, 2002). The field of *ab initio* methods is progressing and special potential energy functions for folding simulations are under development (Xu *et al.*, 1999). Also lattice models and genetic algorithms associated with some modified energy functions have been used for protein structure predictions (Villoutreix, 2002).

## 2.8. Availability of tools and databases

Most protein tools can be used through Web servers or downloaded from the Internet and used in local computers. Most of the tools are free of charge or with a



minimum cost to academic users, while commercial users sometimes have to pay a license fee. An excellent starting point for researchers interested in DNA and protein sequences and structures is the SRS system at: <http://srs.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-page+databanks+-newld>. Another resource for protein modelling is the Oak Ridge National Laboratory at <http://compbio.ornl.gov/structure/resource/>. The two databases used extensively in this work were the Swiss-Prot database (<http://www.expasy.org>) containing protein sequences and the PDB database (<http://www.pdb.org>) containing 3D protein structures.

## 2.9. Experimental and computational data

Nowadays, in the area of protein science, computational methods and experimental approaches are complementary. New experimental techniques rely increasingly on computer tools developed for highly specific purposes. Many researchers use computational tools routinely to study proteins and it is almost impossible to handle the amount of new protein sequences, structures and study their functions without using the power of computational processing. One example is similarity searching using the BLAST program (Altschul *et al.*, 1990; Altschul *et al.*, 1997). This program is used to compare a test nucleotide or protein sequence against other existing nucleotide or protein sequences held in databases. The use of BLAST has become a fundamental tool in biology: in the 16 years since its publication, the original paper describing BLAST has been cited over 20,000 times (ISI Web of Knowledge - <http://portal.isiknowledge.com> – last accessed 20/05/2006).

On the other hand, many results from computational tools need to be checked experimentally in the laboratory, to provide a high level of quality and confidence in the conclusion. For protein secondary and tertiary structure prediction it is usually rewarding to try different tools available to obtain a consensus prediction. Consensus and variations among different predictions may provide clues as to whether the predictions are reliable or not. Whenever any experimental information is available, a user should process the information from the available tools or at least use the information to verify the output results (Xu *et al.*, 1999).

## **Chapter 3 – Structural Bioinformatics of Membrane**

### **Proteins**

Protein modelling tools have achieved substantial advances during the past two decades; more reliable predictive software and user-friendly interfaces have been developed. Many modelling programs are freely available for academic use and can be run on any PC under the Linux operating system. Using the Internet, other modelling programs can be accessed by giving as the input the amino acid sequence and receiving the output structure as a PDB format file. However, these tools are mainly available for the analysis of soluble proteins and are not applicable to membrane proteins. Rules and programs that apply to soluble proteins are rarely appropriate to study membrane proteins (Villoutreix, 2002), creating the need for the development of new algorithms and tools specially designed for this class of proteins. Another observation is a very large gap between the number of globular protein structures and membrane protein structures: Only about 131 3D structures for membrane proteins are available, as shown at the Stephen White Laboratory home page (<http://blanco.biomol.uci.edu/>) (release 18-Jan-2004) (White and Winley, 1999); against 23,914 3D protein structures of all kinds deposited in the PDB database (release 13-Jan-2004). This gap creates many difficulties in the development of new homology-based prediction approaches for membrane proteins.

Despite difficulties in studying the structure of membrane proteins, including the intrinsic problems involved in growing crystals, many computational methods to identify potential integral membrane proteins and predict their topology from amino acid sequence have been developed. The topology of the individual membrane helices contributes towards the overall topology of the protein (Sonnhammer *et al.*, 1998). They have improved significantly in quality, providing more reliable results in terms of accurately determining TM regions (Möller, *et al.*, 2001).

Table 3.1 (taken from Chen and Rost, 2002) shows the available transmembrane region prediction programs and their respective web sites.

**Table 3.1 – Topology predictive tools**

<b>Helical membrane proteins</b>	<b>Web server address</b>	<b>Reference</b>
ALOM	<a href="http://psort.nibb.ac.jp/form.html">http://psort.nibb.ac.jp/form.html</a>	Nakai and Horton, 1999.
DAS	<a href="http://www.sbc.su.se/~miklos/DAS">http://www.sbc.su.se/~miklos/DAS</a>	Cserző <i>et al.</i> , 1997.
HMMTOP	<a href="http://www.enzim.hu/hmmtop">http://www.enzim.hu/hmmtop</a>	Tusnády and Simon, 1998.
MEMSAT	<a href="http://www.pspred.net">http://www.pspred.net</a>	McGuffin <i>et al.</i> , 2000.
KD	<a href="http://fasta.bioch.virginia.edu/fasta/grease.htm">http://fasta.bioch.virginia.edu/fasta/grease.htm</a>	Kyte and Doolittle, 1982.
PHDhtm	<a href="http://cubic.bioc.columbia.edu/predictprotein">http://cubic.bioc.columbia.edu/predictprotein</a>	Rost <i>et al.</i> , 1996.
SOSUI	<a href="http://sosui.proteome.bio.tuat.ac.jp/sosui/frame0E.html">http://sosui.proteome.bio.tuat.ac.jp/sosui/frame0E.html</a>	Hirokawa <i>et al.</i> , 1998.
TMAP	<a href="http://www.mbb.ki.se/tmap/index.html">http://www.mbb.ki.se/tmap/index.html</a>	Persson and Argos, 1994.
TMHMM	<a href="http://www.cbs.dtu.dk/services/TMHMM-2.0">http://www.cbs.dtu.dk/services/TMHMM-2.0</a>	Krogh <i>et al.</i> , 2001.
Tmpred	<a href="http://www.ch.embnet.org/software/TMPRED_form.html">http://www.ch.embnet.org/software/TMPRED_form.html</a>	Hofmann and Stoffel 1993.
TopPred2	<a href="http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html">http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html</a>	Claros and von Heijne, 1994
MPEX	<a href="http://blanco.biomol.uci.edu/mpex/">http://blanco.biomol.uci.edu/mpex/</a>	Jaysinghe <i>et al.</i> , 2000.
<b>β-sheet membrane proteins</b>		
β-strand predictor	<a href="http://www.biocomp.unibo.it">http://www.biocomp.unibo.it</a> (upon request)	Martelli <i>et al.</i> , 2002.

All the listed predictive web sites use different methods and create an output in different formats. Some results are shown as text containing the predicted TM regions and their topology, like ALOM, TopPred2, HMMTOP and MEMSAT programs. TMHMM, KD and PHDhtm give a graphical output, using a plot

representation to show the predicted TM regions. The SOSUI prediction program gives a comprehensive graphical output. It shows the TM region as a diagram with the topology and the amino acids represented as circles placed into the membrane/extra-membrane space; and also an additional helix wheel representation for each TM region is shown.

All the predictive tools are designed to identify potential TM regions and several programs can also predict the overall in-out topology of the protein in the membrane using different methods. Some methods like TMHMM and HMMTOP use a hidden Markov model to describe the architecture of an integral membrane protein. PHDhtm is based on a neural network predictor; MEMSAT uses dynamic programming to optimally thread a polypeptide chain through a set of topology models. TOPPRED identifies putative TM  $\alpha$ -helices from a standard hydrophobicity plot and then chooses the most likely topology based on the positive inside rule (von Heijne, 1992). SOSUI, SPLIT and TMPRED use various different propensity scales (Tusndy and Simon, 2001).

The early prediction methods were based on the amino acids' hydrophobicity determined by various physicochemical measurements. However, looking from the point of view of protein structure formation, parameters obtained by statistical analysis of protein sequence databases are perhaps more reliable than parameters based on hydrophobicity measures only (Tusndy and Simon, 2001). Mller and colleagues (2001), presented an evaluation of the currently best known and most widely used methods for the prediction of membrane spanning regions, TMHMM (based on the Hidden Markov Models in all its three versions) was found to be by far the best in this comparison, followed by MEMSAT.

These predictive tools are fundamental to the study of membrane proteins in terms of finding the TM region and their topology. Nevertheless, there is a great lack of tools capable of predicting the 3D structure of membrane proteins (Chen and Rost, 2002), not only to predict the 3D co-ordinates of individual amino acids, but even to predict the general associations between TM regions from which a predicted structure of a membrane protein can be generated.

One of the few predictive tools that identify the angular orientation of the TM segment is called kPROT (“knowledge-based scale for propensities residue orientation in transmembrane segments”, Pilpel, *et al.*, 1999). Using the kPROT web site (<http://bioinfo.weizmann.ac.il/kPROT/>), the user can submit the protein sequence(s) of a previously identified TM segment and the server predicts the rotational orientation of the helical segment as would be expected if it was embedded in a helical bundle within the membrane; i.e. lipid-exposed vs. protein-buried faces of  $\alpha$ -helices. The helical orientation predictions are done using the kPROT (knowledge-based Propensities for Residue Orientation in TM segments) energy-like scale. The kPROT scale gives the propensities for residue orientation (in terms of the likelihood of being buried or exposed) in transmembrane segments, giving one value for each residue. It was derived from more than 5000 non-redundant Swiss-Prot membrane protein sequences. “The kPROT value for each residue is defined as the logarithm of the ratio of its proportions in single and multiple TM spans” (Pilpel, *et al.*, 1999).

The post-genomic era will demand more tools to be able to predict 3D structures from the primary sequence of membrane proteins, due in large part to the interest of researchers and pharmaceutical companies in developing new drugs. This

thesis describes the development of software designed to predict the TM region association and 3D structure of membrane proteins starting only from sequence information.

## **Chapter 4 – Objectives of the Project**

In the light of all the difficulties related to the determination of new membrane protein structures, the objectives of this project are: (1) to create a visualisation and analysis tool specifically designed for membrane proteins, and (2) to develop a computer program (algorithm and interface) to predict 3D structures of membrane proteins from primary sequences using an association score method.

The project was divided in two distinct phases. The first one was to collect all the relevant data from PDB and Swiss-Prot files in terms of TM regions from the Swiss-Prot database and PDB  $\alpha$ -helix 3D co-ordinates. This sets up the foundation to the knowledge-based algorithm, creating a 20x20 amino acids association matrix, derived from known membrane protein PDB files. From this, it was possible predict inter-helical associations. Two software interfaces were developed for this phase: *TMCompare* (Togawa *et al.*, 2001) and *TMDistance*, which creates and evaluates the 20x20 association matrix, providing a statistical “signature” of membrane protein structure(s).

The second phase involves the prediction of inter-helical associations. This considers detailed inter-helical associations taking into account helix periodicity and the end-on orientation of individual helices. For this phase, *TMRelate* was developed. This piece of software reads as an input the 20x20 association matrix and the



membrane protein sequence file in Swiss-Prot format. The algorithm then calculates the association score between TM regions, producing as an output a graphical 2D end-on view with the associations between TM regions. *TMRelate* also rotates each TM region to find the arrangement with the best score in terms of helical periodicity.

To achieve the final stage of this project several different versions of the predictive software were developed. Versions for 12 TM regions, versions using different scales like kPROT, versions integrating the kPROT scale and the association matrix, all focussed on predicting associations between TM regions from the primary sequence. This project is in the vein of the words of Gibas and Jambeck (2001) who gave a good definition of what many researchers working in bioinformatics are seeking: “The ultimate goal of analytical bioinformaticians is to develop predictive methods that allow scientists to model the function and phenotype of an organism based only on its genome sequence. This is a grand goal, and one that will be approached only in small steps, by many scientists working together”.

The following chapters will describe in detail the development of a suite of bioinformatics tools for the analysis and prediction of membrane protein structure.

## **Chapter 5 - Material and Methods**

The software development is based on a PC platform running Microsoft Windows. All the programs use a graphical interface and are user-friendly. The development environment was Borland Delphi 5.0, an object-oriented language based on the Pascal programming language.

Borland Delphi is an object-oriented, visual programming environment for rapid application development. It is used for developing all kinds of applications and uses sophisticated data access programs. Delphi allows the creation of highly efficient applications, using a comprehensive library of reusable components and a suite of RAD design tools, including application and form templates. These tools simplify application prototyping and development and shorten development time. Another advantage using the Delphi development environment is the incorporation of Windows components into the user interface, allowing, for example, an Internet explorer web browser to be used with the CHIME plugin (MDL Information Systems, Inc. - <http://www.mdli.com/chime/>) embedded into the predictive program. It allows the displaying of protein 3D structure, without the need to use an external 3D rendering program such as Rasmol (Sayle and Milner-White, 1995), Pymol (DeLano, 2002) or Deep View Swiss-Pdb Viewer (Guex and Peitsch, 1997).

All developed programs were tested in the following versions of Microsoft Windows: 95, 98, 98SE, NT, 2000 and XP. The programs can be compiled using any version of Windows operational system; and the created executable program can also be run in any version of Windows. Due to the recent development of Kylix, a Delphi environment for Linux, in the future all the predictive tools developed will be suitable for the Linux platform as well.

The following sections describe hardware, software, databases and tables used during the project development.

### 5.1. Hardware

For the development of the programs two machines were used: (1) A Pentium II – MMX 450 MHz with 64Mb of memory located at Rothamsted Research and (2) a Pentium III 450 MHz with 64Mb located at University of Luton - UK. For the benchmarks, three other machines were used: (1) A Pentium III 600 MHz, with 256Mb of memory, (2) a Pentium Celeron 450 MHz with 128Mb of memory both located at University of Luton - UK and (3) A dual Pentium XEON 1.7Ghz with 2Gb of memory located at Embrapa – Genetic Resources and Biotechnology (Brasilia – Brazil).

### 5.2. Software

The operating systems used for the software development were Microsoft Windows XP; Microsoft Windows NT workstation service pack 5; Windows NT

server service pack 5; Windows 98SE and Windows 95. For the development of a web-based version of *TMCompare* a Microsoft PWS 4.0 was used and for the web server Microsoft IIS 4.0 was used. For the mark-up language HTML and ASP was used. The programming environment/language was the Borland Delphi 5. The 3D structure rendering was made by CHIME plugin version 2.6. For the scripting language, Rasmol (Sayle and Milner-White, 1995) commands were used. The Web Browsers used during the development and testing were Netscape 4.7.x, Microsoft Internet Explorer 5 and Opera 6.05.

### 5.3. Databases

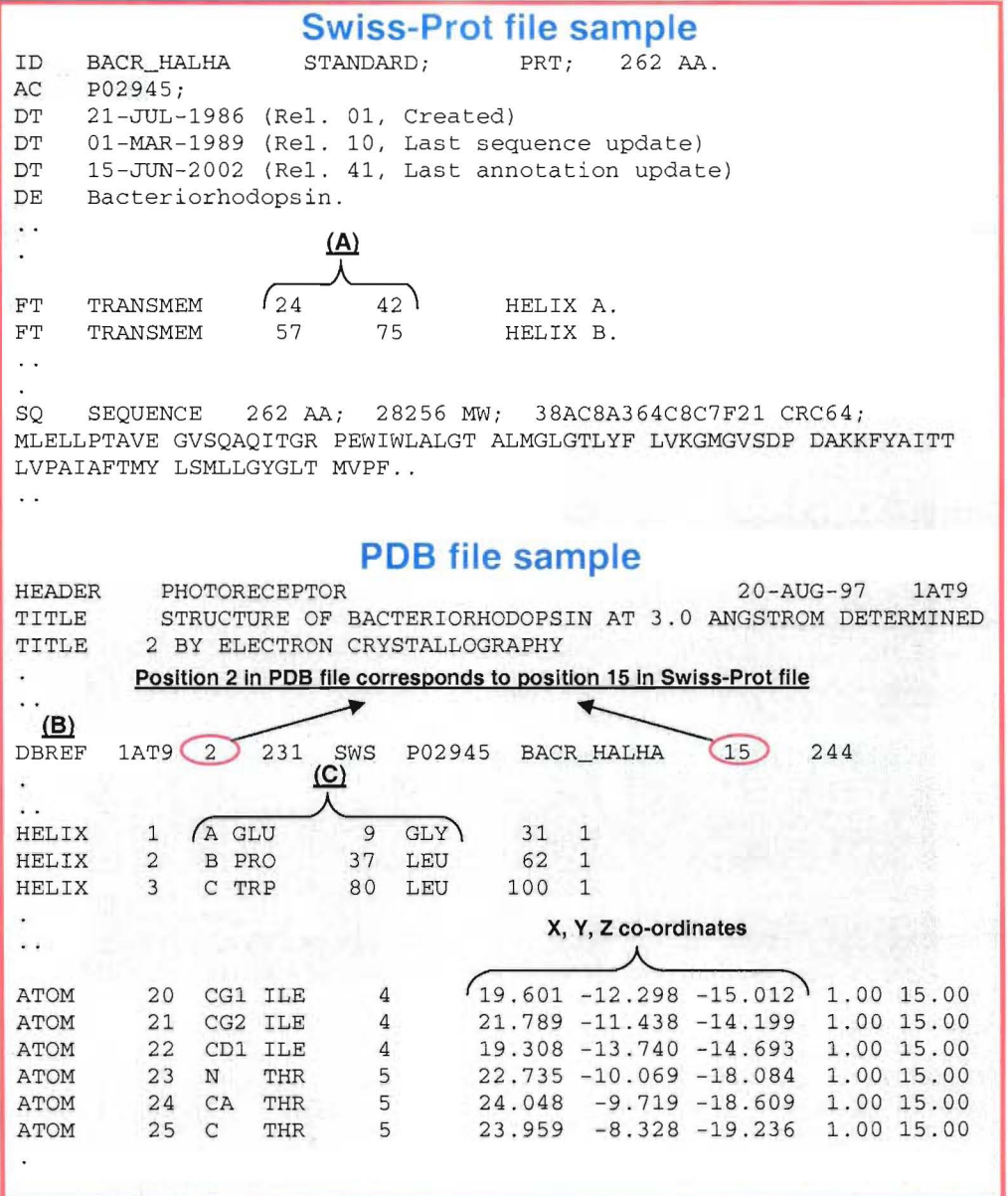
Swiss-Prot (Bairoch and Apweiler, 2000). The last audit version of was release 48.7 of 20-Dec-2005 with 204,086 entries. The programs developed in this project use the following Swiss-Prot annotations:

ID	Contains a brief description
AC	Contains Swiss-Prot accession code
DT	Contains the release information
DE	Contains the protein description
FT	Contains many annotations including 'TRANSMEM' tag which define the transmembrane regions. At the prediction pipeline the transmembrane regions are obtained using HMMTOP algorithm and the predicted transmembrane regions are placed at the FT TRANSMEM tag.
SQ	Contains the protein sequence in one letter code

PDB (Berman *et al.*, 2000). The last audit version of PDB was from 27-Dec-2005 with 34,376 structures. The programs developed in this project use the following PDB annotations:

HEADER	Contains an organism name
TITLE	Contains a brief description
DBREF	Contains an equivalence between the PDB file and protein sequence databases
HELIX	Contains an $\alpha$ -helix region. This is generated automatically by PDB site using the DSSP algorithm (Kabsch and Sander, 1983) although they may be provided by the depositor instead. The $\alpha$ -helix region is also visually checked after running a local DSSP program. In the matrix generating algorithm, this statement is overwritten with the residue details of the TRANSMEM statement from the appropriate Swiss-Prot file.
ATOM	Contains information about each residue in the structure in three-letter code, and contains the atom name, residue number, and XYZ co-ordinates.

Figure 5.1 - Sample of PDB and Swiss-Prot files



This figure shows the PDB (code 1AT9) and Swiss-Prot (accession number P02945) files for bacteriorhodopsin. (A) - Swiss-Prot TM regions (begin and end). (B) - DBREF tag - The PDB code: 1AT9 corresponds to the Swiss-Prot file with accession number P02945. (C) - PDB  $\alpha$ -helix definitions (beginning and end)

#### 5.4. Amino acid colour code

The project adopts a colour code for each residue. The residues are defined according to physical and chemical characteristics; charge (acidic, basic), polarity, hydrophobicity and variphobicity. The amino acid properties were obtained from examination of a number of scales, including the polarity scales of Zimmerman (Zimmerman *et al.*, 1968), and those of Grantham (1974), and the hydrophobicity scales of Kyte and Doolittle (1982), and those of Eisenberg (Eisenberg *et al.*, 1984).

The colour code is used to show the amino acids type used to examine amino acids at similar membrane depth ( $\pm 1.5 \text{ \AA}$ ) on different TM regions by alignment (figure A.6) and in the helix wheel representation (figure 8.1). The colours used in the project are defined in table 5.1.

**Table 5.1 - Colours used to display physical and chemical characteristics of amino acids**

ASP	D	Aspartic acid	Acidic	Red
GLU	E	Glutamic acid	Acidic	Red
ARG	R	Arginine	Basic	Blue
LYS	K	Lysine	Basic	Blue
HIS	H	Histidine	Basic	Blue
ASN	N	Asparagine	Polar	White
GLN	Q	Glutamine	Polar	White
GLY	G	Glycine	Polar	White
SER	S	Serine	Polar	White
CYS	C	Cysteine	Hydrophobic	LightGray
ALA	A	Alanine	Hydrophobic	LightGray
PRO	P	Proline	Hydrophobic	LightGray
THR	T	Threonine	Hydrophobic	LightGray
TYR	Y	Tyrosine	Hydrophobic	LightGray
PHE	F	Phenylalanine	Hydrophobic	DarkGray
ILE	I	Isoleucine	Hydrophobic	DarkGray
LEU	L	Leucine	Hydrophobic	DarkGray
MET	M	Methionine	Hydrophobic	DarkGray
TRP	W	Tryptophan	Hydrophobic	DarkGray
VAL	V	Valine	Hydrophobic	DarkGray

## 5.5. Knowledge based approaches

### 5.5.1. Association matrix

For the prediction of the TM regions, the project uses a 20x20 association matrix. This matrix was built testing all the potential associations between the TM regions of the examined integral membrane proteins, based on the information available from known 3D protein structures contained in the PDB databank repository. The association matrix was created by a module called *TMDistance* (complete description in the chapter 7), which reads the PDB file entries and calculates the distance between residues using their side chain atomic co-ordinates with the closest distance located in different TM regions. Distances less than or equal to a user-selected one are displayed on the matrix counter, so that pairs of residues within the set limit are available for later analysis.

The following criteria were considered in selecting the PDB files used to create the matrix:

- 1) alpha-helical multi-spanning membrane protein;
- 2) membrane protein structures derived from X-ray crystal experimental data with resolution of better than  $\sim 2.5\text{\AA}$ ;
- 3) If there is more than one PDB file from the same family, the one with the best resolution was selected. Also the B-factor and the missing residues were considered at this stage.

The following membrane protein PDB files composed the final selected group for the matrix creation:



Bacterial Rhodopsins				
PDB code	Description	Experiment/ Resolution	Swiss-Prot code	N. of TM regions
1C3W	Bacteriorhodopsin <i>H. salinarum</i>	X-ray 1.55 Å	P02945	7
1E12	Halorhodopsin (HR) <i>H. salinarum</i>	X-ray 1.8 Å	P16102	7
1H2S	Sensory Rhodopsin II with Transducer - <i>N. Pharaonis</i>	X-ray 1.94 Å	P42196	7
G Protein-Coupled Receptors				
1U19	Rhodopsin: Bovine Rod Outer Segment <i>B. taurus</i>	X-ray 2.2 Å	P02699	7
Other Channels				
1YMG	Aquaporin water channel: Bovine lens <i>B. taurus</i>	X-ray 2.2 Å	P06624	6
1FX8	GlpF glycerol facilitator channel <i>E. coli</i>	X-ray 2.2 Å	P11244	8
1U7G	AmtB ammonia channel (mutant) <i>E. coli</i>	X-ray 1.35 Å	P37905	11
Photosynthetic Reaction Centers				
1EYS	<i>T. tepidum</i>	X-ray 2.2 Å	P51762	5
1DXR	<i>R. viridis</i>	X-ray 2.0 Å	P06009	5
1RZH	<i>R. sphaeroides</i>	X-ray 1.8 Å	P02954	5
Photosystems				
1JBO	Photosystem I: <i>S. elongatus</i>	X-ray 2.5 Å	P25896	11
ATPase				
1T5S	E1 state with bound calcium and AMPPC P-type <i>O. cuniculus</i>	X-ray 2.6 Å	P04191	10
2BL2	Rotor of V-type Na <sup>+</sup> -ATPase <i>E. hirae</i>	X-ray 2.1 Å	P43457	4
Respiratory Proteins				
1QLA	Fumarate Reductase Complex <i>W. succinogenes</i>	X-ray 2.2 Å	P17413	5
1KQF	Formate dehydrogenase-N <i>E. Coli</i>	X-ray 1.6 Å	P24185	4
1OKC	Mitochondrial ADP/ATP Carrier: Bovine heart mitochondria <i>B. Taurus</i>	X-ray 2.2 Å	P02722	6
Oxidases				
1XME	Cytochrome C Oxidase, ba3 <i>T. Thermophilus</i>	X-ray 2.3 Å	Q5S379	13
References:				
1C3W (Luecke <i>et al.</i> , 1999), 1E12 (kolbe <i>et al.</i> , 2000), 1H2S (gordeliy <i>et al.</i> , 2002), 1U19 (okada <i>et al.</i> , 2004), 1YMG (Harries <i>et al.</i> , 2004), 1FX8 (Fu <i>et al.</i> , 2000), 1U7G (khademi <i>et al.</i> , 2004), 1EYS (Nogi <i>et al.</i> , 2000), 1DXR (Lancaster <i>et al.</i> , 2000), 1RZH (Xu <i>et al.</i> , 2004), 1JBO (Nield <i>et al.</i> , 2003), 1T5S (Sorensen <i>et al.</i> , 2004), 2BL2 (Murata <i>et al.</i> , 2005), 1QLA (Lancaster <i>et al.</i> , 1999), 1KQF (Jormakka <i>et al.</i> , 2002), 1OKC (Pebay-Peyroula <i>et al.</i> , 2003), 1XME (Hunsicker-Wang <i>et al.</i> , 2005)				

For each protein in the test set, 4 matrices were created using different distance limits (3.0 Å, 3.5 Å, 4.0 Å and 4.5 Å), excluding the protein being tested on a one-out basis.

The following is an example of the association matrix generated in the examination of 1C3W (Bacteriorhodopsin) using a distance cut-off of 3.5 Å:



	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	42	2	9	4	8	3	0	21	4	10	31	0	16	37	5	46	30	14	21	7
ARG	2	2	3	4	0	1	0	1	0	3	3	2	1	1	0	4	3	2	7	1
ASN	9	3	2	4	3	1	3	10	0	5	1	1	10	2	5	12	7	5	7	6
ASP	4	4	4	0	0	0	1	0	0	3	3	0	1	0	3	0	1	1	6	0
CYS	8	0	3	0	0	1	0	2	3	3	1	0	0	5	0	1	5	0	1	2
GLN	3	1	1	0	1	0	1	22	3	1	1	1	0	4	1	38	4	3	4	1
GLU	0	0	3	1	0	1	2	2	3	2	5	1	0	0	3	3	11	0	11	4
GLY	21	1	10	0	2	22	2	60	4	7	27	0	5	32	8	37	16	11	25	7
HIS	4	0	0	0	3	3	3	4	32	6	7	0	6	3	0	9	9	6	7	5
ILE	10	3	5	3	3	1	2	7	6	14	26	0	5	14	1	14	4	7	10	1
LEU	31	3	1	3	1	1	5	27	7	26	22	2	11	48	7	14	6	12	20	7
LYS	0	2	1	0	0	1	1	0	0	0	2	0	0	3	0	2	1	0	4	0
MET	16	1	10	1	0	0	0	5	6	5	11	0	8	9	2	5	3	8	7	5
PHE	37	1	2	0	5	4	0	32	3	14	48	3	9	18	4	15	7	9	18	13
PRO	5	0	5	3	0	1	3	8	0	1	7	0	2	4	0	21	3	2	4	1
SER	46	4	12	0	1	38	3	37	9	14	14	2	5	15	21	16	21	8	5	12
THR	30	3	7	1	5	4	11	16	9	4	6	1	3	7	3	21	2	10	22	8
TRP	14	2	5	1	0	3	0	11	6	7	12	0	8	9	2	8	10	2	13	5
TYR	21	7	7	6	1	4	11	25	7	10	20	4	7	18	4	5	22	13	8	8
VAL	7	1	6	0	2	1	4	7	5	1	7	0	5	13	1	12	8	5	8	12

### 5.5.2. Variability between datasets used in the generation of the association matrices

The variability analyses were carried out in order to gain a better statistical understanding of the composition of the association matrices. These matrices were analysed by examining the percentage of associations between each pair of amino acids. The expectation in comparing one structure or a group of structures, with another is that the higher the variance the greater the differences between the structures will be. The main test was carried out using bacteriorhodopsin, for the reason that there are more 3D structures in the PDB repository with different experimental resolutions. The graphs shown in figure 5.2 and 5.3 illustrate that the experimental resolution is an important issue in selecting the PDB files to build the dataset. The following sections discuss different tests undertaken using different datasets.

#### 5.5.2.1. Bacteriorhodopsin structures

To examine the variability between datasets, eleven structures of bacteriorhodopsin, obtained from different types of experiment and with different resolutions, were selected. The bacteriorhodopsin protein was used for the reason that it is the singly most characterized and most abundant membrane protein in terms of solved structures. The structures were divided by resolution into 2 groups: high (h) and medium (m) as shown in table 5.2.

Table 5.2 – Bacteriorhodopsin protein divided by resolution

PDB code	Experiment	resolution
1P8H (h)	X-ray	1.52 Å
1C3W (h)	X-ray	1.55 Å
1F50 (h)	X-ray	1.7 Å
1C8R (h)	X-ray	1.8 Å
1QHJ (h)	X-ray	1.9 Å
1QKP (m)	X-ray	2.1 Å
1BRX (m)	X-ray	2.3 Å
1BRR (m)	X-ray	2.9 Å
1AT9 (m)	Electron Crystallography	3.0 Å
1BM1 (m)	X-ray	3.5 Å
2BRD (m)	Electron diffraction	3.5 Å

For each protein, the associations between amino acids in different transmembrane alpha-helices were counted and grouped (high and medium). From the 210 possible residue pair associations, no associations were found for 103 of the pairs and were excluded from the analysis. An excel spreadsheet was used to tabulate the data and the variability of each of the 107 pair associations was calculated and plotted as shown in the figures 5.2 and 5.3.

Figure 5.2 – Bacteriorhodopsin variability: High resolution dataset

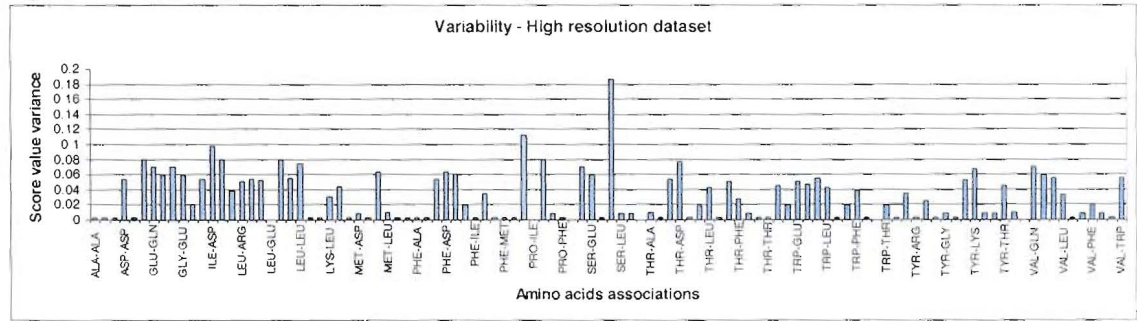
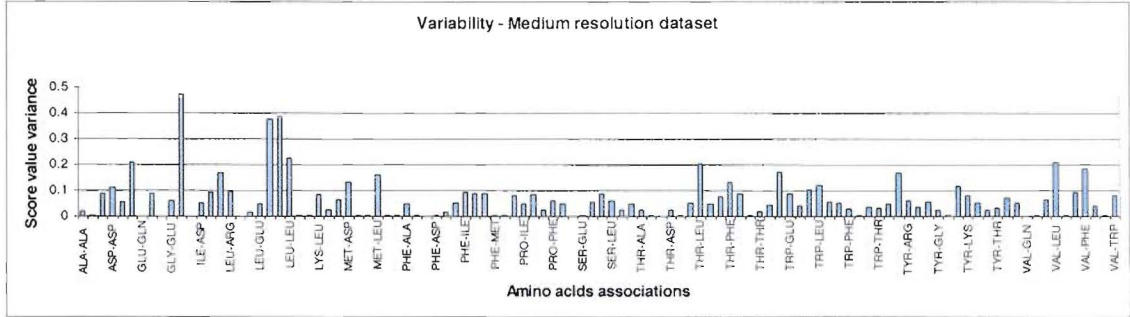


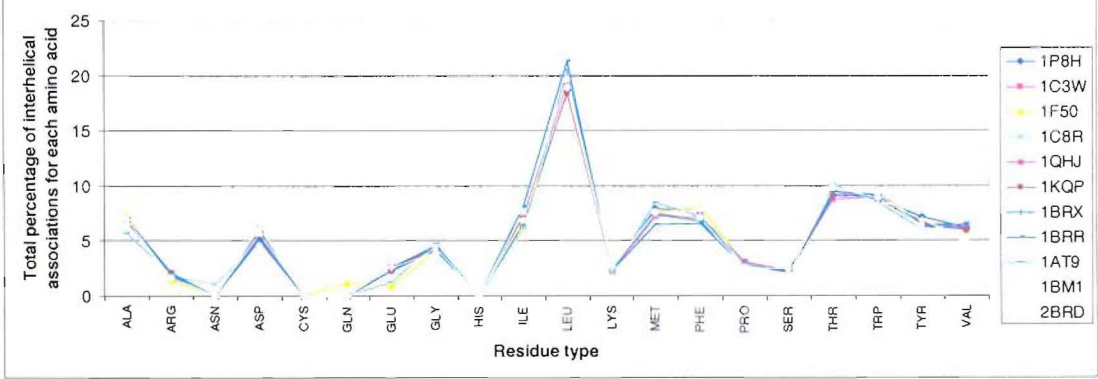
Figure 5.3 - Bacteriorhodopsin variability: Medium resolution dataset



The results show generally low variability, indicating the conserved nature of the associations between amino acids that compose the TM regions from different experiments and resolutions. Comparing the two graphs, the variability of the high resolution group is lower than the medium resolution, indicating the importance of the resolution for selection of the dataset to be used to create the association matrix.

A further analysis was carried out, counting all the inter-helical associations involving each of the 20 amino acids in different bacteriorhodopsin structures, applying a cut-off distance of 4.5 Å. The values for each amino acid was converted into a percentage of the overall number of associations and plotted as shown in the figure 5.4. The different structures share almost identical distribution in terms of number of associations for each amino acid.

Figure 5.4 - Bacteriorhodopsin variability: Inter-helical associations



5.5.2.2. All 7 TM protein structures

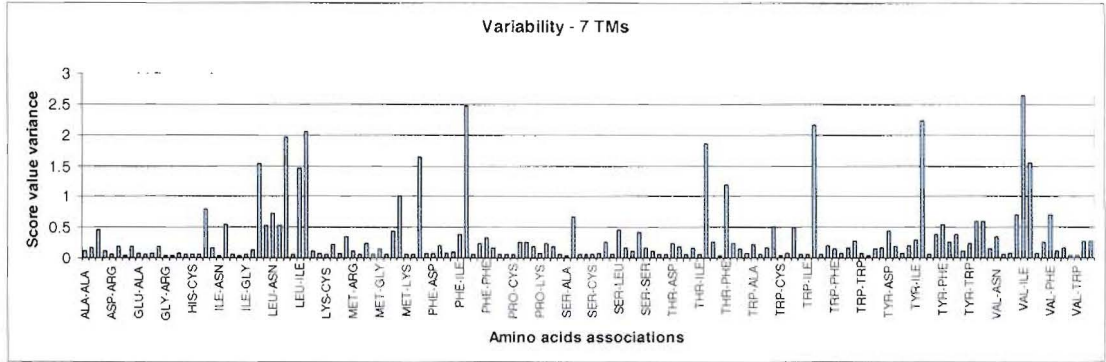
From the routinely used dataset of 17 proteins, four proteins with seven TM regions were used for this analysis as shown in table 5.3:

Table 5.3 – Proteins with Seven TM regions

PDB code	Protein / Experiment	resolution
1C3W	Bacteriorhodopsin / X-ray	1.55 Å
1E12	Halorhodopsin / X-ray	1.80 Å
1H2S	Sensory Rhodopsin II / X-ray	1.94 Å
1U19	Rhodopsin: Bovine rod outer / X-ray	2.20 Å

The same analysis was carried out as previously for bacteriorhodopsin and the results are shown in figure 5.5.

Figure 5.5 – seven TM membrane protein variability



Despite their structural similarity in all being 7 helix bundles, some differences in certain peaks are observed, probably because of differences in their sequences as shown in the following multiple alignment.

CLUSTAL W (1.82) multiple sequence alignment			
1C3W_A	-----TGRPEWIWLALGTALMGLGTL	21	
1E12_A	-----AVRENALLSSSLWVNVALAGIAIL	24	
1H2S_A	-----MVGLTTLFWLGAIGMLVGTL	20	
1U19_A	XMNGTEGPNFYVPFSNKTGVVRSPEAPQYYLAEPWQFSMLAAYMFLIMLGFPINFLTL	60	
	: : *		
1C3W_A	YFLVKGMGVSDPDACKFYAITTLVPAIAFTMYLSMLLGYGLTMVPFGG-----EQNPIY	75	
1E12_A	VFVYMGRITIRPGRPRLIWGATLMIPLVSISSYLGLLSGLTVGMIEMPAGHALAGEMVRSQ	84	
1H2S_A	AFAWAGR DAGSGE-RRYYVTLVGISGIAAVAYVVMALGVGVWPVAERT-----VF	69	
1U19_A	YVTVQHKKLRTPLNYILLNLAVADLFMVFGGFTTTLTSLHGYFVFPGPTGCNLEGGFFATL	120	
	: : :		
1C3W_A	WARYADWLFTTPLLLLDLALLVDADQGTILA--LVGADGIMIGTGLVGALTK-VYSYRFV	132	
1E12_A	WGRYLTWALSTPMILLALGLLADVDLGSFLT--VIAADIGMCVTGLAAAMTTSALLFRWA	142	
1H2S_A	APRYIDWILTTP LIVYFLGLLAGLDSREFGI--VITLNTVVMLAGFAGAMVP--GIERYA	125	
1U19_A	GGEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAF*TWVMALACAAPPLVGWSRYI	180	
	* : : : : : : : : : : *		

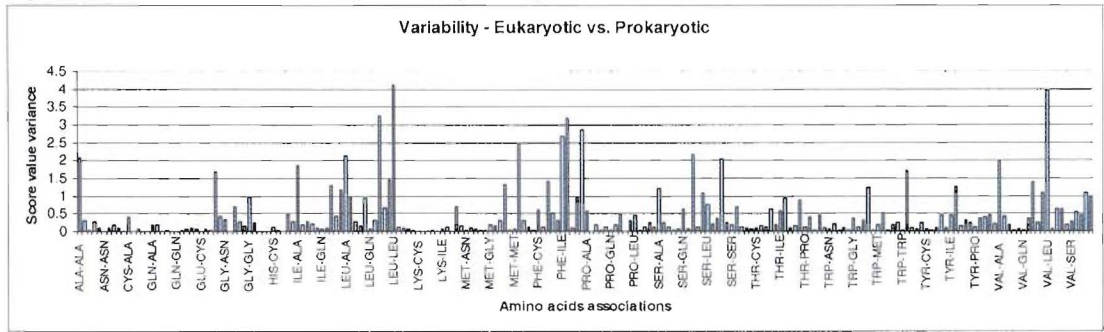


1C3W_A	WWAIST-----AAMLYILYVLF-----FGF	152
1E12_A	FYAISC-----AFFVVVLSALV-----TDW	162
1H2S_A	LFGMGA-----VAFLGLVYYLV-----GPM	145
1U19_A	PEGMQCSCGIDYYTPHEETNNESFVIYMFVVHFIPLIVIFFCYQLVFTVKEAAQQQE	240
	: : : : :	
1C3W_A	SMRP----EVASTFKVLRNVTVVLWSAYP--VVWLGISEG-----AGIVP-LNI	194
1E12_A	AASAS--SAGTAEIFDTRLRVLTVVWLWLGYP--IVWAVGVEG-----LALVQSVGA	208
1H2S_A	TESASQRSSGIKSLYVRLRNLTVILWAIYP--FIWLLGPPG-----VALLT-PTV	192
1U19_A	SATTQKAEKEVTRMVIIMVIAFLICWLPYAGVAFYIFTHQGSDFGPIFMTIPAFFAKTSA	300
	: : : : * * : : *	
1C3W_A	ETLLFMVLDVSAKVGFGILLRSRAIFG-----	222
1E12_A	TSWAYSVLDFAKYVFAFILLRWANNERTVAVAGQTLGTMSDD----	253
1H2S_A	DVALIVYLDLVTKVGGFGFIALDAAATLRAEHGE-----	225
1U19_A	VYNPVIYIMMNKQFRNCMVTTLCCGKNPLGDDEASTTVSKTETSQVAPA	349
	: : : : :	

5.5.2.3. Eukaryotic vs. prokaryotic

The 17 proteins were divided into 2 groups (eukaryotic and prokaryotic). 4660 associations were observed between amino acids in different transmembrane regions. 950 associations were found for the eukaryote structures and 3799 for the prokaryote structures. The variability analysis was carried out and the results are shown in figure 5.6.

Figure 5.6 – Variability – Eukaryotic vs. prokaryotic



The variability between these groups is relatively high, with greater differences observed than for any other comparison undertaken.

5.5.2.4. Photosynthetic reaction centres

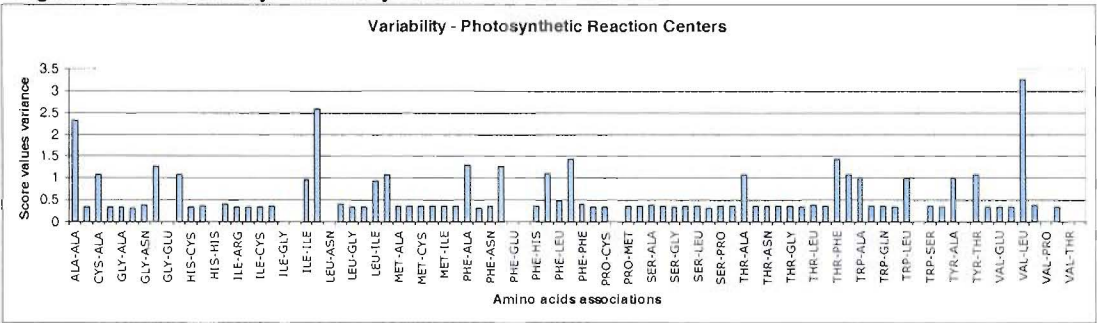
Three structures were analysed as show in table 5.4:

**Table 5.4 – Photosynthetic reactions centres**

PDB code	Protein / Experiment	resolution
1EYS	<i>T. tepidum</i> / X-ray	2.2 Å
1DXR	<i>R. viridis</i> / X-ray	2.0 Å
1RZH	<i>R. Sphaeroides</i> / X-ray	1.8 Å

The results are shown in figure 5.7

**Figure 5.7 – Variability – Photosynthetic reaction centres**



The results show low variability for the most part, with minor differences in certain peaks probably due to sequence differences.

### 5.5.2.5. High resolution vs. medium resolution

The resolution is measured in Å (angstrom) units; the smaller this number is, the higher the resolution and therefore the greater the amount of detail that can be seen (Branden and Tooze, 1999).

The group of 17 proteins used to create the association matrix was divided into 2 groups by the resolution of the crystal structures. The high resolution group was composed of those with resolution equal to or better than 2.0 Å, and those lower than 2.0 Å were considered to be of medium resolution.

The resolution groupings are:

**Table 5.5 – The group of 17 proteins divided into high and medium resolution**

PDB code	Description	Experiment/Resolution
High resolution		
1C3W (P)	Bacteriorhodopsin <i>H. salinarum</i>	X-ray 1.55 Å
1E12 (P)	Halorhodopsin (HR) <i>H. salinarum</i>	X-ray 1.8 Å
1U7G (P)	AmtB ammonia channel (mutant) <i>E. coli</i>	X-ray 1.35 Å
1RZH (P)	<i>R. sphaeroides</i>	X-ray 1.8 Å
1KQF (P)	Formate dehydrogenase-N <i>E. Coli</i>	X-ray 1.6 Å
1H2S (P)	Sensory Rhodopsin II with Transducer – <i>N. Pharaonis</i>	X-ray 1.94 Å
1DXR (P)	<i>R. viridis</i>	X-ray 2.0 Å
Medium resolution		
1U19 (E)	Rhodopsin: Bovine Rod Outer Segment <i>B. Taurus</i>	X-ray 2.2 Å
1OKC (E)	Mitochondrial ADP/ATP Carrier: Bovine heart mitochondria <i>B. Taurus</i>	X-ray 2.2 Å
1YMG (E)	Aquaporin water channel: Bovine lens <i>B. Taurus</i>	X-ray 2.2 Å
2BL2 (P)	Rotor of V-type Na <sup>+</sup> -ATPase <i>E. hirae</i>	X-ray 2.1 Å
1FX8 (P)	GlpF glycerol facilitator channel <i>E. coli</i>	X-ray 2.2 Å
1QLA (P)	Fumarate Reductase Complex <i>W. succinogenes</i>	X-ray 2.2 Å
1EYS (P)	<i>T. tepidum</i>	X-ray 2.2 Å
1XME (P)	Cytochrome C Oxidase, ba3 <i>T. Thermophilus</i>	X-ray 2.3 Å
1JBO (P)	Photosystem I: <i>S. elongates</i>	X-ray 2.5 Å
1T5S (E)	E1 state with bound calcium and AMPPC P-type <i>O. cuniculus</i>	X-ray 2.6 Å

The variability analysis of high vs. medium resolution shown in figure 5.8, showed generally low variability between the two groups, though some differences were observed. The low variability is an indication of consistent distance associations between particular pairs of residues in reliable structures. Another analysis of the variability between high resolution structures and a further group of significantly lower resolution structures showed higher variability, shown in figure 5.9, indicating less consistent distance associations between these groups. For this group, the following proteins were selected:

Table 5.6 – significantly lower resolution structures

PDB code	Description	Experiment/ Resolution
1S5L	Photosystem II from thylakoid membranes of chloroplasts - <i>Synechococcus elongates</i>	X-ray / 3.5 Å
1FFT	Cytochrome c oxidase - <i>Escherichia coli</i>	X-ray / 3.5 Å
1IWO	Calcium ATPase - <i>Oryctolagus cuniculus</i>	X-ray / 3.1 Å
1LOV	Fumarate reductase - <i>Escherichia coli</i>	X-ray / 3.3 Å
1PV7	Transport protein - <i>Escherichia coli</i>	X-ray / 3.6 Å
1Q90	B6F complex - <i>Chlamydomonas reinhardtii</i>	X-ray / 3.1 Å
1BL8	Potassium channel - <i>Streptomyces lividans</i>	X-ray / 3.2 Å
1UAZ	Ion-translocating microbial rhodopsin - <i>Halobacterium sp.</i>	X-ray / 3.4 Å
1PW4	Glycerol-3-phosphate transporter - <i>Escherichia coli</i>	X-ray / 3.3 Å
1IJD	Light-harvesting complexes from cytoplasmic membranes of bacteria - <i>Rhodopseudomonas acidophila</i>	X-ray / 3.0 Å

References:  
1S5L (Ferreira *et al.*, 2004), 1FFT (Abramson *et al.*, 2000), 1IWO (Toyoshima and Nomura, 2002), 1LOV (Mignon *et al.*, 2002), 1PV7 (Abramson *et al.*, 2003), 1Q90 Stroebel *t al.*, 2003), 1BL8 (Doyle *et al.*, 1998), 1UAZ (Enami *et al.*, 2003), 1PW4 (Huang *et al.*, 2003), 1IJD (McLuskey *et al.*, 2001).

Figure 5.8 – Variability – high vs. medium

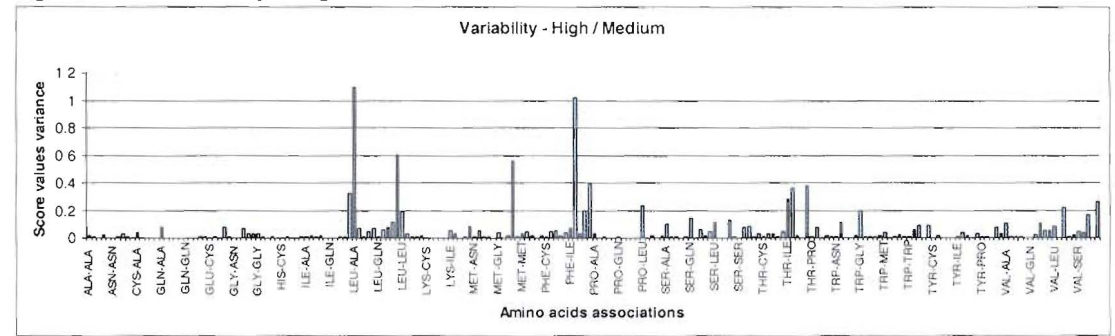
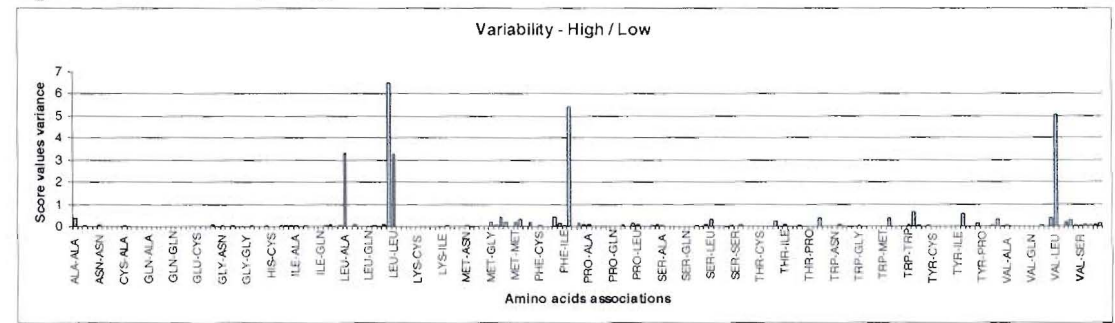


Figure 5.9 – Variability – high vs. low



These analyses suggest that the resolution of membrane protein structures is an important consideration in the selection of structures for incorporation in the standard datasets, but that structures with resolution up to 3.0 Å are generally reliable in possessing consistent inter-helical associations.



### 5.5.3. kPROT scale

The “knowledge-based scale for propensities residue orientation in transmembrane segments (kPROT)” was used in the development of *TMRelate\_K* (chapter 8). The kPROT scale is available at <http://bioinformatics.weizmann.ac.il/kPROT/kPROTScales> and uses the knowledge-based propensities for residue orientation in TM segments derived from information on all  $\alpha$ -helical TM protein sequences in the Swiss-Prot database. It gives a value for each amino acid (table 5.7). The kPROT scale is based on the idea that a higher abundance of a residue in the TM segments of multi-span proteins indicates an enhanced propensity to face the protein's interior. In contrast, a higher abundance of a residue in the TM segments of single-span proteins indicates that it has a higher tendency to be exposed to the lipid phase. In the kPROT scale, the transmembrane helix orientation propensity of each residue is related to the ratio of the two abundances. The kPROT value for residue *i* is defined as:

$$\text{kPROT}^i = \ln \left[ \frac{f_s^i}{f_m^i} \right]$$

Where  $f_s^i$  and  $f_m^i$  are the abundances of the residue in the total set of TM segments of proteins with single and multiple spans, respectively (Pilpel *et al.*, 1999).

**Table 5.7 - The used kPROT scale**

kPROT Scale	
Residue	Value
A	0.0193
C	0.2672
D	-0.8658
E	-0.8551
F	-0.1126
G	-0.1247
H	-0.3423
I	0.1248

K	0.2451
L	0.1908
M	-0.3120
N	-0.6757
P	-0.5092
Q	-0.5367
R	0.1782
S	-0.2141
T	-0.0162
V	0.2281
W	-0.1157
Y	-0.1175

### 5.6. Distances and angles used in the project

During the project development, several different distances between residues were applied. To calculate the distance between inter helical residues, using their side chain co-ordinates, as part of the process of creating the association matrix the user can select from 3.0 Å to 5.0 Å, with 0.5 Å increments. These distances were chosen on the basis of the types of interactions between amino acid residues and the forces controlling the protein structure.

Another distance used was the rise along the  $\alpha$ -helix for each amino acid that is 1.5 Å (Branden and Tooze, 1999). Using this distance, the algorithm calculates the membrane thickness, by multiplying the number of amino acids in the shortest TM region by 1.5 Å. The predictive program also considers the intra-membrane amino acid depth (more details in Chapter 8 – *TMRelate*). For each pair of amino acids in different TM regions, if the designated depth values for the amino acids are less than 1.5 Å apart, the program takes the appropriate value from the 20x20 matrix, and an accumulative score is calculated for the predicted association between each pair of TM regions.

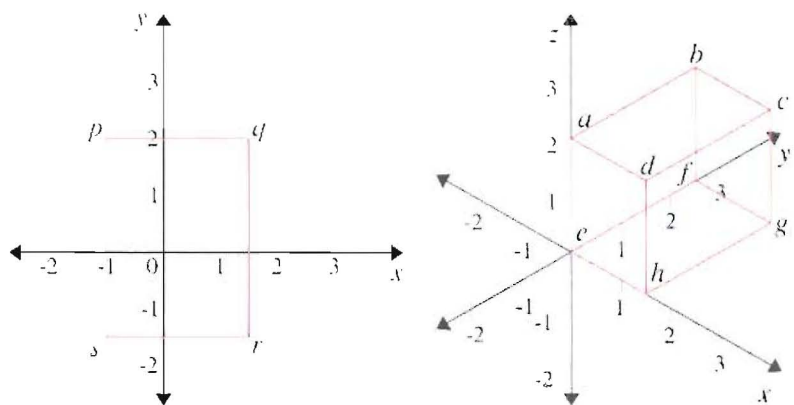
For the rotational score the predictive tool considers the angle range between the 2 residues. This angle is used to simulate the TM region rotation. For each residue an angle value is given. This angle corresponds to the rotational position in the  $\alpha$ -helix. When proteins of known three-dimensional structure are examined, it is found that sequences that form helices tend to have, on average, a strong periodicity of 3.6 residues, the period of the alpha helix, making  $100^\circ$  the difference between each amino acid (Eisenberg *et al.*, 1984). The angle zero is given to the first residue in each  $\alpha$ -helix and an increment of  $100^\circ$  is added for the next residue position. Starting from one side of the membrane, the angle is increased by  $100^\circ$  for each subsequent amino acid, and starting from the other side, the angle is decreased by  $100^\circ$ . During the calculation of the rotational score the program simulates the rotation of each TM region using the given positional angle for each amino acid. If the angle range between the 2 residues on different helices undergoing simulated rotation are equal or less than  $60^\circ$  apart, a score for these two residues is added. The detailed explanation of how these angles are used for the predictive tool is in the section describing *TMRelate* (chapter 8).

### 5.7. Cartesian mathematics

Cartesian coordinates are rectilinear two or three-dimensional coordinates. The two dimension Cartesian coordinate system is commonly defined by two axes, at right angles to each other, forming a plane (an xy-plane). The horizontal axis is labelled x, and the vertical axis is labelled y. In a three dimensional coordinate system, another axis, normally labelled z, is added, providing a third dimension of

space measurement. The axes are commonly defined as mutually orthogonal to each other.

The diagram below shows a two-dimensional Cartesian graph (on the left) and a three-dimensional Cartesian graph (on the right):



Where:

$p = (-1, 2),$ $q = (1.5, 2),$ $r = (1.5, -1.5)$ and $s = (-1, -1.5)$	$a = (0, 0, 2),$ $b = (0, 2.5, 2),$ $c = (1.5, 2.5, 2),$ $d = (1.5, 0, 2),$ $e = (0, 0, 0),$ $f = (0, 2.5, 0),$ $g = (1.5, 2.5, 0)$ and $h = (1.5, 0, 0)$
--	--

5.7.1. The use of Cartesian mathematics

In the development of a piece of software that predicts the 3D structures for membrane proteins based on  $\alpha$ -helical secondary structure, the following principles were considered: (1) For an  $\alpha$ -helix, the angle of rotation around the N- $C_\alpha$  bond is termed phi ( $\phi$ ) and is  $-60^\circ$ , while the angle of rotation around the  $C_\alpha$ - $C'$  bond is termed psi ( $\psi$ ) and is characteristically  $-50^\circ$ . (Ramachandran *et al.*, 1963). (2) An  $\alpha$ -

helix has 3.6 amino acids and a rise of 5.4 Å per turn, giving a rise along the helix of 1.5 Å for each amino acid. This arrangement gives a 100° difference between each amino acid, facilitating 2D representation. The basic 3.6-amino-acid-per-turn motif and the right-handedness of each individual  $\alpha$ -helix in membrane proteins are consistently maintained (Popot and Engelman, 2000). With the number of 3.6 amino acids per turn the two-dimensional Cartesian coordinates were used to build the helix wheel output and the predicted end-on configuration. The helix wheel representation is useful in terms of visualising the relative positions of particular amino acids, and if a combined representation for more than one TM region may be generated, then this provides a graphical representation of putative side chain interactions between TM regions.

The three-dimensional Cartesian coordinates were used in order to build the  $\alpha$ -helix (TM region) 3D structure compatible with the PDB file format. Principles 1 and 2 were used to build the 3D model.

### 5.8. Permutation process

A permutation of objects is an arrangement of those objects in different order; one object is placed in the first position, another in the second position, and so on, until all objects have been placed in all positions. The number of permutations of a set of  $n$  elements is denoted  $n!$  ( $n$  factorial). Thus  $n!$  is the number of ways to count a set of  $n$  elements. For example, for the set of elements {1,2,3} there are just six ways to count these three elements: ( $3! = 6$ )

{1,2,3} {1,3,2} {2,1,3} {2,3,1} {3,1,2} {3,2,1}.

The use of permutation is one of the strengths of this project. It was applied to place all TM regions in all the pre-determined positions. Therefore, by checking all permutations we are able to test every single possibility and obtain the best score based on the 20x20-association matrix.

In the programming process, a string of 10 digits (predict up to 10 TM regions) was used to create the permutations. The string {1234567---} is an example of 7 TM regions using 10 different positions. To predict 11 and 12 TM regions another group of string was used: {123456789ABC}. The letters 'A', 'B' and 'C' represent the TM regions 10, 11 and 12. These letters, which are the same as used in the hexadecimal representation, facilitate the programming task in terms of string manipulation and speeding up the processing time.

To run the predictive tool, the developed software uses a textual file containing all non-repeated permutations. The use of such files (one different for each number of TM regions) prevents the creation of all permutations discarding the repeated positions every time the program runs. For example, looking at the string: {123--4765-}, changing the 4<sup>th</sup> and 5<sup>th</sup> positions results in the same string. In this case one of the repeated string needs to be discarded.

To create the permutation files a PERL program (Christiansen and Torkington, 1997) under the UNIX environment was used. After creating the file with all permutations, the next step was to discard the repeated ones. For this step a UNIX *sort* command with a '-u' (unique) parameter was used and the final permutation file was ready to be used in the *TMRelate* program. Table A.3 shows some examples of the permutation file and table A.4 shows the used permutation file names and its statistics.

## **Chapter 6 - TMCompare**

### **6.1. Introduction**

The development of this project required manipulation and understanding of information related to membrane proteins in terms of protein sequence stored in the Swiss-Prot database (Bairoch and Apweiler, 2000) and protein 3D structures stored in the PDB data bank (Berman *et al.*, 2000). From the outset, it was necessary to know how sequence and structure are related.

Different pieces of information can be extracted from the available PDB files. However, for researchers working in the membrane protein area, there is no information in PDB files relating to where transmembrane (TM) regions begin and end.

From the Swiss-Prot file, required information relating to membrane proteins are found at the TRAMSMEM tag. The definition of the TM region(s) is obtained from laboratory experimentation or computer based prediction methods. But, no information relating to both 3D spatial co-ordinates and transmembrane regions can be obtained from a single source.

So, there was a need to integrate the contents of Swiss-Prot and PDB files in order to study and understand the relationship between membrane protein sequence

and structure, showing the correct position of the TM region in the 3D structure. The Swiss-Prot and PDB files are cross-referenced making it possible to use the information contained to examine the relationship between the sequence and structure of membrane proteins.

*TMCompare* (Togawa *et al.*, 2001) is a Windows based program and was created to examine PDB and Swiss-Prot TM information in an automated and highly visual way in terms of comparing sequence alignments and 3D structures. A web-based version of *TMCompare* has been developed and is available on the Internet at the address: <http://membraneproteins.swan.ac.uk/tmcompare/>.

## 6.2. Description

*TMCompare* allows visual comparison for individual proteins of the annotations of the PDB database (definitions of helical regions) and the Swiss-Prot database (definitions of transmembrane regions).

*TMCompare* shows the sequence and structure information in 2 separate frames.

The upper frame (figure 6.1) is the 3D view of the protein shown by the CHIME plugin, with all the usual functions for molecule rotation, slab plane, zoom and other rendering. Added to the CHIME 3D viewer frame are three buttons. The left most button may be used to colour the molecular structure according to the schemes for annotated PDB helices, the middle button colours the Swiss-Prot TM regions based on the TRANSMEM tag found of the Swiss-Prot files and the right one displays only TM regions. This colour scheme matches the determined order of



colours used in the PDB / Swiss-Prot alignment graphic in the lower frame. The following figure shows the 3D-structure frame and its functionality.

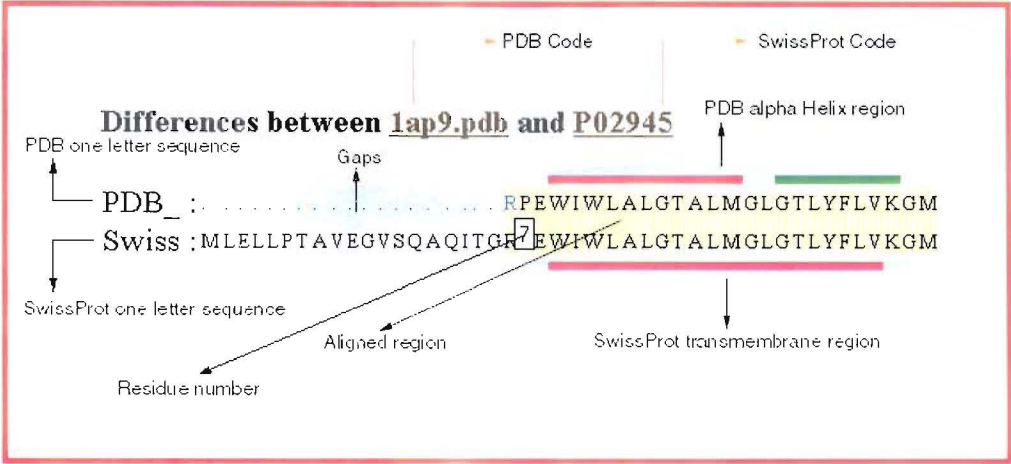
**Figure 6.1 – TMCompare: Structure frame**



This figure shows the three different selections (by each button) in the structure frame. 1) Colouring the selection of the PDB definition using “Helix” tag. 2) Colouring the selection of the TM region as defined in Swiss-Prot file using the “TRANSMEM” tag. 3) Same as (2) but displaying only the TM regions. The structure shown is a *Bacteriorhodopsin* with PDB code 1AP9 and Swiss-Prot accession number P02945.

The lower frame (figure 6.2) displays the sequence coverage of the PDB and Swiss-Prot files sequences being compared. The sequence corresponding to regions of the protein covered by PDB co-ordinates is shown on the top line, while the sequence contained in the corresponding Swiss-Prot sequence file is shown below. The amino acid residues identical in both aligned sequences are coloured yellow, while dots are used to indicate gaps in coverage, usually in the PDB file. A series of colours (up to 35 different ones) are used in a determined order to indicate regions annotated as  $\alpha$  helical structures (HELIX tag) in PDB files and regions annotated as TM regions (TRANSMEM tag) in Swiss-Prot files. Moving the cursor over the sequence displays the residue number. There is an additional option that allows viewing of the text content of the loaded Swiss-Prot and PDB files.

Figure 6.2 – *TMCompare*: Sequence frame details



This figure illustrates the sequence frame created by *TMCompare*

6.3. The algorithm

See complete algorithm description at the appendix II-1.

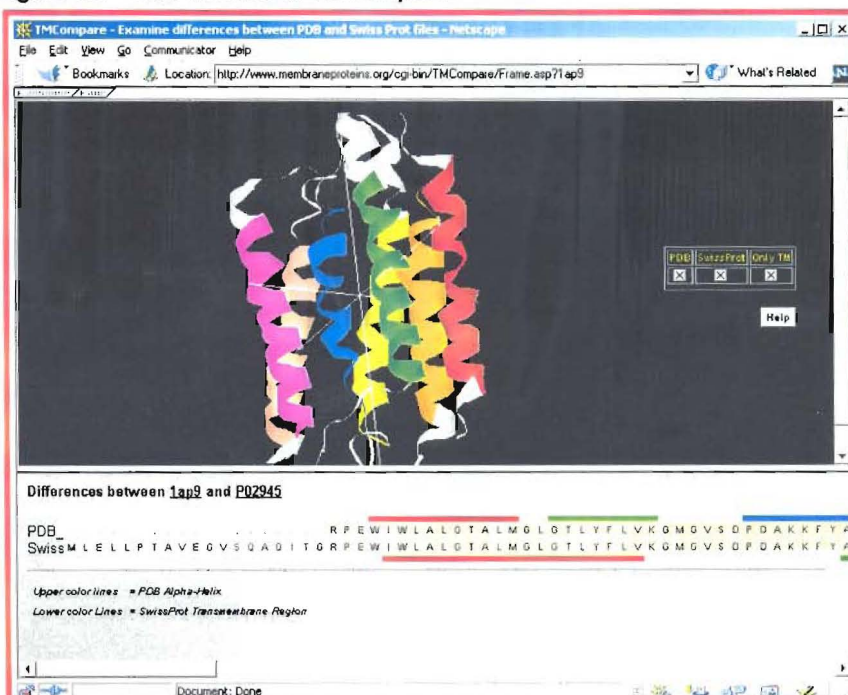
6.4. Software development

For the generation of the *TMCompare* output, Delphi uses a HTML component, introducing a web page into the user interface. It also facilitated the development of a web-based version of *TMCompare*, which is available at the address: <http://membraneproteins.swan.ac.uk/tmcompare/>.

The following figures (6.3 and 6.4) show the *TMCompare* stand-alone and web version running.

Figure 6.3 - Stand-alone version of *TMCompare*

This figure shows the stand-alone *TMCompare* program running. It is working with the *Bacteriorhodopsin* with PDB code 1AP9 and Swiss-Prot accession number P02945.

Figure 6.4 - Web version of *TMCompare*

This figure shows the web version of *TMCompare* running using the CHIME plugin to render the 3D structure. It is working with the same *Bacteriorhodopsin* with PDB code 1AP9 and Swiss-Prot accession number P02945. The web address of *TMCompare* is: <http://membraneproteins.swan.ac.uk/tmcompare/>.



### 6.5. Discussion

The outcome of the described development of *TMCompare* facilitates a better understanding of the relationship between Swiss-Prot protein sequence and PDB protein structure files. The acquired information was very useful in the subsequent development of *TMDistance* module (that creates the association matrix - chapter 7). A paper describing *TMCompare* was published in the December/2001 issue of *Bioinformatics* journal and the web version of the software is available through the web at the address: <http://membraneproteins.swan.ac.uk/tmcompare/>.

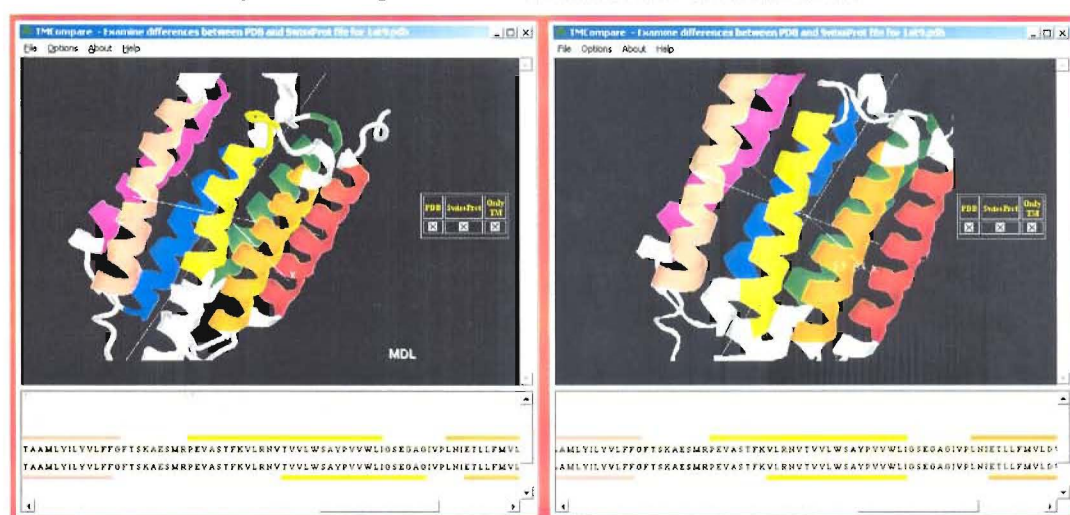
This module provided a very good foundation to the project and it formed the basis for the next module (*TMDistance*) that creates the association matrices.

In terms of its' scientific contribution, *TMCompare* can be used to observe and evaluate PDB and Swiss-Prot annotations. Other researchers working in the membrane protein area can use the web version of *TMCompare* to observe and compare the relationships between TM regions and  $\alpha$ -helices. For example, in testing *TMCompare*, an interesting observation was made: executing it with the PDB code 1AP9 and Swiss-Prot accession number P02945 (for the *Bacteriorhodopsin*) Swiss-Prot version 38, a misalignment was observed in TM regions 3 and 6. However, using the P02945 file from the next release of Swiss-Prot (version 40), the 2 TM misalignments had been corrected. The latest version has correct TM region annotation as shown in the following:

(rel. 38, Last annotation update) 01-JUN-1999					(Rel. 40, Last annotation update) 16-OCT-2001				
FT	TRANSMEM	23	42	HELIX A.	FT	TRANSMEM	24	42	HELIX A.
FT	TRANSMEM	57	76	HELIX B.	FT	TRANSMEM	57	75	HELIX B.
PT	TRANSMEM	95	114	HELIX C.	PT	TRANSMEM	92	109	HELIX C.
PT	TRANSMEM	121	140	HELIX D.	PT	TRANSMEM	121	140	HELIX D.
PT	TRANSMEM	148	167	HELIX E.	PT	TRANSMEM	148	167	HELIX E.
PT	TRANSMEM	191	210	HELIX F.	PT	TRANSMEM	186	204	HELIX F.
PT	TRANSMEM	217	236	HELIX G.	PT	TRANSMEM	217	236	HELIX G.

*TMCompare* can be used to illustrate the variability in the annotations of the Swiss-Prot and PDB databases, and also indicates the potential functionality of *TMCompare* as a useful tool for examining membrane protein sequences and structures (see figure 6.5). *TMCompare* is also useful for identifying any kind of annotation error or omission, like with the PDB “DBREF” tag and the Swiss-Prot “TRANSMEM” tag that defines the TM region.

Figure 6.5 – *TMCompare* running with different versions of Swiss-Prot file



The left image shows *TMCompare* running using Swiss-Prot accession number P02945 (*Bacteriorhodopsin*) Swiss-Prot version 38 and is observed a misalignment in the TM regions 3 and 6 (blue and yellow). The right image is using version 40 of the same accession number, and is observed the correction in the TRANSMEM annotation.

The development of a second version of *TMCompare* has been started. This new version will show the usual differences between the Swiss-Prot and PDB files but also the calculated position of the two faces of the membrane. The “limits” of the membrane are calculated using the last and first  $\alpha$ -carbon atomic co-ordinate from each end of the TM regions as defined by the Swiss-Prot database. Using these co-ordinates, a central point for the each side of the membrane is calculated. From this central point a circle is radially projected. The program creates a PDB file with added

object co-ordinates that correspond to the series of radials surrounding the central point.

#### 6.5.1. TMLimits algorithm

See complete algorithm description at the appendix II-2

## **Chapter 7 – *TMDistance***

### **7.1. Introduction**

In this project, PDB files are used in association with information from the related Swiss-Prot file. The *TMDistance* module uses sequence and structural information from amino acids in TM regions of membrane proteins, to create a 20x20 amino acid association matrix based on the determined membrane protein structures. The created association matrix may be used as a basis to predicting likely associations in membrane proteins of undetermined structure, directly from the primary sequence.

The developed software interface called *TMDistance* uses the information available from known 3D structures available in the PDB databank repository (<http://www.pdb.org>). *TMDistance* creates a 20x20-association matrix by reading the PDB file entries (atomic co-ordinates) and calculating the distance between two residues in different TM regions. If this distance is less than a given (user-selected) distance, then each pair of residues is added to the matrix counter for later analysis. This matrix forms the basis of the association score method described in the next chapter. The *TMCompare* program (Chapter 6) was used to verify TM region

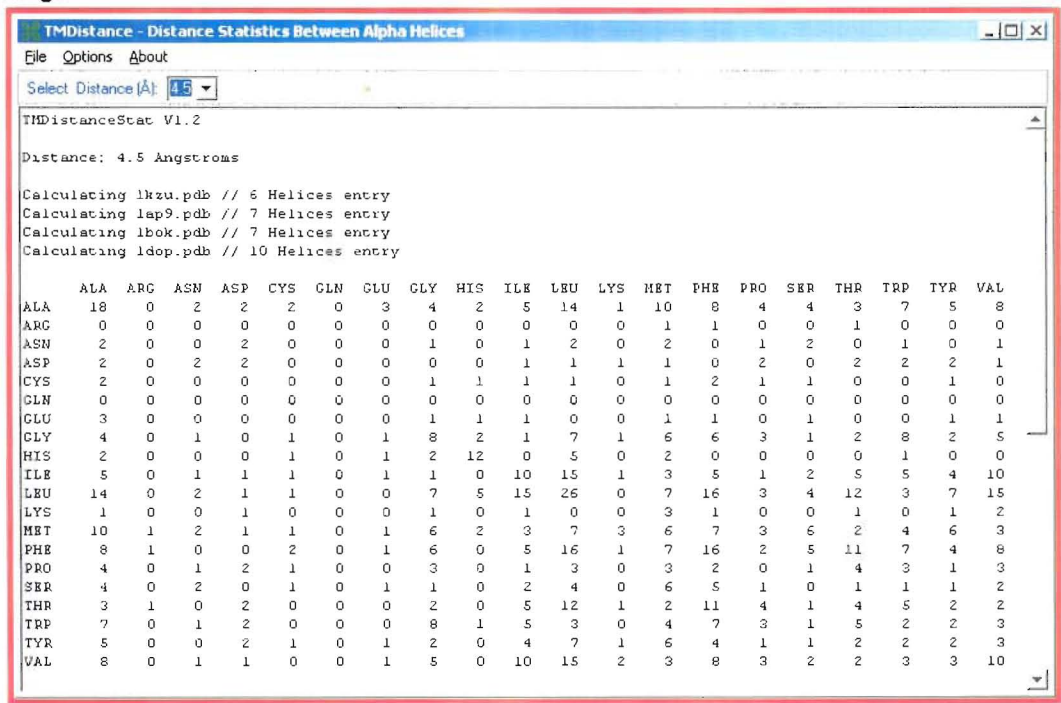
coverage in PDB and Swiss-Prot files. This chapter will describe the way that the association matrix is constructed.

7.2. Description

To use *TMDistance*, the user chooses one or more PDB files, and then initiates the calculation.

The output consists of a matrix containing the numbers of associations between the 20 different amino acids in different TM regions (figure 7.1). Only the closest pair of atoms is considered for the purpose of counting the number of associations within a given distance range, each association within range adding one to the appropriate residue pair score in the matrix.

Figure 7.1 - 20x20 association matrix with the number of associations within given distance range

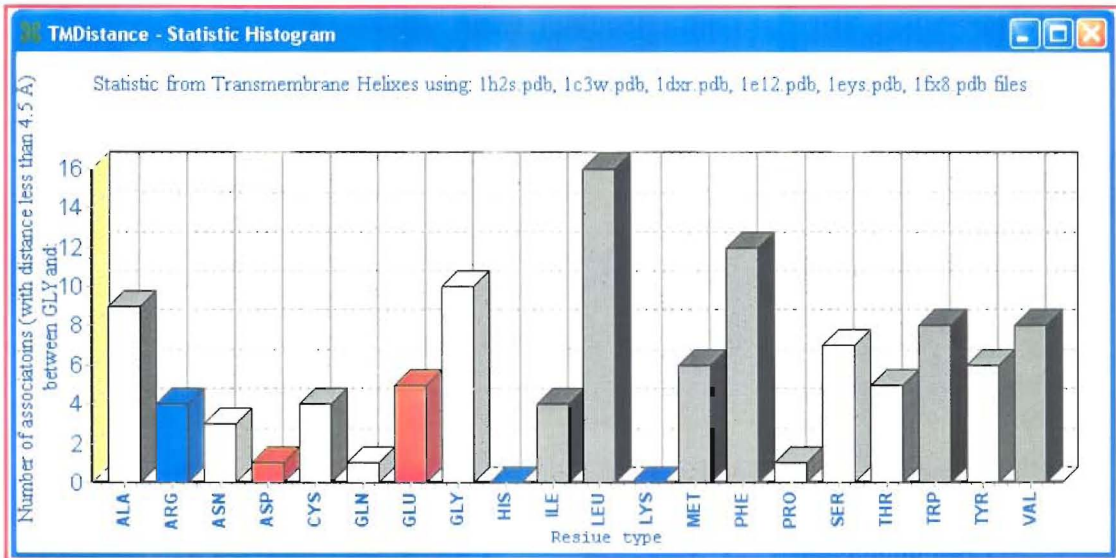


This figure shows a 20x20 association matrix generated by *TMDistance*, the chosen distance was 4.5 Å. This matrix was created based on 4 PDB files with different numbers of TM regions.



The user can also display the 20x20 association matrix as a histogram to examine the number of proximities between each amino acid residue type, and atom type (figure 7.2). This histogram may be used to indicate the most likely amino acid residues to be associated in different TM regions, and the frequency of each amino acid residue association.

**Figure 7.2 – Number of association histogram**

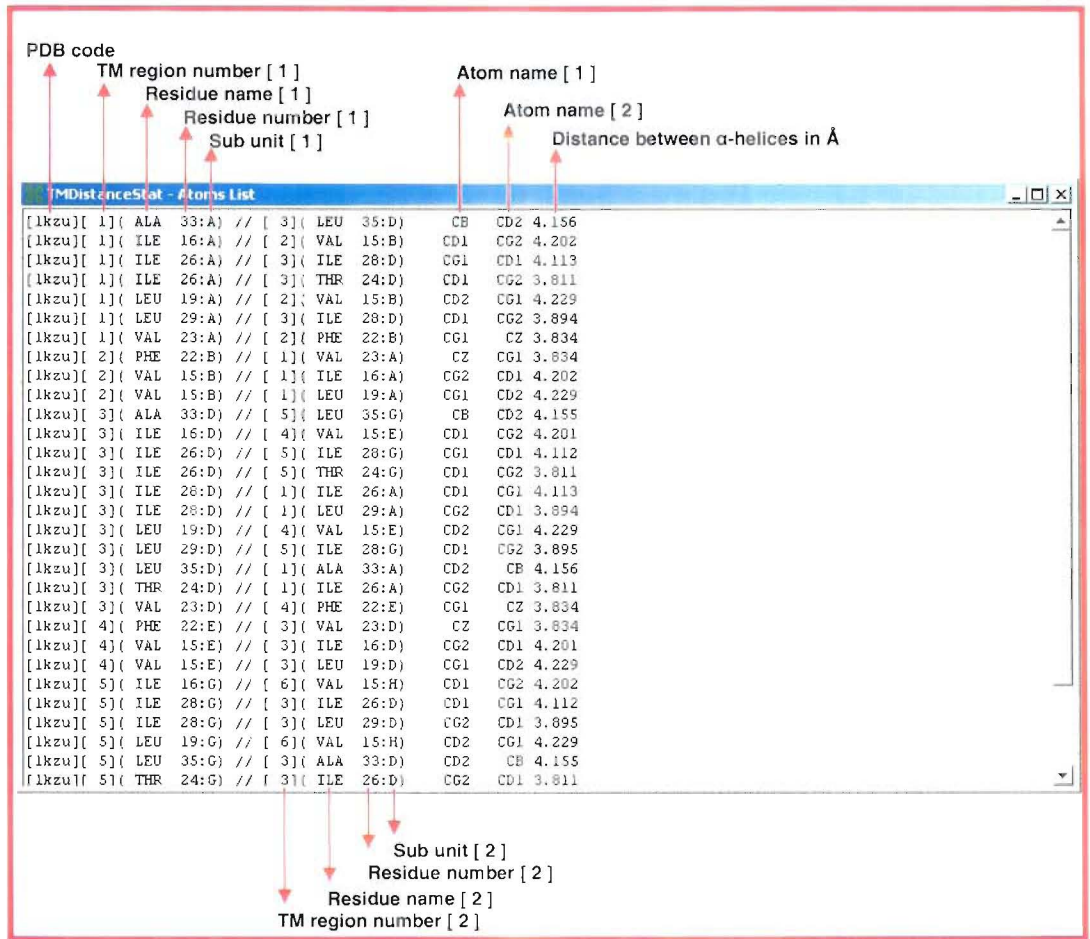


This figure shows a sample histogram produced by *TMDistance*. It shows the number of associations between glycine and the 20 different amino acids with a distance less than 4.5 Å.

By grouping the membrane proteins by specific family of known 3D structure, like the rhodopsin family, cytochrome oxidase, etc, it is possible to create a family-specific association matrix, providing the basis for better quality prediction of associations for related proteins.

A list of all associations is also available as an option in the *TMDistance* as shown in figure 7.3. This list gives complete information about the associations: the residue name, the residue number, the sub-unit (chain) ID, the atom name, the number of the TM region and the distance between the atoms.

Figure 7.3 – Atom list with distances between interacting amino acids from different TM region



This figure shows the complete list with the information relating to amino acid proximities. The numbers in brackets represents the TM regions involved in the association.

7.3. The algorithm

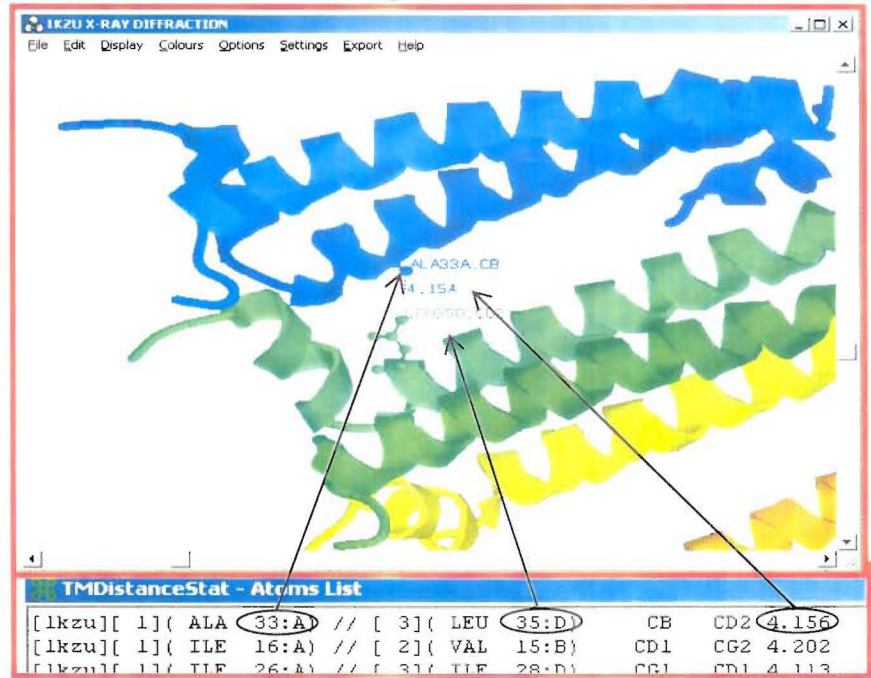
See the complete algorithm description at the appendix II-3.

7.4. Software development

Using object oriented programming the software running time has been constantly improving by optimisation of the code organisation, resulting in a more robust and reliable program.

To routinely evaluate the reliability of the program, a list of all associations (figure 7.3) was used. Rasmol (Sayle and Milner-White, 1995) scripts are used to select the specific associations between residues on different TM regions and to create a visual confirmation of the associations (see figure 7.4).

Figure 7.4 - A screen shot showing the 3D confirmation of distance associations.



This figure shows the confirmation of distance between 2 residues in different TM regions using the Rasmol program to render 3D structure and calculate the distance. The selection was made using the residue number (33:A and 35:D) generated by *TMDistance* program.

7.5. Results

The association matrix generated by the *TMDistance* module provides the basis for the substantive predictive tool described in this thesis.

Analysis of different membrane protein structures aimed at assessing the variability between datasets, suggests that crystallographic resolution is an important consideration in the selection of structures for incorporation in the standard datasets. See detailed description at Chapter 5 (section 5.5.2).

Following the original version of *TMDistance*, many features were subsequently added. The atom list was added to the latest version to check all associations between TM regions and it was also important in the debugging phase. This list is very useful to provide a confirmation of the residue distances in different TM regions, which may also be enhanced visually using molecular viewers packages.

*TMDistance* has been used by other members of the group to develop a database to analyse the packing of GxxxG groove arrangements with branched chain and aromatic amino acids in the TM regions of integral membrane proteins.

## 7.6. Brief discussion

By developing the association matrix module, the understanding of associations and possible interactions between residues in different TM regions becomes clearer. It is possible to look for patterns of data, facilitating the statistical study of the associations between large and small residues (ridge/groove arrangements) like the branched chain amino acids (isoleucine, leucine and valine) or aromatic residues (phenylalanine, tryptophan, tyrosine and histidine) on one hand, and glycine on the other. Many studies involving the interactions between large and small residues in different TM regions in membrane proteins have been undertaken (Senes *et al.*, 2000; Russ and Engelman, 2000) giving a strong motivation for the new development involving patterns of associations.

Using the generated association matrix and the graphics of statistical data obtained in the matrix it became possible to infer which amino acid is more likely to be associated with another.

By generating Rasmol script commands with the created associations and distance list, it is possible to create a graphical representation. This manual script execution may be automated in the next version of *TMDistance*. It will be a useful tool to be combined with statistical and visual approaches. The user may select, for instance, a glycine residue, and the program will give all the associated residues within the selected distance in different TM regions and at the same depth giving a breakdown of possible associations between glycine and all residues in the TM regions of the protein being tested. This procedure is different from simply loading a PDB file into the Rasmol program because *TMDistance* is working on the TM regions as defined in the Swiss-Prot file, rather than on any part of the protein or just those regions defined as alpha helix in the PDB file. Moreover, *TMDistance* provides a structural “signature” of the packing of TM regions in membrane proteins on an individual or family basis, which may be used as a basis to predicting associations between TM regions of unknown structure.

## **Chapter 8 - TMRelate**

### 8.1. Introduction

Due to the known difficulties of obtaining the 3D structures of membrane proteins from experimental data, many different computational methods of predicting the 3D structure from primary sequence have been instigated in the last decade. Many efficient methods are available to predict 2D topology and the transmembrane (TM) regions from the primary sequence as summarised in table 3.1, and lead to the annotations in protein sequence databases like GenBank and Swiss-Prot.

Due to genome projects, more and more sequences from different organisms are elucidated and catalogued every day. Using the predictive tools, the TM regions are easily predicted. Laboratory methods are also used to confirm TM regions and the resulting TM region annotations are deposited in the protein sequence repository for public scientific use.

Following the prediction of general topology of TM regions, prediction of their general positions with respect to each other in terms of an end on configuration is the next problem to be solved. In solving this problem, the possible helix packing involved in assembly of the membrane protein 3D structure must be considered.

Helix-helix packing plays a critical role in maintaining the tertiary structures of helical membrane proteins (Adamian and Liang, 2001).

Using the association matrix described in chapter 7, an *ab initio* method for prediction of the associations between TM regions is presented in this Chapter. This method results in a prediction of the general configuration and packing directly from the primary sequence. Also in this Chapter, the rotational method used to predict the most likely angular position for each TM region and its 3D structure is described.

## 8.2. Description

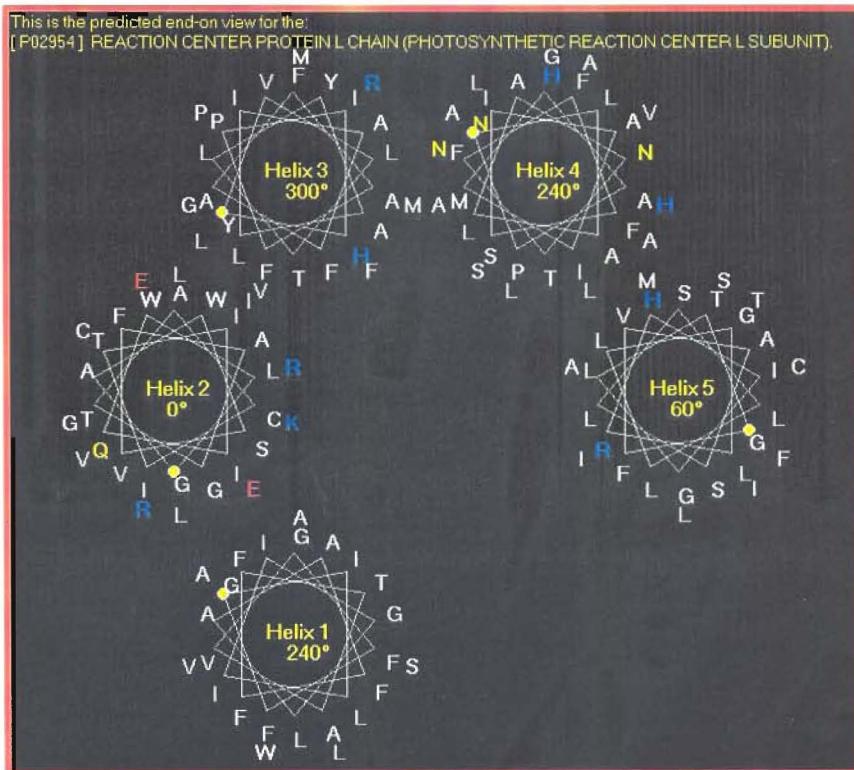
This is the main module for the predictive tool developed in this thesis. *TMRelate* predicts the 3D structure of membrane proteins from their primary sequence (Swiss-Prot format) using a 20x20 association matrix based on a number of determined membrane proteins structures, as described in Chapter 7.

*TMRelate* is loaded with the primary sequence of a chosen membrane protein and, using the 20x20 association matrix, predicts the overall configuration (stage 1). This stage assumes the most compact helix-helix packing possible. It can be summarised by one TM region being surrounded by a maximum of six others (depending on the chosen configuration). In real structures this form of compact packing can be observed in membrane proteins with higher numbers of TM regions e.g. cytochrome oxidase, where several of the central TM regions are surrounded in an unobstructed fashion by 5 or 6 others. Once the best possible configuration is found, in terms of a score for the optimal packing, *TMRelate* rotates each helix to find the optimal score in terms of capacity for association between TM regions. Then



the final 2D prediction is shown as an integrated helix wheel diagram (figure 8.1) (stage 2). Using the information portrayed in the helix wheel representation, the program then calculates the 3D CA (Carbon Alpha) co-ordinates for the predicted membrane protein structure and displays it using the backbone presentation (stage 3). In the algorithm used to build the 3D structure has a fixed inter-helical distance of 8.0 Å is applied. This distance was fixed based on the statistic obtained by the average distance between the associations of adjacent alpha carbon backbones. This statistic uses the same data set (PDB files) described on Chapter 5, giving a total of 27,868 associations and the average distance of 8.2 Å. If the side chain is considered instead of the alpha carbon backbones, the totals of associations are 188,028 with the average distance of 7.9 Å.

**Figure 8.1 – TMRelate: Helix wheel representation**



This figure shows the predicted 2D model in the form of a helix wheel representation. The yellow dot of each wheel represents the first residue of each TM region. The rotational orientation is anti-clockwise, rotating 60° each time.

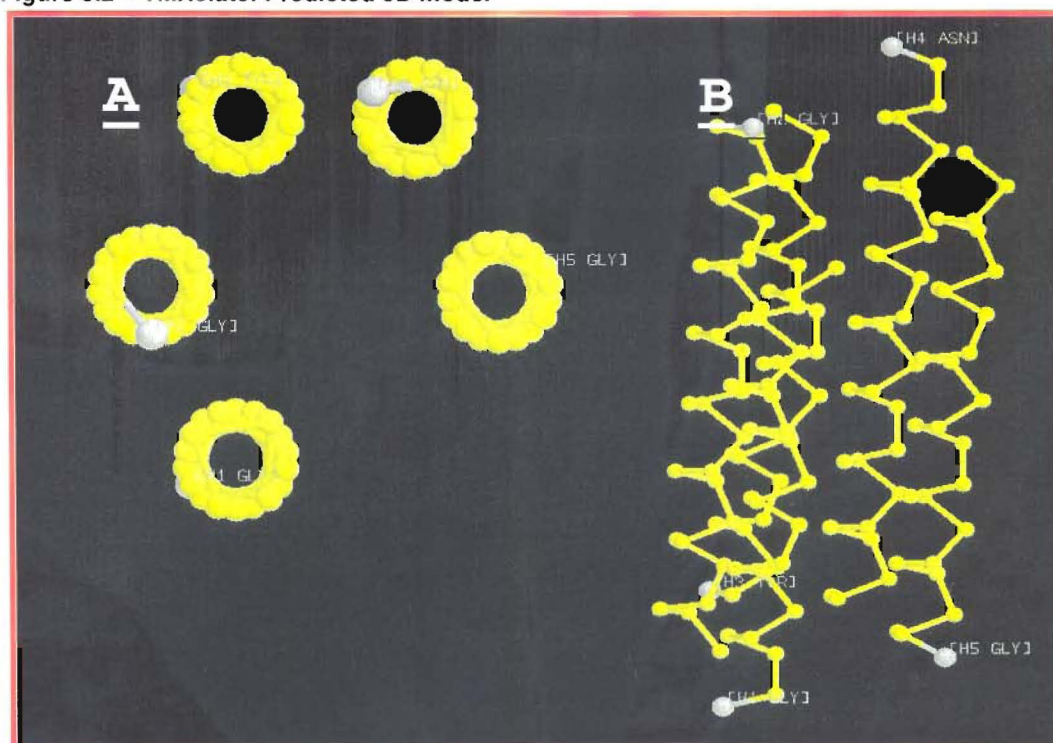


*TMRelate* uses the process of permutation. It places each TM region in every possible position in the pattern pre-defined by the users overall configuration. The top 50 configurations are shown in the output, along with their scores.

For the TM region rotation stage, *TMRelate* works in the same fashion as a car odometer, fixing all but one TM region and rotating one TM region anti-clockwise by  $60^\circ$  each time. After this first TM region has been rotated through a complete  $360^\circ$  rotation, the next TM region is rotated  $60^\circ$ , and so on until the last TM region has undergone a complete rotation. For each arrangement, a score is calculated, and at the end the highest score is taken as the best-predicted 2D model. The graphical output is shown as a helix wheel representation.

In the final stage, a 3D model is generated, and displayed using the CHIME plugin (figure 8.2). This 3D model is calculated based on the helix wheel rotational position.

**Figure 8.2 – *TMRelate*: Predicted 3D model**



This figure shows the predicted 3D model (using the Chime plugin) of the photosynthetic reaction centre protein, L chain – Swiss-Prot code: P02954. **A** shows the end on view. **B** shows the lateral view. The algorithm calculates the 3D co-ordinates to build the  $\alpha$ -helix based on their  $\alpha$ -carbon atoms. The grey sphere indicates the first residue of each  $\alpha$ -helix.

8.3. The algorithm

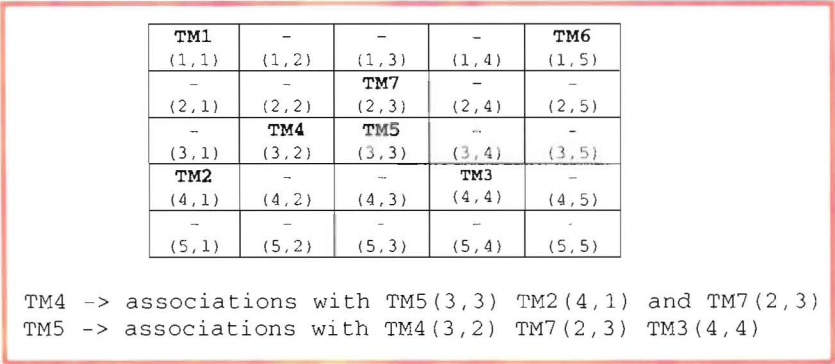
The complete description of the algorithm can be found in appendix II-4.

8.4. The software development

In the final stage of the development of *TMRelate*, many different problems were addressed. In the early stages of the development, the outputs of *TMRelate* were only textual. The obtained results were reasonable, but it was difficult to analyse the information in textual form. It was necessary to convert the textual information into a user-friendly graphical interface, which could be also utilised as a stage in the development of 3D prediction.

The first step was to convey the obtained association between TM regions as a result of optimally placing each TM region by permuting all possible positions. The very first attempt involved using a 5x5 grid and placing each TM region in an appropriate cell. The greatest problem encountered using the 5x5 grid, was the huge number of permutations ( $1.55 \cdot 10^{25}$ ) required to test all possible interactions between the TM regions.

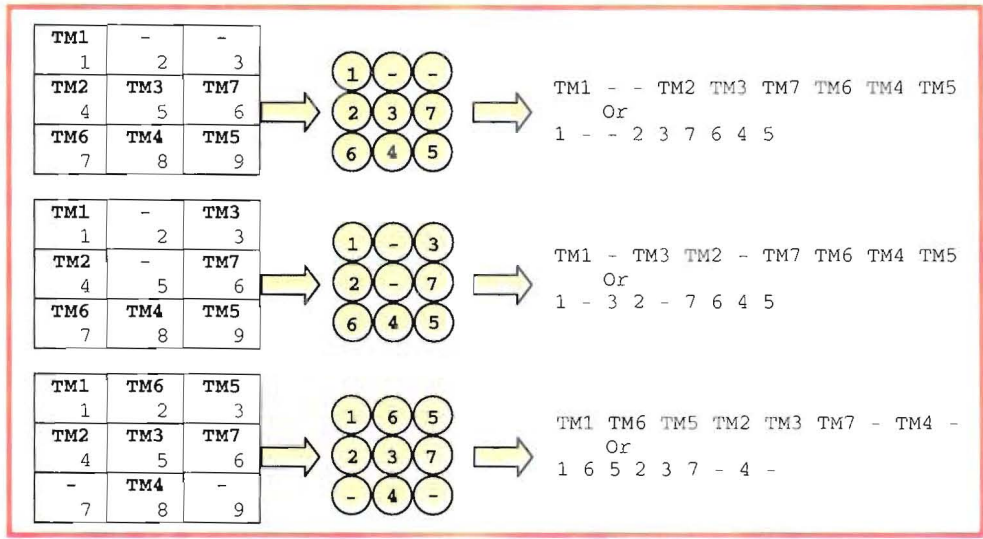
Figure 8.3 - The original 5x5 grid



This figure shows the first attempt to position each TM region into the grid. The first idea was to use a 5x5 grid, and was discarded due to the huge number of permutations.

Considering the high CPU time required for processing a 5x5 grid for a single prediction, a 3x3 grid was employed as shown in the following figure:

Figure 8.4 - Transforming the grid into linear form



This figure shows how the numerical representation was created from the initial grid. Looking at the grid, the numbers represents the grid position and the TM $n$  represents the TM region number. The circles represent the end on view for 9 TM regions. The sequence on right is the corresponding numerical representation.

The program was designed to provide the output with the “circles” configuration shown above. The algorithm makes use of the neighbour table, which indicates which TM region makes contact with which other TM regions (table A.5).

To increase the number of associations between TM regions and mimic more closely how TM regions are packed in reality, the tenth position was introduced, resulting in an end on view configuration as shown in figure 8.5:

Figure 8.5 - The end on configuration for up to 10 TM regions

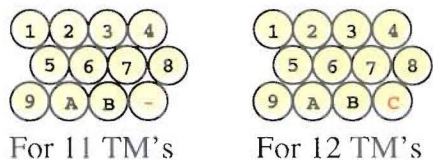


This figure shows the basic configuration for the end on view. A) The first configuration used in the development, and is suitable for 1 to 9 TM regions. B) The final configuration using 10 positions and can work on 1 to 10 TM regions.

The required changes in the algorithm were simple, showing the flexibility of the programming code. The changes in the code were an adjustment in the neighbour table and modification of the drawing position table.

A version of *TMRelate* was also created to work with 11 and 12 TM regions. This was called *TMRelate\_12*. It works in the same way as *TMRelate*, but uses a different end on configuration. In the version for 12 TM's there is no need to have a permutation file because there are no repeated digits. For the case of 11 and 12 TM's the algorithm works by using the following digits and letters to make the permutation:

- 123456789AB- (11 TM's)
- 123456789ABC (12 TM's)



The required changes in the *TMRelate\_12* version were:

- The neighbour contact table, which defines the different interactions between TM regions. It changes because the end on configuration is different for 12 TM's compared with 10 TM's.
- Changes to the internal graphical positioning table used to generate the 2D end on configuration.

8.4.1. *TMRelate\_K*

A complementary development aimed at improving the quality of resulting predictions in terms of which TM region should be buried or facing out was made. This development resulted in the *TMRelate\_K* version. This version differs in terms



of the indices used to determine the predicted packing of TM regions and the angular orientations of TM regions with respect of the other TM regions. To this end, a knowledge-based scale called kPROT (Pilpel, *et al.*, 1999) was used.

Using this scale, *TMRelate\_K* calculates and predicts which TM regions are buried and which ones are facing out toward the lipid bilayer. Adding the amino acid score for each residue (Table 5.7) that composes the TM region, a final value is calculated for the TM region, and an overall score for each possible configuration of the whole protein is calculated. In the kPROT scale, the lower the score, the more buried the TM region is.

For the helix wheel rotation, after the optimal configuration has been obtained, *TMRelate\_K* also uses the 20x20-association matrix. The algorithm is the same as in the original *TMRelate* that scores all possible rotations and stores the arrangements with the highest values.

#### 8.4.2. The algorithm

The complete description of the algorithm is given in appendix II-5.

### 8.5. Results

The results obtained from the program were promising. By running *TMRelate*, using sequences of proteins of determined 3D structure as "controls" the quality of the predictions was assessed. The evaluation process involves counting the number of associations (transmembrane region adjacencies) in the native structure

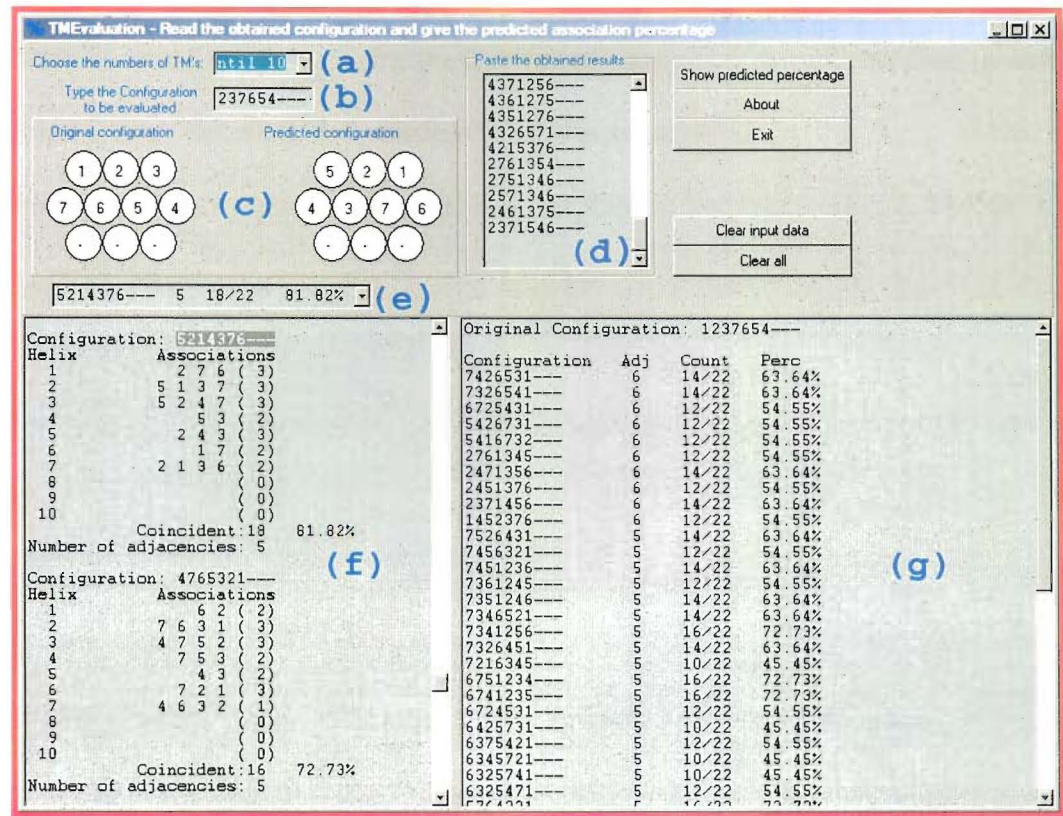
and comparing this configuration with the numbers of coincident associations in the predicted arrangement, calculating this value as a percentage of correctly predicted adjacencies. For bacteriorhodopsin, the percentage of correctly predicted associations between TM regions is 68.4%, for reaction centre protein L chain (photosynthetic reaction centre L subunit) the percentage of correctly predicted associations is 75%. The results vary depending on the initial overall template configuration selected by the user into which the predicted configuration is fitted.

A test was carried out in order to evaluate how well randomly generated configurations would perform. 1000 randomly generated configurations were evaluated and gave an average percentage of about 41%, which remained at that level across different numbers of TM regions.

Using *TMRelate\_K*, which applies the kPROT scale (Pilpel, *et al.*, 1999), the results are even more promising. The percentage of correctly predicted associations using the same *Bacteriorhodopsin* protein sequence and the same configuration is 96.6% (compared with 68.49% with *TMRelate* program) and for the photosynthetic reaction centre is 100% (compared with 75.00% with *TMRelate* program).

To make the evaluation, a program called *TMEvaluation* was created. The input for this program is the actual configuration; comparisons are made with the 50 highest scoring predicted arrangements. The following figure shows the *TMEvaluation* program running:

Figure 8.6 - *TMEvaluation* user interface



This is the *TMEvaluation* user interface running. a) The user chooses from 2 to 12 TM's. b) The user types in the configuration. c) The end on view; on the left the native configuration and on the right the predicted configuration (e). d) The user "pastes" the output from *TMRelate* program. e) The drop down list with the analyzed results (is the same as the list produced in (g). f) The detailed list with the actual and predicted associations for each TM region. g) Full listing of compared arrangements.

## 8.6. Brief discussion

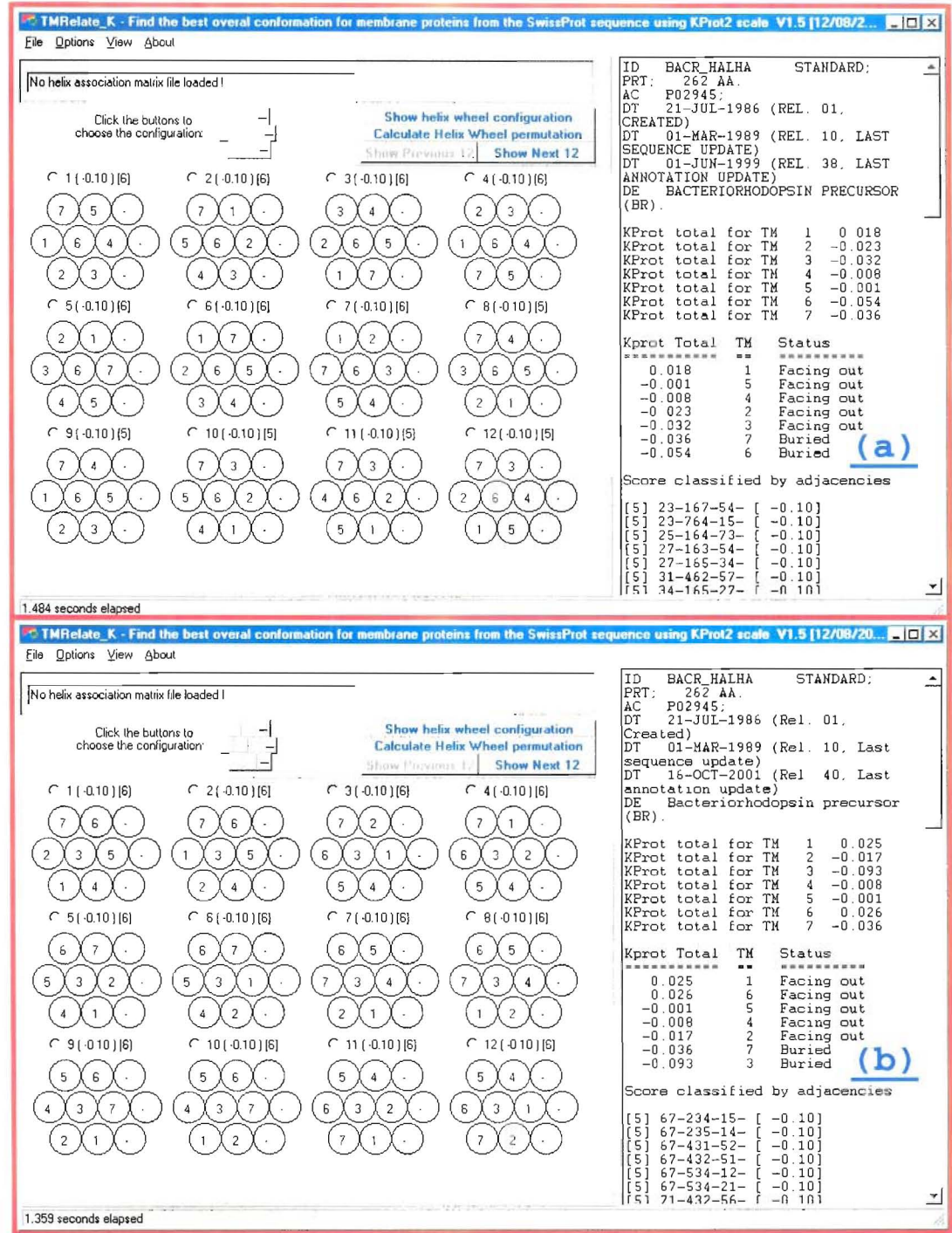
During the development period, 3 updates occurred to the Swiss-Prot database which due to this being the primary input, impacted upon the results obtained for particular proteins. For example, accession number P02945 (*Bacteriorhodopsin*), was updated from version 38 (01-Jun-1999) to 40 (16-Oct-2001) and then in version 41 (15-Jun-2002). The major change was in annotations fields such as new bibliography references and new 3D structure cross-references. For our purposes, the most significant change observed from version 38 to 40, was in





changes 25% of the atoms used by *TMDistance* and subsequently by *TMRelate* at either end of the TM. This significant change in annotation is why these particular results are affected so much.

Figure 8.7 - Differences in results obtained with *TMRelate\_K* due to changes between Swiss-Prot versions.



The first image (a) shows the output for Swiss-Prot version 38 for Bacteriorhodopsin (accession number P02945); the most buried TM region is TM6. For the version 40 file (image b), the most buried TM regions is TM3.

The algorithm allows the testing of the 50 highest scoring arrangements. As expected, minor differences in definitions of TM region composition result in significant changes in predicted arrangements. The problem with this approach for 12 TM proteins was the high CPU processing time required to execute all permutations. During the development, some improvements have been made in the code. To illustrate this, the processing time for 12 TM regions (*Cytochrome C Oxidase* polypeptide I-Beta – Swiss-Prot code: P98002) was 100:00 hours (using a Pentium III 600 KHz machine). After changes in the code the running time dropped to 4:30 hours. The main change in this case was to the neighbour contact table, avoiding the need for calls to the subroutine to manipulate this table.

Using the kPROT scale in the development of another version of *TMRelate* called *TMRelate\_K* proved to be very beneficial. This version seems more accurate in terms of predicting the correct TM associations between TM regions. Maybe in the future this version will be the primary version to be used for predicting the associations between TM regions.

The developed software represents an advance to the field of structural prediction for membrane proteins. With the described approach and pieces of software, the user can predict with reasonable accuracy the relative positions of TM regions from the primary sequence of a membrane protein. The user can also obtain a helix wheel prediction of TM region orientation, and from this, a prediction of the 3D structure. This piece of software can be further improved in terms of giving better quality 3D structural predictions.

In the next version, the module for building the 3D structure will use a database containing the sequences, breaks in the  $\alpha$ -helical structure, kinks and tilts of

the TM regions of all the determined membrane protein structures, gathered using *TMAAlpha* (Sarakinou *et al.*, 2001). The module will predict helix breaks, kinks and extent of tilt brought about by specific sequences according to the information stored in this database.

Another future implementation of the described software is alongside the use of a genetic algorithm (Noushin Minaji *et al.*, unpublished) to evaluate and improve the energetic stability of predicted structures. Noushin Minaji is developing a piece of software that calculates the free energy of a specific predicted membrane protein structure and applies a Genetic algorithm (GA) to manipulate inter-helical distances and orientations and to select candidate structures with the lowest free energy by means of an evolutionary approach.

## Chapter 09 – Evaluation

### 9.1. Associations between TM regions – end on view

To evaluate the developed predictive tools, *TMRelate* and *TMRelate\_K*, seventeen membrane proteins in Swiss-Prot format with corresponding PDB files were used. The evaluation process involved obtaining the percentage of predicted associations between TM regions, compared with the actual determined 3D structure.

The seventeen Swiss-Prot files selected and their corresponding PDB files were as shown in the following list:

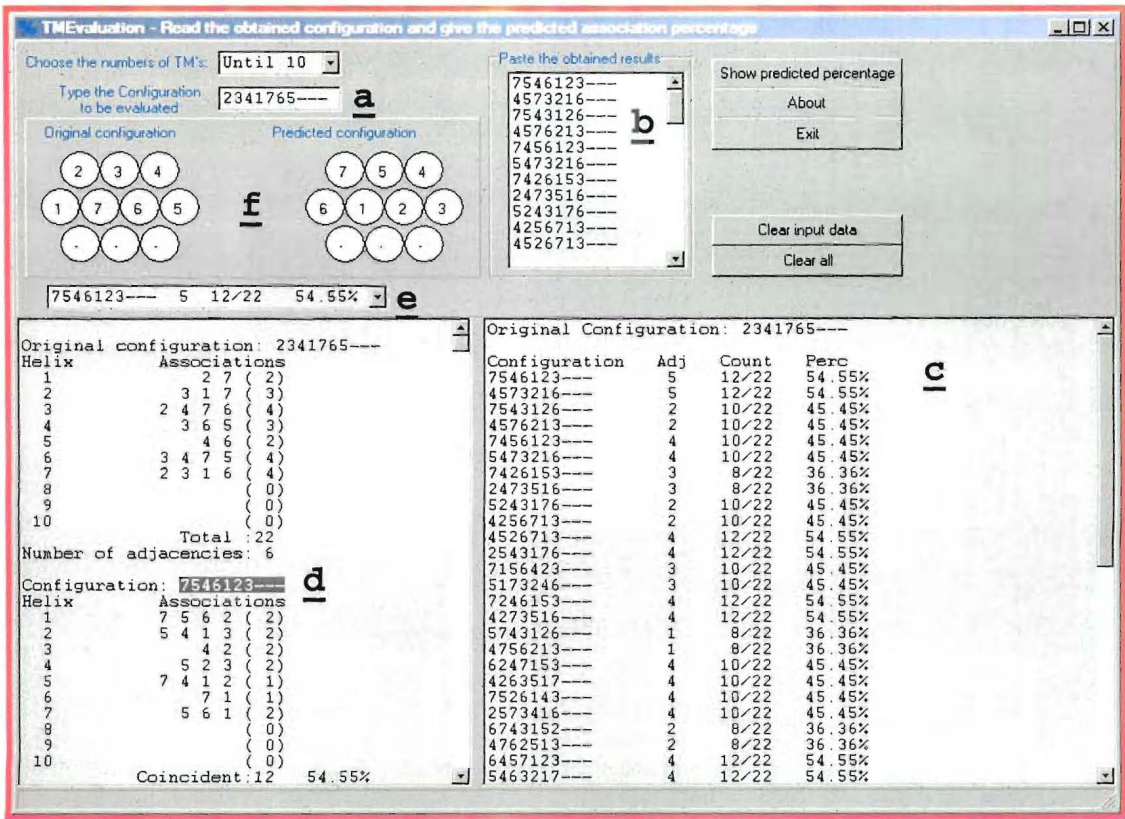
PDB code	Description	N. of TM regions
1) P02945 x 1C3W	Bacteriorhodopsin - <i>H. salinarum</i>	7
2) P16102 x 1E12	Halorhodopsin (HR) - <i>H. salinarum</i>	7
3) P42196 x 1H2S	Sensory Rhodopsin II with Transducer - <i>N. Pharaonis</i>	7
4) P02699 x 1U19	Rhodopsin: Bovine Rod Outer Segment - <i>B. taurus</i>	7
5) P06624 x 1YMG	Aquaporin water channel: Bovine lens - <i>B. taurus</i>	6
6) P11244 x 1FX8	GlpF glycerol facilitator channel - <i>E. coli</i>	8
7) P37905 x 1U7G	AmtB ammonia channel (mutant) - <i>E. coli</i>	11
8) P51762 x 1EYS	Photosynthetic Reaction Center - <i>T. tepidum</i>	5
9) P06009 x 1DXR	Photosynthetic Reaction Center - <i>R. viridis</i>	5
10) P02954 x 1RZH	Photosynthetic Reaction Center - <i>R. sphaeroides</i>	5
11) P25896 x 1JBO	Photosystem I: - <i>S. elongatus</i>	11
12) P04191 x 1T5S	E1 state with bound calcium and AMPPC P-type-O. cuniculus	10
13) P43457 x 2BL2	Rotor of V-type Na <sup>+</sup> -ATPase - <i>E. hirae</i>	4
14) P17413 x 1QLA	Fumarate Reductase Complex - <i>W. succinogenes</i>	5
15) P24185 x 1KQF	Formate dehydrogenase-N - <i>E. Coli</i>	4
16) P02722 x 1OKC	Mitochondrial ADP/ATP Carrier: Bovine heart - <i>B. Taurus</i>	6
17) Q5SJ79 x 1XME	Cytochrome C Oxidase, ba3 - <i>T. Thermophilus</i>	13

9.2. End on view evaluation

The testing of each protein produces four different results based on the four matrices generated with different distance cut-offs, excluding the protein undergoing testing (see Chapter 5).

A program was developed to calculate the percentage of correctly predicted associations between TM regions and this was called *TMEvaluation*. It reads the output from *TMRelate* / *TMRelate\_K* and counts the associations for each TM region for the native and predicted configuration calculating the percentage of correctly predicted associations. Figure 9.1 shows the *TMEvaluation* program running.

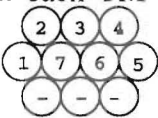
Figure 9.1 – *TMEvaluation* user interface



The user pastes the obtained results from *TMRelate*/*TMRelate\_K* module (**b**) and the program calculates the percentage of correctly predicted associations based on the original configuration (**a**). The given results are: the configuration, number of adjacencies, number of coincident associations and the calculated percentage (**c**). Using the drop-down option (**e**), the user can select one specific configuration and the detailed statistics (**d**) are shown with the corresponding end on view (**f**).



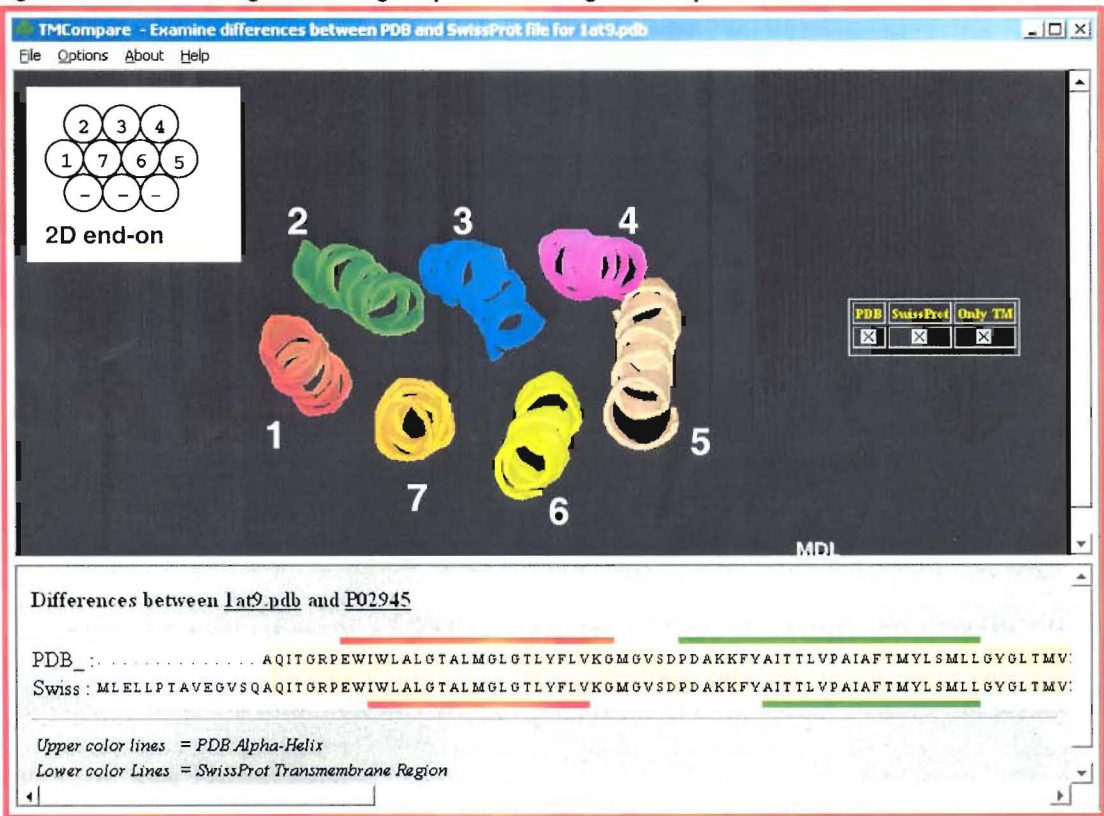
To obtain the end-on view structural associations, the *TMCompare* program was used, by loading the corresponding PDB file and viewing only the TM regions, visualised end-on (with different colours for each TM region). This process facilitates the recording of overall structural positions (figure 9.2). Then each TM region is annotated in the 2D end-on representation in the following way:



The 2D end-on view is converted to the linear format, i.e.: [2341765---] that is used in the *TMEvaluation* program as the input in the field: “configuration to be evaluated” (figure 9.1 - item a).

Some predicted configurations are different from the actual end on view of the structure, but the percentage may still be high, as the evaluation program compares the associations between all TM regions, not matching the end on view *per se*.

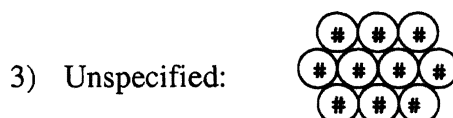
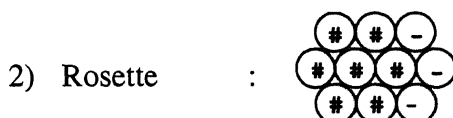
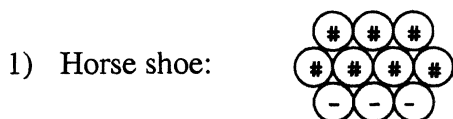
Figure 9.2 – Recording the TM region position using *TMCompare*



*TMCompare* was used to record the relative position of each TM region to form the 2D end-on view. The process links the 3D structure of the TM regions with the 2D end-on view formed in the helix wheel representation.

There are two sets of evaluations: One using the 20x20 association matrix (*TMRelate* program) and another using the kPROT scale (*TMRelate\_K* program). For the *TMRelate* program, the evaluation is based on four different matrices with different distance cut-offs: 3.0 Å, 3.5 Å, 4.0 Å and 4.5 Å. These matrices are created using the PDB files described in Chapter 5 (materials and methods). Proteins were tested on a one-out basis

For both sets of evaluations, three different configurations were chosen to execute *TMRelate* and *TMRelate\_K* as shown below:



The character '#' represents locations occupied by a TM region. The character '-' (dash) represents an empty space. The unspecified one leaves the algorithm to find the configuration with the highest association score. It takes more time for the execution, because all possible TM region positions are tested.

For the evaluation of 12 TM regions, two different configurations with slight differences were used instead of the three configurations used for membrane proteins with 11 or less TM regions.

Table 9.1 and 9.2 show comparative results with the best-predicted percentage, the average percentage and the overall percentage obtained by *TMRelate* and *TMRelate\_K* programs. For the detailed evaluation results see appendix III.



Table 9.1 – Evaluation of *TMRelate* by assessment of the percentage of correctly predicted associations between TM regions (see complete evaluation at appendix III)

Protein	# of TM regions	Best percentage of correctly predicted associations obtained using <i>TMRelate</i> with different distance matrices and configurations				Average percentage using <i>TMRelate</i>			
		3.0Å	3.5Å	4.0Å	4.5Å	3.0Å	3.5Å	4.0Å	4.5Å
Bacteriorhodopsin (P02945 x 1C3W)	7	72.73%	72.73%	75.00%	72.73%	70.69%	63.13%	74.24%	65.91%
Rhodopsin (P02699 x 1U19)	7	63.64%	63.64%	63.64%	63.64%	60.10%	57.32%	60.10%	60.10%
Sensory Rhodopsin II (HR) (P42196 x 1H2S)	7	72.73%	75.00%	66.67%	75.00%	63.13%	71.21%	65.66%	68.18%
Halorhodopsin (P16102 x 1E12)	7	66.67%	66.67%	63.64%	58.33%	59.85%	65.66%	60.01%	57.07%
Aquaporin (P06624 x 1YMG)	6	80.00%	100.00%	80.00%	80.00%	70.00%	93.33%	75.55%	75.55%
Glycerol uptake facilitator protein (Aquaglyceroporin) (P11244 x 1FX8)	8	69.23%	61.54%	61.54%	69.23%	64.10%	58.97%	56.41%	64.10%
Photosynthetic Reaction Center <i>Thermochromatium tepidum</i> (P51762 x 1EYS)	5	75.00%	75.00%	75.00%	75.00%	75.00%	75.00%	75.00%	75.00%
Photosynthetic Reaction Center <i>Rhodospseudomonas viridis</i> (P06009 x 1DRX)	5	75.00%	75.00%	75.00%	75.00%	75.00%	66.67%	75.00%	66.67%
Photosynthetic Reaction Center <i>Rhodobacter sphaeroides</i> (P02954 x 1RZH)	5	75.00%	75.00%	75.00%	75.00%	75.00%	66.67%	66.67%	66.67%
P-type ATPase (P04191 x 1T5S)	10	63.16%	52.63%	52.63%	52.63%	54.38%	49.12%	50.87%	50.87%
Respiratory proteins – Mitochondrial ADP/ADP carrier (P02722 x 1OKC)	6	66.67%	55.56%	66.67%	77.78%	55.55%	46.29%	50.00%	53.70%
Respiratory proteins – Fumarate Reductase complex ( <i>Wolinella succinogenes</i> ) (P17413 x 1QLA)	5	71.43%	71.43%	71.43%	71.43%	66.67%	71.43%	66.67%	71.43%
V-type ATPase (P43457 x 2BL2)	4	80.00%	80.00%	100.00%	100.00%	70.00%	70.00%	93.33%	93.33%
Formate dehydrogenase-N: <i>Escherichia coli</i> (P24185 x 1KQF)	4	80.00%	80.00%	80.00%	80.00%	80.00%	80.00%	80.00%	80.00%
Photosystem I - <i>Thermosynechococcus elonatus</i> (P25896 x 1JBO)	11	68.42%	68.24%	73.68%	68.42%	65.79%	65.79%	68.42%	65.79%
AmtB ammonia channel (mutant): <i>E. coli</i> (P37905 x 1U7G)	11	57.14%	76.19%	71.43%	61.90%	57.14%	76.19%	71.43%	61.90%
Cytochrome c oxidase, ba3: <i>T. Thermophilus</i> (Q5SJ79 x 1XME)	13	63.64%	60.87%	54.55%	54.55%	62.25%	59.98%	53.36%	53.36%
Overall percentage		70.61%	71.14%	70.93%	71.21%	66.15%	66.86%	67.21%	66.44%

Table 9.2 – Evaluation of *TMRelate\_K* by assessment of the percentage of correctly predicted associations between TM regions

Protein	# of TM regions	Best percentage of corrected predicted associations using <i>TMRelate_K</i>	Average percentage using <i>TMRelate_K</i>
Bacteriorhodopsin (P02945 x 1C3W)	7	100.00%	96.67%
Rhodopsin (P02699 x 1U19)	7	72.73%	68.69%
Sensory Rhodopsin II (HR) (P42196 x 1H2S)	7	75.00%	68.68%
Halorhodopsin (P16102 x 1E12)	7	81.82%	68.94%
Aquaporin (P06624 x 1YMG)	6	83.33%	81.11%
Glycerol uptake facilitator protein (Aquaglyceroporin) (P11244 x 1FX8)	8	76.92%	74.35%
Photosynthetic Reaction Center <i>Thermochromatium tepidum</i> (P51762 x 1EYS)	5	100.00%	100.00%
Photosynthetic Reaction Center <i>Rhodopseudomonas viridis</i> (P06009 x 1DRX)	5	100.00%	100.00%
Photosynthetic Reaction Center <i>Rhodobacter sphaeroides</i> (P02954 x 1RZH)	5	100.00%	100.00%
P-type ATPase (P04191 x 1T5S)	10	73.68%	66.67%
Respiratory proteins – Mitochondrial ADP/ADP carrier (P02722 x 1OKC)	6	100.00%	96.29%
Respiratory proteins – Fumarate Reductase complex ( <i>Wolinella succinogenes</i> ) (P17413 x 1QLA)	5	100.00%	100.00%
V-type ATPase (P43457 x 2BL2)	4	100.00%	100.00%
Formate dehydrogenase-N: <i>Escherichia coli</i> (P24185 x 1KQF)	4	100.00%	100.00%
Photosystem I - <i>Thermosynechococcus elongatus</i> (P25896 x 1JBO)	11	57.89%	57.89%
AmtB ammonia channel (mutant): <i>E. coli</i> (P37905 x 1U7G)	11	47.62%	45.24%
Cytochrome c oxidase, ba3: <i>T. Thermophilus</i> (Q5SJ79 x 1XME)	13	56.52%	50.00%
Overall percentage		83.88%	80.85%

### 9.3. Relationship between the number of proteins used to build the matrix and the percentage of correctly predicted TM region adjacencies

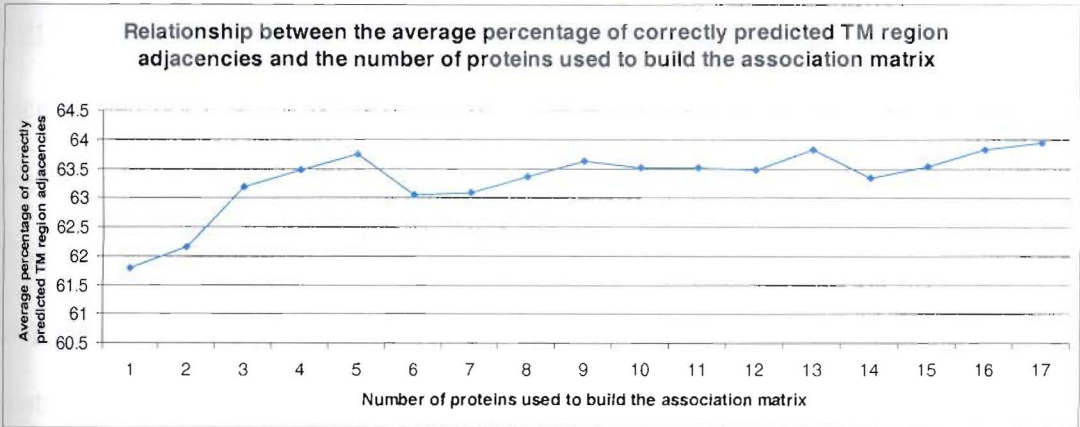
An additional evaluation was undertaken to correlate the number of proteins used to build the association matrix with the percentage of correctly predicted inter-helical associations. This evaluation used the bootstrap method (see glossary).

For this evaluation, from 1 to 17 high resolution membrane protein structures were used to build the association matrix and subsequently predict the TM region adjacency for the 17 proteins, and the accuracy of those predictions was measured. The following steps were used to implement the test:

- I. perform steps *a*, *b*, *c*, and *d* with a number varying from 1 to 17 proteins:
  - a. Pick at random a protein from the 17 proteins and create an association matrix (internal random routine implemented at the pipeline);
  - b. With the created matrix, run *TMRelate* for the 17 proteins;
  - c. Evaluate the percentage of correctly predicted associations for the *TMRelate* results;
  - d. Do steps *a*, *b* and *c* a 100 times.

As a result for each step *I*, 1700 predictions/evaluations were carried out, and the average percentage is shown in figure 9.3.

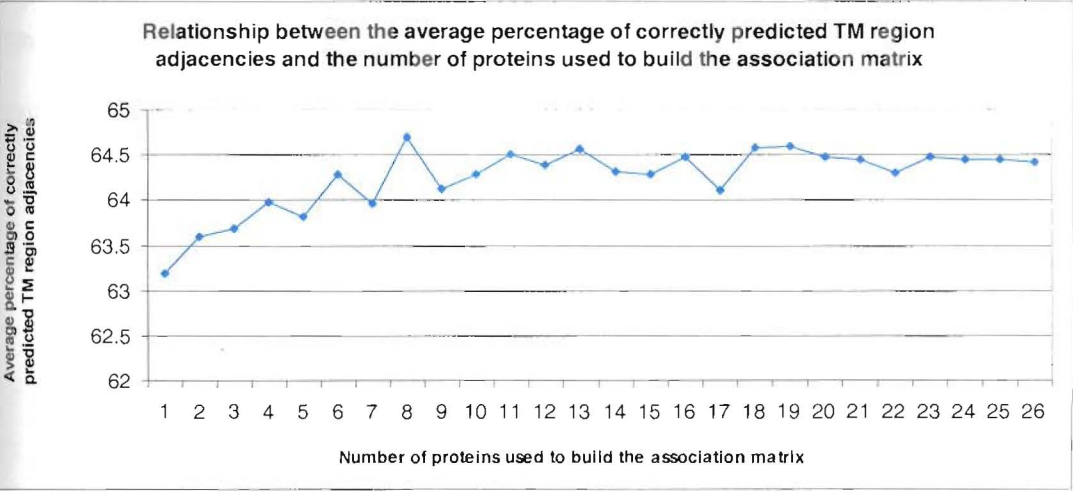
**Figure 9.3 – Relationship between the number of proteins used to build the association matrix and the average percentage of correctly predicted TM region adjacencies**



The trend observed is that the more proteins that are used in the generation of the association matrix, the higher the percentage of correctly predicted TM region adjacency.

Another evaluation (figure 9.4) using 26 proteins (see table 5.5 and 5.6) was made to show the increase in the percentage of correctly predicted TM region adjacencies. The result is about 0.5% higher than using the 17 proteins, showing a small but steady improvement in the accuracy of the predictions with increasingly populated matrices.

**Figure 9.4 – Relationship between the number of proteins used to build the association matrix and the average percentage of correctly predicted TM region adjacencies using up to 26 proteins to build the matrix**





A final bootstrap evaluation using only the 7 helix bundle proteins (1C3W, 1EI2, 1H2S, 1UI9, 1UAZ) was carried out, and the results (figure 9.5) show an increase of 2-3% in correctly predicted TM region adjacencies. This test investigates the potential advantage of applying matrices specific to a given family of proteins. Compared with the other results it gives higher accuracy, and suggests that a prior step of querying test sequences for their number of TM regions, and then applying a matrix constructed only from proteins with that number of TM regions may be advantageous.

**Figure 9.5 – Relationship between the number of proteins used to build the association matrix and the average percentage of correctly predicted TM region adjacencies using only 7 helix bundle proteins**

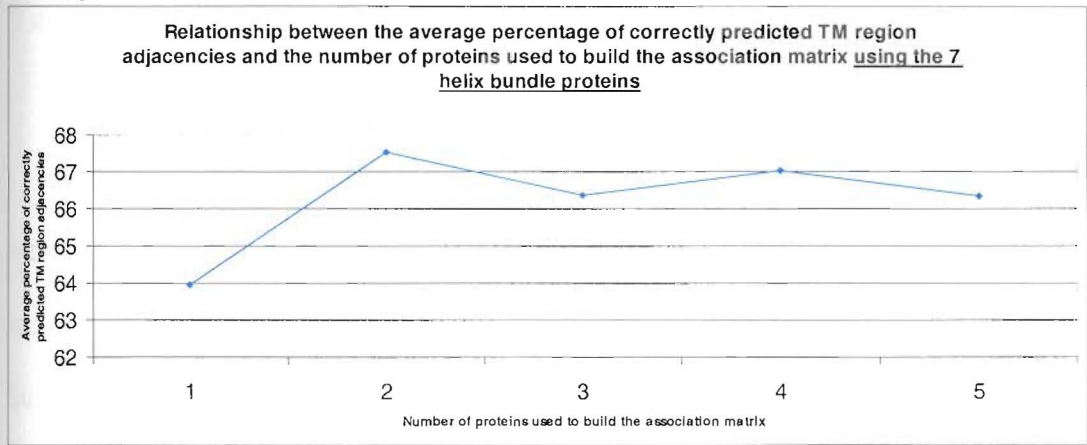
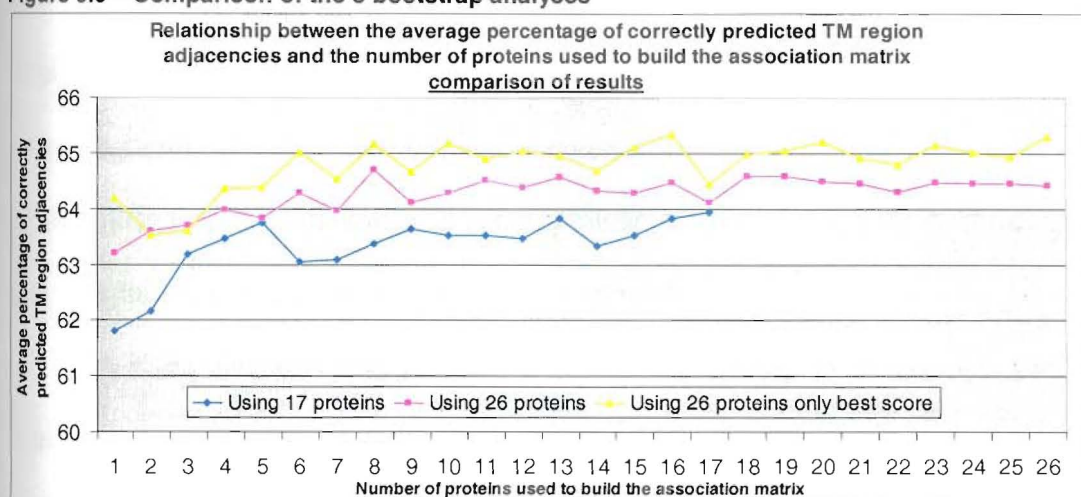


Figure 9.6 shows the comparison between the three bootstrap tests. The results from the best score selected by *TMRelate* (yellow line) seems to give the best results in terms of overall performance. This score, is based on only the first predicted top-scoring configuration on the *TMRelate* output list, rather than an average of the top 50.

Figure 9.6 – Comparison of the 3 bootstrap analyses



## 9.4. 3D model evaluation

For the evaluation of the 3D structural predictions, many different approaches were used without gaining a satisfactory picture of the accuracy of the obtained model. The first evaluation was using **RMSD** (Root mean square deviation). As the built model was a perfect helix without considering any breaks or tilt the expected RMSD value was high (between 8~10 Å). This can be partly explained by the algorithm used to build the 3D structure, where the distance between the transmembrane regions has a fixed value of 8.0 Å. The expectation is that the algorithm can be improved and this value will decrease. At this stage, we are considering the associations between amino acids in different transmembrane regions. The **RMSD** value does not give an evaluation in terms of correct associations, but only a global value. For the **RMSD** calculation, the Swiss-PDB Viewer program was used (Guex and Peitsch, 1997).

Another approach taken was to develop a program based on the elastic similarity score algorithm developed by Holm and Sander (1993). But again, the obtained similarity index does not give a percentage of associations between amino acids in different transmembrane region that were correctly predicted. And so, using only this index, it was difficult to evaluate the model.

A more meaningful approach was to develop a program to calculate the 3D similarity in terms of recorded distance associations between residue pairs compared to the native 3D structure based on the distance table. The program was named *TMEvaluation\_3D* and the input is the distance table created by *TMDistance*. *TMEvaluation\_3D* compares the common structural associations (native and predicted) giving a percentage of coincident associations.

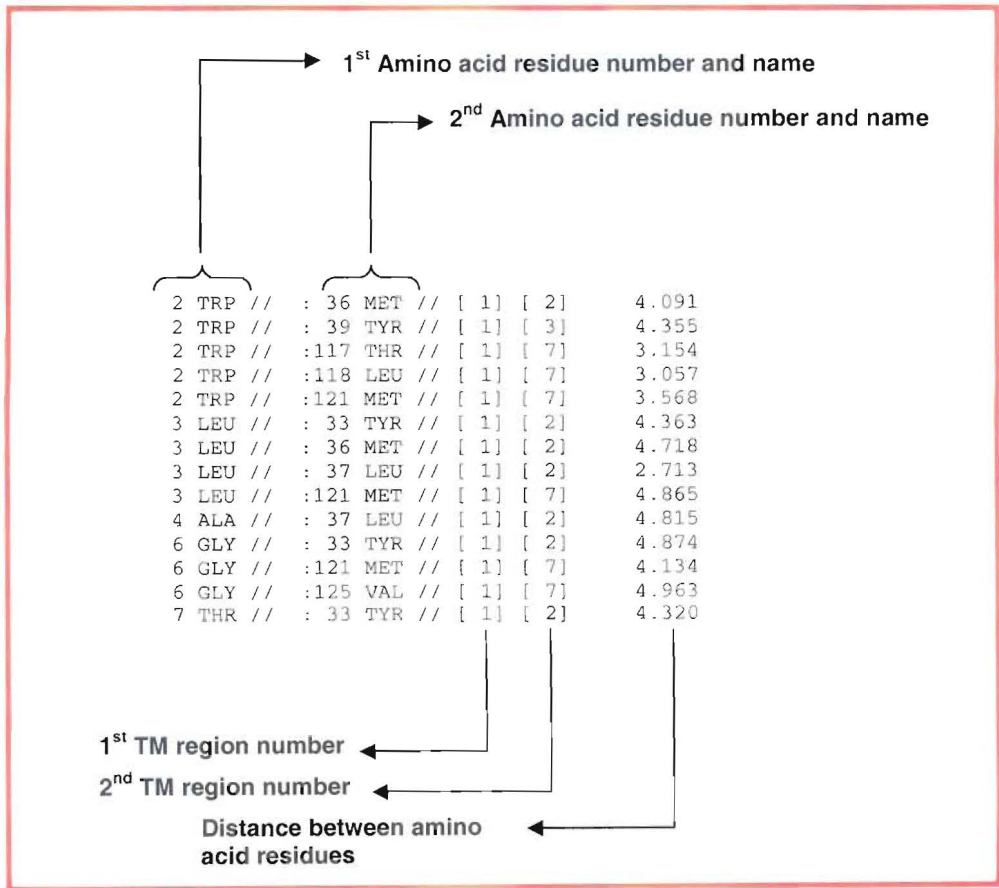
### 9.5. The 3D evaluation pipeline using *TMEvaluation\_3D*

*TMRelate* creates a 3D  $\alpha$ -helical structure in PDB format based on starting coordinates provided by the rotated helix wheel. The created model contains only the  $\alpha$ -helix backbone information (C alpha atom) in the PDB file format. The side chains are created using the Maxsprout program that is a database based algorithm for generating protein backbone and side chain co-ordinates from a C alpha coordinates (Holm and Sander, 1991). The backbone is assembled from fragments taken from known structures. Side chain conformations are optimised in rotamer space using a rough potential energy function to avoid clashes. The Maxsprout service is available at the address: <http://www.ebi.ac.uk/maxsprout/index.html>, where the user inputs the C alpha coordinates and the output is a PDB file with calculated side chains.

The appropriate native structure (PDB file) for each predicted protein was used for the evaluation. Only the TM regions of these PDB structures were used. These regions were identified following the analysis of the respective PDB file using the *TMCompare* program. The file generated by this process was the PDB file, but only with the C alpha coordinates. The reason for subsequently creating the side chains for the stripped native structure using the Maxsprout algorithm was that the same evaluation method could be applied as for the predicted structure.

*TMDistance* reads the PDB file created by the Maxsprout program and produces an output of the distances between the amino acids in different transmembrane regions as shown in the following figure.

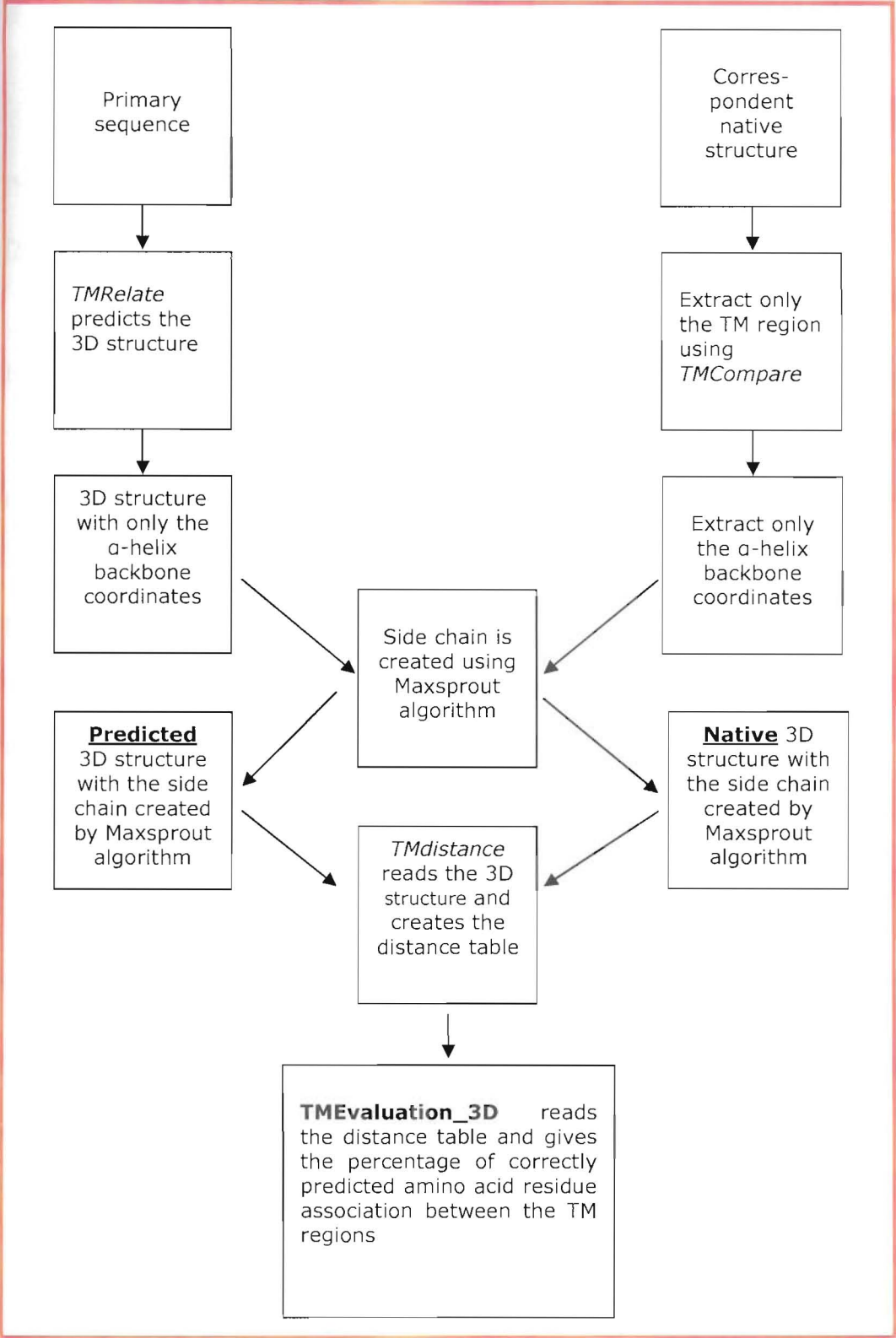
Figure 9.7 – *TMDistance* results with distances between residues in different TM regions



This is the output created by the *TMDistance* program for the 3D evaluation. Using this distance table, *TMEvaluation\_3D* creates a coincident percentage between the modeled and native structures.



Figure 9.8 - 3D evaluation pipeline using *TMEvaluation\_3D* program



This flowchart shows the 3D evaluation process using the *TMEvaluation\_3D* program. The left side shows steps for the predicted structure and the right side for the native structure steps. In the middle are the common tasks and the comparison to find the correct percentage of predicted associations between amino acid residues in different transmembrane regions.

## 9.6. Structural evaluation

The tables on appendix IV show the evaluation of predicted 3D structures. They show the percentage of correctly predicted associations generated by the *TMEvaluation\_3D* program.

For each protein, a cut-off distance of 5.0 Å and 8.0 Å was used to create the distance list. The distances are calculated based on 2 different parameters; one considering all the atoms from each amino acid residue present in the  $\alpha$ -helix and another considering only the C(alpha) atom.

The “#” column represents the number of associations between amino acids in different  $\alpha$ -helices of the TM regions. The “%” column contains the percentage of coincident associations (using side-chain atomic co-ordinates with the closest distance) between the native structure and the predicted one. For the detailed evaluation results see appendix IV.

## 9.7. Test for sensory rhodopsin against a matrix of bacterial rhodopsin

The results shown in table 9.3 are those for the test of sensory rhodopsin using a matrix created from the bacterial rhodopsins. Two other matrices are used for the comparison: one based on the group of 17 proteins routinely used to build the association matrix for the test set (see Chapter 5 – material and methods) and another generated with the 17 protein group excluding the sensory rhodopsin structure. The three matrices are named: ‘B Rhodopsin’ (only bacterial rhodopsin), ‘All set’ (the 17 proteins) and ‘No SR’ (the 17 proteins minus sensory rhodopsin). Three

configurations are used for the test (horse shoe, rosette and unspecified). The table of results is divided into five columns, four giving the correctly predicted percentage of helical associations using three different matrices generated with different distance cut-offs and the final column with the average prediction accuracy. The best result was observed with the unspecified configuration.

As expected, the results show that the percentages of correctly predicted associations between TM regions using the matrix created only with the more closely related bacterial rhodopsin are higher than those attained using all the 17 structures. A lower percentage accuracy was obtained using the matrix that excluded sensory rhodopsin. However, even in this group the results attains a reasonable prediction accuracy of around 60%, which was on average within 10% of the accuracy attained using all the available structures.

Table 9.3 – Test for sensory rhodopsin against a matrix of bacterial rhodopsin.

<u>Horse shoe</u>					
Matrix	3.0 Å	3.5 Å	4.0 Å	4.5 Å	Average
B_Rhodopsin	64.73%	52.72%	54.18%	52.00%	55.90%
All_set	57.09%	49.09%	49.09%	46.54%	50.45%
No_SR	57.09%	48.72%	48.72%	42.54%	49.26%
<u>Rosette</u>					
Matrix	3.0 Å	3.5 Å	4.0 Å	4.5 Å	Average
B_Rhodopsin	82.66%	72.66%	64.33%	64.66%	71.07%
All_set	52.00%	64.33%	56.33%	64.33%	59.25%
No_SR	52.33%	66.33%	56.33%	66.33%	60.33%
<u>Unspecified</u>					
Matrix	3.0 Å	3.5 Å	4.0 Å	4.5 Å	Average
B_Rhodopsin	78.99%	70.99%	70.33%	66.67%	71.74%
All_set	53.98%	66.66%	61.99%	70.33%	63.24%
No_SR	50.33%	70.33%	61.99%	62.99%	61.41%

## **Chapter 10 – Discussion**

### 10.1. The developed software

The central question addressed in the thesis is whether and how a method could be developed to reliably predict the 3D structure of membrane proteins directly from primary sequence. This question has been answered by the development of the software that records and predicts associations between TM regions, testing every possible arrangement within a pattern selected by the users on a permutational basis. In addition, the TM regions can be "rotated" in order to find the best rotational orientation based on the created 20x20 association matrix score. This rotational process is a simple high scoring method, where the best prediction is the highest scored configuration. The next stage of the development will adopt a method similar in ways to that of Engelman and Brunger (Treutlein *et al.*, 1992) making a series of energy minimizations to confirm the lowest energy configuration. The major difference between this approach and previous ones is that in this approach the helix structures are being fitted according to propensity for residue pair formation and the known likelihood of a transmembrane region residue for being buried or exposed. The permutation approach allows the identification of the optimal arrangement

according to these physicochemical and biostatistical properties and requires no previous knowledge of the structure of a given protein.

The  $\alpha$ -helical 3D structure of a membrane protein is then created in atomic coordinate form, and the models derived in this way for proteins of determined structure have been observed to possess interactions between amino acids that are found in reality in those proteins. However, the predicted structures are still far from depicting correct structures. The improvement of the 3D models will be the focus of the next modules to be developed in the project. It will use information based on the helix kink, helix tilt databases and genetic algorithms to determine optimal helix-helix packing evaluated by free energy (force field) calculations.

In this project, all the developed modules are fully integrated. Each module was designed to be complementary to the others but at the same time is able to work as an individual program.

#### 10.1.1. TMCompare

The first developed module was *TMCompare*, which provides a clear visual representation of the way specific PDB and Swiss-Prot files are related. *TMCompare* was originally developed for the purpose of verification of PDB and Swiss-Prot files used in the generation of 20x20 association matrices. A paper describing *TMCompare* was submitted and published in the *Bioinformatics* journal (Togawa *et al.*, 2001). *TMCompare* is in keeping with other protein analysis tools available on the Internet. Similar available tools such as GRASS (Nayal *et al.*, 1999), STING

(Neshich *et al.*, 2003), PDBSum (Laskowski *et al.*, 1997), Protein Explorer (Martz, 2002) among others, also have a user-friendly interface with many embedded scripts that allows the manipulation of the protein 3D structure being studied.

The development of *TMCompare* provided a good foundation for the rest of the project. It was fundamental to the understanding of the methodology, specific file formats and the basis of inter-helical associations.

### 10.1.2. *TMDistance*

The next developed module was *TMDistance*. It reads the known membrane proteins structure file(s) and creates a 20x20-association matrix based on the proximity of the amino acids that make up the TM regions. It is a knowledge-based process, similar to other available tools for the prediction of TM helix localisation and topology prediction. One example is MEMSAT (Jones *et al.*, 1994) that uses statistical tables (log likelihoods) compiled from well-characterised membrane proteins and dynamic programming algorithm, to recognise membrane topology models by expectation maximisation. PHDhtm (Rost *et al.*, 1994), uses information from proteins families and a neural network and DAS (Cserző *et al.*, 1997) is based on the RreM scoring matrix originally introduced to improve alignments for G-protein coupled receptors.

In the final version, the program development evolved from the linear/procedural programming design to object oriented programming, improving the processing time and increasing the reliability and robustness. For its

development, the basic Swiss-Prot and PDB ‘*classes*’ were created. These ‘*classes*’ manipulate the information contained in the PDB and Swiss-Prot files creating an ‘object’ for different purposes. Since these created ‘*classes*’ are reusable, they were integrated into the *TMRelate* module.

Additionally, *TMDistance* allows the searching for patterns of information in terms of ridge and groove arrangements, by looking for the interactions between glycine and aromatic (phenylalanine, tryptophan, tyrosine, histidine) or aliphatic (isoleucine, leucine, valine) amino acids at the same depth. The significance of being able to identify such associations between the TM regions of membrane proteins, especially in terms of particular motifs formed by these amino acids when located four or three residues apart (formation of ridges and grooves), but also in terms of associations between single residues (associations possibly equivalent to partial ridge-groove arrangements), led to the further development of this module to a computational tool to allow more detailed analysis of these arrangements (see section 10.3).

An important aspect when testing *TMDistance* was the selection of the known membrane protein structures to generate the association matrix. Comparative analysis of the existing membrane protein structures has revealed that the crystallographic resolution of the structures is an important issue for the selection of the proteins for the formation of the dataset. The criteria set to choose the proteins were fundamental to the quality and subsequent predictive value of the generated matrix.

### 10.1.3. *TMRelate*

The central module was *TMRelate*. This program represents an advance in the field of prediction of membrane protein 3D structure. No other piece of software has been developed using the same strategy to successfully predict associations between TM regions. From the membrane protein primary sequence, *TMRelate* predicts the associations between whole TM regions using a knowledge-based association matrix and the kPROT scale (Pilpel *et. al.*, 1999). It uses a permutation algorithm to test all possible positions (end-on view) in order to find the arrangement with the best association score. Additionally, *TMRelate* predicts the associations between individual amino acids by rotating all the TM regions and predicting the angular orientation of each TM segment, i.e. to determine which residues are optimally exposed to the lipid phase and which are buried in the interior of the TM bundle. For this process, the algorithm scores every possible association between pairs of amino acids on adjacent TM regions using the created 20x20 association matrix, after considering the buried angle and the calculated depth between each amino acid in different TM segments.

The generated output is given as a helix wheel representation with the highest scoring end-on arrangement rotated angle. Using this angle for each TM region, a 3D-structure co-ordinate with  $\alpha$ -carbon backbone is created. The created 3D structure is far from the ideal model, but it gives a good starting point to 3D structural prediction, and it can be evaluated by formulating benchmark results, as described in Chapter 9 (Evaluation).



An evaluation of the accuracy of the developed piece of software has been made using a set of 17 different membrane proteins with a differing number of TM regions, as assigned in their Swiss-Prot files. To obtain the percentage of correctly predicted associations, the corresponding known high-resolution 3D structures were used (corresponding PDB files with the best resolution) for comparison. For this process, an additional program called *TMEvaluation* was created, while the modules *TMCompare* and *TMDistance* were also used in order to prepare the dataset.

*TMCompare* was used to select the PDB files with corresponding Swiss-Prot files and also to find the correct arrangement associations between TM regions. *TMCompare* was used to visualize and analyse the TM regions in the real structure, which was essential in the analysis of structures like Photosystem I (PDB code 1JBO) with 11 TM regions. Analysis of such structures using a molecular rendering program such as Rasmol and CHIME is a difficult task due to the nature of the structure i.e., 13 different sub-units and no visual information as to where the TM region starts and where it ends. *TMCompare* facilitates this analysis by reading the annotation of the TM regions from Swiss-Prot file and applying it to those amino acids in the structure, selecting and showing only the TM regions.

*TMDistance* was used to create the association matrix used for the predictions where a dataset of 17 known high-resolution structures of membrane proteins has been used (Chapter 5). The final selection of the proteins of determined structure used to create the dataset gives a good general association matrix, because it is derived from a mixture of known structures of high resolution of membrane proteins that belong to different structural families. Using the association matrix, *TMRelate* was executed loading the membrane protein sequence in Swiss-Prot format, and the

results were compared with the corresponding known structure. This process resulted in a percentage of correct associations between TM regions, calculated in an automated way by *TMEvaluation*.

*TMEvaluation* reads the output arrangement from *TMRelate*, counts the number of associations between each TM region and compares it with the correspondent native structure (known 3D structure), calculating the percentage of correctly predicted associations between TM regions. The comparative results from running *TMRelate* and the *TMRelate\_K* program are listed in the evaluation section (Chapter 9). A summary of the evaluation results is shown in table 10.1:

**Table 10.1 – The evaluation of the prediction of TM region associations by *TMRelate***

Protein	# Of TM regions	Average percentage using <i>TMRelate</i>	Average percentage using <i>TMRelate_K</i>
Bacteriorhodopsin (P02945 x 1C3W)	7	73.30%	96.67%
Rhodopsin (P02699 x 1U19)	7	63.64%	68.69%
Sensory Rhodopsin II (HR) (P42196 x 1H2S)	7	72.35%	68.68%
Halorhodopsin (P16102 x 1E12)	7	63.83%	68.94%
Aquaporin (P06624 x 1YMG)	6	85.00%	81.11%
Glycerol uptake facilitator protein (Aquaglyceroporin (P11244 x 1FX8)	8	65.39%	74.35%
Photosynthetic Reaction Center <i>Thermochromatium tepidum</i> (P51762 x 1EYS)	5	75.00%	100.00%
Photosynthetic Reaction Center <i>Rhodospseudomonas viridis</i> (P06009 x 1DRX)	5	75.00%	100.00%
Photosynthetic Reaction Center <i>Rhodobacter sphaeroides</i> (P02954 x 1RZH)	5	75.00%	100.00%
P-type ATPase (P04191 x 1T5S)	10	55.26%	66.67%
Respiratory proteins – Mitochondrial ADP/ADP carrier (P02722 x 1OKC)	6	66.67%	96.29%
Respiratory proteins – Fumarate Reductase complex ( <i>Wolinella succinogenes</i> ) P17413 x 1QLA	5	71.43%	100.00%
V-type ATPase (P43457 x 2BL2)	4	90.00%	100.00%
Formate dehydrogenase-N: <i>Escherichia coli</i> (P24185 x 1KQF)	4	80.00%	100.00%
Photosystem I - <i>Thermosynechococcus elontatus</i> (P25896 x 1JBO)	11	69.69%	57.89%
AmtB ammonia channel (mutant): <i>E. coli</i> (P37905 x 1U7G)	11	66.67%	45.24%
Cytochrome c oxidase, ba3: <i>T. Thermophilus</i> (Q5SJ79 x 1XME)	13	58.40%	50.00%
Overall percentage		70.98%	80.85%

After analysing the results from *TMRelate*, an average of higher than 70% of correct predicted associations between TM regions was observed, giving promising

indications for the approach. Furthermore, the execution of the version of *TMRelate* that uses the kPROT scale (*TMRelate\_K*) resulted in an even better average of 80% correctly predicted associations. The use of the kPROT scale made the software more accurate in terms of predicting correct associations between TM regions. It predicts the buried and exposed sides of each TM region with better accuracy, which is fundamental to the algorithm that identifies the most associated (normally the most buried) TM region, making the prediction more precise than using the association matrix alone.

Considering the obtained results, *TMRelate* can be developed further before it becomes available. The findings from using its two different versions suggest that the final version has to be based on that uses the kPROT scale in order to find the associations between TM regions. The association matrix is useful in the predicting the optimum rotational arrangement i.e. angle, for each TM region (see appendix II-4).

## 10.2. Advances to the field

There are two important aspects to the approach and developed pieces of software: the use of a knowledge-based approach, based on real information to predict the best associations between TM regions in order to build predicted 3D structure; and the optimisation strategy of testing all the possible arrangements and associations by permutation.

*TMRelate* uses statistical information based on the associations between TM regions from known membrane protein structures (the association matrix).

Statistically, the more structures with high resolution of membrane proteins that become available, the better the prediction will become, since the matrices created will be more populated giving better basis to the *ab initio* prediction.

*TMRelate\_K* incorporates another knowledge-based scale, the kPROT scale (Pilpel *et al.*, 1999) that is derived from more than 5000 known membrane protein sequences deposited in the Swiss-Prot databank (Bairoch & Apweiler, 2000). The use of this scale in addition to the developed algorithm provides a useful approach for identifying TM regions, by identifying which are likely to be buried, and those that are likely to be toward the outside, making the prediction more accurate. This algorithm is unique in terms of combining knowledge-based approaches with statistics and mathematics for the prediction of membrane protein associations and the  $\alpha$ -carbon backbone 3D structure. The incorporation of the kPROT scale combined with the algorithm using the buried angle table (table A.7) is clearly advantageous, giving very accurate results in terms of prediction of the most buried TM region, and as shown in table 10.1 the average percentage of corrected predicted association between TM regions are about 10% higher.

Furthermore, the use of a permutation approach gives confidence in arriving at optimised arrangements where all TM regions have been placed in all possible positions for a chosen configuration. However, with this approach, there is a disadvantage in terms of processing time, particularly when the number of TM regions is higher than 12. This is due to the fact that the permutation is based on a factorial and for every additional TM region, there is a many fold increase in the number of calculations needed. Nevertheless, considering the number of TM regions as defined in table 10.2, the associations of a large proportion of membrane proteins

(membrane proteins with less than 13 TM regions) can be predicted in a reasonable time using the developed permutation method.

**Table 10.2 – The number of membrane proteins classified by TM regions in the Swiss-Prot database.**

Number of TM regions	Number of Swiss-Prot entries with TRANSMEM tag. Version 15-Nov-2002	Number of Swiss-Prot entries with TRANSMEM tag. Version 08-Nov-2005
1	5672 (34.41%)	9328 (34.10%)
2	1701 (10.32%)	2593 (9.48%)
3	926 (5.61%)	1380 (5.04%)
4	1427 (8.65%)	2504 (9.15%)
5	666 (4.04%)	1203 (4.40%)
6	1077 (6.53%)	1852 (6.77%)
7	1988 (12.06%)	3501 (12.80%)
8	501 (3.03%)	793 (2.90%)
9	251 (1.52%)	383 (1.40%)
10	572 (3.47%)	1079 (3.94%)
11	436 (2.64%)	712 (2.60%)
12	868 (5.26%)	1376 (5.03%)
13	142 (0.86%)	219 (0.80%)
14	107 (0.64%)	163 (0.59%)
15	29 (0.17%)	57 (0.21%)
16	17 (0.10%)	30 (0.11%)
17	35 (0.21%)	59 (0.21%)
18	3 (0.02%)	13 (0.05%)
19	3 (0.02%)	9 (0.03%)
20	1 (0.01%)	4 (0.01%)
21	–	8 (0.03%)
22	1 (0.01%)	15 (0.05%)
23	1 (0.01%)	3 (0.01%)
24	57 (0.34%)	66 (0.24%)
30	2 (0.02%)	2 (0.01%)
41	–	1 (0.004%)
Total	16483	27353

The table was derived from the analysis of two different versions of the Swiss-Prot database searching for the keyword "TRANSMEM". The first version is from 15-November-2002 and contained 16483 entries. The second version is from 08-Nov-2005 and contained 27353 entries. The number of transmembrane proteins deposited in the Swiss-Prot database has increased about 65% in three years, but the percentage distribution of the number of TM regions is similar between the two versions.

The helix wheel representation created by *TMRelate* gives a graphical output, showing the rotational angle and a position for each amino acid that allows detailed structural scrutiny. The only similar output is found in SOSUI (Hirokawa *et al.*, 1998) a secondary structure prediction tool available on the Internet (<http://sosui.proteome.bio.tuat.ac.jp/sosui/frame0E.html>). This tool gives the predicted TM region in a helix wheel representation, but without any relational

associations between the predicted TM segments, showing only a representation of the unrelated individual helix wheels side by side.

The optimal rotational angle is another feature advanced in the developed piece of software. The method developed by Pilpel and colleagues (1999) for automatic helix orientation prediction using the kPROT scale (<http://bioinfo.weizmann.ac.il/kPROT>), gives a predicted angle for each TM region. In their study, it was observed that using the kPROT scale to predict the angular orientation of each TM segment is better than hydrophobic moments (Eisenberg *et al.*, 1982, 1984; Rees *et al.*, 1989) and methods based on the statistics of known high-resolution structures of integral membrane proteins to derive lipid exposure propensities of the different residues (Cronet *et al.*, 1993; Donnelly *et al.*, 1993). However, the kPROT system does not predict the associations between TM regions; rather it builds the rotational angle for each TM region considering the known general configuration like for the *bacteriorhodopsin* and *glycophorin* family. By contrast, *TMRelate* uses two stages to obtain a full structural prediction; the optimal (highest scoring) configuration based on the association score between TM regions, and the optimal rotational angle for each TM region in relation to all other TM regions based on propensity for being buried or exposed, building a 3D structure  $\alpha$ -carbon backbone for each TM region.

### 10.3. Future work

### 10.3.1. Improvements to the developed software

When considering the many solved high resolution proteins, it is observed that  $\alpha$ -helices are not ideal cylinders fixed by linear hydrogen bonds, but up to 60% are curved or even sharply kinked (Barlow and Thornton, 1998) and 90% of the hydrogen bonds are bifurcated (Preissner *et al.*, 1999). The next step of the development is to improve the 3D model using a helix kink database of helix breaks and kinks developed by a colleague, Dina Sarakinou (Sarakinou *et al.*, 2001) using a program called *TMA $\alpha$* . This piece of software may be used to confirm alpha and beta structure, quantifying percentage alpha composition of individual TM regions and for all those in a given protein. It also calculates helix tilt, including 3D tilt with respect to particular axes, as well as precisely locating points of helix breakage, amino acids located in non-helical regions and the changes in helix tilt and orientation (kinking) that occur at given helix breaks. The resulting information is being used for the construction of a database of series of amino acids involved in helix breaks and the extent and nature of kinking brought about. This database will be used to create the tilt and the bend of a given  $\alpha$ -helix to be constructed, based on the specific sequence of the TM region.

Another approach being used to refine predicted models is the use of a GA (Genetic Algorithm). The research carried out by Noushin Minaji-Moghaddam (Minaji-Moghaddam *et al.*, unpub) uses predicted 3D models as a starting point and applies a GA to create new generations of structures. In the GA, a population of current solutions is maintained. The solutions evolve by mutations and crossovers. In computational terms, the operation consists of exchanging parts of strings between

pairs of solutions, so as to produce new solutions. Through such interactions, good features from one solution can be transferred to the others and further evolved. The solutions are evaluated by calculation of energy, by the force-field approach. The results appear promising, with populations of 3D structures showing gradual improvement by selection.

### 10.3.2. Software availability

*TMRelate* will be available soon on the research group web site (<http://membraneproteins.swan.ac.uk>), providing researchers with an important tool for predicting the associations between TM regions of membrane proteins. The user will submit a sequence in the Swiss-Prot format or a raw sequence in FASTA format, and obtain a representation with the optimal configuration based on the kPROT scale and association matrix. It will run as a batch file returning the prediction by e-mail or by accessing a given temporary web address with the result.



## **Chapter 11 - Conclusion**

Knowledge of membrane protein folding is still rudimentary and the numbers of 3D structures are small in comparison with the soluble proteins. However, in the past 5 years, the number of new unique helix bundle structures has increased to 38 (Bowie, 2005). These new structures will help the researchers working in the membrane proteins area in many ways, but especially the people working in computational prediction methods, for validating and improving existing structural prediction methods and to understand more about how primary sequence, associations between amino acids, and secondary structure are related to tertiary structure, allowing the development of new predictive methods. The reliable prediction of the basis of membrane protein 3D structure, namely the adjacencies and orientations of TM regions directly from amino acid sequence will represent a significant advance for this field.

The software tools developed here have made good ground with respect to predicting the general arrangement of TM regions, but have also provided insight into how difficult it is to develop reliable *in silico* methods to predict membrane protein 3D structure from the primary sequence. For the first stage, dealing with inter-helical associations, the predictive tool is useful for casting light the relationship between the general arrangements of amino acids that compose the TM

regions, and may be used for focusing laboratory experiments to specific residues or regions of a protein. At this stage, the generated 3D structures are highly approximate cylindrical structures, but they may however provide some reasonable ideas about general structure and serve as a baseline for laboratory structural elucidation and investigation of structure/function relationships. The created association matrix can be used as a statistical representation of the associations between amino acids that form the TM region, providing strong motivation for the development of further tools involving patterns of associations. This approach can also further understanding about the determinants of helix assembly of membrane proteins and helix packing.

Considering future development, it is an exciting time. After analysing the results obtained, it is clear that the work carried out in the project can be improved in many ways. Several modules will be incorporated in the very near future, such as evolutionary computing approaches using a genetic algorithm to refine predicted structures based on iterative evaluation of calculated free energy; and understanding the sequence determinants of helix kinking and predicting effects on structure will be an important issue. These kinks enable the small structural adjustments needed to position functional groups precisely, which could facilitate functional diversification of a common architecture (Bowie, 2006). The integration of these considerations and approaches to predicting 3D structure will greatly improve the value of the predictive approach and associated software tools.

## Glossary

**ASP** - An Active Server Page (ASP) is an HTML page that includes one or more scripts (small-embedded programs) that are processed on a Microsoft Web server before the page is sent to the user. An ASP is somewhat similar to a server-side include or a common gateway interface (CGI) application in that all involve programs that run on the server, usually tailoring a page for the user. Typically, the script in the Web page at the server uses input received as the result of the user's request for the page to access data from a database and then builds or customises the page on the fly before sending it to the requestor. ASP is a feature of the Microsoft Internet Information Server (IIS), but, since the server-side script is just building a regular HTML page, it can be delivered to almost any browser.

**BLAST** - (**B**asic **L**ocal **A**lignment **S**earch **T**ool), provides a method for rapid searching of nucleotide and protein databases. Since the BLAST algorithm detects local as well as global alignments, regions of similarity embedded in otherwise unrelated proteins can be detected. Both types of similarity may provide important clues to the function of uncharacterized proteins.

**Bootstrap** - In statistics bootstrapping is a method for estimating the sampling distribution of an estimator by re-sampling with replacement from the original sample.

**Chime** - Chime is a molecular graphics browser plugin that is freeware from MDL information systems. Chime is in part built upon the molecular

graphics rendering and command language in RasMol. However, Chime has several additional significant capabilities, such as the ability to render solvent-accessible molecular surfaces and animations (Martz, 2002).

**Class** - In object-oriented programming, a class is a template definition of the methods and variables in a particular kind of object. Thus, an object is a specific instance of a class; it contains real values instead of variables. The class is one of the defining ideas of object-oriented programming. Among the important ideas about classes are:

- A class can have subclasses that can inherit all or some of the characteristics of the class. In relation to each subclass, the class becomes the super-class.
- Subclasses can also define their own methods and variables that are not part of their super-class.
- The structure of a class and its subclasses is called the class hierarchy.

**Compiler** - A compiler is a special program that processes statements written in a particular programming language and turns them into machine language or "code" that a computer's processor uses. Typically, a programmer writes language statements in a language such as Pascal or C one line at a time using an editor. The file that is created contains what are called the source statements. The programmer then runs the appropriate language compiler, specifying the name of the file that contains the source statements.

**HMM – Hidden Markov model** - The Hidden Markov Model is a finite set of states, each of which is associated with a (generally multidimensional) probability distribution. Transitions among the states are governed by a set of probabilities called transition probabilities. In a particular state an outcome or observation can be generated, according to the associated probability distribution. It is only the outcome, not the state visible to an external observer and therefore states are ``hidden" to the outside; hence the name Hidden Markov Model.

**HTML** - (Hypertext Mark-up Language) is the set of mark-up symbols or codes inserted in a file intended for display on a World Wide Web browser page.

**hydrophobicity** - The property of being water-repellent; tending to repel and not absorb water.

**Linux** - Linux (often pronounced LIH-nuhks with a short "i") is an UNIX-like operating system that was designed to provide personal computer users a free or very low-cost operating system comparable to traditional and usually more expensive UNIX systems. Linux has a reputation as a very efficient and fast-performing system. Linux's kernel (the central part of the operating system) was developed by Linus Torvalds at the University of Helsinki in Finland. To complete the operating system, Torvalds and other team members made use of system components developed by members of the Free Software Foundation for the GNU project. Linux is a remarkably complete operating system, including a graphical user interface, an X Window System, TCP/IP, the Emacs editor, and other components usually found in a comprehensive UNIX system. Although copyrights are held by various creators of Linux's components, Linux is distributed using the Free Software Foundation's copyleft stipulations that mean any modified version that is redistributed must in turn be freely available.

**Lipid Bilayer** - A double-layer of amphipathic lipid molecules arranged with their non-polar hydrocarbon tails facing inward. These bilayers can spontaneously form under certain conditions; the plasma membranes of animal cells are formed mainly from phospholipid (phosphate-containing lipids) bilayers. The structure of the lipid bilayer explains its function as a barrier. Lipids are fats, like oil, that are insoluble in water. There are two important regions of a lipid that provide the structure of the lipid bilayer: the hydrophilic region, also called a polar head region, and the

hydrophobic, or non-polar tail region. The hydrophilic region is attracted to aqueous water conditions while the hydrophobic region is repelled from such conditions. Since a lipid molecule contains regions that are both polar and non-polar, they are called amphipathic molecules (Definition from Wikipedia - [http://en.wikipedia.org/wiki/Lipid\\_bilayer](http://en.wikipedia.org/wiki/Lipid_bilayer)).

**Neural Network** - In information technology, a neural network is a system of programs and data structures that approximates the operation of the human brain. A neural network usually involves a large number of processors operating in parallel, each with its own small sphere of knowledge and access to data in its local memory. Typically, a neural network is initially "trained" or fed large amounts of data and rules about data relationships (for example, "A grandfather is older than a person's father"). A program can then tell the network how to behave in response to an external stimulus (for example, to input from a computer user who is interacting with the network) or can initiate activity on its own (within the limits of its access to the external world).

In making determinations, neural networks use several principles, including gradient-based training, fuzzy logic, genetic algorithms, and Bayesian methods. Neural networks are sometimes described in terms of knowledge layers, with, in general, more complex networks having deeper layers. In feed forward systems, learned relationships about data can "feed forward" to higher layers of knowledge. Neural networks can also learn temporal concepts and have been widely used in signal processing and time series analysis.

Current applications of neural networks include: oil exploration data analysis, weather prediction, the interpretation of nucleotide sequences in biology labs, and the exploration of models of thinking and consciousness.

**PDB** – (Protein data bank) is a weekly updated archive of experimentally determined three-dimensional structures of biological macromolecules, serving a global community of researchers, educators, and students. The

archives contain atomic co-ordinates, bibliographic citations, primary and secondary structure information, as well as crystallographic structure factors and NMR experimental data (Berman *et al.*, 2000).

**Perl** - (Practical Extraction and Reporting Language) is a script programming language that is similar in syntax to the C language and that includes a number of popular UNIX facilities such as SED and awk.

**plugin** - Plug-in applications are programs that can easily be installed and used as part of your Web browser. The Delphi program is using a web browser component to show some html output.

**PSI-BLAST** - (Position specific iterative BLAST) refers to a feature of BLAST 2.0 in which a profile (or position specific scoring matrix, PSSM) is constructed (automatically) from a multiple alignment of the highest scoring hits in an initial BLAST search. The PSSM is generated by calculating position-specific scores for each position in the alignment. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero. The profile is used to perform a second (etc.) BLAST search and the results of each "iteration" used to refine the profile. This iterative searching strategy results in increased sensitivity.

**RSMD - Root mean square deviation** - For a set of data, we can compute the mean, and we can compute the deviation of each piece of data, i.e. how far it is from the mean. The squared deviation is the square of the deviation, and the mean squared deviation is the mean of all these squared deviations.

(Taken from: <http://thesaurus.maths.org/dictionary/map/word/3701>)

**String** - In programming, a string is a contiguous sequence of symbols or values, such as a character string (a sequence of characters) or a binary digit string (a sequence of binary values).

**Variability** - The quality, state, or degree of being variable or changeable. A quantitative measure of the degree to which scores in a distribution are spread out or clustered together. It describes the distribution by giving the distance within the distribution. Also measures how well an individual score represents the entire distribution.

**UNIX** - (often spelled "Unix" in news media) is an operating system that originated at Bell Labs in 1969 as an interactive time-sharing system. Ken Thompson and Dennis Ritchie are considered the inventors of UNIX.

The terms above are taken from: <http://www.whatis.com>



## **Appendix I**

### **TMCompare paper**

For the TMCompare paper please see

Togawa, R.C., Antoniw, J. F. and Mullins, J. G. L. (2001) 'TMCompare: transmembrane region sequence and structure', *Bioinformatics*, 17(12), pp. 1238-1239.

## Appendix II

### Algorithm descriptions

#### *1) TMCompare*

#### The algorithm

Input :        A membrane protein file in PDB format.  
Output:        A visual comparison between PDB and Swiss-Prot files, in a textual  
                  format and 3D structural view.

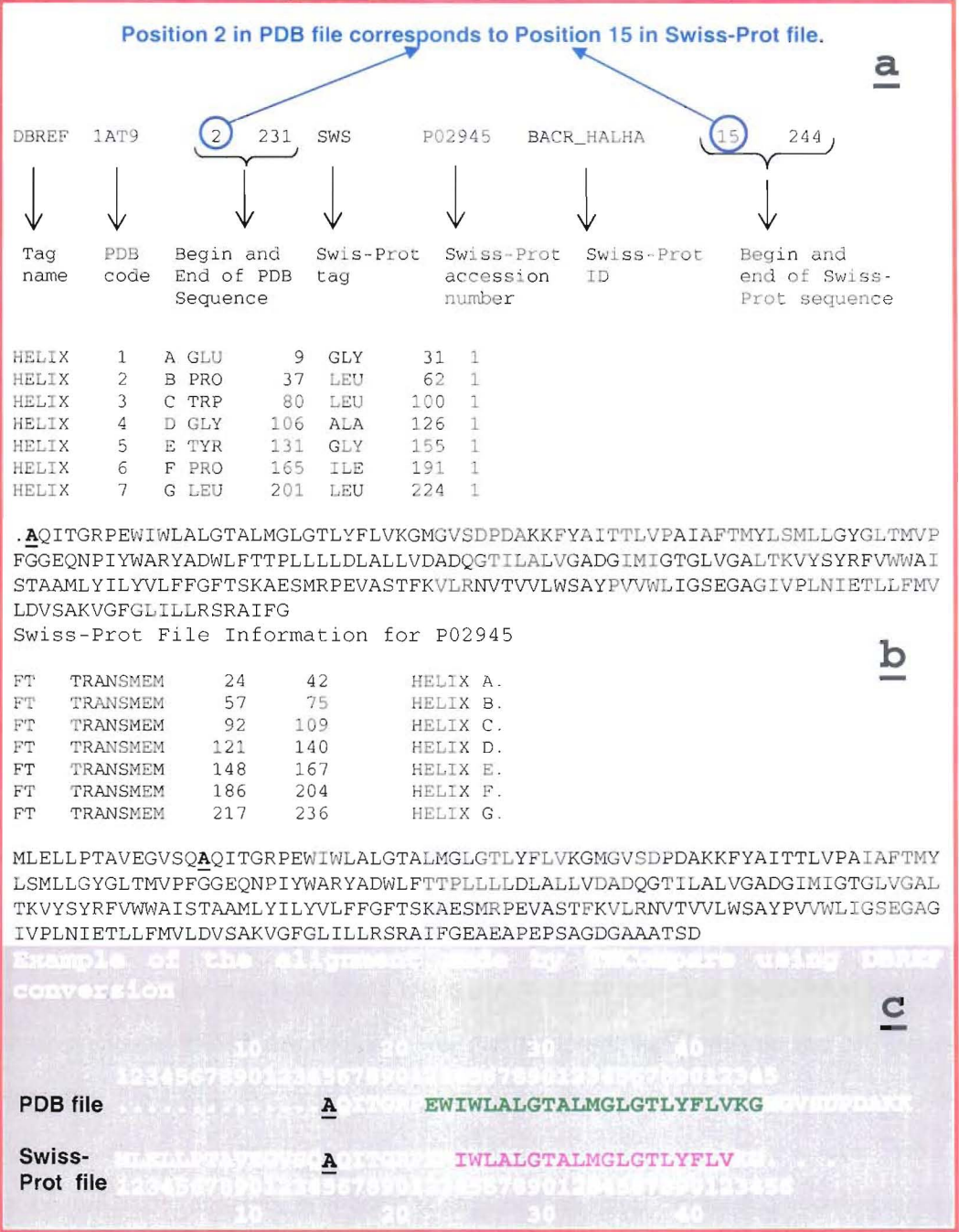
#### 1.1. User interface and definitions

To begin the program, the user operates a file selection menu and chooses the PDB file, and then *TMCompare* begins the processing of the differences between PDB and Swiss-Prot files.

From the PDB file a "DBREF" tag will be used to identify the corresponding Swiss-Prot file. This tag is especially important to the conversion of the sequence

correspondence between the PDB and Swiss-Prot file and may be used to define the new residue position in the PDB file. The figure A.1 shows how *TMCompare* superimposes the logical position (sequence) into the physical position (structure) used by the PDB file.

Figure A. 1 – The “DBREF” tag from the PDB file and its’ use in *TMCompare*



This figure shows the importance of the PDB “DBREF” tag. This tag might appear or not depending on the particular PDB file. In some cases it can have one entry for each sub-unit and subsequently more than one “DBREF” entry is found. It provides the information required to convert the sequence positions (upper blue arrows). **a)** Shows the PDB α-helix definitions and the DBREF tag. **b)** Shows the Swiss-Prot TRANSMEM tag with the TM regions definition. **c)** Shows how *TMCompare* aligns the sequence to identify the correct TM region using the PDB co-ordinates.

The RasMol script with the rendering commands are created "on the fly" when *TMCompare* is executed and activated by the buttons available on the interface. The following RasMol scripts are used to create the different TM region selections displayed in the upper frame. The following script lines are taken from the execution of *TMCompare* running with the PDB code 1AT9 (*Bacteriorhodopsin*).

**Table A.1 – Rasmol script for the 3 defined buttons**

<b>a) Rasmol Script for the left button, for the PDB <math>\alpha</math>-helix display:</b>	
Select all;	Cartoon; color White;
Select 9-31;	Color [255,000,000];
Select 37-62;	Color [000,204,000];
Select 80-100;	Color [000,000,255];
Select 106-126;	Color [255,000,255];
Select 131-155;	Color [255,128,128];
Select 165-191;	Color [255,255,000];
Select 201-224;	Color [255,128,000];
<b>b) Rasmol script for the middle button, for the defined Swiss-Prot TM regions:</b>	
Select all;	Cartoon; color White;
Select 11-29;	Color [255,000,000];
Select 44-62;	Color [000,204,000];
Select 79-96;	Color [000,000,255];
Select 108-127;	Color [255,000,255];
Select 135-154;	Color [255,128,128];
Select 173-191;	Color [255,255,000];
Select 204-223;	Color [255,128,000];
<b>c) Rasmol script for the right button, only for the Swiss-Prot TM regions hiding the structure surrounding these regions</b>	
Select all;	Cartoon off; color White;
Select 11-29;	Color [255,000,000]; Cartoon;
Select 44-62;	Color [000,204,000]; Cartoon;
Select 79-96;	Color [000,000,255]; Cartoon;
Select 108-127;	Color [255,000,255]; Cartoon;
Select 135-154;	Color [255,128,128]; Cartoon;
Select 173-191;	Color [255,255,000]; Cartoon;
Select 204-223;	Color [255,128,000]; Cartoon;

In the item (a) the 'select' command uses sequence ranges, which are taken directly from the PDB file using a 'HELIX' tag. The items (B) and (c) 'select' command uses sequence ranges, which are the result of the conversion using the 'DBREF' tag.

*TMCompare* also downloads the Swiss-Prot file from the Swiss-Prot web site. If the user chooses the "Network Swiss-Prot file" option, it will download the corresponding Swiss-Prot file using the cross-referenced Swiss-Prot accession number (using DBREF tag). If "Local Swiss-Prot file" is chosen, *TMCompare* will work with the local Swiss-Prot file located in the directory in which *TMCompare* is running. Figure A.2 shows the network option.

Figure A. 2 - Option menu in *TMCompare*



This figure shows the menu option for work with local or to download the Swiss-Prot file. The program is using the following web address at the main Swiss-Prot web site for downloading the file: <http://www.expasy.ch/cgi-bin/get-sprot-raw.pl?<Swiss-Prot code>>

## 1.2. Algorithm implementation

### Main loop

Using the loaded PDB file, the algorithm performs a loop searching for the "DBREF" tag. If a "SWS" string was found in the string, it stores the  $\alpha$ -helix positions using the HELIX tag (beginning and end of each  $\alpha$ -helix). Otherwise, it shows only the PDB file information and reads the next "DBREF" entry. After manipulating the PDB file, the program manipulates the Swiss-Prot file. It is necessary to use the corresponding Swiss-Prot accession number (from "DBREF"), and download the Swiss-Prot file from the Internet or local working directory. Then it reads the TM regions from Swiss-Prot file using the "FT TRANSMEM" tag (the transmembrane regions are obtained using HMMTOP algorithm and the predicted transmembrane regions are placed at the FT TRANSMEM tag), and stores this information in the program tables (beginning and end of each TM region). Also, using the DBREF tag, the program generates the alignments for each sequence, see figure A.1.

With all the information necessary to analyse the PDB and Swiss-Prot files, the algorithm creates 2 frames within the user interface *form* using the Web component (upper frame for 3D structure view and lower frame for sequence view). In the sequence frame coloured lines are generated showing the TM and Helix regions with corresponding alignments (see figure 6.2). For the 3D structure frame the program loads the PDB file and creates 3 buttons with chime scripts (see table A.1). The first button will select, colour and show regions of structure defined by the PDB alpha helix definition (Using 'HELIX' tag). The second button will select, colour and show the Swiss-Prot TM region (using the "FT TRANSMEM" tag) using the correspondent PDB X, Y, Z co-ordinates) and the third button will select, colour and show only the TM region (using "FT TRANSMEM" tag) hiding the structure outside the TM regions.



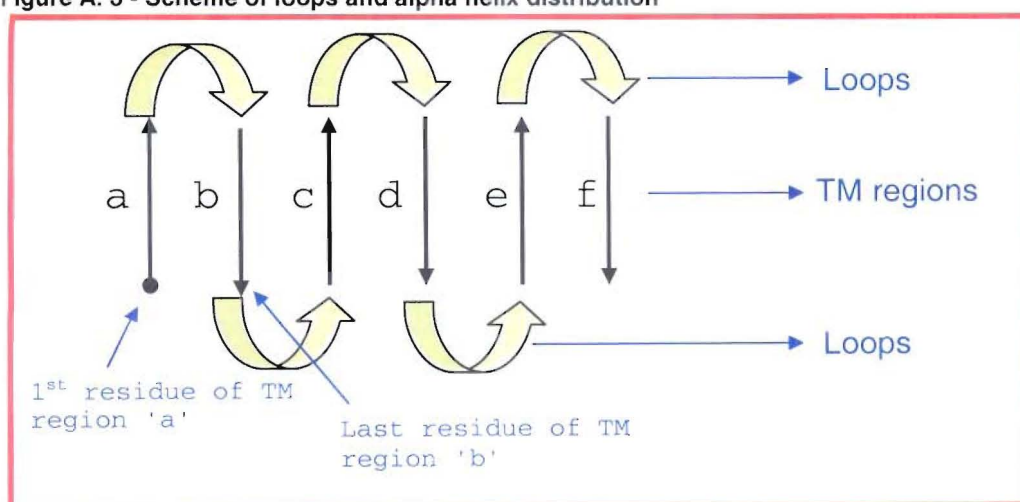
## 2) TMLimits

### 2.1. Definitions

The basis of the described algorithm rests on first finding the average spatial coordinate in terms of X, Y, and Z for the ends of the TM regions at each side of the membrane, indicated by Swiss-Prot TRANSMEM tags. The distance between the end of each individual TM region and the central point is calculated, and then the vector between this central point and the end alpha (CA) of each TM region (i.e. the average rate of change of X, Y, and Z per unit distance from the central point). From these values, and by using defined radii for virtual rim atoms, a circular rim is generated, resulting in the appearance of a circular face surrounding the central point at each side of the membrane.

Special attention is given to the beginning and the end of each side of the TM region to calculate the central point, as shown in the figure A.3:

**Figure A. 3 - Scheme of loops and alpha helix distribution**



This figure shows the direction scheme for each TM region with their corresponding loop. In order to calculate each side of the membrane, TM regions 'a' and 'b' are considered, the first  $\alpha$ -carbon residue co-ordinate of the TM region 'a' needs to be used with the last  $\alpha$ -carbon residue of the TM region 'b' and so on.

For the calculation of the central point and the circle surrounding it, the following definitions are used:

$X_{\text{centre}}$  = average X co-ordinate at centre of circle

$Y_{\text{centre}}$  = average Y co-ordinate at centre of circle

$Z_{\text{centre}}$  = average Z co-ordinate at centre of circle

$X_{\text{rim}}$  = X co-ordinate at edge of circle

$Y_{\text{rim}}$  = Y co-ordinate at edge of circle

$Z_{\text{rim}}$  = Z co-ordinate at edge of circle

$\Delta X$  = Change in X

$\Delta Y$  = Change in Y

$\Delta Z$  = Change in Z

$v\Delta X$  = average change in X per unit distance (Angstrom) (3D vector)

$v\Delta Y$  = average change in Y per unit distance (Angstrom) (3D vector)

$v\Delta Z$  = average change in Z per unit distance (Angstrom) (3D vector)

Calculating the average X, Y and Z co-ordinates at each side of the membrane gives the central point for the membrane protein on a mathematically averaged membrane face. The average change in X, Y and Z, per unit of distance is used to calculate the series of circles surrounding the central point and is calculated by the following formula:

$$(X_{\text{coord of a}} - X_{\text{coord of d}})$$

---


$$\text{Distance between } a_x \text{ and } d_x$$

The formula considers the change in X and TM regions 'a' and 'd' (see figure A.3):

## 2.2. Algorithm implementation

### **Calculating the averaged position of the membrane**

Using the loaded PDB file, and after allowing for positional differences for amino acids defined by the "DBREF" tag (the same algorithm used in *TMCompare*), the algorithm calculates the central point for each membrane face by taking the first and last X, Y, Z  $\alpha$ -carbon co-ordinates for each TM region, and calculating the average for each side. This central point is named  $X_{\text{centre}}$ ,  $Y_{\text{centre}}$  and  $Z_{\text{centre}}$ . The next step is to calculate the change in X ( $\Delta X$ ), change in Y ( $\Delta Y$ ) and change in Z ( $\Delta Z$ ) from this central point to the end of each TM region. This will be calculated by considering the sum of the distance difference between each TM

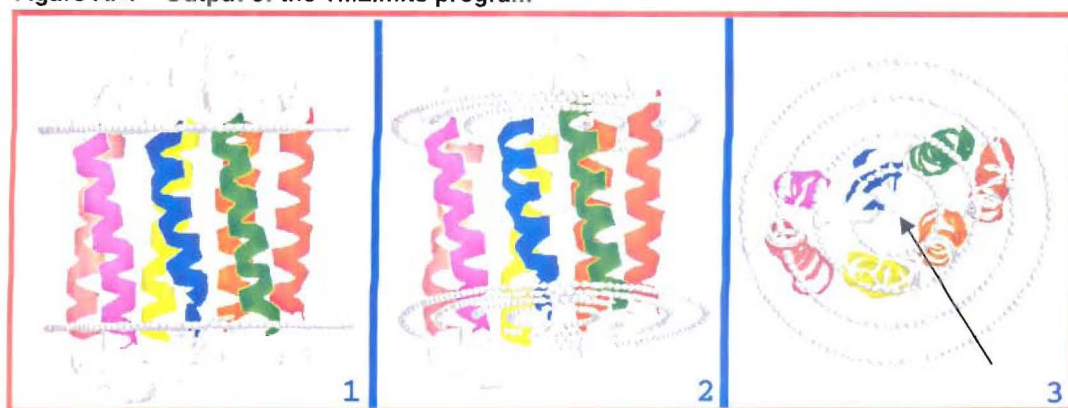


region and the centre for each co-ordinate divided by the true distance between those parts.

Using the central X, Y and Z calculated co-ordinates and the  $\Delta X$ ,  $\Delta Y$  and  $\Delta Z$ ; the algorithm will create the circles surrounding this point by incrementally increasing the radius ( $R$ ). The rim is calculated by multiplying the radius by  $\Delta$  added by the central point. Using these values a 4 times loop (to build 4 circles around the central point) is executed, calculating each point and creating the PDB like output with the 2 new chains (V and W) composing the wall of the membrane.

The final result of *TMLimits* showing the faces of the membrane is presented in the figure A.4.

**Figure A. 4 – Output of the *TMLimits* program**



**This figure shows the output created by *TMLimits*. 1) Lateral view with the membrane wall. 2) The same view but with some rotation. 3) The end-on view; the dot indicated by the arrow is the calculated centre point of the TM regions.**

This development is in the process of being converted for inclusion in the membrane proteins group web server for public use.

The development trend for the Internet applications is increasingly visually oriented. For structural analysis, it is important that the capabilities to manipulate the image (rotate, slab, zoom, etc) are available in order to better understand and more clearly study particular proteins. In this project we gave special attention to an important aspect of the design philosophy: to provide outputs in explicit graphical form. Consequently a user-friendly interface is created which makes the process of investigating new structures of membrane proteins more efficient and reliable.

### 3) TMDistance

#### The algorithm

Input :                    A membrane protein file in PDB format.  
Output:                    An association matrix.

#### 3.1. User interface and definitions

The user selects the distance range (in angstroms) using the selection box and *TMDistance* adopts this parameter for the calculation. The default distance is 3.0 Å. The user can select one or more PDB files to create the 20x20 association matrix, by pressing the 'ctrl' key in the file selection menu.

The TM regions extracted from the Swiss-Prot file ("TRANSMEM" tag) will be used instead of the PDB "HELIX" tag. For this purpose a "DBREF" tag from the PDB file is used for the conversion (see figure A.1). If the corresponding Swiss-Prot file is not in the working directory, *TMDistance* downloads the Swiss-Prot file from the Swiss-Prot web site (<http://www.expasy.ch>).

The internal mathematical structure of the created matrix is a bi-dimensional array with 400 positions (matrix [i,j]), corresponding to the 20x20 for the different amino acids. 'i' and 'j' vary from 1 to 20 and correspond to the amino acids in alphabetical order as follows:

[1] := 'ALA'	[11] := 'LEU'
[2] := 'ARG'	[12] := 'LYS'
[3] := 'ASN'	[13] := 'MET'
[4] := 'ASP'	[14] := 'PHE'
[5] := 'CYS'	[15] := 'PRO'
[6] := 'GLN'	[16] := 'SER'
[7] := 'GLU'	[17] := 'THR'
[8] := 'GLY'	[18] := 'TRP'
[9] := 'HIS'	[19] := 'TYR'
[10] := 'ILE'	[20] := 'VAL'

## 3.2. Algorithm implementation

### Generation of the association matrix

To create the association matrix, the following steps are executed for the chosen PDB file(s). For each PDB file, the algorithm searches for the "DBREF" tag entry. Once the tag is found, it searches for "SWS" string to find the appropriate Swiss-Prot accession number. If it is necessary the program downloads and saves the corresponding Swiss-Prot file in the working directory. If it is necessary *TMDistance* converts the amino acid sequence numbers between PDB and Swiss-Prot files using the information contained in the DBREF tag (see figure A.1). Then the algorithm creates and saves a temporary PDB file with the corresponding TM region from the Swiss-Prot file saved into the new PDB file (keeping the new HELIX tag) to be used in the next step. Using the created PDB file, the program reads the spatial co-ordinates for the atom in each TM region. With each residue pair in different TM regions, if the distance between the two residues is less than a user-selected distance (calculated by the distance formula), the relevant residue-pair is added to the internal bi-dimensional array (matrix counter). After all the PDB file(s) are read, *TMDistance* creates the matrix output with the average distances represented in the internal bi-dimensional array.

## 4) TMRelate

### The algorithm

- Input :        A 20x20 association matrix and a membrane protein sequence file in the Swiss-Prot format.
- Output:        Predicted associations between TM regions in a graphical output and a predicted 3D model for the TM regions of the whole structure.

## 4.1. User interface and definitions

To begin using *TMRelate*, the user chooses the appropriate overall configuration. (i.e. the positions in which each possible combination of TM regions are to be tested) by pressing the buttons. The program gives an error if the number of TM regions in the Swiss-Prot file is less than the number made available in the selected configuration. To find how many TM regions are in the Swiss-Prot file, the algorithm counts the "TRANSMEM" tags.

The next required input is the loading of the association matrix, previously created and saved by the *TMDistance* module. This matrix is stored in an internal bi-dimensional array for ready use.

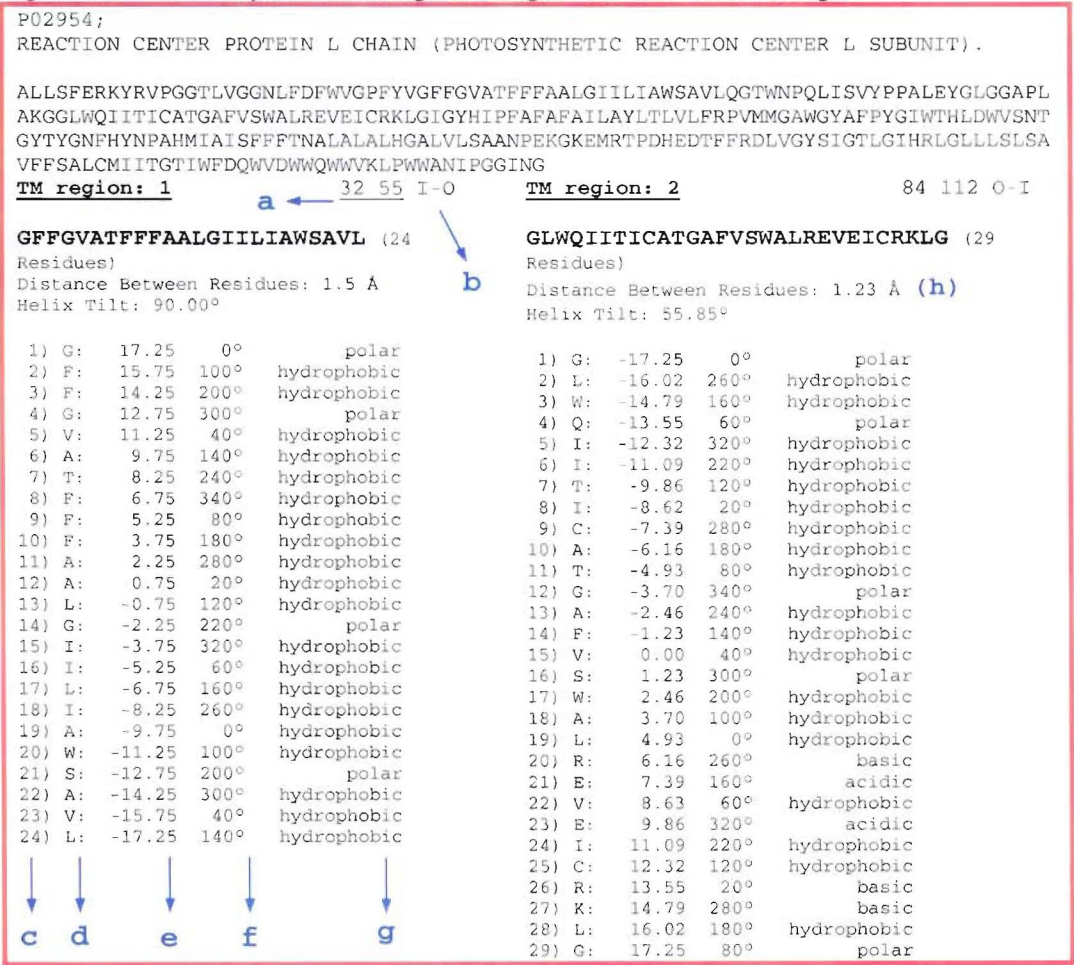
The final input required is the Swiss-Prot file containing the membrane protein sequence and defined TM regions information. Then *TMRelate* will start to predict the associations between available TM regions.

*TMRelate* will extract, calculate or associate the following information from the Swiss-Prot file:

- The Swiss-Prot Identification: Extracted using the 'ID' tag from the Swiss-Prot file.
- The Swiss-Prot accession number: Extracted using the 'AC' tag from the Swiss-Prot file.
- The Swiss-Prot last updated information: Extracted using the 'DT' tag from the Swiss-Prot file.
- The Swiss-Prot description: Extracted using the 'DE' tag from the Swiss-Prot file.
- The amino acid residues sequence: Extracted using the 'SQ' tag from the Swiss-Prot file.
- The amino acid residues sequence length: Calculated by counting the number of amino acid in the sequence.
- The transmembrane (TM) region: Extracted using the 'FT TRANSMEM' tag from the Swiss-Prot file.
- The number of the first and last amino acid of each TM region: Extracted using the 'FT TRANSMEM' tag from Swiss-Prot file.
- The length of each TM region: Calculated as indicated below:
  - $TMLength := (\text{sequential number of the last residue} - \text{sequential number of the first residue}) + 1$
- The lowest TM length: Obtained by calculating the length of each transmembrane region, and then selecting the lowest.
- The sequence loops: Extracted from the amino acid sequence. In the program, every amino acid outside the TM region is considered loop.

- The membrane thickness: Calculated by multiplying the lowest TM region length by 1.5 (the rise along the  $\alpha$ -helix for each amino acid is 1.5 Å).
- The N-terminus position based on the inside-positive rule (Von Heijne, 1992): Obtained by the ratio between the positive and negative charges of amino acids found on the extra-membrane loops. If the positive:negative ratio on the inside is higher than that on the outside, the N-terminus is inside of the membrane. If the positive:negative ratio on the inside is lower than that on the outside, the N-terminus is outside the membrane.
- Designation of an angle for each residue (this angle will be used to simulate the TM region rotation): For each residue is given an angle value. This angle corresponds to the rotational position in the  $\alpha$ -helix. An  $\alpha$ -helix has 3.6 amino acids per turn, making 100° the difference between each amino acid. The angle zero is given for the first residue in each TM region and an increment of 100° is added to the next residue position. Starting from one side of the membrane, the angle is increased by 100° for each subsequent amino acid, and starting from the other side, the angle is decreased by 100° (see figure A.5).

Figure A. 5 - An example of the designated angle values for each TM region



The associated angle value for each residue must be between 0° and 360°. The increase or decrease of the angle and the angle range control is executed by the following code:

```

If (helix number is even) then
  Begin
    Angle := Angle - 100
    If (Angle < 0) then
      Angle := Angle + 360
    End
  Else if (helix number is odd) then
    Begin
      Angle := Angle + 100
      If (Angle >= 360) then
        Angle := Angle - 360
      End
    End
  End

```

- Calculation and association of the distance between each residue (The calculated distance between each residue is used to simulate the depth of each residue in the membrane): In the  $\alpha$ -helix each residue receives a distance attribute, and it varies according to the TM region length. It is calculated by dividing the thickness of the membrane by the number of amino acids in each TM region. The middle of the membrane is considered zero. For one side the value increases incrementally for each amino acid by the obtained distance. For the other side the value decreases by the obtained distance until it reaches the predicted edge of the membrane as defined by the membrane thickness value (figure A.5 (h)).
- Calculation of the average theoretical 2D TM region tilt: Calculated by the following formula:

$$\text{Helix TILT} := \text{Arcsine} \left( \frac{(\text{Number of amino acids in shortest TM region})}{(\text{Number of amino acids in TM region of interest})} \right) \times 180 / \pi$$

To calculate the association scores for each pair of TM regions, *TMRelate* considers the intra-membrane amino acid depth. For each pair of amino acids in different TM regions, if the designated depth values for the amino acids are less than 1.5 Å, the program will take the appropriate value from the 20x20 matrix. Then an accumulative score will be calculated for the predicted association between each pair of TM regions. The higher this score the more likely the TM regions are to be compatible for association.

*TMRelate* will create a list of associations with correspondent TM pair score as show in the table A.2:



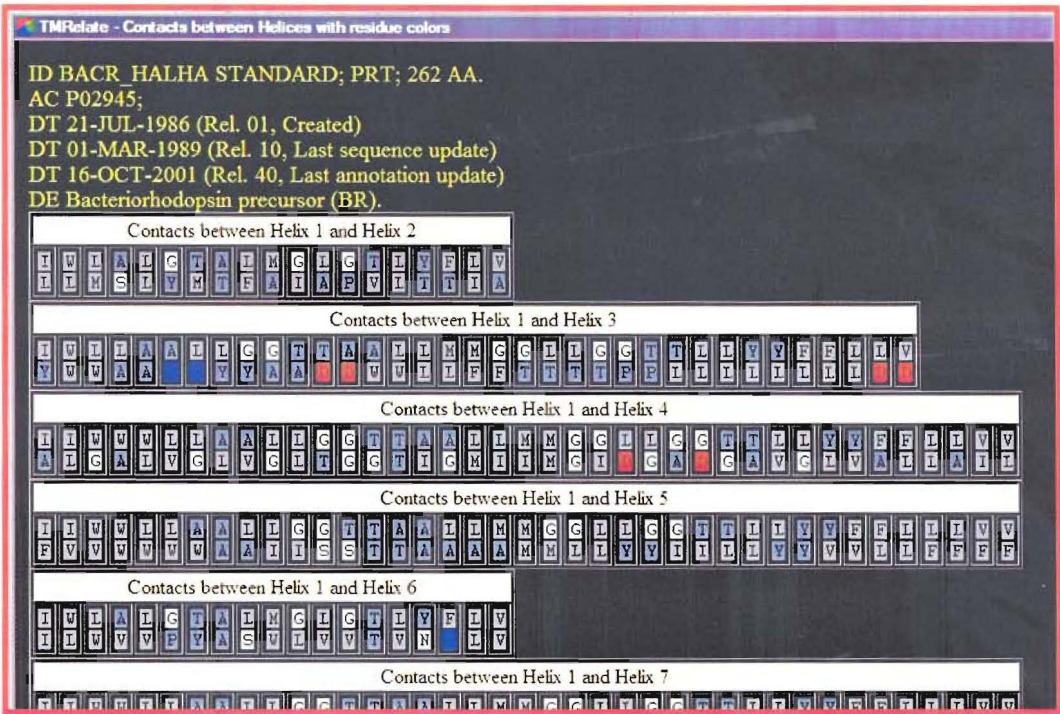
Table A.2 - The obtained score for each pair of TM regions

Scores between TM regions	
TM region number	Score
6 and 2	14.970
6 and 7	15.120
6 and 5	15.430
6 and 1	15.460
4 and 7	15.480
6 and 4	15.630
6 and 3	15.660
4 and 3	16.780
2 and 4	16.850
5 and 7	17.190
4 and 5	17.450
2 and 5	17.870
2 and 3	17.900
7 and 2	17.910
5 and 1	18.050
4 and 1	18.240
2 and 1	18.850
5 and 3	18.920
7 and 3	19.410
7 and 1	19.700
1 and 3	20.170

This table shows an example of the scores between TM regions obtained from *TMRelate* output. The predicted protein is *bacteriorhodopsin*) – Swiss-Prot ID: P02945. After considering all the possible pairing of amino acids in different TM regions, the algorithm finds the score for each pair of TM regions using the association matrix, and places them in ascending order.

*TMRelate* can also analyse the pairs of amino acids in possible association between TM regions using an option that allows examination of the alignment using the amino acid colour code. This coloured alignment (figure A.6) can be used to assess the compatibility of TM regions in terms of physical and chemical consistency. This alignment is generated using the same routine through which the association score was created. When the algorithm finds pairs of amino acids of similar depth, it adds the matrix score and shows the association using the amino acid colour code. The colour code used in the output is described in the materials and methods chapter.

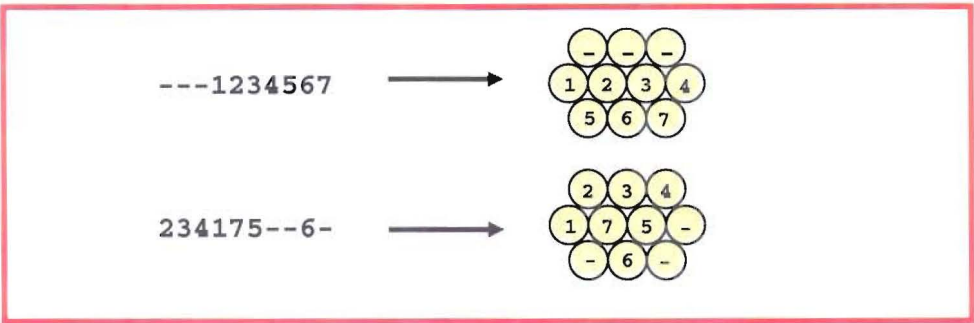
Figure A. 6 - Residues of similar membrane depth ( $\pm 1.5 \text{ \AA}$ ) on different TM regions by alignment using colour coding



This is a sample of TM regions using the amino acid colour code. The loaded protein is *Bacteriorhodopsin with Swiss-Prot accession number: P02945*

The algorithm uses a permutation concept, calculating all possible scores for each TM region in each position. The permutation combines the scores between pairs of adjacent TM regions. A 10-digit string list is used to generate the permutations. This string represents the position in "end on view" of the predicted membrane protein (figure A.7).

Figure A.7 - 10-digit string and the corresponding end on view configuration

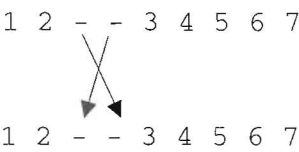


This figure shows the 10-digit string used in the algorithm. The dash (-) represents the non-used position, and the number represents the TM region. The digits on the left shows how the algorithm is operating internally, and the circles with numbers on the right, shows how the user views the end on configuration.



The algorithm uses a sequential text file containing the permutation list. The reason for this is to avoid the extensive processing time required to build the permutation list and to discard the repeated ones.

The permutation needs to discard repeated combinations, such as the following:



In the example above, when the 3<sup>rd</sup> and 4<sup>th</sup> positions are changed around, the resulting permutation is the same. To create the permutation file, a 'Perl' program from the web site: [http://www.rocketaware.com/perl/perifaq4/How do I permute N elements of a.htm](http://www.rocketaware.com/perl/perifaq4/How%20to%20permute%20N%20elements%20of%20a%20list.html) (Christiansen & Torkington , 1997) based on a Unix platform was used and to discard the repeated position a *Unix sort* command with the '*unique*' parameter was used.

Table A.3 shows the permutation file with different numbers of TM regions:

Table A.3 - An example of the permutation file

2 TM regions	5 TM regions	7 TM regions	8 TM regions
-----12	-----12345	---1234567	--12345678
-----21	-----12354	---1234576	--12345687
-----1-2	-----12435	---1234657	--12345768
-----12-	-----12453	---1234675	--12345786
..	..	..	..
.	.	.	.
.	.	.	.
2---1-----	5432---1--	765432---1	87654312--
2--1-----	5432--1---	765432--1-	8765432--1
2-1-----	5432-1----	765432-1--	8765432-1-
21-----	54321-----	7654321---	87654321--

This table shows a sample of the permutation files for 2, 5, 7 and 8 transmembrane regions. A dash (-) is used to indicate positions that are not in use for a TM region.

For each number of TM regions the following files were created:

Table A.4 - The permutation file size statistics

File Name	Number of permutations (10!)	Number of unique permutations	File size
uniq_perm_2_10.txt	3,628,800	90	2 Kb
uniq_perm_3_10.txt	3,628,800	720	9 Kb
uniq_perm_4_10.txt	3,628,800	5,400	60 Kb
uniq_perm_5_10.txt	3,628,800	32,240	355 Kb
uniq_perm_6_10.txt	3,628,800	151,200	1.772 Kb
uniq_perm_7_10.txt	3,628,800	604,800	7.088 Kb
uniq_perm_8_10.txt	3,628,800	1,814,400	21.263 Kb
uniq_perm_9_10.txt	3,628,800	Not in use 3,628,800	42.525 Kb
uniq_perm_10_10.txt	3,628,800	Not in use 3,628,800	42.525 Kb

This table shows the number of unique permutations and the file size. For each number of TM regions there is a different file.

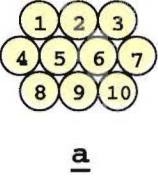
The uniq\_perm\_9\_10.txt and uniq\_perm\_10\_10.txt files are not in use, because in these two situations, no repeated permutations are generated. For the 10<sup>th</sup> digit the letter 'A' was used as a hexadecimal numbering.

- 123456789- <- No repeated permutation for this string
- 123456789A <- Using the letter 'A' for the 10<sup>th</sup> digit

To get a score for each TM region association, the algorithm uses the neighbour association table. For the 10-digit configuration, the following association table is in use:

Table A.5 - The neighbour association table

TM region	Associated TM region
1	2,4,5
2	1,3,5,6
3	2,6,7
4	1,5,8
5	1,2,4,6,8,9
6	2,3,5,7,9,10
7	3,6,10
8	4,5,9
9	5,6,8,10
10	6,7,9

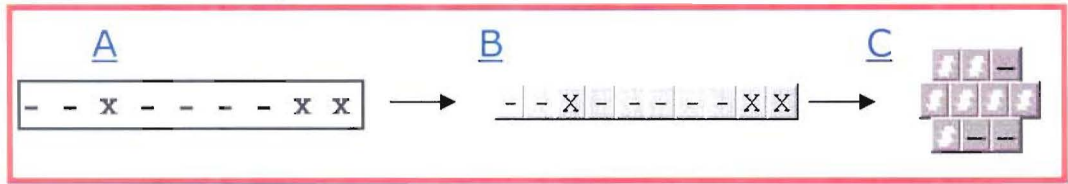


a

In this table the circles (a) illustrate the end on view for each TM region. Considering the first line: TM region 1 has associations with TM regions 2,4,5. The second line: TM region 2 has associations with TM regions 1,3,5,6 and so on.

Another important feature implemented during the development was the user end on configuration buttons, which evolved from the linear textual configuration to the end on view buttons:

Figure A. 8 - The development of the configuration buttons



This figure shows how the configuration buttons evolved from the original idea. A) The user needs to type the configuration in the box. B) The user clicks on the button and chooses the required configuration. C) The user selects the appropriate configuration using the end-on graphic.

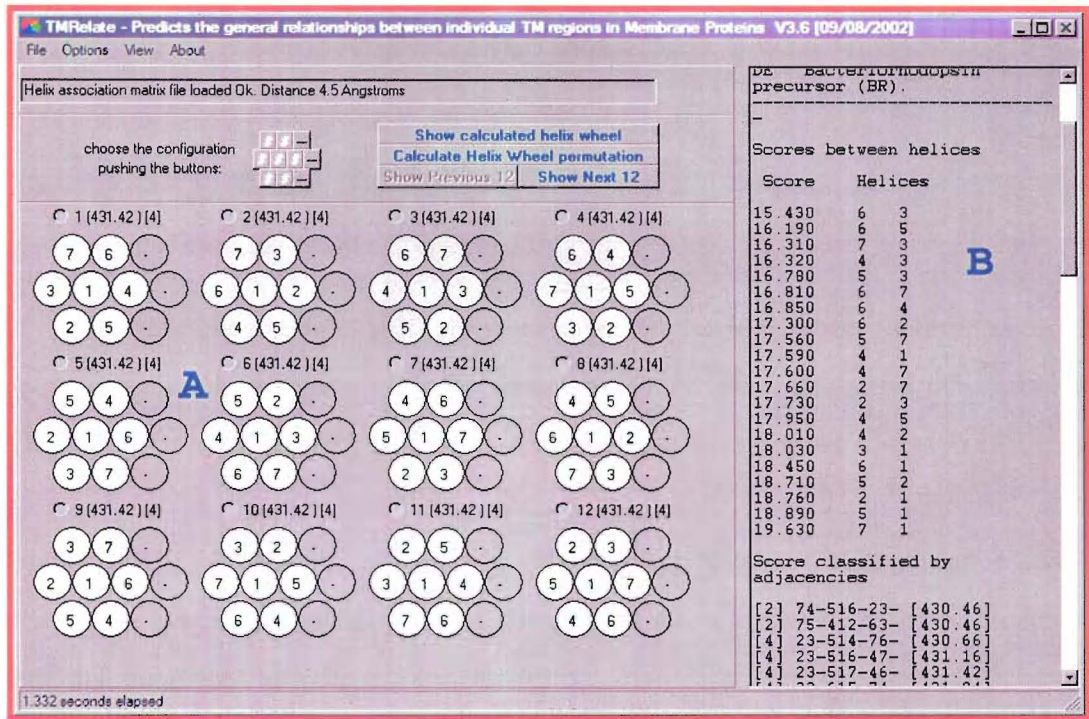
Improvements were also made to the structure of the permutation file. The use of the letter 'A' for the 10<sup>th</sup> digit made it possible to work with no spaces between each digit. The original permutation file used one space between digits, i.e.: [1 2 3 4 5 6 7 8 9 10]. Now each permutation is represented as: [123456789A]. This change reduced the file size by about 40% by merely cutting off the spaces. Also, *TMRelate* was modified to no



longer use the 9 and 10 TM permutation files, instead creating its own internal permutation for these conditions, again with non-repeated digits.

*TMRelate* displays the top 50 scores, and creates a navigation button for viewing the next 12 and previous 12 in the list as shown in figure A.9.

Figure A. 9 – *TMRelate*: The user interface



This figure shows the *TMRelate* program running. A) The predicted end on view for the *Bacteriorhodopsin* with Swiss-Prot accession number: P02945. B) The textual information related to the scores between TM regions. TM regions 1 and 7 have the highest score, indicating the TM region most likely to be associated.

## 4.2. The helix Wheel output

*TMRelate* creates a helix wheel representation using the chosen configuration. For this step the algorithm rotates each of the TM regions by 60° at a time, and for each rotation a score is calculated. The rotation works like an odometer, in which each TM region has a complete turn. After this turn, the next TM region is rotated by 60° until all TM regions have completed one whole turn. For each rotational position, the combinations of all the TM Regions are scored. Again the score calculation is based on the 20x20 association matrix. In the calculation of the score for each pair of TM regions, 2

Where RA is decreased by 100 for each consecutive residue, and:

$r = \text{radius of the helix} = (3.817719/3.6478)*(((3.817719*3.6))/\text{PI}())^2$ , again obtained from sampling helical structures.

In the event that the helix has a starting end on rotation (EOR) calculated by the present helix wheel program, then the equation for x is:

$$x = (\text{SIN}((\text{EOR} + \text{RA}) * \text{PI}() / 180)) * r$$

Similarly, for z:

$$z = \text{SIN}((\text{EOR} + \text{RA} - 90) * \text{PI}() / 180) * r$$

The only difference between the x and z equations is the prior subtraction of 90 from the rotation angle.

## 4.4. Algorithm implementation

### **Main Loop**

In this step, the algorithm loads a Swiss-Prot file and tests if the number of TM regions match with the end-on view chosen by the user. In the next step the routine to Calculate the association score is executed (described in the next paragraph). The algorithm will execute a permutation procedure if the number of TM regions is higher than 8, otherwise, the algorithm loads a correspondent permutation file to be used in the program (see table A.4). A loop to read the permutation list and calculate the association score will then be executed. This routine uses the neighbour table and calculates the *TM\_region\_pair\_score\_array*. A score between each pair of matched TM regions is calculated, and at the end the result is accumulated in the *intermediate\_total* counter. For each permutation, this total is added in the *sorted\_output\_list* with 50 positions (that is the top 50 scores). At the end of all permutations, using the *sorted\_output\_list*, the program shows the top 50 best score configuration in descending order as shown in figure A.9.

### **Calculate the association score**

This routine calculates the association score between different TM regions using the 20x20-association matrix. Using the loaded Swiss-Prot file, the algorithm works with each TM region and tries to match residues in each of the TM regions. If the assigned depth between residues in different TM regions is equal or less than 1.5Å, then, the association score between these two residues obtained from

the matrix is added and accumulated to the *TM\_region\_pair\_score\_array* (see table A.2 as an example). The routine returns the array containing the score for each pair of TM regions (*TM\_region\_pair\_score\_array*).

### **Rotating the TM regions**

This routine rotates the TM regions, searching for the best score after trying all possible rotational positions. Again this routine uses the 20x20 association matrix. For each rotational position, the algorithm performs a loop for each residue in each different TM region. If the depth between residues in different TM region is equal to or less than 1.5 Å and the angle range between the 2 residues is equal to or less than 60°, then the association score between these two residues obtained from the matrix is added and accumulated to the *Helix\_wheel* variable. The algorithm holds the highest score of the rotated configuration with the respective angles. At the end of testing all possible rotational positions, the arrangement highest rotational score is depicted as a helix wheel representation. (figure 8.1).

### **Creating $\alpha$ -helix 3D structure**

This routine builds the 3D structure based on the arrangement with the highest rotational score. It uses the rotational angles found by the *Rotating\_the\_TM\_regions* routine. Using the definitions described above it builds the CA backbone for each TM region. Special attention is required regarding the direction of the TM region, i.e. when it is passing from inside to outside and vice-versa. For each residue, the position of the alpha carbon (CA) is calculated.

## **5) TMRelate K**

### **The algorithm**

Input :        A 20x20 association matrix and a membrane protein sequence file in the Swiss-Prot format.

Output:        Predicted associations between TM regions in a graphical output and a predicted 3D model for the TM regions of the whole structure.

## 5.1. User interface and definitions

The basic algorithm is the same as described for *TMRelate*. It uses the same permutation algorithm, except for the associations between TM regions, where it uses the kPROT scale. The algorithm calculates the aggregate the kPROT score for each TM region and uses this value to find the optimal configuration (helix packing).

To define the helix packing for the predicted membrane protein using kPROT scale, the algorithm identifies how many TM regions are buried (TM region that is in the interior of the membrane protein) and exposed (TM region that is exposed to the lipid) depending on the number of TM regions in the protein. For example, for the *Bacteriorhodopsin*, protein with seven TM regions, the algorithm considers two TM regions buried and five exposed. Table A.6 shows the numbers (buried and exposed) used by the algorithm.

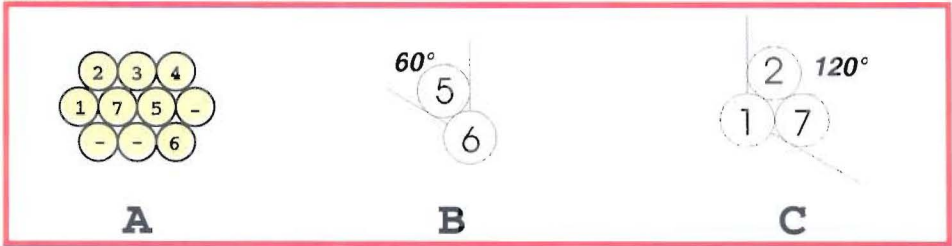
**Table A.6 – *TMRelate\_K* algorithm: helix packing definition**

Number of TM regions in Membrane protein	Number of TM region(s) 'buried'	Number of TM regions 'exposed'
3	1	2
4	1	3
5	1	4
6	1	5
7	2	5
8	2	6
9	2	7
10	3	7
11	3	8
12	3	9

**Variation in the overall number of buried and exposed TM regions, depending on the numbers of TM regions in the protein.**

To predict the helix packing the algorithm calculates a score using the kPROT scale and gives a weighting based on the number of associations for each TM region. Each association between TM regions contributes 60° to the extent of "buriedness". Looking at figure A.10, TM region 6 has one association with TM 5, and the algorithm considers it as 60° buried. For TM region 1 there are 2 associations, and 120° buried and so on.

Figure A.10 – Buried angle



(A) An example of an end on configuration. (B) Detail for the association between TM 6 and TM 5: the buried angle is 60°. (C) Detail for the association between TM 1,2 and 7: the buried angle is 120°.

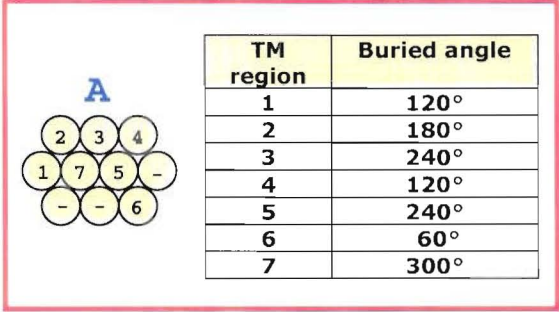
The following table shows the buried angle for each number of TM region/TM region associations:

Table A.7 - The buried angle

Number of TM region(s) associations	Buried angle
0	0°
1	60°
2	120°
3	180°
4	240°
5	300°
6	360°

The buried angle depending on the number of TM region associations

Figure A.11 – Example how the algorithm consider the buried angle



Taking the configuration above (A), the buried angle for each TM region is as shown in the right hand table.

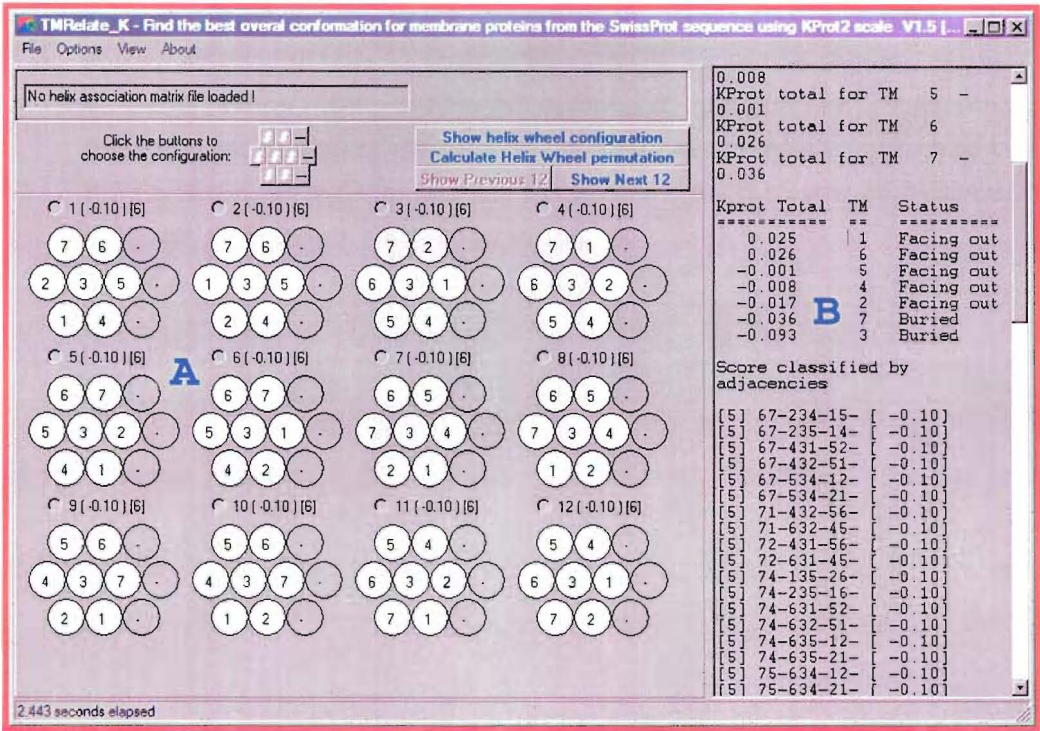
The angles shown in the table A.11 would be used for the score calculation. The rationale is to use a buried angle range depending on the number of possible associations each TM region can have. The buried angle provides a higher weighting for TM regions that have more associations, leading to a higher weighted contribution from the kPROT aggregate scores. The buried angle is used as in the formula below.



$$\text{kPROTHelixScore} := (\text{Buried angle}/360) * \text{kPROT Score For This TM region}$$

The following output is obtained from the *TMRelate\_K* program:

Figure A.12 - The *TMRelate\_K* user interface



This figure shows the *TMRelate\_K* program executing a prediction for the *Bacteriorhodopsin* protein with Swiss-Prot accession number: P02945. A) The numbered circles represent the end on view of the TM regions. B) On the indicated textual information, *TMRelate\_K* shows the aggregated kPROT scores for the TM regions of the predicted protein. The buried TM regions are predicted to be 3 and 7.

## 5.2. Algorithm implementation

### kPROT calculation

The algorithm is similar to *TMRelate*, in terms of loading the Swiss-Prot file, and extraction of information as described in section 4.1 The deviation begins when the program calculates the kPROT score for each TM region. To find the total kPROT score for each TM region, the program reads the amino acid residues in each TM region and adds the corresponding kPROT score from the scale (table 5.7), at the end of this loop, the aggregate score for each TM region is given. The next step is to find the highest scoring configuration using the aggregate kPROT scores for each TM region. A loop to read the entire permutation list and calculate the association score is executed. This routine uses the weighted neighbour table (table A.7) and the calculated kPROT aggregate for each TM region. A score for each TM region is calculated according to the following equation:

$\text{kPROT\_TM\_Score} := (\text{Buried angle}/360) * \text{kPROT Score For This TM region}$

Then the routine accumulates the partial score into *intermediate\_total* counter variable. For each permutation, this total is added into the *sorted\_output\_list* with 50 configurations (with the top 50 scores). At the end of all permutations, using the *sorted\_output\_list*, the program shows the top 50 highest score configurations as shown in figure A.12. The steps taken to build the helix wheel representation and the 3D structure are the same as described in the *TMRelate* algorithm.

### **Appendix III**

*TMRelate* end on view evaluation

**Bacteriorhodopsin**

Swiss-Prot code used to run TMRelate : P02945  
 PDB Code used to evaluate the model : 1C3W  
 Number of transmembrane regions : 7

**Prediction 1 - "horse shoe" configuration:**

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	16/22 coincident associations 72.73%	16/22 coincident associations 72.73%	16/22 coincident associations 72.73%	16/22 coincident associations 72.73%
2341765---	4326517---	7143256---	4326517---	5126437---

**Prediction 2 - "rosette" configuration:**

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	16/24 coincident associations 66.67%	14/24 coincident associations 58.33%	18/24 coincident associations 75.00%	16/24 coincident associations 66.67%
17-236-45-	76-453-21-	76-312-45-	24-315-76-	56-413-27-

**Prediction 3 - unspecified configuration:**

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	16/24 coincident associations 66.67%	14/24 coincident associations 58.33%	18/24 coincident associations 75.00%	14/24 coincident associations 58.33%
17-236-45-	13-754-62-	74-315-26-	-24-315-76	73-214-65-

**Rhodopsin**

Swiss-Prot code used to run TMRelate : P02699  
 PDB Code used to evaluate the model : 1U19  
 Number of transmembrane regions : 7

**Prediction 1 - with "horse shoe" configuration:**

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	14/22 coincident associations 63.64%	14/22 coincident associations 63.64%	14/22 coincident associations 63.64%	14/22 coincident associations 63.64%
7651234---	5324761---	1275634---	6351247---	3654217---

**Prediction 2 - with "rosette" configuration:**

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	14/24 coincident associations 58.33%	14/24 coincident associations 58.33%	14/24 coincident associations 58.33%	14/24 coincident associations 58.33%
76-135-24-	47-635-21-	74-532-16-	74-536-12-	75-316-42-

**Prediction 3 - with unspecified configuration:**

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	14/24 coincident associations 58.33%	12/24 coincident associations 50.00%	14/24 coincident associations 58.33%	14/24 coincident associations 58.33%
76-135-24-	15-724-36-	74-321-65-	75-431-26-	16-532-74-



Sensory Rhodopsin II (HR)

Swiss-Prot code used to run TMRelate : P42196  
 PDB Code used to evaluate the model : 1H2S  
 Number of transmembrane regions : 7

Prediction 1 - with "horse shoe" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	16/22 coincident associations 72.73%	14/22 coincident associations 63.64%	14/22 coincident associations 63.64%	12/22 coincident associations 54.55%
7651234---	5347621---	2173456---	6542713---	3762154---

Prediction 2 - with "rosette" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	14/24 coincident associations 58.33%	18/24 coincident associations 75.00%	16/24 coincident associations 66.67%	18/24 coincident associations 75.00%
76-135-24-	76-425-31-	24-315-76-	32-417-56-	45-216-37-

Prediction 3 - with unspecified configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	14/24 coincident associations 58.33%	18/24 coincident associations 75.00%	16/24 coincident associations 66.67%	18/24 coincident associations 75.00%
76-135-24-	75-126-34-	23-417-56-	23-714-65-	76-315-24-

Halorhodopsin

Swiss-Prot code used to run TMRelate : P16102  
 PDB Code used to evaluate the model : 1E12  
 Number of transmembrane regions : 7

Prediction 1 - with "horse shoe" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	12/22 coincident associations 54.55%	14/22 coincident associations 63.64%	14/22 coincident associations 63.64%	12/22 coincident associations 54.55%
7651234---	2364517---	2516347---	6517342---	6517324---

Prediction 2 - with "rosette" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	16/24 coincident associations 66.67%	16/24 coincident associations 66.67%	14/24 coincident associations 58.33%	14/24 coincident associations 58.33%
76-135-24-	42-735-61-	17-456-23-	73-156-42-	76-413-25-

Prediction 3 - with unspecified configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	14/24 coincident associations 58.33%	16/24 coincident associations 66.67%	14/24 coincident associations 58.33%	14/24 coincident associations 58.33%
76-135-24-	76-451-23-	17-456-23-	74-312-56-	14-752-36-



Aquaporin

Swiss-Prot code used to run TMRelate : P06624  
 PDB Code used to evaluate the model : 1YMG  
 Number of transmembrane regions : 6

Prediction 1 - with "horse shoe" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	08/10 coincident associations 80.00%	10/10 coincident associations 100.00%	08/10 coincident associations 80.00%	08/10 coincident associations 80.00%
312---5-64	254---6-31	312---5-64	452---1-63	452---1-63

Prediction 2 - with "rosette" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	06/12 coincident associations 50.00%	12/12 coincident associations 100.00%	08/12 coincident associations 66.67%	08/12 coincident associations 66.67%
25-1-4-36-	65-2-4-13-	25-1-4-36-	14-3-5-62-	63-2-1-54-

Prediction 3 - with unspecified configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	08/10 coincident associations 80.00%	08/10 coincident associations 80.00%	08/10 coincident associations 80.00%	08/10 coincident associations 80.00%
312---4-65	26--15-34-	316-52--4-	624-15--3-	315-624---

Glycerol uptake facilitator protein(Aquaglyceroporin).

Swiss-Prot code used to run TMRelate : P11244  
 PDB Code used to evaluate the model : 1FX8  
 Number of transmembrane regions : 8

Prediction 1 - with "horse shoe" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	16/26 coincident associations 61.54%	16/26 coincident associations 61.54%	16/26 coincident associations 61.54%	16/26 coincident associations 61.54%
26513784--	71463285--	58436127--	85436127--	85436127--

Prediction 2 - with "rosette" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	16/26 coincident associations 61.54%	16/26 coincident associations 61.54%	16/26 coincident associations 61.54%	18/26 coincident associations 69.23%
26513784--	71463285--	25836147--	51486372--	83156427--

Prediction 3 - with unspecified configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	18/26 coincident associations 69.23%	14/26 coincident associations 53.85%	12/26 coincident associations 46.15%	16/26 coincident associations 61.54%
26513784--	736-125-48	714268-35-	81-265-437	785-361-42




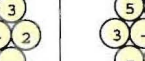
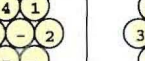


Photosynthetic Reaction Center*Thermochromatium tepidum*Swiss-Prot code used to run TMRelate : P51762

PDB Code used to evaluate the model : 1EYS

Number of transmembrane regions : 5

Prediction 1 - with "horse shoe" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%
				
2351--4---	4325--1---	5431--2---	5413--2---	5413--2---

Prediction 2 - with "rosette" configuration:

Structural position	Highest scoring config. obtained using <u>3.0Å</u> cut-off matrix	Highest scoring config. obtained using <u>3.5Å</u> cut-off matrix	Highest scoring config. obtained using <u>4.0Å</u> cut-off matrix	Highest scoring config. obtained using <u>4.5Å</u> cut-off matrix
	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%
35-2-4-1--	34-2-5-1--	54-1-3-2--	45-1-3-2--	45-1-3-2--

Prediction 3 - with unspecified configuration:






Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%
2351--4---	43-152---	43-15--2--	452-13----	452-13----

Photosynthetic Reaction Center*Rhodospseudomonas viridis*Swiss-Prot code used to run TMRelate : P06009

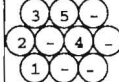

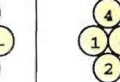
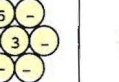
PDB Code used to evaluate the model : 1DRX

Number of transmembrane regions : 5




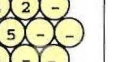

Prediction 1 - with "horse shoe" configuration:

Structural position	Highest scoring config.obtained using 3.0Å cut-off matrix	Highest scoring config.obtained using 3.5Å cut-off matrix	Highest scoring config.obtained using 4.0Å cut-off matrix	Highest scoring config.obtained using 4.5Å cut-off matrix
	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%
				
2351--4---	5423--1---	5124--3---	5413--2---	5413--2--

Prediction 2 - with "rosette" configuration:

Structural position	Highest scoring config. obtained using <u>3.0Å</u> cut-off matrix	Highest scoring config. obtained using <u>3.5Å</u> cut-off matrix	Highest scoring config. obtained using <u>4.0Å</u> cut-off matrix	Highest scoring config. obtained using <u>4.5Å</u> cut-off matrix
	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%
				
35-2-4-1--	45-2-3-1--	45-1-3-2--	45-1-3-2--	45-1-3-2--

Prediction 3 - with unspecified configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	6/8 coincident associations 75.00%	4/8 coincident associations 50.00%	6/8 coincident associations 75.00%	4/8 coincident associations 50.00%
				
2351--4---	54--12--3-	42-15--3--	54-312----	54-213----



Photosynthetic Reaction Center*Rhodobacter sphaeroides*Swiss-Prot code used to run TMRelate : P02954

PDB Code used to evaluate the model : 1RZH

Number of transmembrane regions : 5

Prediction 1 - with "horse shoe" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%
2351--4---	5413--2---	5413--2---	5413--2---	5413--2---

Prediction 2 - with "rosette" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%	6/8 coincident associations 75.00%
35-2-4-1--	45-1-3-2--	45-1-3-2--	45-1-3-2--	54-3-1-2--

Prediction 3 - with unspecified configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	6/8 coincident associations 75.00%	4/8 coincident associations 50.00%	4/8 coincident associations 50.00%	4/8 coincident associations 50.00%
2351--4---	43-15--2--	53-214----	54-213----	54-213----

P-type ATPaseSwiss-Prot code used to run TMRelate : P04191

PDB Code used to evaluate the model : 1T5S

Number of transmembrane regions : 10

Prediction 1 - with "horse shoe" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	20/38 coincident associations 52.63%	18/38 coincident associations 47.37%	18/38 coincident associations 47.37%	18/38 coincident associations 47.37%
29A1687435	A768592431	A213976458	7A65418392	7A65418392

Prediction 2 - with "rosette" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	18/38 coincident associations 47.37%	18/38 coincident associations 47.37%	20/38 coincident associations 52.63%	20/38 coincident associations 52.63%
129468A357	A768592431	A218793654	7A65418392	7A65418392

Prediction 3 - with unspecified configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	24/38 coincident associations 63.16%	20/38 coincident associations 52.63%	20/38 coincident associations 52.63%	20/38 coincident associations 52.63%
3417562A89	A768592431	A548793621	7A65418392	7A65418392




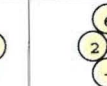
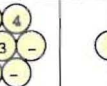


Respiratory proteins - Mitochondrial ADP/ADP carrierSwiss-Prot code used to run TMRelate : P02722



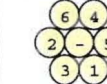
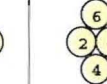
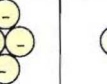
PDB Code used to evaluate the model : 1OKC

Number of transmembrane regions : 6

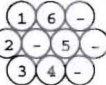
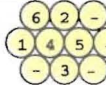
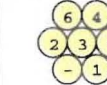
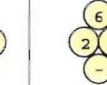
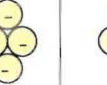
Prediction 1 - with "horse shoe" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	12/18 coincident associations 66.67	10/18 coincident associations 55.56%	12/18 coincident associations 66.67%	14/18 coincident associations 77.78%
				
123654----	526341----	634215----	654213----	534612----

Prediction 2 - with "rosette" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	6/12 coincident associations 50.00%	4/12 coincident associations 33.33%	4/12 coincident associations 33.33%	4/12 coincident associations 33.33%
				
16-2-5-34-	65-2-3-14-	64-2-5-31-	65-2-1-43-	65-2-1-43-

Prediction 3 - with unspecified configuration:

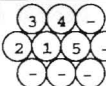
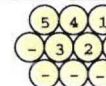
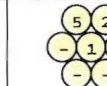
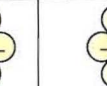
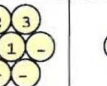
Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	6/12 coincident associations 50.00%	6/12 coincident associations 50.00%	6/12 coincident associations 50.00%	6/12 coincident associations 50.00%
				
16-2-5-34-	62-145-3-	64-235--1-	63-215--4-	62-134-5--

Respiratory proteins - Fumarate Reductase complex  
(Wolinella succinogenes)Swiss-Prot code used to run TMRelate : P17413

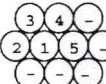

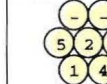
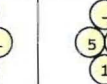
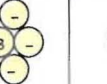
PDB Code used to evaluate the model : 1QLA

Number of transmembrane regions : 5


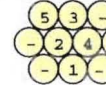
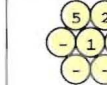
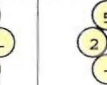
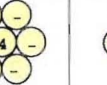
Prediction 1 - with "horse shoe" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	10/14 coincident associations 71.43%	10/14 coincident associations 71.43%	10/14 coincident associations 71.43%	10/14 coincident associations 71.43%
				
34-215----	541-32----	524-13----	523-41----	54-123----

Prediction 2 - with "rosette" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	10/14 coincident associations 71.43%	10/14 coincident associations 71.43%	10/14 coincident associations 71.43%	10/14 coincident associations 71.43%
				
34-215----	---532-14-	---523-14-	---523-14-	54-12--3--

Prediction 3 - with unspecified configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	8/14 coincident associations 57.14%	8/14 coincident associations 71.43%	8/14 coincident associations 57.14%	8/14 coincident associations 71.43%
				
34-215----	53--24--1-	523-14----	53-21--4--	523-14----



V-type ATPase

Swiss-Prot code used to run TMRelate : P43457  
 PDB Code used to evaluate the model : 2BL2  
 Number of transmembrane regions : 4

Prediction 1 - with "horse shoe" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	8/10 coincident associations 80.00%	8/12 coincident associations 80.00%	10/10 coincident associations 100.00%	10/10 coincident associations 100.00%
-24-13----	43-21-----	43-21-----	12--34----	12--34----

Prediction 2 - with "rosette" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	8/10 coincident associations 80.00%	8/10 coincident associations 80.00%	8/10 coincident associations 80.00%	8/10 coincident associations 80.00%
24-13-----	43--12-----	43-21-----	---12--34-	24-13-----

Prediction 3 - with unspecified configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	8/10 coincident associations 50.00%	8/10 coincident associations 50.00%	10/10 coincident associations 100.00%	10/10 coincident associations 100.00%
24-13-----	43--12-----	43-21-----	43--21-----	24-13-----

Formate dehydrogenase-N: Escherichia coli

Swiss-Prot code used to run TMRelate : P24185  
 PDB Code used to evaluate the model : 1KQF  
 Number of transmembrane regions : 4

Prediction 1 - with "horse shoe" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	8/10 coincident associations 80.00%	8/10 coincident associations 80.00%	8/10 coincident associations 80.00%	8/10 coincident associations 80.00%
14-23-----	-12--43---	-12--43---	-12--43---	-12--43---

Prediction 2 - with "rosette" configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	8/10 coincident associations 80.00%	8/10 coincident associations 80.00%	8/10 coincident associations 80.00%	8/10 coincident associations 80.00%
14-23-----	---12--43-	---12--43-	----12-34-	----12-34-

Prediction 3 - with unspecified configuration:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	8/10 coincident associations 80.00%	8/10 coincident associations 80.00%	8/10 coincident associations 80.00%	8/10 coincident associations 80.00%
14-23-----	-----13-42	----12--43	----12-34	----12--43

Photosystem I - *Thermosynechococcus elontatus*Swiss-Prot code used to run TMRelate\_12 : P25896

PDB Code used to evaluate the model : 1JBO

Number of transmembrane regions : 11

Prediction 1 - One position fixed:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	24/38 coincident associations 63.16%	24/38 coincident associations 63.16%	24/38 coincident associations 63.16%	24/38 coincident associations 63.16%
AB12965378-4	AB52943168-7	7845A9326B-1	9A138B2476-5	78936A421B-5

Prediction 2:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	26/38 coincident associations 68.42%	26/38 coincident associations 68.42%	28/38 coincident associations 73.68%	26/38 coincident associations 68.42%
AB12965378-4	-84569327BA1	78436921-BA5	79458A32-6B1	-9B16A327845

AmtB ammonia channel (mutant): *E. coli*Swiss-Prot code used to run TMRelate\_12 : P37905

PDB Code used to evaluate the model : 1U7G

Number of transmembrane regions : 11

Prediction 1:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	24/42 coincident associations 57.14%	32/42 coincident associations 76.19%	30/42 coincident associations 71.43%	26/42 coincident associations 61.90%
-9AB68547132	-1968247A35B	-24B358A1769	-6719834AB52	-6719852AB34

Prediction 2:

Structural position	Highest scoring config. obtained using 3.0Å cut-off matrix	Highest scoring config. obtained using 3.5Å cut-off matrix	Highest scoring config. obtained using 4.0Å cut-off matrix	Highest scoring config. obtained using 4.5Å cut-off matrix
	24/42 coincident associations 57.14%	32/42 coincident associations 76.19%	30/42 coincident associations 71.43%	26/42 coincident associations 61.90%
-9AB68547132	913A6824-7B5	1769358A-24B	AB529834-671	AB719834-652



Cytochrome c oxidase, ba3: *T. Thermophilus*

Swiss-Prot code used to run TMRelate\_12 : Q5SJ79  
 PDB Code used to evaluate the model : 1XME  
 Number of transmembrane regions : 13

Prediction 1:

Structural position	Highest scoring config.obtained using 3.0Å cut-off matrix	Highest scoring config.obtained using 3.5Å cut-off matrix	Highest scoring config.obtained using 4.0Å cut-off matrix	Highest scoring config.obtained using 4.5Å cut-off matrix
	28/44 coincident associations 63.64%	26/44 coincident associations 59.09%	24/44 coincident associations 54.55%	24/44 coincident associations 54.55%
75436A2189BC	C2871465B39A	28A7C965B134	A43251C8769B	A43B51C27698

Prediction 2:

Structural position	Highest scoring config.obtained using 3.0Å cut-off matrix	Highest scoring config.obtained using 3.5Å cut-off matrix	Highest scoring config.obtained using 4.0Å cut-off matrix	Highest scoring config.obtained using 4.5Å cut-off matrix
	28/46 coincident associations 60.87%	28/46 coincident associations 60.87%	24/46 coincident associations 52.17%	24/46 coincident associations 52.17%
D54376A289BC	D75A846B932C	65784DC23AB9	A2D4BC563879	ABD43C562879

*TMRelate K* end on view evaluation

Bacteriorhodopsin

Swiss-Prot code used to run TMRelate\_K : P02945  
 PDB Code used to evaluate the model : 1C3W  
 Number of transmembrane regions : 7

Prediction 1 - "horse shoe" configuration:

Structural position : 2341765---

Best predicted result : 7361245---

Result with 20/22 coincident associations 90.91%

Prediction 2 - "rosette" configuration:

Structural position : 17-236-45-

Best predicted result : 12-734-65-

Result with 24/24 coincident associations 100.00%

Prediction 3 - unspecified configuration:

Structural position : 17-236-45-

Best predicted result : 17-236-45-

Result with 24/24 coincident associations 100.00%

Rhodopsin

Swiss-Prot code used to run TMRelate\_K : P02699  
 PDB Code used to evaluate the model : 1U19  
 Number of transmembrane regions : 7

Prediction 1 - with "horse shoe" configuration:

Structural position : 7651234---

Best predicted result : 5214376---

Result with 16/22 coincident associations 72.73%

Prediction 2 - with "rosette" configuration:

Structural position : 76-135-24-

Best predicted result : 65-174-23-

Result with 16/24 coincident associations 66.67%

Prediction 3 - with unspecified configuration:

Structural position : 76-135-24-

Best predicted result : 65-174-23-

Result with 16/24 coincident associations 66.67%




Sensory Rhodopsin II (HR)

Swiss-Prot code used to run TMRelate\_K : P42196  
 PDB Code used to evaluate the model : 1H2S  
 Number of transmembrane regions : 7

Prediction 1 - with "horse shoe" configuration:

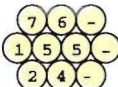
Structural position : 7651234--- 

Best predicted result : 7561432--- 

Result with 16/22 coincident associations 72.73%

Prediction 2 - with "rosette" configuration:


Structural position : 76-135-24- 

Best predicted result : 76-153-24- 

Result with 18/24 coincident associations 75.00%

Prediction 3 - with unspecified configuration:

Structural position : 76-135-24- 

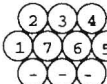
Best predicted result : -6--752143 

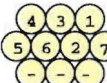
Result with 14/24 coincident associations 58.33%

Halorhodopsin

Swiss-Prot code used to run TMRelate\_K : P16102  
 PDB Code used to evaluate the model : 1E12  
 Number of transmembrane regions : 7

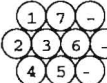
Prediction 1 - with "horse shoe" configuration:


Structural position : 7651234--- 

Best predicted result : 4315627--- 

Result with 18/22 coincident associations 81.82%

Prediction 2 - with "rosette" configuration:


Structural position : 76-135-24- 

Best predicted result : 67-541-32- 

Result: Result with 18/24 coincident associations 75.00%

Prediction 3 - with unspecified configuration:

Structural position : 76-135-24- 

Best predicted result : -765-14-32 

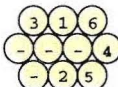
Result: Result with 12/24 coincident associations 50.00%

Aquaporin

Swiss-Prot code used to run TMRelate\_K : P06624  
 PDB Code used to evaluate the model : 1YMG  
 Number of transmembrane regions : 6

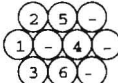
Prediction 1 - with "horse shoe" configuration:

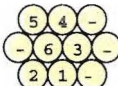
Structural position : 312---5-64 

Best predicted result : 316---4-25 

Result with 8/10 coincident associations 80.00%

Prediction 2 - with "rosette" configuration:

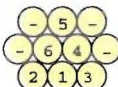
Structural position : 25-1-4-36- 

Best predicted result : 54--63-21- 

Result with 10/12 coincident associations 83.33%

Prediction 3 - with unspecified configuration:

Structural position : 312---4-65 

Best predicted result : -5--64-213 


Result with 8/10 coincident associations: 80.00%

Glycerol uptake facilitator protein  
(Aquaglyceroporin).

Swiss-Prot code used to run TMRelate\_K : P11244  
 PDB Code used to evaluate the model : 1FX8  
 Number of transmembrane regions : 8

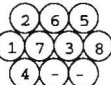
Prediction 1 - with "horse shoe" configuration:

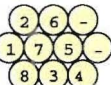
Structural position : 26517384-- 

Best predicted result : 65827341-- 

Result with 20/26 coincident associations 76.92%

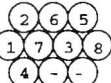
Prediction 2 - with "rosette" configuration:

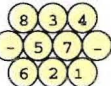
Structural position : 26513784-- 

Best predicted result : 26-175-834 

Result with 18/26 coincident associations 69.23%

Prediction 3 - with unspecified configuration:

Structural position : 26513784-- 

Best predicted result : 834-57-621 

Result with 20/26 coincident associations 76.92%

Photosynthetic Reaction Center  
*Thermochromatium tepidum*

Swiss-Prot code used to run TMRelate\_K : P51762  
PDB Code used to evaluate the model : 1EYS  
Number of transmembrane regions : 5

Prediction 1 - with "horse shoe" configuration:

Structural position : 2351--4---

Best predicted result : 145-23----

Result with 8/8 coincident associations 100.00%

Prediction 2 - with "rosette" configuration:

Structural position : 35-2-4-1--

Best predicted result : 1--24--35-

Result with 8/8 coincident associations 100.00%

Prediction 3 - with unspecified configuration:

Structural position : 2351--4---

Best predicted result : 1--24--35-

Result with 8/8 coincident associations: 100.00%

Photosynthetic Reaction Center  
*Rhodospseudomonas viridis*

Swiss-Prot code used to run TMRelate\_K : P06009  
PDB Code used to evaluate the model : 1DRX  
Number of transmembrane regions : 5

Prediction 1 - with "horse shoe" configuration:

Structural position : 2351--4---

Best predicted result : 145-23----

Result with 8/8 coincident associations 100.00%

Prediction 2 - with "rosette" configuration:

Structural position : 35-2-4-1--

Best predicted result : 1--24--35-

Result with 8/8 coincident associations 100.00%

Prediction 3 - with unspecified configuration:

Structural position : 2351--4---

Best predicted result : -35-24--1-


Result with 8/8 coincident associations: 100.00%

Photosynthetic Reaction Center  
*Rhodobacter sphaeroides*

Swiss-Prot code used to run TMRelate\_K : P02954  
PDB Code used to evaluate the model : 1RZH  
Number of transmembrane regions : 5

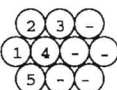
Prediction 1 - with "horse shoe" configuration:

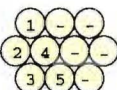
Structural position : 2351--4--- 

Best predicted result : 145-23---- 

Result with 8/8 coincident associations 100.00%

Prediction 2 - with "rosette" configuration:

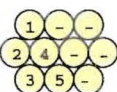
Structural position : 35-2-4-1-- 

Best predicted result : 1--24--35- 

Result with 8/8 coincident associations 100.00%

Prediction 3 - with unspecified configuration:

Structural position : 2351--4--- 

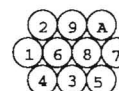
Best predicted result : 1--24--35- 

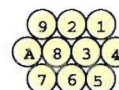
Result with 8/8 coincident associations: 100.00%

P-type ATPase

Swiss-Prot code used to run TMRelate\_K : P04191  
PDB Code used to evaluate the model : 1T5S  
Number of transmembrane regions : 10

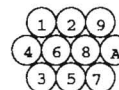
Prediction 1 - with "horse shoe" configuration:


Structural position : 29A1687435 

Best predicted result : 921A834765 

Result with 26/38 coincident associations 68.42%

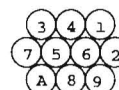
Prediction 2 - with "rosette" configuration:

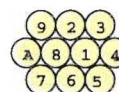
Structural position : 129468A357 

Best predicted result : 921A834765 

Result with 28/38 coincident associations 73.68%

Prediction 3 - with unspecified configuration:

Structural position : 3417562A89 

Best predicted result : 923A814765 

Result with 22/38 coincident associations: 57.89%



Respiratory proteins - Mitochondrial ADP/ADP carrier

Swiss-Prot code used to run TMRelate\_K : P02722  
 PDB Code used to evaluate the model : 1OKC  
 Number of transmembrane regions : 6

Prediction 1 - with "horse shoe" configuration:

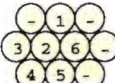
Structural position : 123654----  


Best predicted result : 623154----  


Result with 16/18 coincident associations 88.89%

Prediction 2 - with "rosette" configuration:


Structural position : 16-2-5-34-  


Best predicted result : -1-326-45-  


Result with 12/12 coincident associations 100.00%

Prediction 3 - with unspecified configuration:

Structural position : 16-2-5-34-  

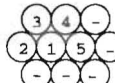

Best predicted result : --1-326-45  


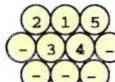
Result with 12/12 coincident associations: 100.00%

Respiratory proteins - Fumarate Reductase complex  
(*Wolinella succinogenes*)

Swiss-Prot code used to run TMRelate\_K : P17413  
 PDB Code used to evaluate the model : 1QLA  
 Number of transmembrane regions : 5

Prediction 1 - with "horse shoe" configuration:


Structural position : 34-215----  


Best predicted result : 215-34----  


Result with 14/14 coincident associations 100.00%

Prediction 2 - with "rosette" configuration:


Structural position : 34-215----  


Best predicted result : -2--13-54-  


Result with 14/14 coincident associations 100.00%

Prediction 3 - with unspecified configuration:

Structural position : 34-215----  

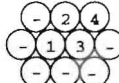
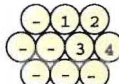

Best predicted result : --2--13-54  


Result with 14/14 coincident associations: 100.00%

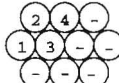
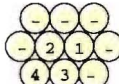
V-type ATPaseSwiss-Prot code used to run TMRelate\_K : P43457

PDB Code used to evaluate the model : 2BL2


Number of transmembrane regions : 4

Prediction 1 - with "horse shoe" configuration:Structural position : -24-13----  
Best predicted result : -12--34---  


Result with 10/10 coincident associations 100.00%

Prediction 2 - with "rosette" configuration:Structural position : 24-13-----  
Best predicted result : ----21-43-  


Result with 10/10 coincident associations 100.00%

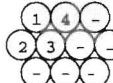

Prediction 3 - with unspecified configuration:Structural position : 24-13-----  
Best predicted result : ----34-12-  


Result with 10/10 coincident associations: 100.00%

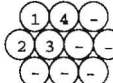

Formate dehydrogenase-N: *Escherichia coli*Swiss-Prot code used to run TMRelate\_K : P24185

PDB Code used to evaluate the model : 1KQF

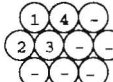
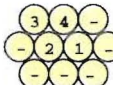
Number of transmembrane regions : 4

Prediction 1 - with "horse shoe" configuration:Structural position : 14-23-----  
Best predicted result : -12-43----  


Result with 10/10 coincident associations 100.00%

Prediction 2 - with "rosette" configuration:Structural position : 14-23-----  
Best predicted result : -2--13--4-  


Result with 10/10 coincident associations 100.00%

Prediction 3 - with unspecified configuration:Structural position : 14-23-----  
Best predicted result : 34-21-----  


Result with 10/10 coincident associations: 100.00%

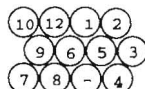


Photosystem I - *Thermosynechococcus elontatus*

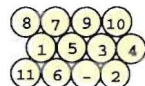
Swiss-Prot code used to run TMRelate\_K\_12 : P25896  
 PDB Code used to evaluate the model : 1JBO  
 Number of transmembrane regions : 11

Prediction 1 - One fixed position:

Structural position : AB12965378-4



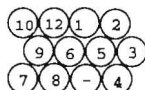
Best predicted result : 879A1534B6-2



Result with 22/38 coincident associations 57.89%

Prediction 2:

Structural position : AB12965378-4



Best predicted result : B698153A-742



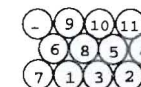
Result with 22/38 coincident associations 57.89%

AmtB ammonia channel (mutant): *E. coli*

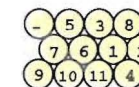
Swiss-Prot code used to run TMRelate\_K\_12 : P37905  
 PDB Code used to evaluate the model : 1U7G  
 Number of transmembrane regions : 11

Prediction 1 - One fixed position:

Structural position : -9AB68547132



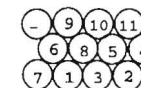
Best predicted result : -53876129AB4



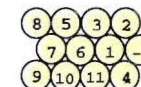
Result with 20/42 coincident associations 47.62%

Prediction 2:

Structural position : -9AB68547132



Best predicted result : 8532761-9AB4



Result with 18/42 coincident associations 42.86%

Cytochrome c oxidase, ba3: *T. Thermophilus*Swiss-Prot code used to run TMRelate K 12 : Q5SJ79

PDB Code used to evaluate the model : 1XME

Number of transmembrane regions : 13

Prediction 1 - TM1 to TM12:

Structural position : 75436A2189BC



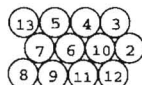
Best predicted result : CB3162754A98



Result with 20/46 coincident associations 43.48%

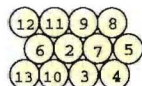
Prediction 2 - TM2 to TM13:

Structural position : D54376A289BC



C432659178AB

Best predicted result : CB986275DA34



Result with 26/46 coincident associations 56.52%

**Appendix IV**

3D evaluation

## Comparison of actual and predicted structures

**Evaluation using *TMEvaluation\_3D* program at different distance ranges**

### 1C3W – Bacteriorhodopsin

Files used to create the distance table	5Å				8Å			
	All		CA		All		CA	
	#	%	#	%	#	%	#	%
1c3w_CA_ss.pdb (native structure)	402		4		1548		308	
Model_1C3W_12-734-65- _ss.pdb	764	9.95%	52	0.00%	2626	45.61%	772	19.48%
Model_1C3W_7631245--- _ss.pdb	624	14.43%	66	0.00%	2452	45.99%	708	15.58%

The “#” column represents the number of associations between amino acids in different  $\alpha$ -helices of the TM regions.  
The “%” column contains the percentage with coincident associations between the native structure and the predicted one.

### 1U19 – Rhodopsin

Files used to create the distance table	5Å				8Å			
	All		CA		All		CA	
	#	%	#	%	#	%	#	%
1u19_CA_ss.pdb (native sctructure)	428		14		1766		326	
Model_1U19_65-174-23- _ss.pdb	830	11.68%	76	0.00%	3178	36.13%	938	9.20%
Model_1U19_5214376--- _ss.pdb	834	13.08%	84	0.00%	3072	40.20%	920	16.56%

**1H2S – Sensory Rhodopsin II (HR)**

Files used to create the distance table	5Å				8Å			
	All		CA		All		CA	
	#	%	#	%	#	%	#	%
1h2s_CA_ss.pdb (native structure)	518		16		2048		398	
Model_1H2S_76-153-24- _ss.pdb	894	15.44%	70	0.00%	3418	36.82	1096	23.62%
Model_1H2S_7561432--- _ss.pdb	972	13.13%	74	0.00%	3320	33.50%	994	13.57%

**1E12 – Halorhodopsin**

Files used to create the distance table	5Å				8Å			
	All		CA		All		CA	
	#	%	#	%	#	%	#	%
1e12_CA_ss.pdb (native structure)	442		10		1892		372	
Model_1E12_67-541-32- _ss_scwrl.pdb	982	15.38%	78	0.00%	3460	40.17%	990	16.67%
Model_1E12_4315627--- _ss_sccomp.pdb	840	9.50%	80	0.00%	3142	37.84%	998	15.05%

**1YMG – Aquaporin**

Files used to create the distance table	5Å				8Å			
	All		CA		All		CA	
	#	%	#	%	#	%	#	%
lymg_CA_ss.pdb (native structure)	278		40		1146		344	
Model_1YMG_54--63-21- _ss.pdb	526	8.63%	56	0.00%	2124	30.02%	690	9.30%
Model_1YMG_316---4- 25_ss_sccomp.pdb	332	2.16%	44	0.00%	1336	27.05%	408	5.23%

**1FX8 – Glycerol uptake facilitator protein (Aquaglyceroporin)**

Files used to create the distance table	5Å				8Å			
	All		CA		All		CA	
	#	%	#	%	#	%	#	%
1fx8_CA_ss.pdb (native structure)	460		50		1852		498	
Model_1FX8_26-175- 834_ss.pdb	682	2.61%	62	0.00%	2596	9.40%	804	4.42%
Model_1FX8_65827341-- _ss.pdb	738	2.61%	74	0.00%	2528	12.42%	710	3.61%



**1EYS – Photosynthetic Reaction Center *Thermochromatium tepidum***

Files used to create the distance table	5Å				8Å			
	All		CA		All		CA	
	#	%	#	%	#	%	#	%
1eys_CA_ss.pdb (native structure)	210		14		896		222	
Model_1EYS_1--24--35--_ss.pdb	662	0.00%	50	0.00%	2230	0.45%	638	0.00%
Model_1EYS_145-23----_ss.pdb	698	0.00%	54	0.00%	2368	0.67%	696	0.00%

**1DXR – Photosynthetic Reaction Center *Rhodopseudomonas viridis***

Files used to create the distance table	5Å				8Å			
	All		CA		All		CA	
	#	%	#	%	#	%	#	%
1dxr_CA_ss.pdb (native structure)	112		4		436		106	
Model_1DXR_1--24--35--_ss.pdb	546	26.79%	42	0.00%	1962	45.41%	592	9.43%
Model_1DXR_145-23----_ss.pdb	640	16.07%	46	0.00%	2036	42.20%	622	7.55%

**1RHZ – Photosynthetic Reaction Center *Rhodobacter apharoides***

Files used to create the distance table	5Å				8Å			
	All		CA		All		CA	
	#	%	#	%	#	%	#	%
1rzh_CA_ss.pdb (native structure)	208		24		842		216	
Model_1RZH_1--24--35- _ss.pdb	534	4.81%	48	16.67%	2038	7.36%	644	2.38%
Model_1RZH_145-23---- _ss.pdb	612	5.77%	64	16.67%	2188	4.75%	660	1.66%

**1T5S – P-type ATPase**

Files used to create the distance table	5Å				8Å			
	All		CA		All		CA	
	#	%	#	%	#	%	#	%
1t5s_CA_ss.pdb (native structure)	614		24		2502		510	
Model_1T5S_921A834765.pdb	1188	10.10%	116	0.00	4444	27.34%	1340	5.10%

**1OKC – Respiratory proteins – Mitochondrial ADP/ADP carrier**

Files used to create the distance table	5Å				8Å			
	All		CA		All		CA	
	#	%	#	%	#	%	#	%
1okc_CA_ss.pdb (native structure)	174		2		698		162	
Model_1OKC_623154---- _ss.pdb	404	6.90%	36	0.00%	1516	34.96%	444	6.17%
Model_1OKC_-1-326-45- _ss.pdb	332	6.90%	22	0.00%	1340	35.53%	378	7.41%

**1QLA – Respiratory Proteins – Fumarate Reductase complex (*Wolinella succinogenes*)**

Files used to create the distance table	5Å				8Å			
	All		CA		All		CA	
	#	%	#	%	#	%	#	%
1qla_CA_ss.pdb (native structure)	234		6		924		170	
Model_1QLA_215-34---- _ss.pdb	574	16.24%	54	0.00%	1862	40.48%	600	14.12%
Model_1QLA_-2--13-54- _ss.pdb	486	12.82%	40	0.00%	1824	39.39%	544	9.41%

### 2BL2 – V-type ATPase

Files used to create the distance table	5Å				8Å			
	All		CA		All		CA	
	#	%	#	%	#	%	#	%
2bl2_CA_ss.pdb (native structure)	224		40		966		302	
Model_2BL2_-12--34--- _ss.pdb	270	22.32%	24	0.00%	1140	56.94%	366	31.79%
Model_2BL2_----21-43- _ss.pdb	270	12.50%	14	0.00%	1160	47.20%	364	13.91%

### 1KQF – Formate dehydrogenase-N *Escherichia coli*

Files used to create the distance table	5Å				8Å			
	All		CA		All		CA	
	#	%	#	%	#	%	#	%
1kqf_CA_ss.pdb (native structure)	164		2		712		122	
Model_1KQF_-2--13--4- _ss.pdb	404	0.55%	24	0.00%	1280	36.52%	396	8.20%
Model_1KQF_-12-43---- _ss.pdb	392	0.00%	36	0.00%	1366	46.35%	436	21.31%

**1JBO – Photosystem I *Thermosynechococcus elontatus***

Files used to create the distance table	5Å				8Å			
	All		CA		All		CA	
	#	%	#	%	#	%	#	%
1jbo_CA_ss.pdb (native structure)	540		62		2126		592	
Model_1JBO_879A1534B6- 2_ss.pdb	1392	4.81%	112	0.00%	5236	19.29%	1486	5.41%
Model_1JBO_B698153A- 742_ss.pdb	1758	5.93%	124	0.00%	5820	22.01%	1692	5.07%

**1U7G – AmtB ammonia channel (mutant) *Escherichia coli***

Files used to create the distance table	5Å				8Å			
	All		CA		All		CA	
	#	%	#	%	#	%	#	%
1u7g_CA_ss.pdb (native structure)	808		90		3634		934	
Model_1U7G_8532761- 9AB4_ss.pdb	1616	0.00%	142	0.00%	5598	0.06%	1732	0.00%
Model_1U7G_- 53876129AB4_ss.pdb	1666	0.00%	156	0.00%	5842	0.00%	1784	0.00%

**1XME – Cytochrome c oxidase, ba3 *T. Thermophilus***

Files used to create the distance table	5Å				8Å			
	All		CA		All		CA	
	#	%	#	%	#	%	#	%
1xme_CA_1_12ss.pdb	742		38		3070		628	
Model_1XME_1_12CB3162754A98.pdb	1578	1.35%	122	0.00%	5946	15.37%	1650	2.23%
1xme_CA_2_13ss.pdb	674				2840		568	
Model_1XME_2_13BA875164C923.pdb	1514	5.34%	136	0.00%	5740	20.99%	1728	7.04%



## **Bibliography**

- Abramson, J., Riistama, S., Larsson, G., Jasaitis, A., Svensson-Ek, M., Laakkonen, L., Puustinen, A., Iwata, S. and Wikstrom, M. (2000). The structure of the ubiquinol oxidase from *Escherichia coli* and its ubiquinone binding site. *Nat. Struct. Biol.*, **7**:910-917.
- Abramson, J., Smirnova, I., Kasho, V., Verner, G., Kaback, H.R. and Iwata, S. (2003). Structure and mechanism of the lactose permease of *Escherichia coli*. *Science*, **301**:603-604.
- Adamian, L. and Liang, J. (2001). Helix-helix packing and interfacial pairwise interactions of residues in membrane proteins. *J. Mol. Biol.*, **311**:891-907.
- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2002). *Molecular Biology of the Cell* (4<sup>th</sup> edition), Garland Science, New York, NY.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, **215**:403-410.
- Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**:3389-3402.
- Attwood, T.K. and Parry-Smith, D.J. (1999). *Introduction to Bioinformatics*, Prentice Hall, Harlow, Essex.
- Bairoch, A. and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**:45-48.
- Barlow, D.J. and Thornton, J.M. (1998). Helix geometry in proteins. *J. Mol. Biol.*, **201**:601-619.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Res.*, **28**:235-242.

- Bonneau, R., Ruczinski, I., Tsai, J. and Baker, D. (2002). Contact order and *ab initio* protein structure prediction. *Protein Sci.*, **11**:1937-1944.
- Bonneau, R. and Baker, D. (2001). *Ab Initio* protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.*, **30**:173-189.
- Bowie, J.U. (2005). Solving the membrane protein folding problem. *Nature*, **438**:581-589.
- Boyd, D., Schierle, C. and Beckwith, J. (1998). How many membrane proteins are there? *Protein Sci.*, **7**:201-205.
- Branden, C. and Tooze, J. (1999). *Introduction to Protein Structure* (2<sup>nd</sup> edition), Garland Publishing, New York, NY.
- Chen, C.P. and Rost, B. (2002). State-of-the-art in membrane protein prediction. *Applied Bioinformatics*, **1**:21-35.
- Chothia, C., Levitt, M. and Richardson, D. (1981). Helix to helix packing in protein. *J. Mol. Biol.*, **145**:215-250.
- Chou, P.Y. and Fasman, G.D. (1974). Prediction of protein conformation. *Biochemistry*, **13**:222-245.
- Christiansen and Torkington. (1997). How do I permute n elements of a list?  
Available:  
[http://www.rocketaware.com/perl/perlfaq4/How\\_do\\_I\\_permute\\_N\\_elements\\_of\\_a.htm#How\\_do\\_I\\_permute\\_N\\_elements\\_of\\_a](http://www.rocketaware.com/perl/perlfaq4/How_do_I_permute_N_elements_of_a.htm#How_do_I_permute_N_elements_of_a), last accessed 02/05/2006.
- Claros, M.G. and von Heijne, G. (1994) TopPred II: An improved software for membrane protein structure predictions. *CABIOS*, **10**:685-686.
- Cronet, P., Sander, C. and Vriend, G. (1993). Modelling the transmembrane seven helix bundle. *Protein Eng.*, **6**:59-64.
- Cserző, M., Wallin, E., Simon, I., von Heijne, G. and Elofsson, A. (1997). Prediction of transmembrane  $\alpha$ -helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Eng.*, **10**:673-676.
- Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M. and Barton, G.J. (1998). Jpred: A consensus secondary structure prediction server. *Bioinformatics*, **14**:892-893.
- Dahl, S.G. and Sylte, I. (2005). Molecular modelling of drug targets: the past, the present and the future. *Basic Clin. Pharmacol. Toxicol.*, **96**:151-155.
- Danielli, J.F. and Davson, H. (1935). A contribution to the theory of permeability of thin films. *J. Cell Comp. Physiol.*, **5**:495-508.

- de Groot, B.L., Engel, A. and Grubmuller, H. (2003). The structure of the aquaporin-1 water channel: a comparison between cryo-electron microscopy and X-ray crystallography. *J. Mol. Biol.*, **325**:485-493.
- DeLano, W.L. (2002). The PyMOL Molecular Graphics. System DeLano Scientific, San Carlos, CA, USA. <http://pymol.sourceforge.net/>. Last accessed: 02/05/2006.
- Dewji, N.N. and Singer, S.J. (1997). The seven transmembrane spanning topography of the Alzheimer disease-related presenilin proteins in the plasma membranes of cultured cells. *Proc. Natl. Acad. Sci. USA*, **94**:14024-14030.
- Donnelly, D., Overington, J.P., Ruffle, S.V., Nugent, J.H.A. and Blundell, T.L. (1993). Modelling  $\alpha$ -helical transmembrane domains: the calculation and use of substitution tables for lipid-facing residues. *Protein Sci.*, **2**:55-70.
- Doyle, D.A., Morais Cabral, J., Pfuetzner, R.A., Kuo, A., Gulbis, J.M., Cohen, S.L., Chait, B.T. and MacKinnon, R. (1998). The structure of the potassium channel: molecular basis of  $K^+$  conduction and selectivity. *Science*, **280**:69-77.
- Eilers, M., Chekar, S.C., Shieh, T., Smith, S.O. and Fleming, P.J. (2000). Internal packing of helical membrane proteins. *Proc. Natl. Acad. Sci. USA*, **11**:5796-5801.
- Eisenberg, D., Weiss, R.M. and Terwilliger, T.C. (1982). The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature*, **299**:371-374.
- Eisenberg, D., Weiss, R.M. and Terwilliger, T.C. (1984). The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA*, **81**:140-144.
- Enami, N., Okumura, H. and Kouyama, T. (2003). X-ray crystallographic studies of archaerhodopsin. *J. Protosci.*, **9**:320-322.
- Ferreira, K.N., Iverson, T.M., Maghlaoui, K., Barber, J. and Iwata, S. (2004). Architecture of the photosynthetic oxygen-evolving center. *Science*, **303**:1831-1838.
- Fischer, D. and Eisenberg, D. (1996). Protein fold recognition using sequence-derived predictions. *Protein Sci.*, **5**:947-955.
- Frishman, D. and Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins*, **23**:566-579.
- Frishman, D. and Argos, P. (1996). Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng.*, **9**:133-142.

- Fu, D., Libson, A., Miercke, L.J., Weitzman, C., Nollert, P., Krucinski, J. and Stroud, R.M. (2000). Structure of a glycerol-conducting channel and the basis for its selectivity. *Science*, **290**:481-486.
- Garavito, R.M. (1998). Membrane protein structures: the known world expands. *Curr. Opin. Biotechnol.*, **9**:344-349.
- Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D. and Bairoch, A. (2005). Protein identification and analysis tools on the ExPASy server, in John M. Walker (ed): *The Proteomics Protocols Handbook*, Humana Press, Totowa, NJ. pp. 571-607.
- Gibas, C. and Jambeck, P. (2001). *Developing Bioinformatics Computer Skills*, O'Reilly & Associates, Inc, Sebastopol, CA.
- Gordeliy, V.I., Labahn, J., Moukhametzianov, R., Efremov, R., Granzin, J., Schlesinger, R., Buldt, G., Savopol, T., Scheidig, A.J., Klare, J.P. and Engelhard, M. (2002). Molecular basis of transmembrane signalling by sensory rhodopsin II-transducer complex. *Nature*, **419**:484-487.
- Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science*, **185**:862-864.
- Guex, N. and Peitsch, M.C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modelling. *Electrophoresis*, **18**:2714-2723.
- Guex, N., Diemand, A. and Peitsch, M.C. (1999). Protein modelling for all. *TiBS*, **24**:364-367.
- Harries, W.E., Akhavan, D., Miercke, L.J., Khademi, S. and Stroud, R.M. (2004). The channel architecture of aquaporin 0 at a 2.2 Å resolution. *Proc. Natl. Acad. Sci., USA*, **101**:14045-14050.
- Henderson, R. and Unwin, P.N.T. (1975). Three-dimensional model of purple membrane obtained by electron microscopy. *Nature*, **157**:28-32.
- Hildebrand, J.G.S. and Shepherd, G.M. (1997). Mechanisms of olfactory discrimination: converging evidence for common principles across phyla. *Annu. Rev. Neurosci.*, **20**:595-631.
- Hirokawa, T.S., Boon-Chieng, S.S. and Mitaku, S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**:378-379.
- Hirs, C.H.W., Moore, S. and Stein, W.H. (1960). The sequence of the amino acid residues in performic acid-oxidized ribonuclease. *J. Biol. Chem.*, **235**:633-647.

- Hofmann, K. and Stoffel, W. (1993). TMBASE – a database of membrane spanning protein structure and topology. *J. Magn. Reson.*, **144**:150–155.
- Holm, L. and Sander, C. (1991). Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. *J. Mol. Biol.*, **218**:183–194.
- Holm, L. and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**:123–138.
- Huang, Y., Lemieux, M.J., Song, J., Auer, M. and Wang, D.N. (2003). Structure and mechanism of the glycerol-3-phosphate transporter from *Escherichia coli*. *Science*, **310**:616–620.
- Hunsicker-Wang, L.M., Pacoma, R.L., Chen, Y., Fee, J.A. and Stout, C.D. (2005). A novel cryoprotection scheme for enhancing the diffraction of crystals of recombinant cytochrome ba3 oxidase from *Thermus thermophilus*. *Acta Crystallogr. D. Biol. Crystallogr.*, **61**:340–343.
- Jaysinghe, S., Hristova, K. and White, S.H. (2001). Mptopo: A database of membrane protein topology. *Protein Sci.*, **10**:455–458. <http://blanco.biomol.uci.edu/mpex>. Last accessed: 02/05/2006.
- Jones, D.T., Taylor, W.R. and Thornton, J.M. (1994). A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry*, **33**:3038–3049.
- Jones, D.T. (1998). Do transmembrane protein superfolds exist? *FEBS Lett.*, **423**:281–285.
- Jones, D.T. (1999a). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**:195–202.
- Jones, D.T. (1999b) GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **287**: 797–815.
- Jormakka, M., Tornroth, S., Byrne, B. and Iwata, S. (2002). Molecular basis of proton motive force generation: structure of formate dehydrogenase-N. *Science*, **295**:1863–1868.
- Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **12**:2577–2637.
- Karchin, R., Karplus, K. and Haussler, D. (2002). Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, **18**:147–159.

- Kelley, L.A., MacCallum, R.M. and Sternberg, M.J.E. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**:499-520.
- Khademi, S., O'Connell, J. 3<sup>rd</sup>, Remis, J., Robles-Colmenares, Y., Miercke, L.J. and Stroud, R.M. (2004). Mechanism of ammonia transport by Amt/MEP/Rh: structure of AmtB at 1.35 Å. *Science*, **305**:1587-1594.
- Kihara, D. and Kenehisa, M. (2000). Tandem clusters of membrane proteins in complete genome sequences. *Genome Res.*, **10**:731-743.
- Kolbe, M., Besir, H., Essen, L.O. and Oesterhelt, D. (2000). Structure of the light-driven chloride pump halorhodopsin at 1.8 Å resolution. *Science*, **288**:1390-1396.
- Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**:567-580.
- Kyte, J. and Doolittle, R.F. (1982). A simple method for displaying the hydrophathic character of a protein. *J. Mol. Biol.*, **157**:105-132.
- Lancaster, C.R., Kroger, A., Auer, M. and Michel, H. (1999). Structure of fumarate reductase from *Wolinella succinogenes* at 2.2 Å resolution. *Nature*, **402**:377-385.
- Lancaster, C.R., Bibikova, M.V., Sabatino, P., Oesterhelt, D. and Michel, H. (2000). Structural basis of the drastically increased initial electron transfer rate in the reaction center from a *Rhodospseudomonas viridis* mutant described at 2.00-Å resolution. *J. Biol. Chem.*, **275**:39364-39368.
- Laskowski, R.A., Hutchinson, E.G., Michie, A.D., Wallace, A.C., Jones, M.L. and Thornton, J.M. (1997). PDBsum: A Web-based database of summaries and analyses of all PDB structures. *TiBS*, **22**:488-490.
- Lesk, A.M. (2001). *Introduction to Protein Architecture*, Oxford University Press, New York, NY.
- Luecke, H., Schobert, B., Richter, H.T., Cartailler, J.P. and Lanyi, J.K. (1999). Structure of bacteriorhodopsin at 1.55 Å resolution. *J. Mol. Biol.*, **291**:899-911.
- Maggio, E.T. and Ramnarayan, K. (2001). Recent developments in computational proteomics. *Trends in Biotechnology*, **19**:266-272.
- Martz, E. (2002). Protein Explorer: easy yet powerful macromolecular visualization. *TiBS*, **27**:107-109.
- McGuffin, L.J., Bryson, K. and Jones, D.T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, **16**:404-405.



- McLuskey, K., Prince, S.M., Cogdell, R.J. and Isaacs, N.W. (2001). The crystallographic structure of the B800-820 LH3 light-harvesting complex from the purple bacteria *Rhodopseudomonas acidophila* strain 7050. *Biochemistry*, **40**:8783-8789.
- Martelli, P.L., Fariselli, P., Krogh, A. and Casadio, R. (2002). A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. *Bioinformatics*, **18**:S46-S53.
- Mignon, P., Steyaert, J., Loris, R., Geerlings, P. and Loverix, S. (2002). A nucleophile activation dyad in ribonucleases. A combined X-ray crystallographic/ab initio quantum chemical study. *J. Biol. Chem.*, **277**:36770-36774.
- Montelione, G.T. and Anderson, S. (1999). Structural genomics: Keystone for a human proteome project. *Nat. Struct. Biol.*, **6**:11-12.
- Möller, S., Croning, M.D.R. and Apweiler, R. (2001). Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics*, **17**:646-653.
- Moult, J., Fidelis, K., Zemla, A. and Hubbard, T. (2001). Critical assessment of methods of proteins structure prediction (CASP) Round IV. *Proteins*, **45**(suppl 5):2-7.
- Murata, K., Mitsuoka, K., Hirai, T., Walz, T., Agre, P., Heymann, J.B., Engel, A. and Fujiyoshi, Y. (2000). Structural determinants of water permeation through aquaporin-1. *Nature*, **407**:599-605.
- Murata, T., Yamato, I., Kakinuma, Y., Leslie, A.G. and Walker, J.E. (2005). Structure of the rotor of the V-Type Na<sup>+</sup>-ATPase from *Enterococcus hirae*. *Science*, **308**:642-644.
- Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**:536-540.
- Nakai, K. and Horton, P. (1999). PSORT: a program for detecting the sorting signals of proteins and predicting their sub cellular localization. *TiBS*, **24**:34-35.
- Nayal, M., Hitz, B.C. and Honig, B. (1999). GRASS: a server for the graphical representation and analysis of structures. *Protein Sci.*, **8**:676-679.
- Nelson, D.L. and Cox, M. (2004). *Lehninger Principles of Biochemistry* (4<sup>th</sup> edition), W. H. Freeman Company, New York, NY.
- Neshich, G., Togawa, R.C., Mancini, A.L., Kuser, P.R., Yamagishi, M.E., Pappas, G. Jr., Torres, W.V., Fonseca, e Campos, T., Ferreira, L.L., Luna, F.M., Oliveira, A.G., Miura, R.T., Inoue, M.K., Horita, L.G., de Souza, D.F., Dominiquini, F.,

- Alvaro, A., Lima, C.S., Ogawa, F.O., Gomes, G.B., Palandrani, J.F., dos Santos, G.F., de Freitas, E.M., Mattiuz, A.R., Costa, I.C., de Almeida, C.L., Souza, S., Baudet, C. and Higa, R.H. (2003). STING Millennium: A web-based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. *Nucleic Acids Res.*, **31**:3386-3392.
- Nield, J., Rizkallah, P.J., Barber, J. and Chayen, N.E. (2003). The 1.45 Å three-dimensional structure of C-phyocyanin from the thermophilic cyanobacterium *Synechococcus elongatus*. *J. Struct. Biol.*, **141**:149-155.
- Nogi, T., Fathir, I., Kobayashi, M., Nozawa, T., and Miki, K. (2000). Crystal structures of photosynthetic reaction center and high-potential iron-sulfur protein from *Thermochromatium tepidum*: thermostability and electron transfer. *Proc. Natl. Acad. Sci. USA*, **97**:13561-13556.
- Okada, T., Sugihara, M., Bondar, A.N., Elstner, M., Entel, P., and Buss, V. (2004). The retinal conformation and its environment in rhodopsin in light of a new 2.2 Å crystal structure. *J. Mol. Biol.*, **342**:571-583.
- Palczewski, K., Kumasaka, T., Hori, T., Behnke, C., Motoshima, H., Fox, B., Trong, I.L., Teller, D., Okada, T., Stenkamp, R., Yamamoto, M. and Miyano, M. (2000). Crystal structure of rhodopsin: a G protein-coupled receptor. *Science*, **289**:739-745.
- Papiz, M.Z., Prince, S.M., Howard, T., Cogdell, R.J. and Isaacs, N.W. (2003). The structure and thermal motion of the B800-850 LH2 complex from *Rps. acidophila* at 2.0 Å resolution and 100K: new structural features and functionally relevant motions. *J. Mol. Biol.*, **326**:1523-1538.
- Paulsen, I.T., Nguyen, L., Sliwinski, M.K., Rabus, R. and Saier, M.H.Jr. (2000). Microbial genome analyses: comparative transport capabilities in eighteen prokaryotes. *J. Mol. Biol.*, **301**:75-100.
- Pebay-Peyroula, E., Dahout-Gonzalez, C., Kahn, R., Trezeguet, V., Lauquin, G.J. and Brandolin, G. (2003). Structure of mitochondrial ADP/ATP carrier in complex with carboxyatractyloside. *Nature*, **426**:39-44.
- Persson, B. and Argos, P. (1994). Prediction of transmembrane segments in proteins utilising multiple sequence alignment. *J. Mol. Biol.*, **237**:182-192.
- Pearson, W.R. and Lipman, D.J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, **85**:2444-2448.
- Petersen, T.N., Lundegaard, C., Nielsen, M., Bohr, H., Bohr, J., Brunak, S., Gippert, G.P. and Lund, O. (2000). Prediction of protein secondary structure at 80% accuracy. *Proteins*, **41**:17-20.

- Pierce, K.L., Premont, R.T. and Lefkowitz, R.J. (2002). Seven-transmembrane receptors. *Nat. Rev. Mol. Cell Biol.*, **9**:639-650.
- Pilpel, Y., Ben-tal, N. and Lancet, D. (1999). KPROT: A knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. *J. Mol. Biol.*, **294**:921-935.
- Popot, J-L. and Engelman, D.M. (2000). Helical membrane protein folding, stability and evolution. *Annu. Rev. Biochem.*, **69**:881-922.
- Preissner, R., Goede, A. and Frommel, C. (1999). Spare parts for helix-helix interaction. *Protein Eng.*, **12**:825-832.
- Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V. (1963) Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.*, **7**:95-99.
- Remm, M. and Sonnhammer, E.L.L. (2000). Classification of transmembrane protein families in the *Caenorhabditis elegans* genome and identification of human orthologs. *Genome Res.*, **10**:1679-1689.
- Rost, B. and Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **2**:584-599.
- Rost, B., Sander, C. and Schneider, R. (1994). PHD – an automatic server for protein secondary structure prediction. *CABIOS*, **10**:53-60.
- Rost, B., Fariselli, P. and Casadio, R. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.*, **7**:1704-1718.
- Ryle, A.P., Sanger, F., Smith, L.F. and Kitai, R. (1955). The disulphide bonds of insulin. *Biochemical Journal*, **60**:541-556.
- Russ, W.P. and Engelman, D.M. (2000). The GxxxG Motif: A framework for transmembrane helix-helix association. *J. Mol. Biol.*, **296**:911-919.
- Sánchez, R., Badretdinov, A.Y., Feyfant, E. and Sali, A. (1997). Homology protein structure modelling. *Transactions Amer. Cryst. Assoc.*, **32**:81-91.
- Sali, A. and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**:779-815.
- Sarakinou, K.S., Antoniow, J.F., Togawa, R.C. and Mullins, J.G.L. (2001). TMAAlpha: A software tool for analysis of transmembrane region alpha helices. *Biochem. Soc. Trans.*, **29**:36.
- Sarramegna, V., Muller, I., Milon, A. and Talmont, F. (2006). Recombinant G protein-coupled receptors from expression to renaturation: a challenge towards structure. *Cell. Mol. Life Sci.*, **63**:1149-1164.

- Sayle, R.A. and Milner-White, E.J. (1995) RasMol: Biomolecular graphics for all, *TiBS*, **20**:374.
- Simons, K.T., Bonneau, R., Ruczinski, I. and Baker, D. (1999). *Ab initio* protein structure prediction of CASP III targets using ROSETTA. *Proteins*, **37**:171-176.
- Singer, S.J. and Nicolson, G.L. (1972). The fluid mosaic model of the structure of cell membranes. *Science*, **175**:720-731.
- Schulz, G.E. (2000).  $\beta$ -barrel membrane proteins. *Current Opinion in Structural Biology*, **10**:443-447.
- Senes, A., Gerstein, M. and Engelman, D.M. (2000). Statistical analysis of amino acid patterns in transmembrane helices: The GxxxG motif occurs frequently and associations with  $\beta$ -branched residues at neighbouring positions. *J. Mol. Biol.*, **296**:921-936.
- Sonnhammer, E.L.L., von Heijne, G. and Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **6**:175-82.
- Sorensen, T.L., Moller, J.V. and Nissen, P. (2004). Phosphoryl transfer and calcium ion occlusion in the calcium pump. *Science*, **304**:1672-1675.
- Stadel, J.M., Wilson, S. and Bergsma, D.J. (1997). Orphan G protein-coupled receptors: a neglected opportunity for pioneer drug discovery. *Trends Pharmacol. Sci.*, **22**:162-165.
- Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Redaschi, N., Stoehr, P., Tuli, M., Tzouvara, K. and Vaughan, R. (2002). The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **30**:21-26.
- Stroebel, D., Choquet, Y., Popot, J.L. and Picot, D. (2003). An atypical haem in the cytochrome b(6)f complex. *Nature*, **426**:413-418.
- Stultz, C.M., White, J.V. and Smith, T.F. (1993). Structural analysis based on state-space modelling. *Protein Sci.*, **2**:305-314.
- Sui, H., Han, B.G., Lee, J.K., Walian, P. and Jap, B.K. (2001). Structural basis of water-specific transport through the AQP1 water channel. *Nature*, **414**:872-878.
- Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**:4673-4680.

- Togawa, R.C., Antoniwi, J.F. and Mullins, J.G.L. (2001). TMCompare: transmembrane region sequence and structure. *Bioinformatics*, **17**:1238-1239.
- Toyoshima, C. and Nomura, H. (2002). Structural changes in the calcium pump accompanying the dissociation of calcium. *Tanpakushitsu Kakusan Koso*, **46**:1374-1380.
- Treutlein, H.R., Lemmon, M.A., Engelman, D.M. and Brünger, A.T. (1992). The glycoporphin A transmembrane domain dimer: Sequence-specific propensity for a right-handed supercoil of helices. *Biochemistry*, **31**:12726-12733.
- Tsukihara, T., Aoyama, H., Yamashita, E., Tomizaki, T., Yamaguchi, H., Shinzawa-Itoh, K., Nakashima, R., Yaono, R. and Yoshikawa, S. (1996). The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8Å. *Science*, **272**:1136-1144.
- Tusnády, G.E. and Simon, I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J. Mol. Biol.*, **283**:489-506.
- Tusnády, G.E. and Simon, I. (2001). Topology of membrane proteins. *J. Chem. Inf. Comput. Sci.*, **41**:364-368.
- Tusnády, G.E., Dosztányi, Z. and Simon, I. (2004). Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics*, **17**:2964-2972.
- Ubarretxena-Belandia, I. and Engelman, D.M. (2001). Helical membrane proteins: diversity of functions in the context of simple architecture. *Current Opinion in Structural Biology*, **11**:370-376.
- Valadon, P. (2002). RasTop - Molecular Visualization Software. RasTop 2.0.2 Released on October 1st, 2002. <http://www.geneinfinity.org/rastop/> Last accessed: 02/05/2006.
- Villoutreix, B.O. (2002). Structural bioinformatics: methods, concepts and applications to blood coagulation proteins. *Current Protein and Peptide Science*, **3**:341-364.
- Vogt, J. and Schulz, G.E. (1999). The structure of the outer membrane protein OmpX from *Escherichia coli* reveals possible mechanisms of virulence. *Structure Fold. Des.*, **7**:1301-1309.
- Von Heijne, G. (1992). Membrane protein structure prediction: hydrophobicity analysis and the "positive inside" rule. *J. Mol. Biol.*, **225**:487-494.
- Von Heijne, G. (1997). Principles of membrane protein assembly and structure. *Prog. Biophys. Mol. Biol.*, **66**:113-139.

- Yeagle, P.L. and Lee, A.G. (2002). Membrane protein structure. *Biochim. Biophys. Acta*, **2**:143.
- Wallin, E. and von Heijine, G. (1998). Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.*, **7**:1029-1038.
- Werten, P.J., Remigy, H.W., de Groot, B.L., Fotiadis, D., Philippsen, A., Stahlberg, H., Grubmuller, H. and Engel, A. (2002). Progress in the analysis of membrane protein structure and function. *FEBS Lett.*, **529**:65-72.
- Westhead, D.R. and Thornton, J.M. (1998). Proteins structure prediction. *Curr. Opin. Biotechnol.*, **9**:383-389.
- White, S.H. and Wimley, W.C. (1999). Membrane protein folding and stability: Physical principles. *Annu. Rev. Biophys. Biomol. Struct.*, **28**:319-365.
- Wise, A., Jupe, S.C. and Rees, S. (2004). The identification of ligands at orphan G-protein coupled receptors. *Annu. Rev. Pharmacol. Toxicol.* **44**:43-66.
- Wong, S.K. (2003). G protein selectivity is regulated by multiple intracellular regions of GPCRs. *Neurosignals*, **12**:1-12.
- Xu, D., Xu, Y. and Uberbacher, E.C. (1999). Computational tools for protein modelling. *Current Protein and Peptide Science*, **1**:1-21.
- Xu, Q., Axelrod, H.L., Abresch, E.C., Paddock, M.L., Okamura, M.Y. and Feher, G. (2004). X-Ray structure determination of three mutants of the bacterial photosynthetic reaction centers from *Rb. sphaeroides*; altered proton transfer pathways. *Structure*, **12**:703-715.
- Xu, Y. and Xu, D. (2000). Protein threading using PROSPECT: design and evaluation. *Proteins*, **20**:343-354.
- Zimmerman, J.M., Eliezer, N. and Simha, R. (1968). The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.*, **21**: 170-201.