Computational Studies of Protein Sequence and Structure

Rong-I Hong

Laboratory of Molecular Biophysics
and
The Queen's College, Oxford.

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Oxford.

Trinity Term 1999.



Computational Studies of Protein Sequence and Structure.

Rong-I Hong

Laboratory of Molecular Biophysics and The Queen's College, Oxford.

Submitted for the degree of Doctor of Philosophy, Trinity Term, 1999.

ABSTRACT

This thesis explores aspects protein function, structure and sequence by computational approaches. A comparative study of definitions of protein secondary structures was performed. Disagreements in assignment resulting from three different algorithms were observed. The causes of inaccuracies in structure assignments were discussed and possibilities of projecting protein secondary structures by different structural descriptors were tested. The investigation of inconsistent assignments of protein secondary structure led to a study of a more specific issue concerning protein structure/function relationships, namely *cis/trans* isomerisation of a peptide bond. Surveys were carried out at the level of protein molecules to detect the occurrences of the *cis* peptide bond, and at the level of protein domains to explore the possible biological implications of the occurrences of the structural motif.

Research was then focussed on α-helical integral membrane proteins. A detailed analysis of sequences and putative transmembrane helical structures was conducted on the ABC transporters from different organisms. Interesting relationships between protein sequences, putative α-helical structures and transporter functions were identified. Applications of molecular dynamics simulations to the transmembrane helices of a specific human ABC transporter, cystic fibrosis transmembrane conductance regulator (CFTR), explored some of these relationships at the atomic resolution. Functional and structural implications of individual residues within membrane-spanning helices were revealed by these simulations studies.

ACKNOWLEDGEMENTS

The research described in this thesis has been carried out under the overall supervision of Professor L. N. Johnson. I am deeply grateful for her guidance and constant encouragement during the course of my D.Phil. work at Laboratory of Molecular Biophysics. The supervision of Dr. G. J. Barton at the early stage of my work is acknowledged. I would like to thank Dr. M. J. Adams for her supervision on the work of protein *cis/trans* isomerisation. Special thanks are due to Dr. M. S. P. Sansom for his excellent supervision, invaluable suggestions and continuous support and encouragement for the major part of my research work at Oxford.

Support from my colleagues in Dr. Sansom's group is gratefully acknowledged. I would like to thank Drs. I. Shrivastava, G. Smith, W. B. Fischer, Miss K. M. Ranatunga, Miss L. R. Forrest and Mr. M Ulmschneider for their valuable discussions and suggestions.

The financial support from the Ministry of Education, Taiwan ROC, which has made all this possible is highly appreciated. I also wish to express my thanks to The Wellcome Trust Centre for Human Genetics and to the Oxford Supercomputing Centre for the CPU time provided.

On a more personal note, my most heartfelt thanks go to my Mother, Father and family in Taiwan for everything they have done for me. I am also indebted to my wife, Charlene, for her constant encouragement, understanding and daily support while she is also busy with her D.Phil. research.

CONTENETS

***CHAPTER 1* INTRODUCTION.....1**

1.1 TERTIARY STRUCTURES OF PROTEIN. 2

1.1.1 Determinations of Protein Structures 2

1.1.2 Experimentally Solved Protein Structures 5

1.2 PRIMARY SEQUENCES OF PROTEIN. 7

1.2.1 The Number of Known Protein Sequences Exceeds the Number of Known Protein Structures 7

1.2.2 Genomic Era 9

1.2.3 Integral Membrane Protein. 11

1.3 COMPUTATIONAL BIOLOGY AND BIOINFORMATICS.....13

1.3.1 Protein Sequence Analysis. 13

1.3.2 Predictions of Protein Secondary Structure. 13

1.3.3 Predictions of Protein Tertiary Structure. 14

1.3.4 Modellings of Protein Conformation. 15

1.3.5 Classifications of Protein Structure.....18

1.4 THESIS PERSPECTIVE AND OVERVIEW. *19*

***CHAPTER 2* ASSIGNMENTS OF PROTEIN SECONDARY STRUCTURES. 21**

2.1 INTRODUCTION. 22

2.1.1 Protein Secondary Structures. 22

2.1.2 Descriptors for Protein Secondary Structures. 23

2.1.3 Criteria for Defining Protein Secondary Structure 26

2.1.4 Algorithms of Protein Secondary Structure Assignments 28

2.2 METHODOLOGY.....33

2.2.1 Protein Data Sets. 33

2.2.2 Protein Secondary Structure Assignments.....34

2.2.3 Comparison of Secondary Structure Assignment at the Amino Acid Residue Level.....36

2.2.4 Comparison of Secondary Structure Assignment at Segment Level. 36

2.2.5 Calculations of Protein Geometric Parameters. 37

2.3 RESULTS AND ANALYSIS. 37

2.3.1 Assignment Results of Each Program. 37

2.3.2 Comparison of All Three Programs. 39

2.3.3 The Relation between Protein Structure Resolution and Protein Secondary Structure Assignment. 42

2.3.4 Lengths of Helix and Strand. 47

2.3.5 Comparison of Length Distributions in Different Secondary Structure .50

2.3.6 Applications of Structural Descriptor. 54

2.4 DISCUSSION. 60

***CHAPTER 3* DISORDER IN PROTEIN STRUCTURE AND FUNCTION: THE *CIS* PEPTIDE BOND 63**

3.1 INTRODUCTION. 64

3.1.1 Disorders in Protein Structure and Function. 64

3.1.2 *Cis* Peptide Bonds. 66

3.2 SURVEY ON PROTEIN STRUCTURE DATABASE. 71

3.2.1 Distributions of the *Cis* Peptide Bond 71

3.2.2 ϕ, ψ Conformations of the *cis* Peptide Bond 73

3.2.3 Temperature Factor Analysis of the *Cis* Peptide Bond 75

3.2.4 Accessibility of the *Cis* Peptide Bond 77

3.3 ANALYSIS ON PROTEIN DOMAIN STRUCTURES 79

3.3.1 Assumption 80

3.3.2 Speculation 81

3.3.3 Observation 89

3.3.4 Hypothesis 96

3.4 DISCUSSION. 99

***CHAPTER 4* ABC TRANSPORTERS: FROM GENOME SEQUENCES TO TRANSMEMBRANE HELICAL STRUCTURES 102**

4.1 INTRODUCTION. 104

4.1.1 Integral Membrane Proteins 104

4.1.2 ABC Transporters. 107

4.1.3 Aims of This Study 110

4.2 METHODOLOGY.....112
4.2.1 Sequence Database.....112
4.2.2 Searching for Similar Sequences.....116
4.2.3 Multiple Sequence Alignment.....118
4.2.4 Prediction of Transmembrane Helices.....121
4.2.5 Periodicity Analysis.....127
4.2.6 α-Helical Structure Model Building.....129
4.2.7 Methods Integrated in This Study.....131
4.3 RESULTS.....132
4.3.1 Extraction and Prediction of ABC Transporters.....132
4.3.2 Characterisations of Each Genome.....134
4.3.3 Periodicity Analysis.....159
4.3.4 Models.....174
4.4 DISCUSSION.....176
***CHAPTER 5* MD SIMULATIONS OF CFTR TM HELICES.....182**
5.1 INTRODUCTION.....183
5.1.1 Cystic Fibrosis Transmembrane Conductance Regulator.....183
5.1.2 Applications of Molecular Dynamics to Integral Membrane Protein...190
5.1.3 Aims of This Study.....192
5.2 METHODOLOGY.....193
5.2.1 Overview of Methods.....193
5.2.2 Predictions of Transmembrane Helices.....195
5.2.3 Building Helical Models.....196
5.2.4 Construction of Helix/Bilayer/Water Simulation Systems.....198
5.2.5 MD Simulations.....200
5.3 ANALYSIS.....201
5.3.1 Progress of Simulations.....201
5.3.2 Dynamics of Transmembrane Helix Backbone.....204
5.3.3 Fluctuations of Transmembrane Helices.....209
5.3.4 Dynamics of Secondary Structures.....214

5.3.5 Dynamics of Helical Structures 219
5.3.6 Dynamics of Hydrogen Bonds 222
5.3.7 Dynamics of Helical Conformations I: Tilt Angles 225
5.3.8 Dynamics of Helical Conformations II: Kink Angles 230
5.3.9 Interactions at the Interface 237
5.4 DISCUSSION. 241
5.4.1 Significance of MD Simulations 241
5.4.2 Biological Implications of This Study 242
5.4.3 Limitations of the MD Simulations. 245
***CHAPTER 6* CONCLUDING REMARKS AND FURTURE DIRECTIONS 246**
6.1 SUMMARY. 247
6.2 FUTURE DIRECTIONS. 248
6.2.1 Protein Secondary Structure Assignment 248
6.2.2 Cis Peptide Bonds and Disordered Protein Structures 249
6.2.3 ABC Transporters. 249
6.2.4 Cystic Fibrosis Transmembrane Conductance Regulator 251
6.3 FINAL REMARKS 253
APPENDICES **..... A-1**
REFERENCES **..... R-1**

Chapter 1

Introduction

1 Protein Function, Structure, and Sequence.

Proteins are probably the most complicated molecules that exist in nature. They are abundant in all organisms and are fundamental for a variety of biological processes. The complexity and biological importance protein possesses attract researchers from many different disciplines; including chemistry, physics and computation. These studies have significant impacts on biology, medicine and industry as well as on our daily life.

The versatile range of protein functions is generally believed to be specified by the enormous variety of their structures (Pertuz, 1992). However, the tertiary structure of a protein has been proposed to be determined simply by the chemistry of its amino acid sequence and its existing environment (Anfinsen, 1973). This idea has been widely accepted and is sometimes referred as the Anfinsen dogma.

With the thinking flow of function, structure and sequence, this chapter will start with discussions on protein structure analysis, section 1.1, and followed by the discussion of protein sequence analysis, section 1.2. Then, how the vast amount of information on protein sequences and structures have revolutionised the way of biological research is being conducted will be covered, section 1.3. Finally, an outline of this thesis will be shown, section 1.4.

1.1 Tertiary Structures of Protein.

1.1.1 Determinations of Protein Structures.

It is of great interest to determine protein structures in order to have a clear understanding and a deeper insight on the functioning mechanisms of protein.

Toward this goal, many efforts have been focused on developing methods to solve protein structures.

X-ray Crystallography.

X-ray crystallography has long been used to determine protein structures (Kendrew et al., 1958) and provided useful structural information on proteins (Drenth, 1994).

To determine a protein structure using X-ray crystallography, it is necessary to have well-ordered crystals of protein molecules. The crystal is then irradiated with X-rays. The X-rays are diffracted from a series of reflective planes, composed with a regular arrangement of molecules in a crystal, and the diffraction patterns are recorded. An electron density map can be derived from these diffraction patterns. A three-dimensional structure is then obtained from the map by modelling the protein sequence into the electron density. Most protein structures solved and stored in the Protein Data Bank (Bernstein et al., 1977) are determined by X-ray crystallography.

The low solubility of integral membrane proteins, which is caused by their hydrophobic nature, presents extra difficulties to produce well-ordered protein crystals. Even though membrane proteins can be co-crystallised with lipids or detergents, there are still only few membrane protein structures available, see section 1.1.2.

Other diffraction studies such as neutron diffraction and electron diffraction, have been applied to determine protein structures as well (Chance and al., 1994). Neutron diffraction is useful for determining the location of hydrogen atoms in a

single crystal, and has been used to study the rotational orientation of transmembrane α-helices in bacteriorhodopsin (Samatey et al., 1994).

Nuclear Magnetic Resonance Spectroscopy (NMR)

Nuclear magnetic resonance (NMR) spectroscopy provides an alternative way to X-ray crystallography for protein structure determinations (Bax and Lerner, 1986). In NMR experiments, the protein sample may be solid or in solution. Thus, it has the advantage over crystallographic techniques since protein molecules are in their native form rather than in the crystal form. However, the physical principles limit the application of this technique to molecules of less than 35 to 40 kD.

NMR techniques have been applied widely in globular proteins to study not only their structures but also their dynamics (Campbell, 1996). For membrane proteins, it is possible to perform solid state NMR experiments with membrane proteins in bilayers (Smith and Peersen, 1992; Opella, 1997). With this application, the possible secondary structures and conformational changes of small TM peptides can be deduced (Watts et al., 1995).

Electron Microscopy

The electron microscopy technique uses focused electron beams as probes to detect protein structures on a layer of protein crystals, and can be seen as an analog of electron crystallography (Heymann et al., 1997). It is especially useful for the determination of membrane protein structures. However, it does not always provide enough atomic resolutions for protein structure determination, and needs to combine with other experimental techniques; for example, the structure of the multidrug resistance P-glycoprotein has been solved with this technique, in

combination with image analysis, to the resolution of 25 Å (Rosenberg et al., 1997).

1.1.2 Experimentally Solved Protein Structures.

Globular Proteins

When a protein structure is experimentally determined, its atomic coordinates are deposited and stored in the Protein Data Bank (PDB) (Bernstein et al., 1977). During the time of this thesis being written, there were 9,352 protein coordinate entries in the PDB. The vast majority of protein structures are of globular proteins. Not suprisingly, many structural analysis databases have been developed on the basis of these protein structures (Murzin et al., 1995; Holm and Sander, 1996; Orengo et al., 1997; Siddiqui, 1997), section 1.3.5.

Integral Membrane Proteins

Even though there are almost 10,000 protein structures experimentally solved, the number of structure-determined transmembrane proteins is still much less than 100. The small number of structure-solved transmembrane proteins reveals difficulties for studying membrane proteins.

There are two main difficulties in structural determination of membrane proteins compared to globular proteins: obtaining enough protein to work with and producing a sample that is suitable for conventional protein structure determination techniques, *i.e.* to be soluble and crystallised. For most membrane proteins, overproduction in an expression system has not been as successful as for globular proteins, due to the lack of suitable expression systems (Schertler, 1992). In addition, because of the hydrophobic nature of transmembrane proteins, they are known to be hard to extract, purify, solubilise and crystallise.

A list of experimentally determined transmembrane protein structures is summarised, (Tab. 1.1).

Protein/Source	Method/ Resolution (Å)	Folds of the transmembrane region	Reference
bacteriorhodopsin *H. halobium* *H. salinarium*	X-ray 2.5 EC 3.0	7 TM α-helices	(Pebay-Peyroula et al., 1997) (Mitsuoka et al., 1999)
Photosynthetic reaction centre *R. viridis* *R. sphaeroides*	X-ray 2.3 X-ray 3.1	3 TM subunits, 11 TM α-helices	(Deisenhofer et al., 1995) (Allen et al., 1987)
light harvesting complex plant (LH)	EC 3.4	3 TM α-helices	(Kuhlbrandt et al., 1994)
light harvesting complex *R. acidophila* (LH2)	X-ray 2.5	2 concentric rings of 9 TM α-helices	(McDermott et al., 1995)
photosystem I *S. elongatus*	X-ray 4.0	trimer, each subunit 31 TM α-helices	(Krauss et al., 1993)
cytochrome c oxidase *P. denitrificans* bovine heart	X-ray 2.8 X-ray 2.8	4 subunits, 12 TM α-helices	(Iwata et al., 1995) (Tsukihara et al., 1996)
cytochrome bc1 (cytochrome c reductase)	X-ray 3.0	TM region mainly α-helical	(Iwata et al., 1998)
KscA	X-ray 3.2	tetramer, each subunit 2 TM α-helices	(Doyle et al., 1998)
porins *R. capsulatus* *R. blastica* OmpF and PhoE *E. coli*	X-ray 1.8 X-ray 2.0 X-ray 3.0	trimers, TM domain 16 stranded β-barrels	(Weiss et al., 1991) (Kreusch et al., 1994) (Cowan et al., 1992)

EC: electron crystallography
TM: transmembrane
X-ray: X-ray crystallography

Table 1.1 Summary of membrane proteins of known structures. In the majority of structures obtained at high resolution so far, the transmembrane domain is α-helical with the exception of the porins.

1.2 Primary Sequences of Protein.

Though experimental techniques, such as X-ray crystallography and NMR analysis, are now available for solving protein structures at the atomic resolution. The primary structures of proteins can be determined more easily by direct amino acid sequencing or derived from DNA sequencing. This indicates an alternative way to understand the protein structure-function paradigm, since the tertiary structure of a protein is determined by its one-dimensional amino acid sequence and useful biological information can be deduced from the primary amino acid sequences (Anfinsen, 1973).

1.2.1 The Number of Known Protein Sequences Exceeds the Number of Known Protein Structures.

While there are almost 10,000 protein structures already solved, the number of known protein sequences in release 37.0 of SWISS-PRO database has already reached 77,977 sequence entries (Bairoch and Apweiler, 1998). And the release 9 of TrEMBL database, which contains translations of all coding sequences in the EMBL Nucleotide Sequence Database, is comprised of 221,422 sequence entries (Bairoch and Apweiler, 1998). The number of available protein sequences overwhelmingly exceeds the number of experimentally-determined protein structures, (Fig. 1.1).

The number of entries in the Protein Data Bank is currently growing by over 100 per month. However, the protein sequence databases are growing even more rapidly and now include the sequences of the entire genomes of several different organisms. Many additional genomes will be solved in the near future. The

discrepancy between the number of known amino acid sequences and the number of structure-solved proteins is still enlarging rapidly with time.

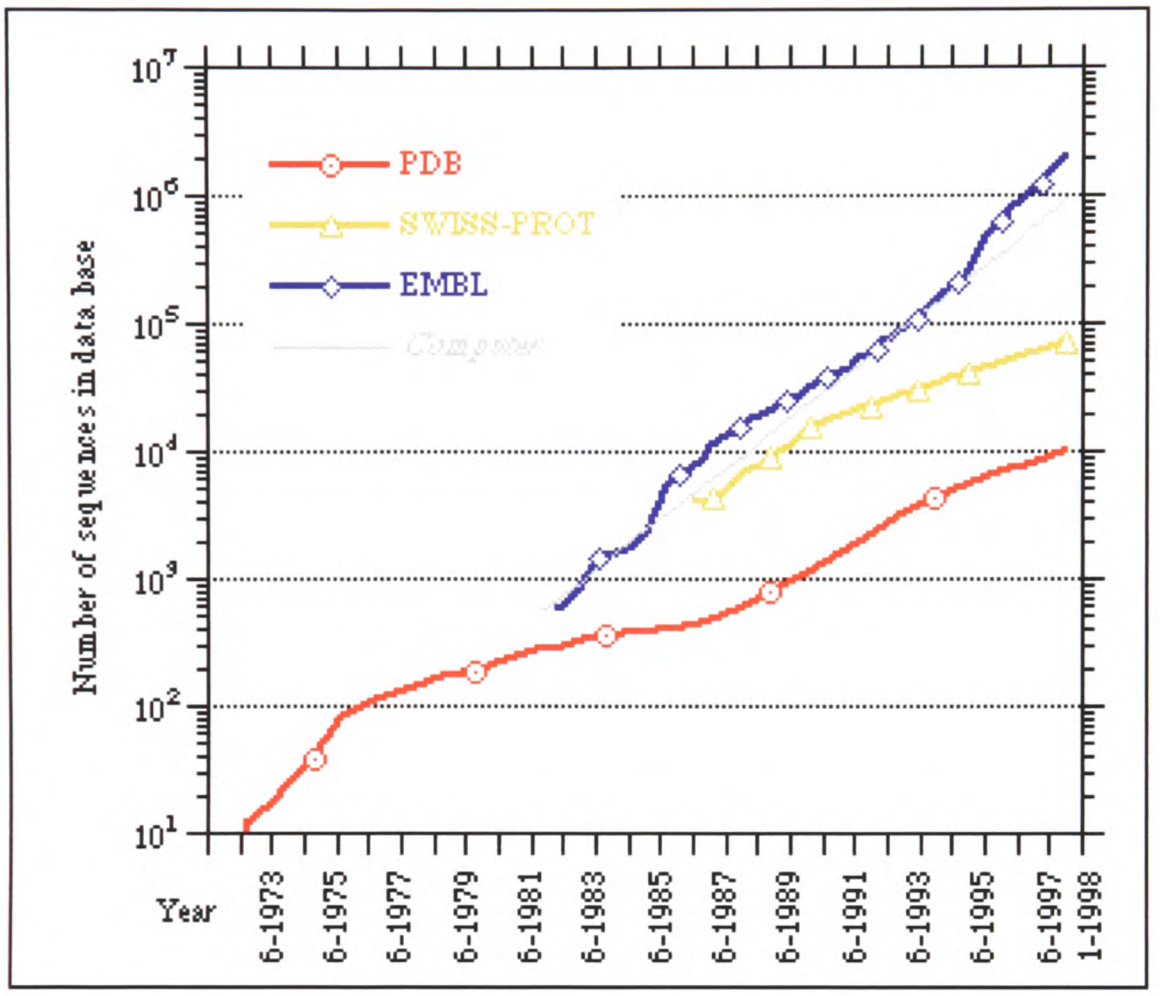


Figure 1.1 Rate of growth in databases of biomolecules. The explosion of biomolecular data is illustrated by three representatives: the protein structure database PDB (Bernstein et al., 1977) (red line, dashed circles), the protein sequence database SWISS-PROT (Bairoch and Apweiler, 1998) (brown line, open triangles), and the nucleotide sequence database EMBL (Stoesser et al., 1997) (blue line, open diamonds). The number of symbols (circles, triangles, diamonds) roughly reflects the number of releases of the respective database. For comparison, the growth of computer speed is also shown (grey thin line, assumptions: speed doubles every 18 months). Adapted from (Rost, 1998) with permission.

1.2.2 Genomic Era.

Detailed catalogues of proteins in different organisms are being compiled at an astonishing speed by different large-scale genome sequencing projects. The results not only enable us to collect a large number of sequences, but also provide glimpses of the inventory of gene products in entire organisms. Our understanding of cell biology is being rapidly advanced with the progress of these genome projects. With abundance of protein sequence data at hand, fundamental relationships between protein functions, structures and sequences are being further studied with a combination of different means; for example, coalescence between the protein structure and function (Murzin, 1996), integration between protein sequence and function (Brent and Finley Jr, 1997), and amalgamation between protein sequence and structure (Rost, 1998).

Several genomes have been completely sequenced during the last few years, see Chapter 4. Attempts have been made to link genomic sequences to the encoded proteins. Recent advances in employing protein two-dimensional gel electrophoresis, image analysis and protein identification, such as direct sequencing or mass spectrometry, have opened a *de novo* way of rapid characterisation of genes and gene-products at a large-scale genomic level (Humphery-Smith et al., 1997; Wilkins et al., 1997). One example comes from the sequencing-completed *S. cerevisiae* genome, from which the organisations and functions of protein-encoded genes are being progressively characterised (Hodges et al., 1999). A new term in English has been created to address this progress, proteome, which indicates the **prote**ins expressed by a gen**ome** (Wasinger et al., 1995).

Knowledge of relationships between protein structure and function remains a primary focus for structural biology. This intention has been benefited from the mass of available protein sequences as well. Similar structures frequently appear in proteins of different genomes (Gerstein, 1997; Gerstein and Levitt, 1997), and it has been estimated that around 10% of proteins in sequencing-completed genomes share structural similarities with proteins in the PDB (Rost, 1998), (Fig. 1.2).

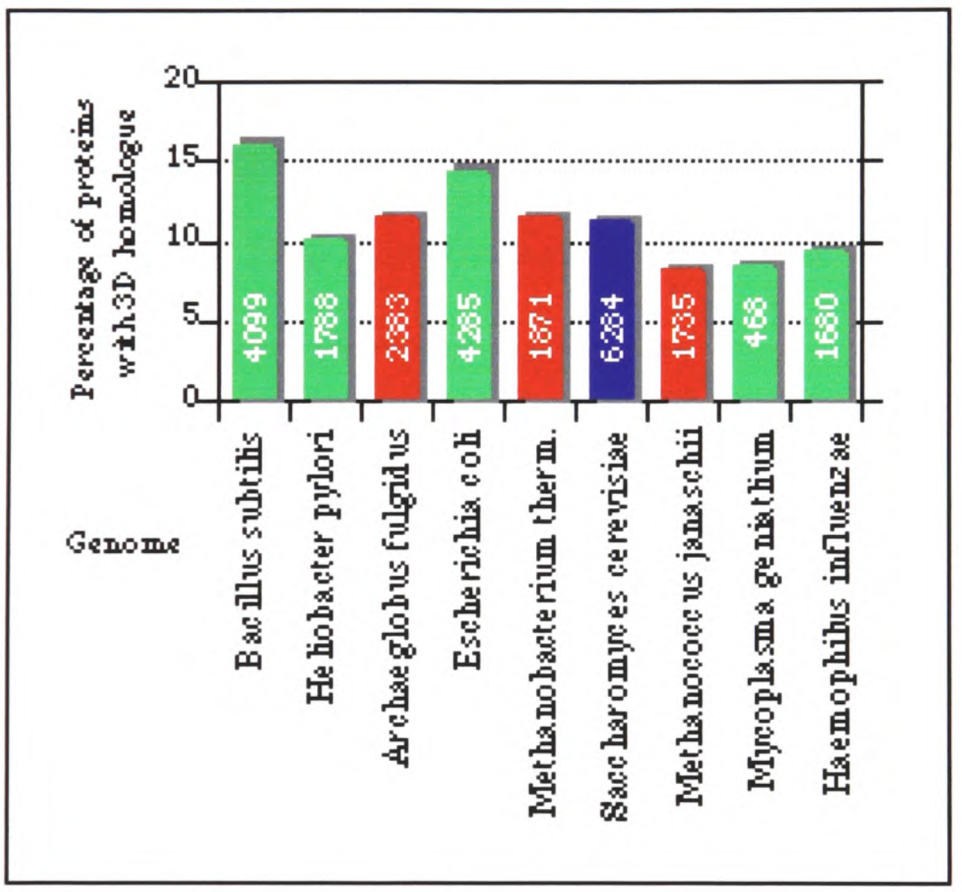


Figure 1.2 Percentage of proteins in genomes with homologues of known structure. Numbers of proteins for which a sequence homologue of known structures exists in PDB as percentage of all identified proteins are shown. The total numbers of proteins used are given in the bars. The eukaryote is shown in blue, prokaryotes are shown in green, and the archae are in red. The structure is inferred by homology modelling. Adapted from (Rost, 1998) with permission.

Extending this idea of homology modelling, section 1.3.3, to a genomic level delivers experimentally determined protein structure models for other proteins, whose structures are yet not known, in different genomes (Sanchez and Sali, 1997; Sali, 1998). This emerging strategy of classifying protein structures in genomes is referred as structural genomics, and has an optimistic object; to find out proteins with *de novo* structures and eventually narrow down the gap between proteins of known sequence and proteins of known structure (Sali, 1998; Rost, 1998).

The idea of structural genomics has been practised on many organisms already; for example, *Mycoplasma genitalium* (Fischer and Eisenberg, 1997; Huynen et al., 1998), *Mycoplasma pneumoniae* (Grandori, 1998), *Saccharomyces cerevisia* (Sanchez and Sali, 1998; Hegyi and Gerstein, 1999). These provide a structure-based approach to study the problems of protein functions.

1.2.3 Integral Membrane Protein.

Studies of the membrane protein have been facilitated with the abundance of protein sequences as well. As mentioned earlier, section 1.1.2, methods commonly employed for studying soluble protein structures, X-ray crystallography and NMR analysis, are with great difficulties to be applied to integral membrane proteins. Alternative methods for studying membrane proteins on the basis of protein primary sequences have been developed.

In the early 1980's, *i.e.* more than 10 years before the genome era, a hypothetical mechanism had been proposed to explain the recurrant appearance of membrane-buried proline residues based on observations of amino acid compositions of membrane protein sequences (Brandl and Deber, 1986). Preferences of specific amino acids on certain positions of the transmembrane (TM) α-helices have been

detected by statistical analysis of TM segments of human type I single spanning membrane proteins (Landolt-Marticorena et al., 1993). Applications of sophisticated statistical analysis to a dataset of putative transmembrane proteins have revealed the details of average lengths, amino acid compositions and distributions of specific amino acids in the TM α-helical segments of membrane proteins (Arkin and Brunger, 1998).

With protein sequence data of many genomes available, efforts have been spent on determining the amount of the genome that corresponds to membrane proteins and their characteristics. For example, percentages of proteins in different genomes corresponding to membrane proteins have been estimated around 20% to 30% based on hydropathy analysis (Boyd et al., 1998; Wallin and von Heijne, 1998).

Attempts have also been focused on detecting the preferred arrangements and distributions of transmembrane helical domains using both hydropathy and statistical analysis. It has been suggested that no particular integral membrane protein families, *i.e.* each containing the same number of transmembrane α-helices, are found in the genomes of *M. genitalium*, *H. influenza*, *M. jannaschii* and *S. cerevisiae* (Arkin et al., 1997). However, another study suggests that there are preferences for membrane proteins of 6 and 12 TM helices in unicellular organisms, and membrane proteins of 7 TM helices for *C. elegans* and *H. sapiens* (Wallin and von Heijne, 1998).

The disagreement reveals the difficulties of studying membrane proteins and characterising the membrane-spanning domains.

1.3 Computational Biology and Bioinformatics.

There are needs for developing tools to characterise the vast amount information of protein structures, protein sequences and genomes. A variety of computational methods (computational biology), and information technologies (bioinformatics) have been implemented to address this problem at different levels.

1.3.1 Protein Sequence Analysis.

The primary goal of protein sequence analysis is to identify similarities between sequences whose structures and functions are unknown and a set of protein sequences whose structures or functions have been elucidated. This implies two things, how to obtain sequences with certain degrees of similarity from a dataset for proteins of known structures or functions, and how to clarify the reliabilities or subtle differences between the query sequences and matched sequences; in other words, database searching and multiple sequence alignments.

Details about the theories and developments of existing protein sequence databases, database searching and protein multiple sequence alignments are presented at sections 4.2.1, 4.2.2 and 4.2.3 respectively.

1.3.2 Predictions of Protein Secondary Structure.

Protein secondary structure elements are regarded as the building blocks of a protein structure, details see chapter 2. Thus, it is of interest to predict protein secondary structures from their primary sequences.

Since the first attempt for predicting protein secondary structures by statistical analysis (Chou and Fasman, 1974), many methods have been developed to improve the accuracy of prediction results by considering chemical and physical

characteristics of amino acids; for example, Wako's method (Wako and Blundell, 1994), DSC (King and Sternberg, 1996), and PREDATOR (Frishman and Argos, 1997a). Delicate mathematical models have been implemented in designing prediction algorithms; such as machine learning for GOLEM (Muggleton et al., 1992), and neural networks for PHD (Rost and Sander, 1993). In general, methods include evolutionary information using multiple sequence alignments tend to have more accurate prediction results (Barton, 1995; Frishman and Argos, 1997b).

Methods mentioned above are largely designed for predicting globular proteins. With the extra constraints provided by the lipid bilayer to limit possible fold arrangements, secondary structure predictions for membrane protein are thought to be easier. Details of membrane protein structure predictions will be presented at section 4.2.4.

1.3.3 Predictions of Protein Tertiary Structure.

Based on the Anfinsen dogma, predictions of protein three-dimensional structure in principle are achievable. This goal has become something of a Holy Grail for biologists. However, with the fact that the number of possible structures for each protein is astronomically large, it is widely believed that predictions of protein three-dimensional structure directly from the amino acid sequence remain impossible.

However, comparative modelling from the experimental-solved structures of evolutionary related proteins, *i.e.* homologues, provides a powerful approach for protein tertiary structure predictions (Sternberg, 1996; Srinivasan et al., 1996). In other words, protein structures and functions can be deduced from related protein

sequences whose structures are solved (Sanchez and Sali, 1997), which is also known as homology modelling.

Approaches based on the assembly of fragments from known protein structures to build up a structure for the query protein sequence have been widely applied in many programs; for example, COMPOSER (Srinivasan and Blundell, 1993). In addition, employing three-dimensional distance constraints or other restraints derived from experiments has been proven to be especially helpful to model a protein structure from distantly related homologues (Srinivasan et al., 1996), and has been implemented in programs like MODELLER (Sali and Blundell, 1993).

Progress of structural genomics, section 1.2.2, is mainly dependent on the developments of comparative modelling on a genomic scale.

1.3.4 Modellings of Protein Conformation.

Molecular dynamics are the extension of classical mechanics into the atomic level, details see appendix, and have been applied widely in structural biology (Karplus and Petsko, 1990). Due to the huge number of possible conformations of each peptide unit and the finite power of current computers, it is not possible to use molecular dynamics to obtain a structure from only its primary amino acid sequence. This is an aspect of this Levinthal paradox (Holm and Sander, 1996; Karplus, 1997).

However, it is possible to use molecular dynamics in conjunction with experimental data to explore protein structures. For example, atom-atom distances determined experimentally using NMR can be included in the molecular dynamics simulation as restraints on atoms' movement in order to restrict the number of

possible conformations. With this combination, protein structures can be obtained (Brunger and Karplus, 1991). The technique can also be applied to refine protein structures obtained from X-ray crystallography studies (Brünger et al., 1987).

This technique is not only used extensively in protein structure predictions with facilitation of extra information (Karpen and Brooks III, 1996), but also especially useful to study dynamical systems; for example, examining movements of protein loop regions (Derreumaux and Schlick, 1998), and tracing protein motions (Gerstein and Krebs, 1998).

It has been proven to be of great help to elucidate mechanism of protein folding as well. A recent MD simulations study has been performed on a system of a 36-residue peptide and approximately 3,000 water molecules for 1 μs, which is the time scale close to the requirement of folding small proteins (Duan and Kollman, 1998). The simulation shows a mechanism for the protein to find a folding intermediate, an important step in proceeding to its fully folded state. This work took approximately 2 months of CPU time on a 256-CPU computer, and has been commented as providing "a glimpse of the Holy Grail" (Berendsen, 1998).

Molecular dynamics simulations have broad applications for studies of transmembrane proteins. For example, a parallel bundle of TM α-helices surrounding a central pore of the nicotinic acetylcholine receptor has been modelled using simulated annealing *via* restrained molecular dynamics (Kerr et al., 1994).

In addition, this technique has an advantage to provide insights into interactions and dynamics between membrane-spanning peptide, lipid bilayer, water molecules

and solutes at the atomic resolutions, which are still beyond the capacity of "wet" experiments (Biggin and Sansom, 1999). The interactions between individual TM α-helices of bacteriorhodopsin and bilayer of dimyristoyl-phosphocholine have been examined by molecular dynamics calculations (Woolf, 1997; Woolf, 1998). Different helical bundles, consisting of different number of alamethicin TM helices, have been explored in an 1-palmitoyl-2-oledylphosphocholine bilayer system *via* molecular dynamics simulations to investigate their stability and conformational dynamics (Tieleman et al., 1999a; Tieleman et al., 1999b).

In conjunction with experimental data derived from spectroscopy studies, *e.g.* NMR, or biochemical studies, *e.g.* site-directed mutagenesis, molecular dynamics deliver reliable chacterisations of protein structures and conformations (Brunger and Karplus, 1991). Apart from immense applications of NMR and MD simulations toward globular proteins, similar applications have been extended to membrane proteins. Integration of structural data deduced form site-directed mutagenesis and molecular dynamics simulations has been demonstrated to model the conformations of nicotinic acetylcholine receptor at the closed state (Sansom et al., 1995). A structure model for the pentameric transmembrane domain of phospholamban has been suggested on the basis of global conformational search methods coupled with molecular dynamics simulations and mutagenesis results (Adams et al., 1995). In addition, combinations of spatial restraints derived from spectroscopic analysis and global conformation searching using molecular dynamics have been employed to obtain the structures of the M2 channel from influenza A (Kukol et al., 1999).

1.3.5 Classifications of Protein Structure.

As the numbers of protein structures increase, various methods have been developed for the classification of protein structures. Since protein secondary structure elements are the constructing elements for protein architectures, these methods are largely devised on categorising the relative orientations and topologies of the protein secondary structure elements in a three-dimensional space, *i.e.* protein folds or domains.

By measuring sequence and structural similarities using dynamic programming, a protein structural domain database, 3Dee, was designed to store protein domain definitions and to categorise protein similarities into different levels. (Siddiqui, 1997)

An algorithm, Dali, has been devise to compare protein structure in a two-dimensional manner using distance matrices of Cα atoms of two proteins to detect similar patterns, and employing a Mote-Carlo procedure to optimise the preliminary results detected (Holm and Sander, 1993). Based on this algorithm, a database of structurally aligned protein fold families was constructed, FSSP (Holm and Sander, 1996).

SSAP is an algorithm developed to perform rapid structural alignments using comparison of inter-atomic distance vectors with the two structures being compared (Orengo et al., 1992). Based on this algorithm and facilitation of human inspections for certain cases, a database of hierarchical domain classifications of protein structure was built, CATH (Orengo et al., 1997). There are five levels within the hierarchy of CATH, which are class, architecture, topology, homology and sequence.

The SCOP (structural classification of proteins) database describes structural and evolutionary relationships between proteins of known structure (Murzin et al., 1995). Automated methods were employed at the level of detecting protein sequence identity. At the level of detecting protein structure similarity, applications of exports' knowledge provide a high degree of reliability of this database for sorting protein structures. It is organised in the order class, fold, superfamily, family and domain.

The idea of protein domains has been used extensively for different surveys in both chapter 2, for 3Dee, and chapter 3, for SCOP.

1.4 Thesis Perspective and Overview.

This thesis focuses on studying some of the relationships between protein sequence, structure and function by computational methods.

As protein secondary structure elements are regarded as the elementary building blocks of protein tertiary structures, detailed studies of protein secondary structure assignments are presented in Chapter 2. This chapter includes comparisons of different algorithms for defining protein secondary structures, statistical analysis of regular secondary structure elements in the Protein Data Bank, quality analysis of protein structures and their influences on assigning secondary structures.

During the survey of protein structure quality, many peptide bonds are found in the *cis* conformations. Since *cis* peptide bonds are seen as "disordered" structures, a survey of *cis* peptide bonds at both protein molecule level and domain level is reported in Chapter 3. Efforts have then been focused on correlating protein functions with this unusual conformation. Information from both a protein domain

study and a literature search were performed. Finally, a hypothesis about the *cis* peptide bond and its potential structural and functional role is proposed.

More interests are focused on transmembrane proteins from chapter 4, since proline residue, a helix destabiliser, is frequently detected during the study of *cis* peptide bonds. To make maximum use of abundant genomic data available on the public domains, intensive studies on a family of transmembrane proteins, the ABC transporter, have been conducted. Different computational methods are employed, which include homologous searching, sequence analysis, phylogenetic analysis, predictions of transmembrane helices, and *ab inito* modelling of transmembrane helices using simulated annealing/molecular dynamics. Correlation between protein sequences, folds of the transmembrane helical domain, and possible biological functions are reported at the end of Chapter 4.

Limitations of sequence analysis and structure predictions based on primary sequences are observed in Chapter 4. With more sophisticated computational methods, the problem of studying protein structures computationally can be refined. However, more computational time is required. Molecular dynamics simulations are applied on one of the most interesting protein from the ABC transporters, cystic fibrosis transmembrane conductance regulator (CFTR). Investigations in chapter 5 have been conducted to examine the biophysical behaviours of the individual transmembrane helices of CFTR in a solvated lipid bilayer system. Interesting results are detected between computational data and experimental data.

Thus, the overall goal of this thesis is trying to solve some biological problems *in silico*.

Chapter 2

Assignments of Protein Secondary Structures

2 Preface to Chapter 2

The biological function of a protein is determined by its three-dimensional structure. An accurate assignment of protein secondary structures is useful for understanding the protein tertiary structure. Thus, it is of interest to examine the assignment results from different programs, and to explore potential conformational descriptors for projecting protein secondary structures.

This chapter is divided into 4 sections. Section 2.1 starts with introductions to protein secondary structure assignment and different geometric parameters to describe protein conformations. Section 2.2 covers protein data sets and programs used for secondary structure assignment in this study. Section 2.3 looks into details of different assignments and different ways to describe protein conformations. Section 2.4 discusses the biological implications of this study.

2.1 Introduction.

2.1.1 Protein Secondary Structures.

The existence of regular spatial arrangements of main chain atoms in a protein molecule was first proposed from theoretical studies in the concepts of the α-helix (Pauling et al., 1951), and the β-strand (Pauling and Corey, 1951). These regularities within protein structures have been referred as secondary structure elements (SSE), and have been found in many protein structures solved by X-ray crystallography or NMR analysis. The α-helix was first verified from high-resolution atomic structure of myoglobin by X-ray crystallography (Kendrew et al., 1960). Later, the β-strand was found to exist in the X-ray structure of lysozyme

(Blake et al., 1965). On the basis of regular secondary structure elements, a protein three-dimensional structure can be simplified as being built up with the spatial organisation of its secondary structure elements in a three-dimensional space.

Accurate assignments of secondary structure elements, unambiguous locations of these elements within the protein molecule, and delineation of secondary structure elements are essential in describing protein conformation, following experimental determinations of a protein structure.

Assignments of protein secondary structure elements have important impacts in many aspects. For instance, it is believed that knowledge of assigning protein secondary structure through its structural coordinates and amino acid sequence will provide a structural basis for *de novo* designing of proteins and for molecular modelling of protein structures. In addition, definitions of protein domains and classifications of protein structures are heavily dependent on accurate assignments of secondary structure elements. Furthermore, accurate assignments of secondary structures will provide necessary templates to examine the quality of secondary structure prediction. In the context of evolution, a comparative study of protein folds has suggested that spatial arrangements of secondary structure elements are more conserved than the protein tertiary structure (Mizuguchi and Go, 1995).

2.1.2 Descriptors for Protein Secondary Structures.

Backbone Dihedral Angles

In a polypeptide chain, the main chain N-Cα and Cα-C bonds are relatively free to rotate. These rotations are represented by the torsion angles ϕ and ψ respectively which have conventionally been used to describe the backbone of a polypeptide

chain. The ϕ dihedral angle is defined by the atoms of C_{i-1}, N_i, $C\alpha_i$, C_i, and the ψ dihedral angle is defined by N_i, $C\alpha_i$, C_i, N_{i+1}; where *i* represents the residue number (Srinivasan et al., 1975), (Fig. 2.1 A).

Systematic analysis of ϕ and ψ dihedral angles of small polypeptides reveals that distribution of (ϕ, ψ) can be classified on the basis of peptide bond conformations. The preferred (ϕ, ψ) angles of a particular secondary structure can be easily observed on a Ramachandran plot (Ramachandran et al., 1963), which provides a method of defining protein secondary structure elements.

There are drawbacks of using (ϕ, ψ) dihedral angles solely to define secondary structure elements. First, small errors in the (ϕ, ψ) values can lead to incorrect assignments (Levitt and Greer, 1977). In addition, the atomic coordinates of the peptide group needed to calculate (ϕ, ψ) are not always known accurately in many earlier determinations of protein structure by X-ray crystallography.

ω Torsional Angles

The ω torsional angle is defined by the atoms of $C\alpha_i$, C_i, N_{i+1} and $C\alpha_{i+1}$ (Srinivasan et al., 1975), (Fig. 2.1 A). This angle is less dependent on the types of secondary structures; however, it determines the *cis/trans* conformation of a peptide bond. It occurs predominantly in the *trans* conformation, *i.e.* around 180^{o}. A detailed study of the ω angles and *cis/trans* isomerisations of the peptide bonds will be presented in Chapter 3.

Cα Pseudo-dihedral Angles

One alternative way for compensating the drawback of some missing main-chain atoms is to consider Cα atoms only for protein secondary structure assignments. A

pseudo-dihedral angle, φ, has been proposed to describe the local conformation of protein residues (Srinivasan et al., 1975). The pseudo-dihedral angle φ is defined by four consecutive Cα atoms; C_{i-1}, C_i, C_{i+1} and C_{i+2}, (Fig. 2.1 B). This pseudo-dihedral angle φ also has an advantage for describing secondary structure elements, since φ considers conformation degrees of freedom involving four consecutive residues which span more residues than those defining the (ϕ, ψ) dihedral angles.

A method based on contemplating the φ pseudo-dihedral angles and triple scalar products of three unit vectors, which are based on the four consecutive Cα atoms, to project the conformation of protein secondary structure elements has been suggested. This method shows that it is possible to reduce the dimensionalities of using combination of (ϕ, ψ) dihedral angles to a pair of protein secondary structure descriptors, which are defined by four sequential Cα atoms (Smith et al., 1997).

Pseudo-torsional OCCO Backbone Angles

It has been proposed that a pseudo-torsional angle, OCCO, can be briefly described as a function of ϕ and ψ dihedral angles (Richardson and Richardson, 1989), which reveals that this pseudo-torsional angles may be used as simplified representations of ϕ and ψ angles. Geometrically, this OCCO angle is defined by consecutive atoms of O_{i-1}, C_{i-1}, C_i and O_i, (Fig. 2.1 C).

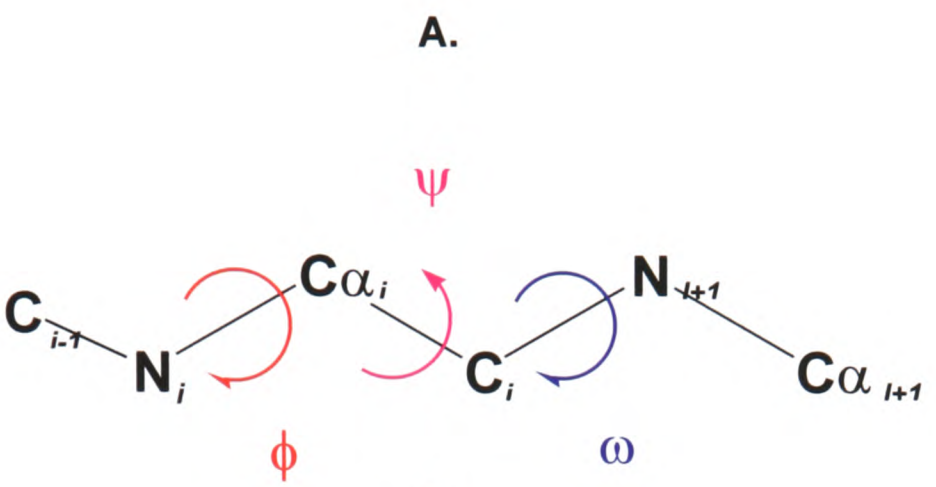
In practice, this pseudo-torsional backbone angle, OCCO, has been employed as a single descriptor of protein secondary structure. It has been reported that the secondary structure of proteins can be described adequately by the OCCO pseudo-torsional backbone angle (Laiter et al., 1995).

2.1.3 Criteria for Defining Protein Secondary Structure.

Many efforts have been spent on various simplified conformational or geometrical parameters for projecting residue conformations and for describing protein secondary structures (Doniach, 1989; Skolnick and Kolinski, 1990).

It is generally believed that hydrogen bonds between peptide amide and carbonyl groups stabilise protein structures. In addition, side chain hydrogen bonds have been proposed to play a role in the formation and stabilisation of secondary structure of soluble proteins (Bordo and Argos, 1994). Thus, hydrogen bonds are generally used as a major criterion to define protein secondary structures.

However, it is sometimes difficult to identify the existence of a possible hydrogen bond, such as in short irregular pieces of secondary structure or at the ends of longer secondary structure elements (Levitt and Greer, 1977). For example, the end points of helices in lysozyme, one of the first protein structures solved by X-ray crystallography, were reported to be located more precisely by considering the (ϕ, ψ) dihedral angles (Phillips, 1967).



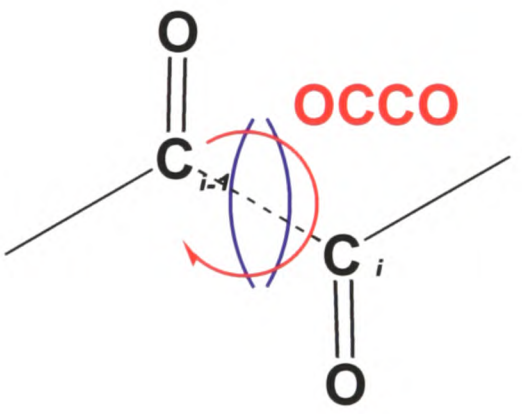


Figure 2.1 Definitions of different geometric parameters. (A) (ϕ, ψ) and ω dihedral angles, (b) φ pseudo-dihedral angles, and (C) OCCO pseudo-torsional angles. Atoms between two carbonyl groups are represented with a blue bracket, and a pseudo bond is represented with a dashed line between the two carbon atoms of C=O groups.

Other measurements, such as inter Cα torsion angles (Srinivasan et al., 1975) or inter Cα distances (Phillips, 1970; Greer, 1976) have been used for identifying secondary structures as well.

Several algorithms for defining secondary structure elements based on experimentally determined atomic coordinates have been developed. The variety of methods used to identify the secondary structure elements reflects the difficulties of assigning a secondary structure state to a residue. Different programs, based on different criteria, do not always agree in their assignments for the same protein.

2.1.4 Algorithms of Protein Secondary Structure Assignments.

The first automatic method to define protein secondary structure was developed on the basis of both hydrogen bond patterns and inter Cα distances to locate secondary structure elements (Levitt and Greer, 1977). Since then, many methods have been developed based on a variety of definitions, for example, DSSP uses the hydrogen bond pattern to assign each residue a secondary structure state (Kabsch and Sander, 1983); DEFINE utilises the inter-Cα atom distance matrices (Richards and Kundrot, 1988); P-curve identifies secondary structures by considering the helicoidal parameters of each peptide unit (Sklenar et al., 1989); STRIDE takes hydrogen bond patterns and torsional angle reliance, which is derived from dihedral angles, into consideration (Frishman and Argos, 1995). The criteria used by these automatic programs to identify protein secondary structure are summarised, (Tab. 2.1).

DSSP secondary structure assignment program

DSSP is probably the most widely used method for secondary structure assignment. HELIX records and SHEET records in the structure files of PDB database may be also generated by using this algorithm (PDBmanual, 1996).

A hydrogen bond between residue *i* and residue *j* is defined by the interaction between the main chain carbonyl and amide groups of these two residues. The overall main chain hydrogen bonds, for both neighbouring and distant residues, are examined. Each residue is assigned a hydrogen bond status; turn, bridge or neither of them.

An amino acid residue *i* will be assigned as the hydrogen bond pattern of *n*-turn at residue *i*, when a single hydrogen bond emerging between residue *i* and residue *i* + *n*; where *n* equals to 3, 4 or 5. When two or more than two consecutive *n*-turns are found, one secondary structure element of helix is located within this *n*-turn segment, (Fig. 2.2).

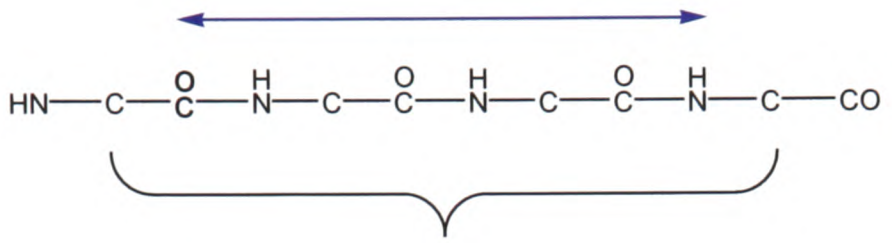
A bridge occurs between two residues of two non-overlapping stretches. A parallel bridge or an anti-parallel bridge is assigned depending on the hydrogen bond pattern. A parallel bridge between residue *i* of one stretch and residue *j* of another stretch is assigned when two hydrogen bonds are found between residues $(i-1, j)$ and residues $(j, i+1)$, or hydrogen bonds between residues $(j-1, i)$ and residues $(i, j+1)$. An anti-parallel bridge between residue *i* and residue *j* is assigned when two hydrogen bonds are found between residues (i, j) and residues (j, i), or hydrogen bonds between residues $(i-1, j+1)$ and $(j-1, i+1)$. If one or more consecutive bridge of identical type emerges, the secondary structure state of the strand is assigned to each corresponding residue (Fig. 2.2).

Methods	**Criteria**				
	H-bond Patterns	Dihedral Angles	Inter Cα Distance	Inter Cα Torsional Angle	Helicodial Parameter
Levitt *et al*			yes	yes	
DSSP	yes				
DEFINE			yes		
P-Curve					yes
STRIDE	yes	yes			

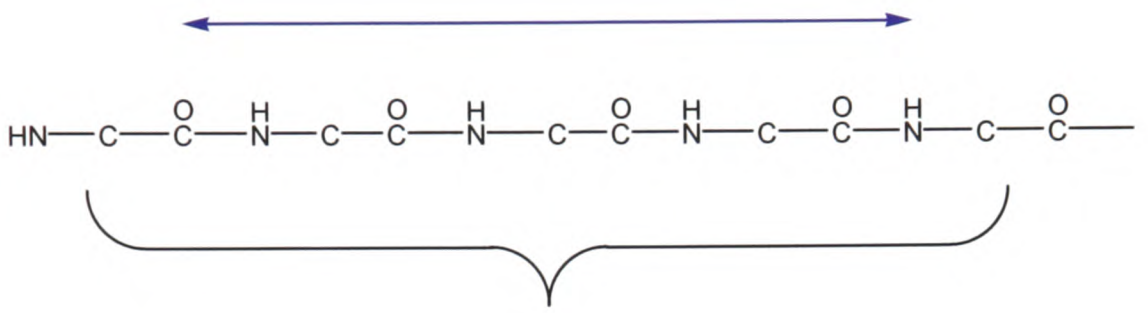
Table 2.1 Summary of programs for protein secondary structure assignments and their criteria.

DEFINE secondary structure assignment program

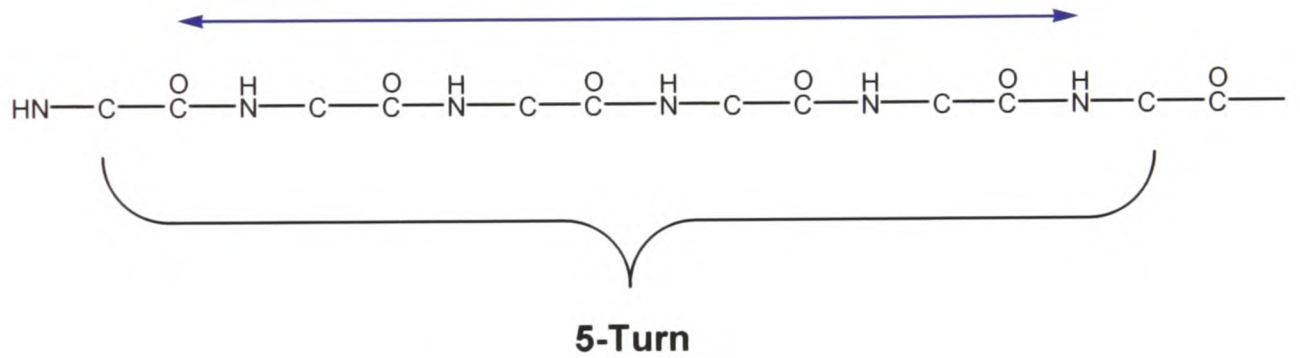
The first step used by DEFINE to locate the secondary structure elements is to search the actual protein structure against idealised Cα atom distance matrices of different secondary structure types. The differences between the distance matrices of actual protein structure and the model ones then are evaluated for agreement within preset limits. All secondary structure elements located at this point are under the assumption that the helices and strands are straight. However, in reality strands and helices can be curved and sharply bent. So these secondary structure elements require further curvature checks.



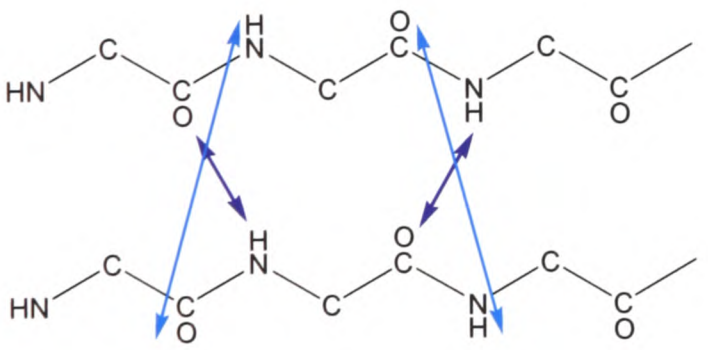
3-Turn



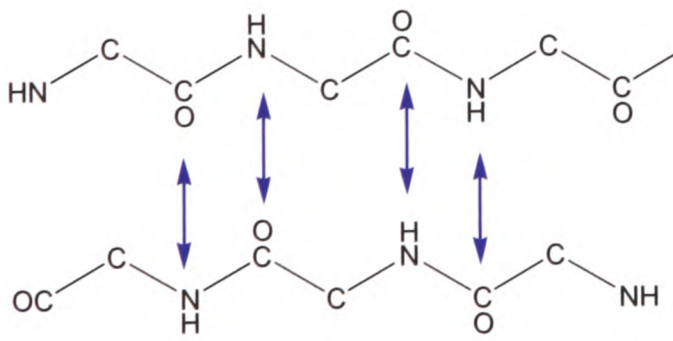
4-Turn



5-Turn



Parallel Bridge



Antiparallel Bridge

Figure 2.2 Elementary patterns of turn and bridge used in DSSP assignments. Atoms in a turn are indicated with a bracket. Hydrogen bonds are denoted by either blue or light blue lines.

The preliminary assignments of secondary structures are further processed by checking the geometric relations of the elements. The axis of each secondary structure segment is determined from spatial coordinates of Cα atoms within the segment. Curvature checks of these secondary structure elements are performed by comparing the directions and locations of axes. After these searching and checking procedures, these secondary structure elements are then well located and defined.

STRIDE secondary structure assignment program

STRIDE is a recently proposed algorithm to assign secondary structure. In order to approximate our intuitive definition of helices and strands, STRIDE takes both hydrogen bonds and statistically derived backbone torsion angle information, into consideration.

An empirical energy function is employed to calculate the hydrogen bond energy. This empirical energy function is derived from the analysis of many experimental data on hydrogen bond geometries in crystal structures of poly-peptides, peptides, amino acids and small organic compounds (Boobbyer et al., 1989; Wade et al., 1993). The probability distributions of torsional angles of helix or strand are calculated by dividing the numbers of residues in a given (ϕ, ψ) zone, defined as helix state or strand state in the PDB records, by the total number of residues with dihedral angles falling within the given zone in a data sample chosen by the authors.

When two consecutive hydrogen bonds emerge between residue pairs (*i, i + 4*) and (*i + 1, i + 5*), the amino acid residues from residue (*i + 1*) to residue (*i + 4*) will be assigned as helical state. If torsional angle probabilities of residue *i* and residue (*i*

+ 5) fall into the calculated torsional angle probability distribution of helix, these two edge residues will be also assigned a helical state.

At least two consecutive hydrogen bonded β-bridges belonging to the hydrogen bond pattern of a strand is the minimal requirement for a strand. Then the hydrogen bond strength and statistical propensity to be a strand of each amino acid residue is examined. The type of bridge, parallel or anti-parallel, is determined by weighting both hydrogen bond energy and torsional angle probabilities of related amino acids.

Comparison of different algorithms

An intensive survey and comparison of three different methods, DSSP, P-CURVE (Sklenar et al., 1989) and DEFINE was carried out by Colloc'h et al (Colloc'h et al., 1993). Colloc'h reported that there were only 63% of residues being assigned the same secondary structure state by these methods. They also noticed that residues being assigned as the same secondary structure state tended to be predicted correctly more often by the methods of Chou and Fasman (Chou and Fasman, 1974) and GOR (Garnier et al., 1978).

2.2 Methodology.

2.2.1 Protein Data Sets.

The protein structures used in this study were divided into two clusters: protein molecules and protein domains.

The set of protein molecules was composed of every protein structure in the March 1997 release of the PDB database, PDB_all. However, the protein sequence redundancy in the PDB database is of potential to bias statistical results, due to the

skewed sampling population. To cope with this problem, a non-redundant set of protein domains was used.

This non-redundant set was derived from the representative members of protein family in the 3Dee protein domain database, 3Dee_D (Siddiqui, 1997). There were 1,234 different protein fold families in the 3Dee database at the time of this study.

2.2.2 Protein Secondary Structure Assignments.

The secondary structure of each residue was assigned using the programs DSSP (Kabsch and Sander, 1983), DEFINE (Richards and Kundrot, 1988) and STRIDE (Frishman and Argos, 1995).

Detailed secondary structure information is given by each program; for example, DSSP has 9 possible secondary structure states, DEFINE has 6 possible secondary structure states and STRIDE has 8 possible secondary structure states. In order to facilitate comparisons of algorithms, assignment results were reduced to one of the three secondary structure situations; α-helix, β-strand and coil (*i.e.* everything else). Details of structure summaries for each program, the percentages of different secondary structure states in the March 1997 released PDB and how it was reduced for simplicity are illustrated, (Tab. 2.2).

Program	Structure	Description	3-state Reduction	Percentage
DSSP	H	4-helix	H	28
	G	3-helix	C	4
	I	5-helix	C	0
	E	extended strand	E	21
	B	residue in isolated β-bridge	C	1
	T	H-bonded turn	C	11
	S	bend	C	9
	_	not defined	C	21
	?	information missing	C	5
DEFINE	H	4-helix	H	33
	G	3-helix	C	1
	E	extended strand	E	29
	S	bend	C	1
	_	not defined	C	33
	?	information missing	C	3
STRIDE	H	4-helix	H	29
	G	3-helix	C	4
	I	5-helix	C	0
	E	extended strand	E	23
	B or b	residue in isolated β-bridge	C	1
	T	turn	C	22
	C	coil	C	18
	?	information missing	C	3

Table 2.2 Notations of secondary structure states derived from different programs, their definitions, and methods of reduction. H denotes helix, E denotes strands and C denotes coil. The percentage of each structure state assigned by programs in the 3Dee_D set is shown.

2.2.3 Comparison of Secondary Structure Assignment at the Amino Acid Residue Level.

The assigned structure information from the 3Dee_D for the DSSP, DEFINE and STRIDE assignments onto each residue of each protein or each protein domain chain were compared on a residue by residue basis. Since the secondary structure assignment was reduced to one of the three states, for three different programs this comparison process yielded (3 x 3 x 3) different combinations of secondary structure states. Each residue was assigned to one of the 27 possible states and recorded.

A residue was assigned in a helix, *i.e.* ``H'', by all three programs would be defined as a consensus helix residue. A consensus strand residue was defined in the same way, when it was assigned in a strand, *i.e.* ``E'', by all the three programs. These consensus assignments were then recorded for further analysis.

2.2.4 Comparison of Secondary Structure Assignment at Segment Level.

At this point, each residue had 4 assignments from DSSP, DEFINE, STRIDE and the consensus assignment. This information, of these secondary structure assignments for all residues within a protein domain chain, was stored in a linear way. When a segment of consecutive ``H'' residues appeared within the string of secondary structure assignments, this ``H'' segment would be defined as a helix. Similarly, when contiguous ``E'' notations appeared in the string of structure assignments, this ``E'' segment would be regarded as a strand. This segment-searching process located the secondary structure elements in each protein or protein domain chain defined by DSSP, DEFINE, STRIDE and consensus assignments.

Once the helix or the strand had been defined and located, the numbers and lengths of helix and strand were counted within a protein and within entire data sets.

These segment-searching process, segment-counting process and residue-counting process gave the details about distributions of residues in different secondary structures and distributions of segment lengths of different secondary structure elements.

2.2.5 Calculations of Protein Geometric Parameters.

Backbone Dihedral Angles

The backbone dihedral angles (ϕ, ψ), Cα pseudo-dihedral angles (φ), and the pseudo-torsional OCCO backbone angles were calculated from atomic coordinates in the PDB files according to their definitions, section 2.1.2.

2.3 Results and Analysis.

2.3.1 Assignment Results of Each Program.

Details of assignment results for both data sets are shown, (Tab. 2.3).

For the proteins in the March 1997 released PDB database, DSSP and STRIDE gave similar percentages of residues in regular secondary structures; at the ranges of 28% to 29% for helices and 21% to 23% for strands. Around 50% of residues were assigned as coils. However, DEFINE assigned more residues into regular secondary structures; 33% for helices and 29% for strands. No more than 40% of residues in the whole PDB database were assigned as coil by DEFINE, (Tab. 2.4).

For the non-redundant set of protein domains, DSSP and STRIDE also gave similar percentages of residues in regular secondary structures; at the ranges of

29% to 30% for helices and 19% to 20% for strands. Around 50% of residues were assigned as coils. As expected, DEFINE still assigned more residues into helices, 35%, and strands, 27%.

It was found that more residues were assigned as helices or strands by DEFINE than DSSP and STRIDE, (Tab. 2.4). This finding indicated that DEFINE tends to assign more residues into regular secondary structures than other programs, which could be an artifact of not taking hydrogen bonds into consideration.

It was also noted that no matter which program was used, the redundant PDB_all set have less resides in helices and more residues in strands, in comparison with the non-redundant 3Dee_D set. The differences of assignment results were less than 2%. However, the over expressions of strand status by different programs in the PDB_all set reflects that the β-strand conformations could be more abundant than the α-helical conformations in the PDB database. This is similar to the finding of a previous study, which indicated that all β-folds were the most common in protein domains (Siddiqui, 1997).

This biased sampling population of redundancy had been erased in the non-redundant set of 3Dee_D, which resulted in this set having a higher percentage of residues assigned as helices and lower percentage of residues assigned as strands.

Data Set		PDB_all		3Dee_D	
Method	**State**	**Number**	**Percentage (%)**	**Number**	**Percentage (%)**
DSSP	helix	509549	28	54612	29
	strand	388028	21	36251	19
	coil	915358	51	97130	52
DEFINE	helix	600134	33	65433	35
	strand	517517	29	51237	27
	coil	695284	38	71323	38
STRIDE	helix	532519	29	57082	30
	strand	414730	23	38473	21
	coil	865686	48	92438	49

Table 2.3 Numbers and percentages of residues assigned as helix, strand and coil in the two protein sets by DSSP, DEFINE and STRIDE.

2.3.2 Comparison of All Three Programs.

Detailed comparisons of residue numbers and percentages for the 27 possible states of secondary structure combinations were shown at Table 2.4. Each combination consisted of three individual secondary structure state assigned by three programs in the order of DSSP, DEFINE and STRIDE.

There were 72.4% of residues in the whole PDB database assigned to the same secondary structure states by all these programs, which consisted of 26.1% for the combination of helix-helix-helix, 15.9% for the combination of strand-strand-strand and 30.4% for the combination of coil-coil-coil in the PDB_all set. Similar figures were found in the non-redundant set of 3Dee_D, which had 71.7% residues

assigned into the same secondary structure conformations totally. An average of more than 70% of residues were categorised into these three combinations.

There is no surprise that the major discrepancy of assignments for both sets came from the combinations of Coil-Strand-Coil and Coil-Helix-Coil, due to the overestimates of helices and strands by DEFINE. In addition, the combination of Strand-Coil-Strand was also the main source of disagreements on assignments. This reflects that some residues assigned as strands by DSSP and STRIDE could still be assigned as coil by DEFINE. The mis-interpretations of strands as coils are likely a result of β-strands being more clearly located by their hydrogen bond patterns rather than by their conformations or inter Cα atom distances.

It was noted that percentages of residues in Helix-Helix-Helix combinations were 26.1% for PDB_all set and 26.9% for 3Dee_D set; and the percentages of residues in Strand-Strand-Strand combinations were 15.9% for PDB_all set and 14.4% for 3Dee_D set. Residues in these two combinations may be regarded as core residues of secondary structure assignment.

There was no significant discrepancy of the overall distributions of residues for helix, strand and coil defined by different programs in both data sets studied here. Other combinations of disagreed assignments are comparatively insignificant, at the range of 1% or less. Approximately, 30% of residues were assigned as helix, 25% of residues were assigned as strand and 45% of residues were assigned as coil. This might be a clue that each secondary structure element in nature has a limited distribution range.

Data Set	PDB_all		3Dee_D	
State DSSP, DEFINE, STRIDE	**Number**	**Percentage (%)**	**Number**	**Percentage (%)**
HHH	473878	26.1	50502	26.9
HHE	0	0	0	0
HHC	9012	0.5	695	0.4
HEH	592	0.0	72	0.0
HEE	0	0	0	0
HEC	332	0.0	23	0.0
HCH	21399	1.2	2794	1.5
HCE	0	0	0	0
HCC	4336	0.2	526	0.3
EHH	0	0	0	0
EHE	1289	0.1	124	0.0
EHC	134	0.0	7	0.0
EEH	0	0	0	0
EEE	288496	15.9	27097	14.4
EEC	3745	0.2	279	0.2
ECH	0	0	0	0
ECE	91983	5.1	8546	4.6
ECC	2381	0.1	198	0.1
CHH	28395	1.6	2852	1.5
CHE	293	0.0	18	0.0
CHC	87133	4.8	11235	6.0
CEH	444	0.0	55	0.0
CEE	16047	0.9	1315	0.7
CEC	207861	11.5	22396	11.9
CCH	7811	0.4	807	0.4
CCE	16622	0.9	1373	0.7
CCC	550752	30.4	57079	30.4

Table 2.4 Numbers and percentages of residues assigned into different combinations of secondary structure by three programs. The assignments notations are in the order of DSSP, DEFINE and STRIDE.

2.3.3 The Relation between Protein Structure Resolution and Protein Secondary Structure Assignment.

To study possible effects of protein structure resolution on the agreement of secondary structure assignments, all proteins in the PDB database were clustered into different subset by their structural resolution. There were 14 subsets in this PDB_all set.

As expected, there were more protein structure entries at the lower structural resolution ranges. In the PDB_all set, the range of structural resolution for the PDB_all sets was from 1.8 Å to 3.0 Å. It is detected the entries fell into subsets that increased as the structural resolution decreased, (Fig. 2.3 A).

A general pattern of higher percentages of agreements on secondary structure assignments with higher structure resolution was found for the assignments in total for helix, strand and coil, (Fig. 2.3 B). Examining this result in detail on three structural states revealed interesting patterns. The percentages of residues assigned as a helix by all three methods decreased from 32.2% to 27.4% while structural resolutions decreased, which revealed a trend of better assignment agreements at higher resolution, (Fig. 2.3 C). However, the assignments of strand and coil showed an ambiguous pattern. For the strand assignments, less residues were assigned as a strand with resolution range of 1.8 Å to 2.2 Å, and 16% of residues were assigned as a strand approximately by all three methods at resolutions lower that 2.2 Å, (Fig. 2.3D). It was also found that more residues were assigned as a coil at lower resolutions than higher resolutions, (Fig. 2.3 E).

Only the residues of helices showed the tendency that increased structural resolution resulted in higher percentages of residues to be assigned as the same

secondary structure states by different programs. This trend was not observed for either strands or coils. Even though the overall agreement of secondary structure assignments was detected to rise with the higher resolution; however, it was largely contributed from the higher agreement in assignment of helix. A question has arisen, if the trends of more residues assigned as a helix by all three methods at higher resolutions could be explained as a result of better structures, less ambiguities in structure assignments; or just simply the nature of the PDB that higher percentages of helical structures exist in the protein structures solved at high resolutions.

To answer this question, distributions of regular secondary structures were examined on the basis of individual program to examine each protein structure in the PDB. It was found that the percentages of residues in the helical conformations were higher in the protein structures at higher resolution and gradually decreased with the structural resolutions decreasing no matter which assignments were used, (Fig. 2.4 A, 2.4 C and 2.4 E). In contrast, the percentages of residues in the strand conformations were lower in the protein structures of higher structural resolutions and gradually increased with the structural resolutions decreasing no matter which assignments were used, (Fig. 2.4 B, 2.4 D and 2.4 F). It was noted that patterns of percentages of residues in helices or strands were similar for DSSP, DEFINE and STRIDE.

This finding reflected the nature of the whole PDB data set. It also explained why the percentages of residue being assigned in strand or coil decreased while the resolution improved. Comparing the pattern for percentages of residues assigned as helices by all three programs, (Fig. 2.3 C), with the patterns for residues

assigned as helices by DSSP (Fig. 2.4 A), DEFINE (Fig. 2.4 C) and STRIDE (Fig. 2.4 E), they appeared to be very similar. In the same way, high degree of similarities was observed between patterns for percentages of residues assigned as strands by all three programs, (Fig. 2.3 D), and patterns for residues assigned as strand by individual programs, (Fig. 2,4 B, D, and F).

It has been reported that protein structure with better resolution has better topology and stereochemistry (Morris et al., 1992). All these three programs locate secondary structure elements and identify secondary structure based on the experiment-determined atomic coordinates of a protein molecule. In this study, a trend of decreasing percentages of resides of helices by all three programs with decreasing structural resolutions is identified. However, the finding of higher percentage of agreements on secondary structure assignments with higher structure resolution does not apply for strands and coils. In addition, it is impossible to differentiate if this trend of higher percentage of agreements on helical structure assignments with higher structure resolution is a result of higher percentages of residues in helical conformations in protein structures at higher resolution in the PDB. Thus, whether protein structures with better resolution will result in better agreements on secondary structure assignments by different programs remains inconclusive.

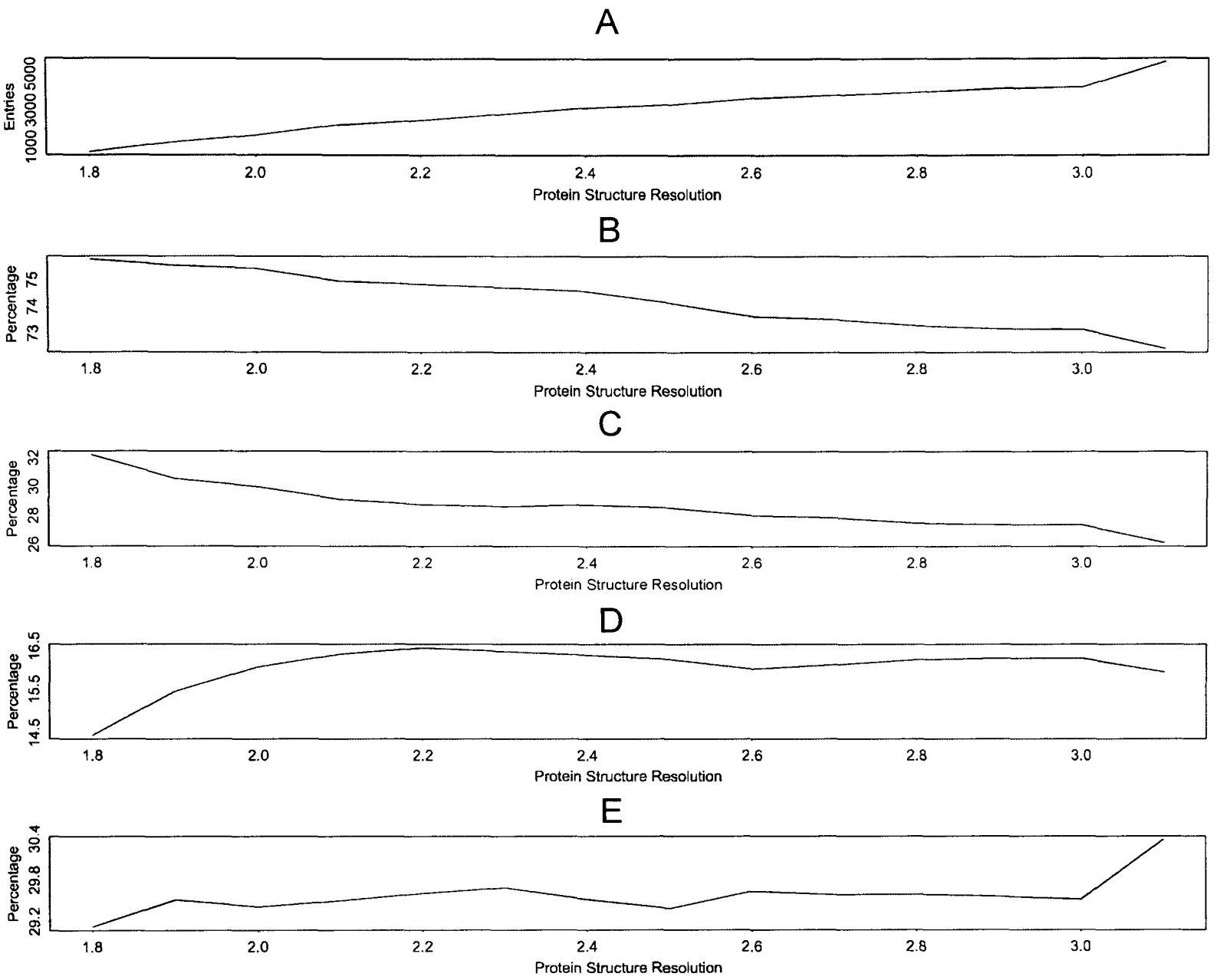


Figure 2.3 Relationships between protein structure resolutions (Å) and agreements of secondary structure assignments in the PDB_all set. (A) numbers of protein *vs.* resolution, (B) agreement on helix, strand and coil status rises with better resolution, (C) percentage of residues with consensus assignments of helix rises with better resolution, (D) percentage of residues with consensus assignments of strand decreases with better resolutions, (E) percentage of residues with consensus assignments of coil decreases with better resolutions.

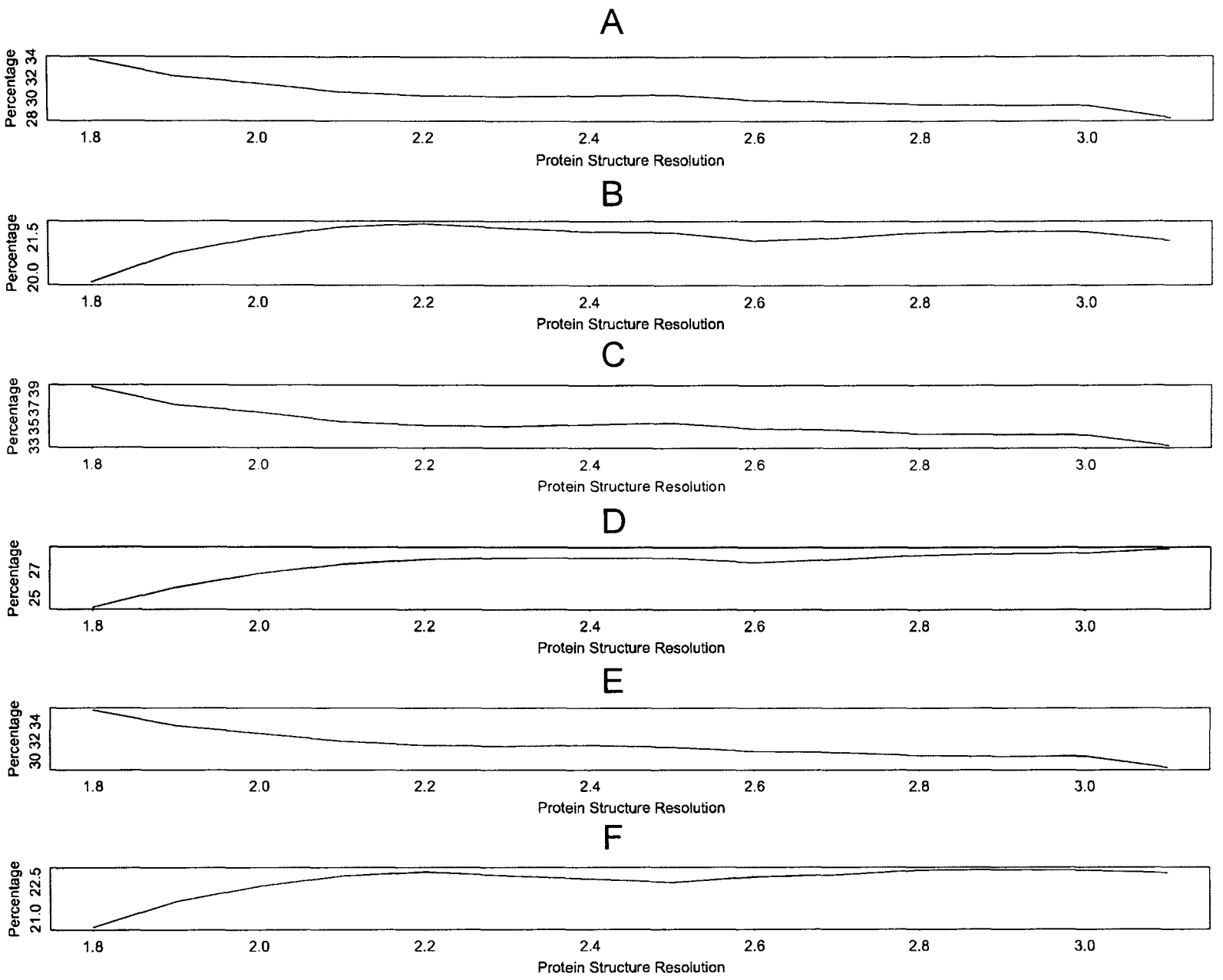


Figure 2.4 Percentages of residues in regular secondary structures by different assignments at different structure resolutions (Å) in the PDB_all set. (A) helix state by DSSP, (B). strand state by DSSP, (C) helix state by DEFINE, (D) strand state by DEFINE, (E) helix state by STRIDE, (F) strand state by STRIDE.

2.3.4 Lengths of Helix and Strand.

Details of the average lengths of the helical, (Tab. 2.5), and strand, (Tab. 2.6), segments are shown. No significant differences were observed in the average lengths of either helices or strands derived from two different sets. In general, a helical segment was estimated to be 11 residues long and a strand was estimated to be 6 residues long. This result is similar to results from other studies. Previous studies reveal that an α-helix is on average 10 or 11 residues long (Srinivasan, 1976); and a β-strand is on average 6.5 residues long (Sternberg and Thornton, 1977). The new results from this study confirm the previous finding about the average length of secondary structure element of helix. However, the average length of the secondary structure element of strand is shorter than the previous result.

It is noted that DEFINE always gives the longest average length of helix in different data sets, (Tab. 2.5). In addition, it is found in the case of strand, DEFINE still assigns longer strand than other programs, (Tab. 2.6).

As mentioned earlier, section 2.3.1, DEFINE can tolerate more structural torsion than by DSSP or STRIDE. It is of no wonder that DEFINE does not only assign more residues into helix or strand, but also assigns longer helices or strands than the other two programs. It was also noted that DSSP always gives the shortest average length of helix and the shortest average length of strand in both whole protein molecules or protein domains, (Tab. 2.5 and 2.6),.

Both DSSP and STRIDE mainly employ the hydrogen bonds to locate secondary structure elements. The major differences between them are STRIDE applies an

empirical energy function to calculate the hydrogen bond energy between two residues, and identifies the terminal residues of secondary structure element through reliance on their torsional angles. Taking torsional angle information, which is complementary to hydrogen bonding information, into consideration may not be sufficient to determine accurately the termini of helices and strands. The procedures prevent premature termination of helical or strand segments. These procedures also avoid long helices with missing hydrogen bonds in the middle can be split into two separate helices in spite of the completely acceptable overall geometry. These may explain why the average length of helix in DSSP assignments is shorter than that of STRIDE assignments.

Data Set	PDB_all		3Dee_D	
Method	**mean**	**median**	**mean**	**median**
DSSP	10.88	10	10.97	10
DEFINE	13.55	12	13.11	12
STRIDE	12.01	11	12.22	11

Table 2.5 Average numbers of residues per helix segment by different programs.

Data Set	PDB_all		3Dee_D	
Method	**mean**	**median**	**mean**	**median**
DSSP	5.24	5	5.23	5
DEFINE	6.43	6	6.24	5
STRIDE	5.41	5	5.37	5

Table 2.6 Average numbers of residues per strand segment by different programs.

2.3.5 Comparison of Length Distributions in Different Secondary Structure

The distribution of the number of residues per helix in the PDB_all data set is shown as a zigzag-shaped curve whatever the program used, (Fig. 2.5). DSSP assignment has maxima at 4, 8 and 11 residues. DEFINE assignment has peaks at 4, 6, 10 and 13 residues. STRIDE assignment has peaks at 6, 9 and 11 residues. The large width of the distribution indicates that no length is preferred.

It seems that there is no correlation between the jagged shapes and the number of complete turns. However, DSSP assignment and STRIDE assignment all had peaks at 11 residues corresponding to 3 full turns. It has been reported that there is no real significance within these zigzag curves (Schultz and Schirmer, 1979). A sharp peak of 4 residues per helix was noted in the DSSP assignment. This has been observed by the authors of this method (Kabsch and Sander, 1983). Actually, these 4-residue helical segments are the shortest helices by DSSP assignments. This is likely caused by using hydrogen bonding patterns of turn for defining helices.

The set of PDB_all had a pointed curve in the distribution of number of residues per strand, (Fig. 2.5). The majority of numbers of residues in a strand by DSSP assignment was ranging from 2 to 5 residues. A maximum peak appeared at the position of two residues for the DSSP strand assignments. The result from STRIDE strand assignment has a large range in length from 2 residues to 7 residues. There was a maximum peak at 2 residues for the STRIDE assignment of strand, which is similar to results from DSSP assignments. The length of a strand by DEFINE assignment was ranging from 4 residues to 7 residues in most cases. However, there was a distinctively high peak at the length of 4 residues, which

could be an artifact inherent in DEFINE algorithm. Since DEFINE locates secondary structure elements without considering hydrogen bonds, this could result in that DEFINE is more sensitive to detect the short non-hydrogen bonded β-strand occurring in loops.

The distinct peak of DSSP helix assignment at 4 residues and the distinct peak of DEFINE strand assignment at 4 residues are neutralised in the assignments of consensus secondary structure elements, (Fig. 2.6). This provides a hint that protein structures should be analysed by at least two or three methods for locating secondary structure elements in order to gain different insights on the protein structures and their secondary structure element arrangements.

A previous study on a non-redundant set of protein structures shows a similar pattern of length distributions of helices and strands (Colloc'h et al., 1993). Similar length distributions of regular secondary structure elements from different protein sets, regardless of their underlying different characteristics, reveal the possible limited distributions on secondary structure arrangements in the protein universe.

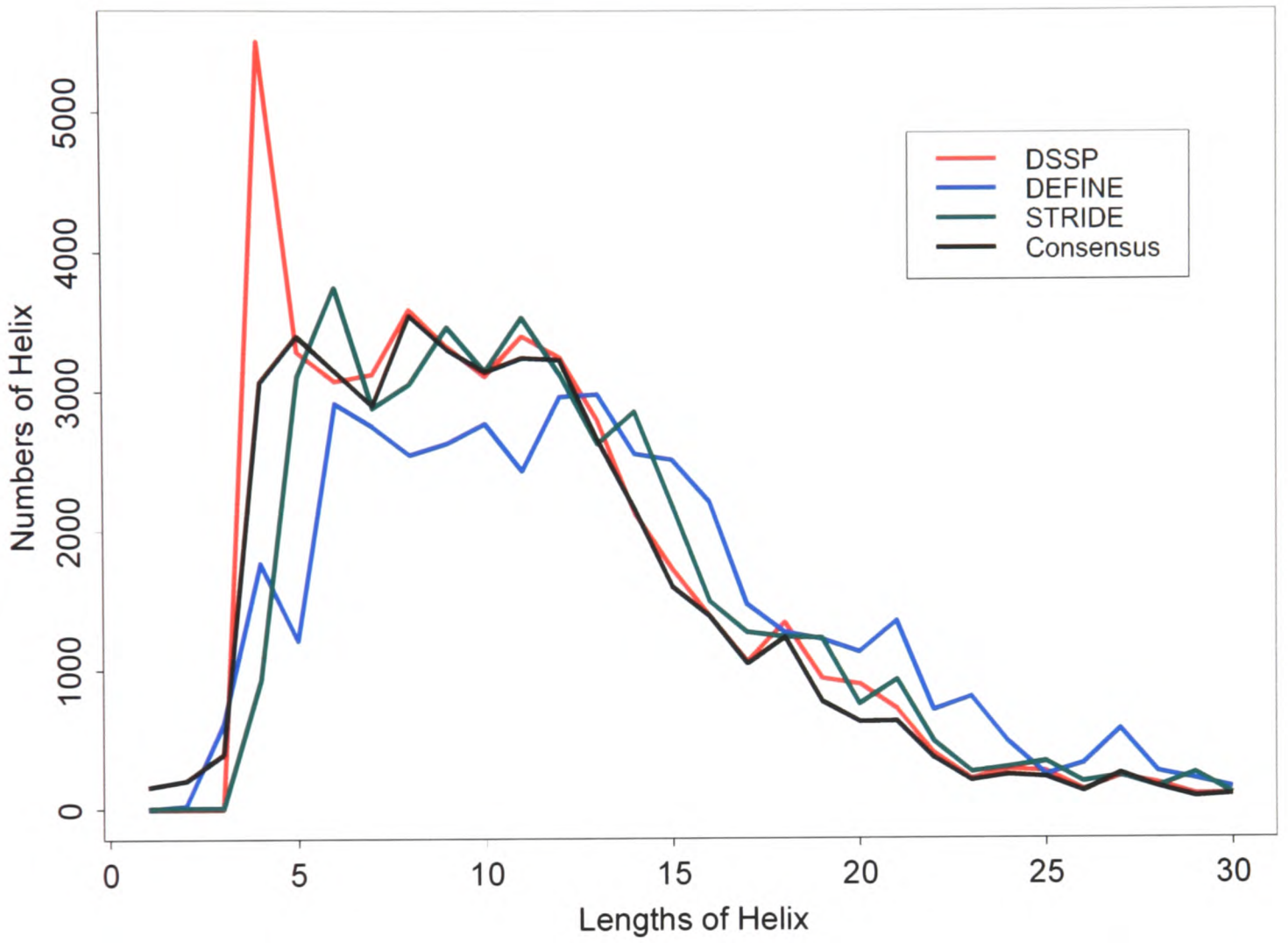


Figure 2.5 Length distributions of helix segments in the PDB_all set. Helices were assigned using DSSP, DEFINE, STRIDE and the consensus assignments.

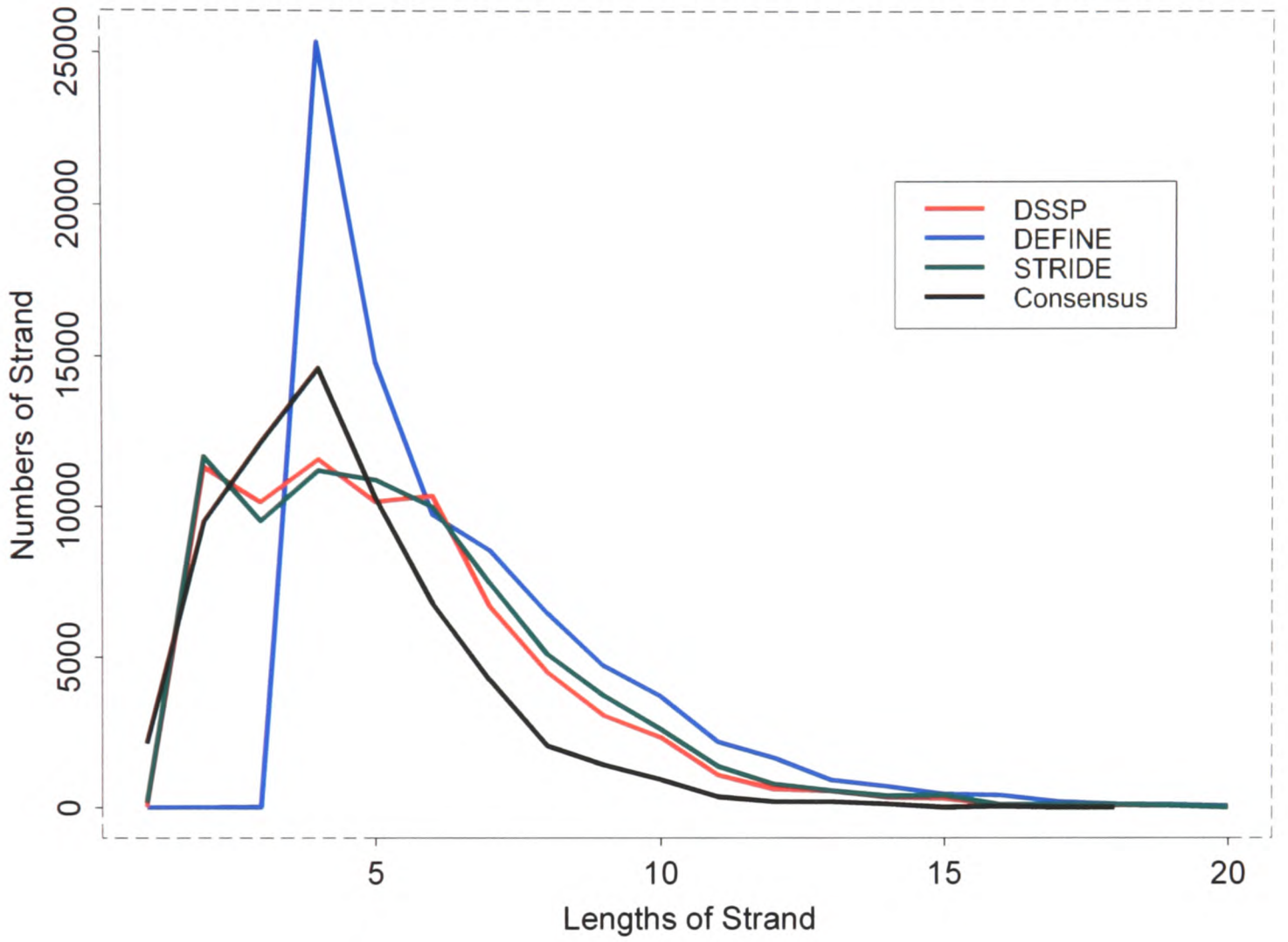


Figure 2.6 Length distributions of strand segments in the PDB_all set. Helices were assigned using DSSP, DEFINE, STRIDE and the consensus assignments.

2.3.6 Applications of Structural Descriptor.

To examine relationships between protein geometry and secondary structure, geometric parameters, backbone dihedral angles (ϕ, ψ), Cα pseudo-dihedral angles (φ), and the pseudo-torsional OCCO backbone angles were calculated for residues in the 3Dee_D set. Three categories of residues were compared, all resides in this set (187,993 entries), residues with consensus assignments of helix (50,502 entries), and residues with consensus assignments of strand (27,097 entries).

The (ϕ, ψ) dihedral angles have long been used as a main criterion to define protein secondary structures (Ramachandran et al., 1963). The majority of (ϕ, ψ) angles for all residues fall into the regions of less steric collisions between side chains and main chains, (Fig. 2.7 A); even though there are still outliers in the disallowed regions. It can be more easily observed from a three-dimensional view that most residues are within the 3 major sterically allowed regions, (Fig. 2.8 A). The (ϕ, ψ) dihedral angle distributions for residues with consensus assignments of helix show an unified pattern, (Fig. 2.7 B) and are largely contained at the ranges of $-107^{o} < \phi < -32^{o}$ and $-70^{o} < \psi < 3^{o}$, (Fig. 2.8 B). For the residues with consensus assignments of strand, one major and one minor region are observed for their (ϕ, ψ) distributions, (Fig, 2.7 C); however, the majority of residues are within the regions of $-180^{o} < \phi < -50^{o}$ and $85^{o} < \psi < 180^{o}$, (Fig. 2.8 C).

The distribution of the Cα pseudo-dihedral angles (φ) for all residues in the 3Dee_D set appears continuous with one major peak and two minor peaks observed, (Fig. 2.9 A). However, the major peak, $25^{o} < \varphi < 75^{o}$, corresponds to the residues with consensus assignments of helix, (Fig. 2.9 B). The φ angle

distributions of residues with consensus assignments of strand fall into two separated parts; the regions of $-180^{o} < \varphi < -78^{o}$, and the regions of $140^{o} < \varphi < 180^{o}$, (Fig. 2.9 C). The obvious preferences of φ distributions by helices and strands reveal a possibility of projecting protein secondary structures by the $C\alpha$ pseudo-dihedral angles.

A non-regular distribution is identified for the pseudo-torsional OCCO backbone angles of all residues in the 3Dee_D set, (Fig. 2.10 A). However, the residues with consensus assignments of helix, a distinct peak is observed for the OCCO angle distributions with a range of $-67^{o} < \mathsf{OCCO} < -5^{o}$, (Fig. 2.10 B). For the residues with consensus assignments of strand, the OCCO backbone angles distributions show a broad peak at the region of $-180^{o} < \mathsf{OCCO} < -60^{o}$, (Fig. 2.10 C). As mentioned in Section 2.1.2, the OCCO angle can be regarded as a function of the (ϕ, ψ) dihedral angle. Unlike the Ramachandran plot with certain strically allowed regions for ϕ and ψ angles, no particular angles are not allowed in the OCCO angle distributions. However, the OCCO angles are still able to describe protein secondary structures adequately.

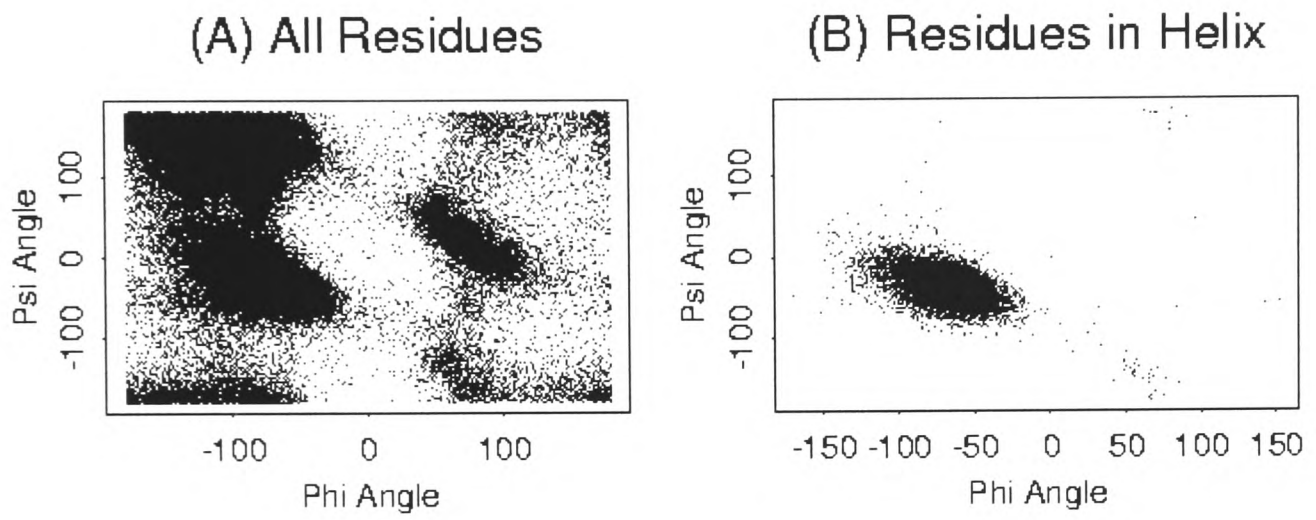
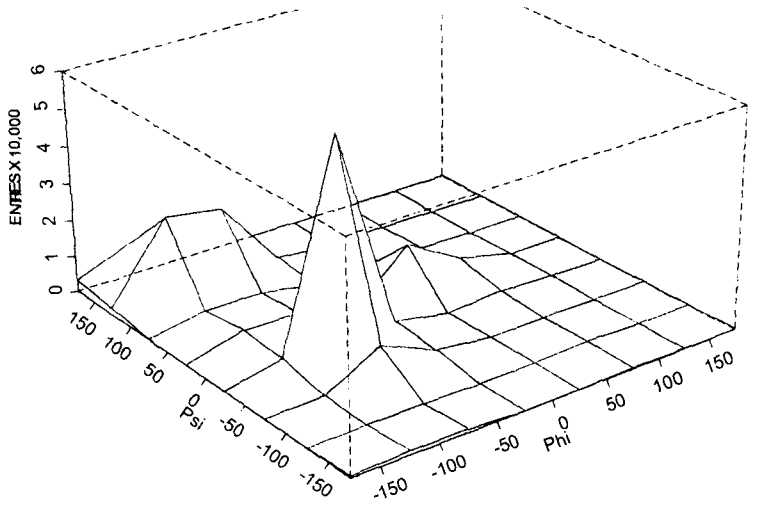
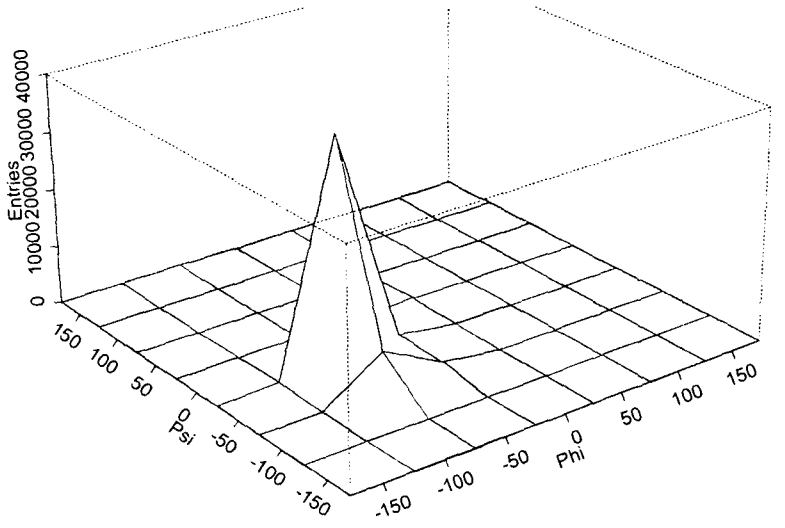


Figure 2.7 (ϕ, ψ) distributions of (A) all residue, (B) residues with consensus assignments of helix, and (C) residues with consensus assignments of strand in the 3Dee_D data set.

A



B



C

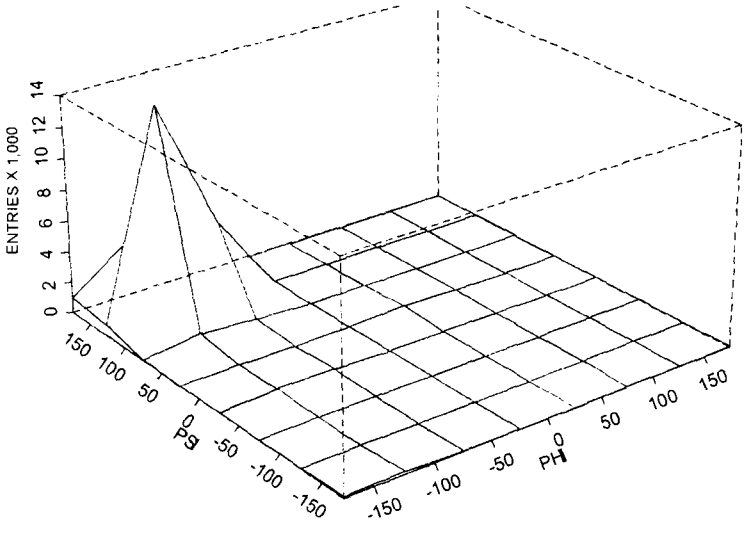


Figure 2.8 (ϕ, ψ) distributions in three-dimensional view of (A) all residues, (B) residues with consensus assignments of helix, and (C) residues with consensus assignments of strand in the 3Dee_D data set.

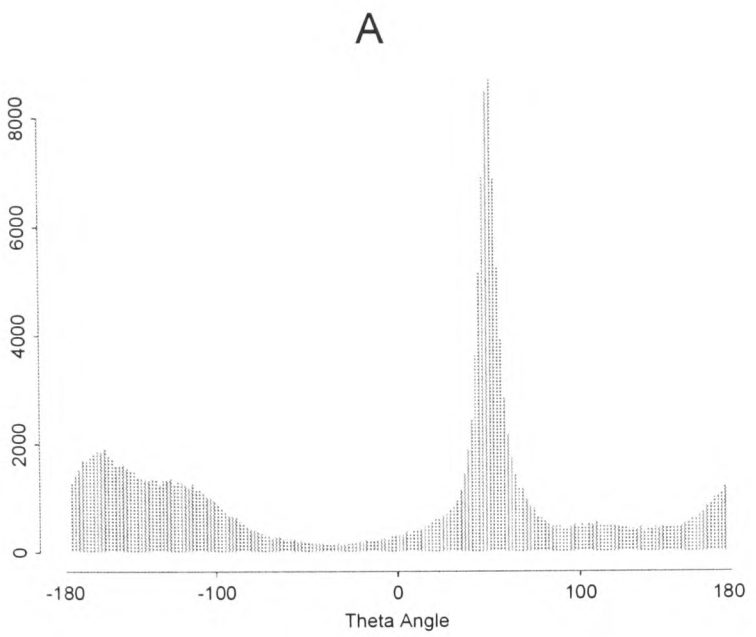


Figure 2.9 φ angle distributions (A) all residues, (B) residues with consensus assignments of helix, and (C) residues with consensus assignments of strand in the 3Dee_D data set.

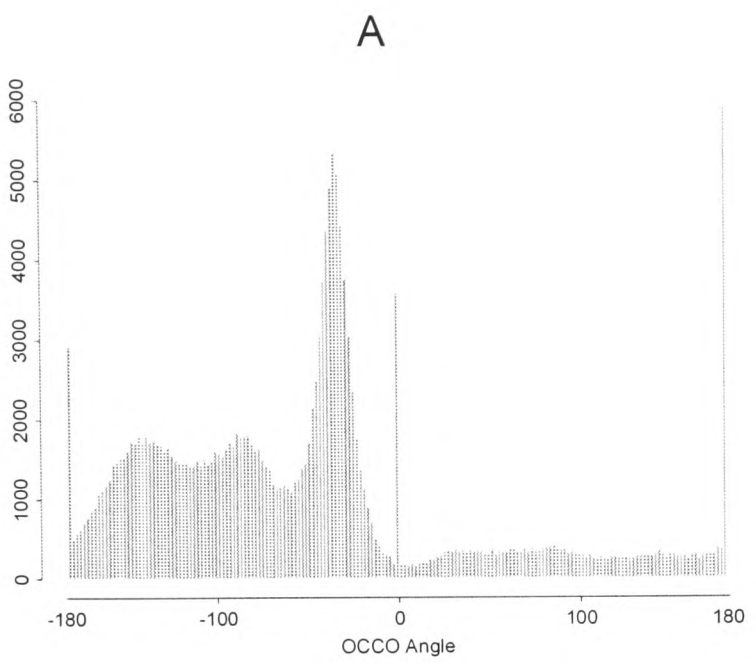


Figure 2.10 OCCO angle distributions (A) all residues, (B) residues with consensus assignments of helix, and (C) residues with consensus assignments of strand in the 3Dee_D data set.

2.4 Discussion.

Assignments of protein secondary structures do not only include examinations of interaction patterns within residues near by, but also involve identification of conformational propensity in geometry of each local amino acid residue. Many efforts have been focussed on this issue based on the hydrogen bond patterns or geometric conformations (Levitt and Greer, 1977; Kabsch and Sander, 1983; Richards and Kundrot, 1988; Frishman and Argos, 1995). However, it has been noted that assignment results from different algorithms do not always agree well, due to different criteria employed (Colloc'h et al., 1993).

The most obvious interactions within local amino acid residues are the hydrogen bonds between pairs of residues. The first step of locating secondary structure elements by the hydrogen bond pattern is to find proper criteria for acceptable hydrogen bonds. How to define the acceptable geometry for hydrogen bonds and to set the precise number of hydrogen bonds in order to specify a secondary structure element are the two major variables to be concerned. In experiments, clear lists of hydrogen bonds require atomic coordinates to be well resolved and refined.

In addition, the conformation of a residue may vary considerably within a secondary structure element, which is not perfectly regular; for example, the residues within a β-strand show a greater variability of (ϕ, ψ) torsional angles than residues within an α-helix. Furthermore, ambiguous data from either X-ray crystallography or NMR spectrum, and the potential experimental errors increase the complexity of correlating the peptide conformations with the protein secondary structures.

Locating the secondary structure elements of helix or strand spanning the whole protein molecule is a real task. The first problem to meet is how to handle experimental data and possible experimental errors properly in order to identify every secondary structure segment accurately. The following problem is that how to clearly define the edges of each fragment, which has been described as ``trivial'' but ``difficult'' (Richardson and Richardson, 1989).

However, assignment of protein secondary structure means no more but to identify and recognise the patterns of consecutive amino acid residues with their propensities for one of the secondary structure states, based on a variety of criteria used. It is no more than a pattern recognition problem, which can be modified and implemented into automatic programs. Even though, agreements between different programs of secondary structure assignments are low, it does not mean one of the algorithms is necessarily better than another. It reflects the difficulties of locating secondary structure elements objectively. For example, DSSP will be useful for researchers interested in hydrogen bonding, and DEFINE will be useful for locating the supersecondary structures.

A comparison study of three different methods has been conducted in this study. The purpose of this study is not to find out which algorithm outperforms others, but to indicate that people should choose different algorithms for different purposes. Different assignments from different programs could be an artifact inherent in each assignment algorithm. For example, assignment results from DSSP and STRIDE share more similarities than results from DEFINE, since both DSSP and STRIDE take the hydrogen bond information into consideration but DEFINE does not. These assignment results compared in this study include

percentages of residues in each secondary structure states, lengths of segments of helices and strands, and the length distributions.

Programs for defining protein secondary structures use structural information stored in the Protein Data Bank (Bernstein et al., 1977). In theory, structures with better resolution will lead to less ambiguous assignments of secondary structures. However, the overall survey of protein structures in the PDB reveals a result of contradiction. The idea of higher resolution higher agreements on structure assignments only applies for helices, but not for strands and coils. In addition, the possibility that the higher percentages of residues with consensus assignments of helix in protein structures with higher resolutions could be merely a reflection of more helical structures existing in the protein structures at higher resolutions in the PDB cannot be ruled out.

Some other structure descriptors are with potential to project secondary structures. The possibilities of using Cα atoms only to locate protein secondary structure elements have been discussed (Labesse et al., 1997). It has been reported that the OCCO angle is not only useful for projecting protein secondary structure conformations, but the OCCO profiles can be useful for rapid identification of protein fold families (Laiter et al., 1995). By taking information of both the hydrogen bond patterns and geometric conformations into consideration, the precise delineations of protein secondary structure elements could be better achieved (Frishman and Argos, 1995).

Chapter 3

Disorder in Protein Structure and Function:

The *Cis* Peptide Bond

3 Preface to Chapter 3

A research interest for studying irregular protein structures arose due to the ambiguities in defining protein secondary structure elements identified in Chapter 2. Thus, a study was designed for looking into disorder in protein structure and function, with an emphasis on peptide bonds in a *cis* conformation.

This chapter is divided into 4 sections. Section 3.1 starts with a discussion of disorder in protein structure and function, followed by a general introduction to the chemistry and biology of *cis* peptide bonds. Section 3.2 provides an overall survey of occurrences, types and distributions of *cis* peptide units in protein structures in the Protein Data Bank. Section 3.3 discusses the distributions of *cis* peptide units in the SCOP protein domain database and then looks into some protein families which contain *cis* peptide bonds. Section 3.4 covers potential implications of the *cis* peptide unit in protein structures, and their possible roles in regulating protein structures and functions.

3.1 Introduction.

3.1.1 Disorders in Protein Structure and Function.

It is generally believed that a protein folds into a particular three-dimensional structure, which is determined by its amino acid sequence (Anfinsen, 1973), in order to perform its biological function (Richardson and Richardson, 1989). However, some unfolded or disordered protein regions have been suggested to play important roles in several biological functions. For example, the disordered

region in bovine trypsinogen has been identified to facilitate biological control by providing a protease digestion site (Bode et al., 1978).

Domain movement has been proposed to be one of the mechanisms to regulate inter-molecular interactions of proteins, such as substrate uptake and protein-protein contacts. And it has been proven that less rigidly folded, flexible or unfolded regions in proteins are essential for domain movements (Huber and Bennett Jr, 1983). Disordered regions are also important for protein traffics. For the topoisomerase I, the unfolded and disordered region of the N-terminus is essential for nuclear import and targeting to the transcriptionally active loci. For topoisomerase II, the random-coiled-like region of the C-terminus is necessary for its nuclear localization (Shaiu et al., 1999).

These findings reveal that there are amino acid sequences that do not fold completely into specific three-dimensional structures or do not have rigid folds in order to be functional. The disordered structure may occur in a large region; for example, the disordered region of *Drosophila* topoisomerase I is composed of up to 430 amino acid residues at the N-terminus of this protein. However, disordered structures may also occur in local regions, which exhibit dynamic equilibria between different conformations; for example, the *cis/trans* isomerisation of a peptide bond. In general, *cis* peptide bonds are not favored in protein structure determinations because of their thermodynamic instability compared to *trans* peptide bonds, see section 3.1.2.

It is thus of functional interest to explore the structural significance of the *cis* peptide unit and the potential biological roles of *cis/trans* isomerisation of peptide

bonds. This study of disordered protein structures mainly focuses on *cis* peptide units and *cis/trans* isomerisation of the peptide bond.

3.1.2 *Cis* Peptide Bonds.

The chemistry of *cis/trans* isomerisation

The peptide bond occurs predominantly in the *trans* conformation in polypeptides and proteins (Ramachandran and Sasisekharan, 1968). The preponderance of the *trans* peptide bond can be explained in terms of fewer steric repulsion between non-bonded atoms in comparison with potential steric repulsion in the *cis* conformation, (Fig. 3.1). Theoretical calculations of the conformational energy of tripeptide units suggest an expected frequency of 0.1% *cis* configurations for an Ala-Ala peptide bond, which corresponds to an enthalpy difference of 4.0 kcal/mol (Ramachandran and Mitra, 1976).

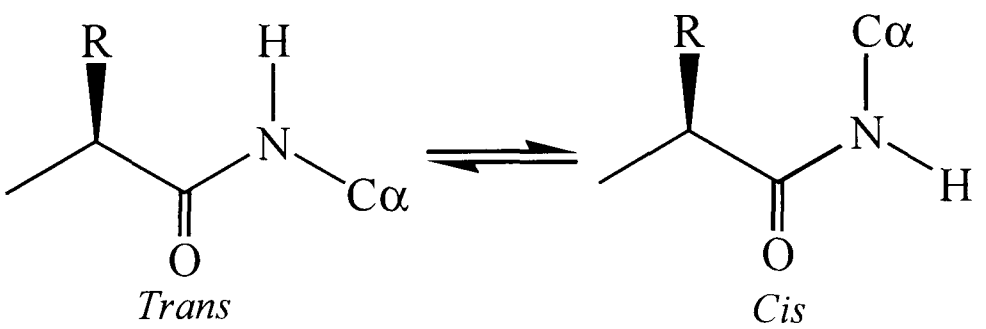


Figure 3.1 *Cis/trans* isomerisation of a peptide bond. R denotes the side chain of the N-terminal residues in *cis* peptide unit.

Thus, the occasional occurrence of a peptide bond in the *cis* conformation was noted as an unusual conformation or sometimes even as a disordered structure. However, a previous study revealed that the *cis* peptide bond is rare but not uncommon in protein structures, and the proline residue is more frequently found in the *cis* conformation (Stewart et al., 1990). The relatively higher intrinsic tendency of proline to form the *cis* conformation can probably be explained by its unique cyclic structure which reduces the repulsion between atoms, (Fig. 3.2).

A theoretical calculation of the conformational energy of tripeptide units suggests an expected frequency of 30% *cis* configurations for an Ala-Pro peptide bond. This number corresponds to the energy barrier of 0.5 kcal/mol (Ramachandran and Mitra, 1976).

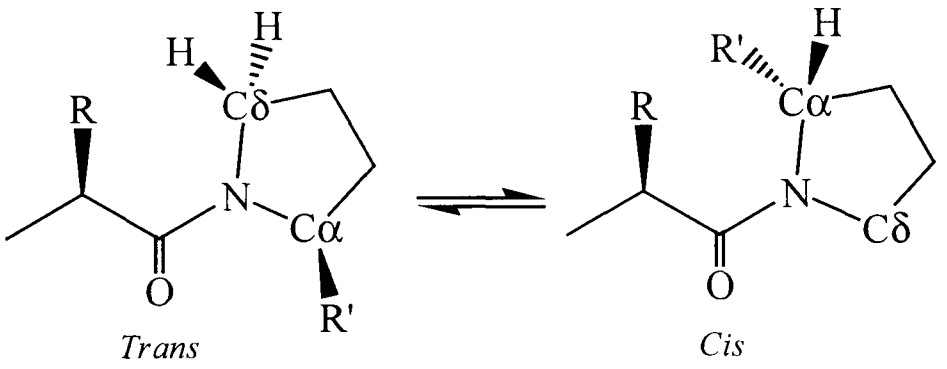


Figure 3.2 *Cis/trans* isomerisation of a peptidyl prolyl bond. R denotes the side chain of the N-terminal residue in a *cis* imide unit, and R' denotes the residues following the proline residue.

Experimental studies of side-chain effects on peptidyl prolyl *cis/trans* isomerisation of short peptides suggest that the conformation of a peptidyl prolyl bond is mainly determined by local effects in the protein structure, (Reimer et al., 1998). Thus, it can be generally speculated that *cis* peptide bonds can exist in protein structures under exceptional situations; such as protein molecule contacts during the protein crystallization, special folding demands in nature for a specific protein structure and unusual structural demands in order to carry out the biological function.

However, the energy differences between the *cis* and *trans* conformations of imide bonds were estimated to be ca. 20 kcal/mol in different protein structures (Stein, 1993). In addition, the rate constant of the isomerisation between two configurations has been estimated to be ca. 10^{-3} S^{-1} at 4°C for the peptides of Ace-Ala-Xaa-Pro-Ala-Lys-NH_2 (Reimer et al., 1998). The relaxation time of *cis/trans* isomerisation reactions has been estimated to range from ten to hundreds of seconds at 25°C by the NMR measurement (Grathwohl and Wuthrich, 1981). Thus, this high energy difference and slow reaction time present extra concerns regarding the role of isomerisation in biological functions.

The biology of *cis/trans* isomerisation.

It was suggested that the fast- and slow- folding populations of protein molecules differed in the *cis/trans* isomeric states of one or more peptidyl prolyl bonds (Brandts et al., 1975). The *cis/trans* isomerisation of peptidyl prolyl bonds has been identified to be a rate-limiting step in the folding process of the RNase A (Schmid and Baldwin, 1978). Furthermore, a class of enzymes, the peptidyl prolyl *cis/trans* isomerase (PPIase), is identified to be able to catalyse protein folding by

accelerating the isomerisation of peptidyl prolyl bonds (Schmid, 1993). The *in vitro* refolding process of a number of proteins has also been proven to be catalysed by the PPIase, such as ribonuclease T1, immunoglobulins and carbonic anhydrase (Lang et al., 1987; Lin et al., 1988; Fischer and Schmid, 1990; Freskgard et al., 1992).

However, other *in vivo* roles of PPIase have been suggested. For example, two categories of PPIase, the FK506-binding protein family (FKBP) (Siekierka et al., 1989) and the cyclophilin (CyP) (Fischer et al., 1989), have been characterized and proposed to play a role in structural regulation of ion channels. In particular, the FK506-binding protein, or FKBP12, has been reported to be copurifed during the sucrose density gradient centrifugation and co-localize to the terminal cisternae of the sarcoplasmitic reticulum with the ryanodine receptor (RyR), four of which constitute an intracellular Ca^{2+} channel of the sarcoplasmic and endoplasmic reticula (Jayaraman et al., 1992). It has been proved that FKBP12 modulates the Ca^{2+} release channel gating by stabilizing the ryanodine receptors probably through the improvement of cooperativity among the four RyR subunits (Brillantes et al., 1994). The details of this cooperative mechanism remain unknown; however, it could involve the FKBP12-assisted folding process or FKBP12-catalysed isomerisations of the proline residue during the RyR tetramerisation within cell membrane.

It is also shown that inhibitions of the cyclophilin will reduce the functional expression of homo-oligomeric receptors containing nicotinic acetylcholine receptor subunit $\alpha 7$ without reducing $\alpha 7$ polypeptide synthesis. The same effect could be generalized to other homo-oligomeric receptors, such as 5-

hydroxytryptamine type 3 receptor, but not to the hetero-oligomeric ion channels. This is possibly achieved with the mechanism of CyP-dependent subunit homo-oligomerization, a reaction in which CyP might catalyse the isomerisation of conserved proline residues in different homo-oligomeric receptors (Helekar et al., 1994).

In addition to the modulation of ion channel by the *cis/trans* isomerisation of a peptidyl prolyl bond, other possible roles for the buried prolines in transporter proteins have been proposed, since proline residues are commonly found in transmembrane helices of many integral membrane proteins (Brandl and Deber, 1986). For example, proline has been suggested to play a role in facilitation of conformational changes via *cis/trans* isomerisation of peptidyl prolyl bond and provision of cation liganding sites by the carbonyl oxygen of the Xaa-Pro peptide bond. These suggestions have been tested by a molecular modeling study, which indicates that proline-exposed carbonyl oxygens are capable of liganding cations (Sansom, 1992). More detailed discussions about the membrane-buried proline residue will be presented in Chapter 5.

These studies suggest that *cis* peptide bonds and the *cis/trans* isomerisation of peptide bonds are important for both protein structure and regulation of biological function. Thus, this study was designed to explore the potential biological roles for *cis* peptide bonds and especially the structural and functional implications of peptidyl prolyl *cis/trans* isomerisation.

3.2 Survey on Protein Structure Database.

A survey was performed on the 1998 January release of Protein Data Bank (PDB) (Bernstein et al., 1977) in order to detect the occurrence of *cis* peptide bonds in protein structures. The ω torsion angle for each peptide bond was derived from the spatial coordinates of the atoms of O_i, C_i, N_{i+1} and $C\alpha_{i+1}$, (Fig. 2.1); where *i* represents the residue number, stored in each PDB structure file. A *cis* peptide bond is defined as $-90° < \omega < 90°$; a *trans* peptide bond is defined ($-180° <= \omega <= -90°$) or ($90° <= \omega <= 180°$). The peptide bonds in *cis* conformation were extracted from the PDB and sorted. The structural descriptors such as ϕ angle, ψ angle and B-factor for residues at either side of each *cis* peptide bonds were derived and recorded for further analysis.

3.2.1 Distributions of the *Cis* Peptide Bond.

Distributions and types of *cis* peptide bond were examined in this study; 6609 peptide bonds were identified in the *cis* conformation. Among those *cis* peptide bonds, there were 5737 occurrences of the imide bond and 872 occurrences of the amide bond, (Tab. 3.1). The frequent occurrence of the *cis* conformation in both imide bonds and amide bonds reveals the potential importance of the *cis* peptide unit in protein structures.

Different amino acid types preceding the proline residue in a *cis* imide bond were examined. The entries of *cis* peptidyl prolyl bond ranged from 23 for Cys-Pro to 685 for Tyr-Pro. The distributions of the N-terminal amino acid residue of *cis* imide units are shown, (Fig. 3.3). The combination of Tyr-Pro has the highest

occurrence than other combinations. An earlier study based on a non-redundant set of protein structures showed a similar result (Stewart et al., 1990).

Type of Peptide Bond	Number of *Cis* Conformation
Xaa-PRO	5737
Xaa-nonPRO	872
Total	6609

Table 3.1 Occurrences of *cis* peptide bonds in Protein Data Bank. Xaa denotes the amino acid residue preceding proline.

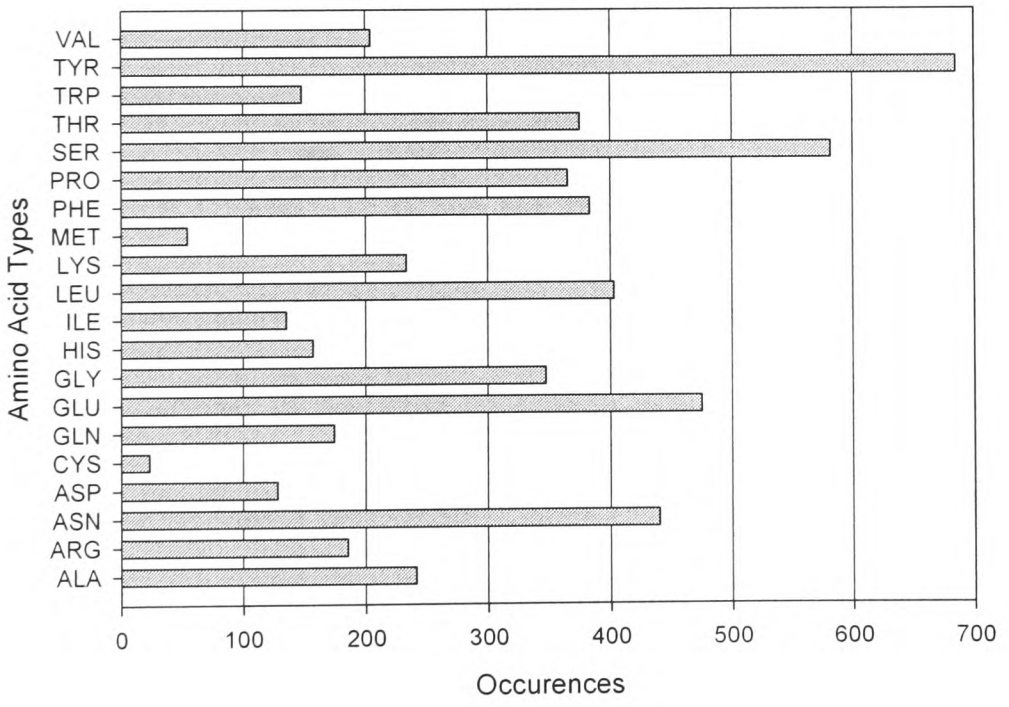


Figure 3.3 Distributions of different amino acid types preceding the proline residue in *cis* imide bonds.

3.2.2 ϕ, ψ Conformations of the *cis* Peptide Bond.

The ϕ, ψ conformations for both *cis* imide and amide peptide bonds were calculated and analysed in this study in order to find out any preferences of secondary structure for the *cis* peptide bonds.

The (ϕ, ψ) dihedral angles of the *cis* imide bond occur predominantly in the theoretical region of β structures on the Ramchandran plot, (Fig. 3.4 A). This finding is expected, since proline does not have the amide hydrogen atom which acts as a hydrogen donor to form the backbone hydrogen bond in order to build up a helical structure. *Cis* amide bonds do not show strong preferences for the β structure like *cis* imide bonds do. However, there are still more than 50% of amide peptide bonds located in the theoretical region of β structures on the Ramchandran plot, (Fig. 3.4 B).

There are 113 *cis* amide bonds with unusual conformations in the region of $40° < \phi < 80°$ and $-130° < \psi < -180°$, (Fig. 3.4 B). It is noted that 107 out of 113 peptide bonds have a glycine on their N-terminus, which explains the positive ϕ angles these peptide bonds adopt. In addition, it is also noted that 102 out of 113 peptide bonds are Gly-Ser. The high occurrences of the *cis* Gly-Ser peptide bond identified in this survey are likely caused due to redundancy of PDB.

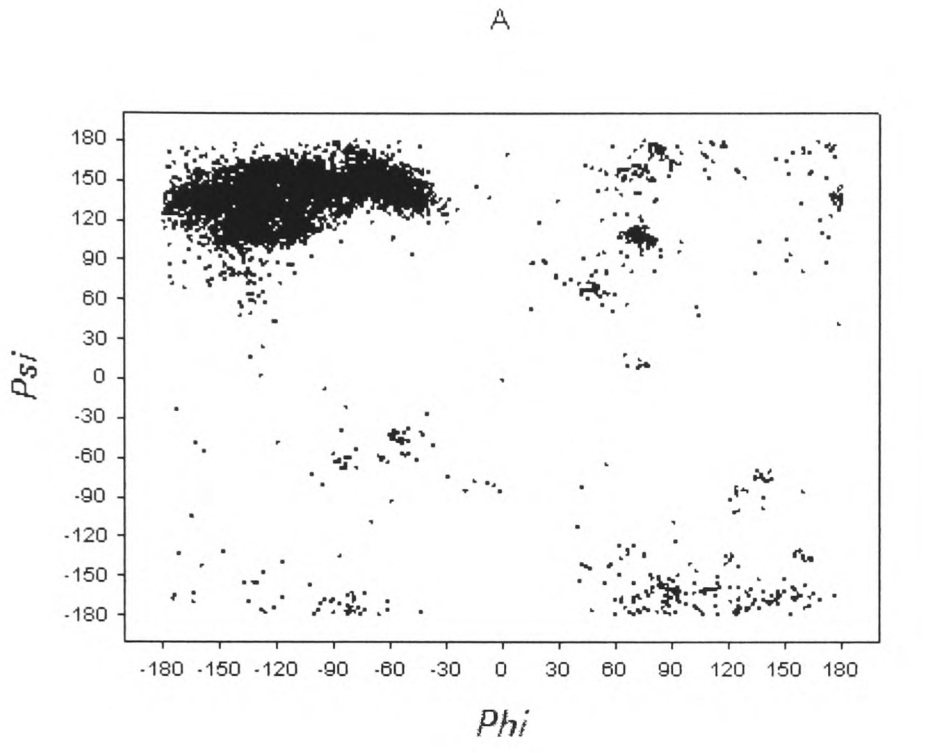


Figure 3.4 (ϕ, ψ) distributions for (A) *cis* imide bonds and (B) *cis* amide bonds.

3.2.3 Temperature Factor Analysis of the *Cis* Peptide Bond.

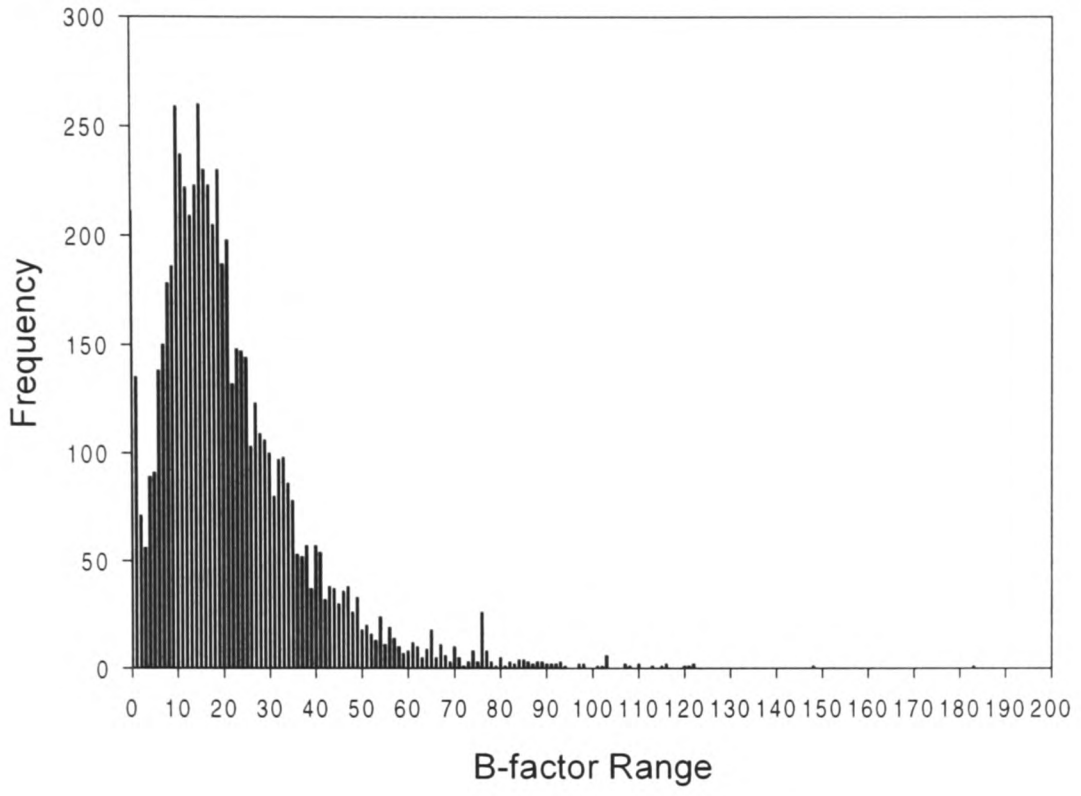
Since there are more steric repulsions between non-bonded atoms in the *cis* peptide bonds, thus it would be interesting to investigate whether the *cis* peptide bond is under higher conformational constraint and in a less stable configuration. One way to measure this dynamic disorder of protein structures is by assessing the B factor, which reflects the temperature-dependent vibration of the atoms in protein structures and structural constraints in protein crystals. Generally speaking, the B factor in proteins ranges from a few $Å^2$ to 30 $Å^2$ in well-ordered structures (Drenth, 1994). In this study, the temperature factors of the *cis* peptide bond were extracted from the PDB files and analysed.

The B-factors of Cα atoms for both the N- and C-terminal residues of the *cis* peptide unit were extracted from the PDB files directly. The extracted B-factors were sorted into different categories at intervals of 1 $Å^2$ and the number of peptide bonds falling into each interval were counted.

It is observed that B-factors of the majority of N-terminal Cα atoms in *cis* peptide bonds fall in the range from 0 to 30 $Å^2$, (Fig. 3.5 A). A very similar pattern is also observed for the distributions of B-factors of the C-terminal Cα atoms, (Fig. 3.5 B). This is as expected since the N- and C- terminal Cα atoms are at the same peptide unit. And it has been suggested that two bonded atoms must have B values which differ by no more 10 $Å^2$ (Daune, 1999).

The distributions of B-factors for both N- and C- terminal Cα atoms in *cis* peptide units suggest that a *cis* peptide bond is not particularly unstable in comparison with the overall protein structures.

A



B

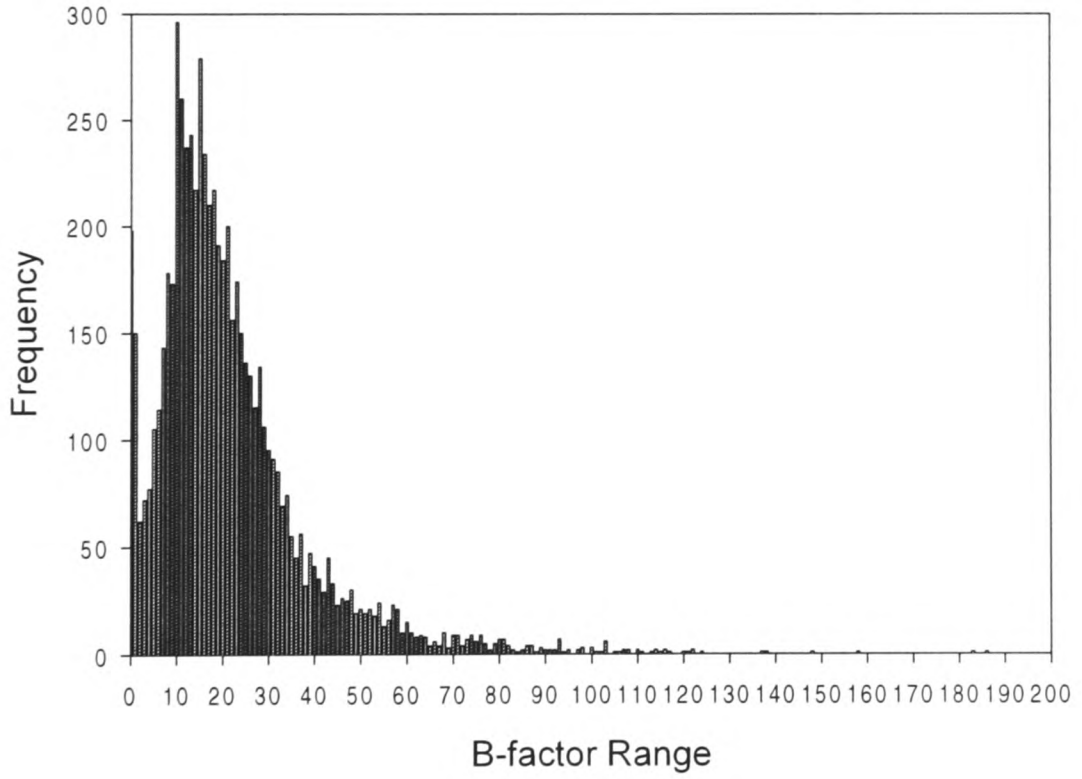


Figure 3.5 Distributions of B factors for the (A) N-terminal Cα atoms; and (B) C-terminal Cα atoms of *cis* peptide bonds. The B factor is expressed with the unit of $Å^2$.

3.2.4 Accessibility of the *Cis* Peptide Bond.

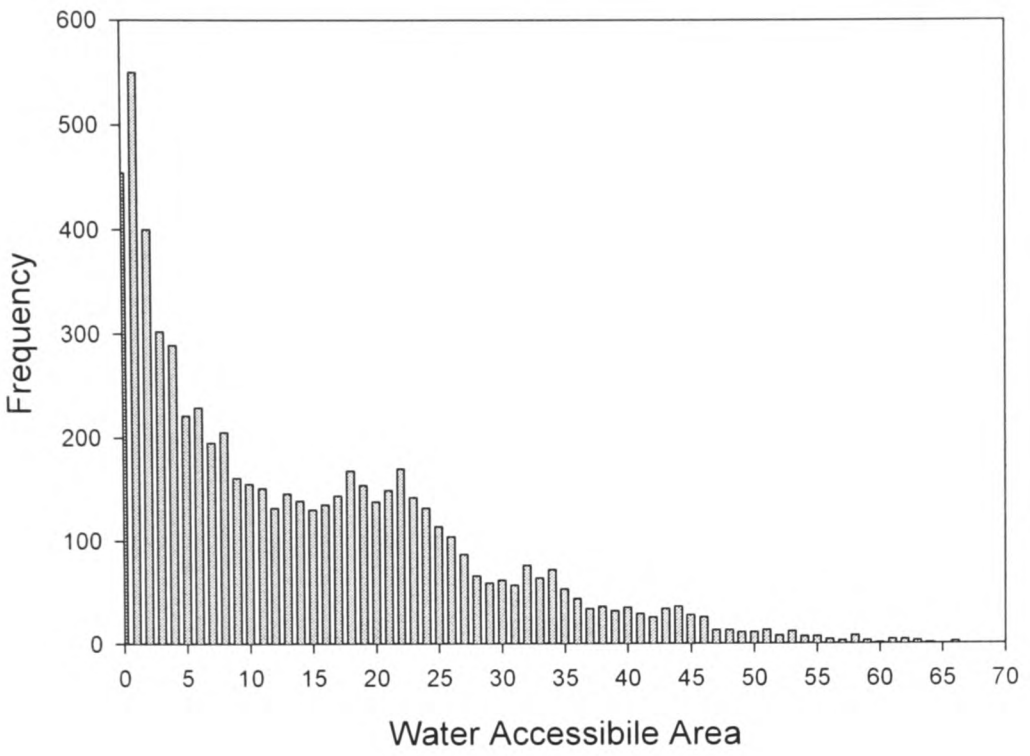
One way to detect the location of a *cis* peptide bond within a protein structure is to calculate its accessible surface area in order to deduce how surface exposed a residue is. A higher accessible surface area implies that a residue resides on the exterior surface; in contrast, a lower accessible surface area implies the residue is embedded in the interior of the protein structure (Ponder and Richards, 1987).

In this study, the aqueous accessible surface area for each residue was calculated by the What If package (Vriend, 1990), and the water molecule is used as a probe with radius set at 1.4 Å. The calculated total accessible area for both the N- and C-terminal residues in *cis* peptide bonds were sorted into different categories at interval of 1 $Å^2$ and the number of amino acid residues falling into each interval were counted.

The majority of N-terminal residues in *cis* peptide bonds are identified to have water accessible areas no larger than 35 $Å^2$, (Fig. 3.6 A). It is also observed that most C-terminal residues in *cis* peptide bonds have water accessible areas less than 35 $Å^2$, (Fig. 3.6 B).

The low accessible surface areas identified for both residues in a *cis* peptide unit suggest that *cis* peptide bonds are likely to occur in interiors rather than exteriors of a protein structure.

A



B

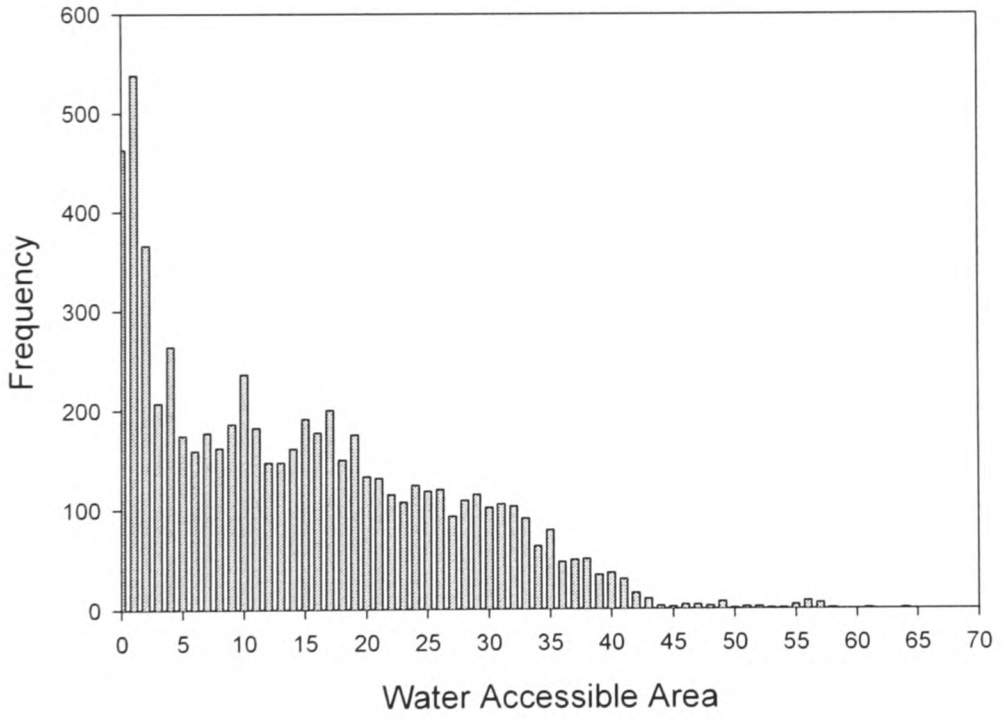


Figure 3.6 Distributions of total accessible area for the (A) N-terminal residues; and (B) C-terminal residues of *cis* peptide bonds. The water accessible area is expressed in units of $Å^2$.

3.3 Analysis on Protein Domain Structures.

The analysis of peptide bond conformations in the PDB revealed that *cis* peptide bonds occur at a significant frequency. This yielded an attempt to explore if *cis/trans* isomerisation may play a role in functional regulation of proteins. To achieve this goal, an overall survey of *cis* peptide bonds at the level of protein domains was performed. Protein domains have been described as compact, local, and semi-independent folding units (Wetlaufer, 1973), and are frequently identified as functional units for protein molecules to execute their biological functions (Murzin et al., 1995; Murzin, 1996; Orengo et al., 1997).

Protein domain definitions in the SCOP database were employed, since this database is constructed with criteria of combining experts' knowledge and automated methods (Murzin et al., 1995). The SCOP database is probably the most widely recognised protein domain database. In SCOP version 1.37, based on the October 1997 release of PDB, there were 13,073 protein domains defined from 6,497 protein structures. The survey of the SCOP domain database revealed that there were 3399 peptide bonds in the *cis* conformation in 2483 protein domains and 1567 protein structures, (Tab. 3.2).

Hierachy	Number and Percentage with *Cis* Peptide Unit
Structure	1567 (24%)
Domain	2483 (19%)
Peptide Unit	3399

Table 3.2 The occurrence of *cis* peptide units in SCOP protein domain database.

3.3.1 Assumption.

Assuming that the transition of *cis/trans* conformations may play a role in regulating protein function, the first concern which comes into consideration is the high energy difference, about 20 kcal/mol, between the *cis* and *trans* configurations. In other words, whether the biological process to be regulated is significant enough to compensate the energy differences between these two isomers. The second concern comes from the structural change due to the isomerisation. The mechanical movement caused by a *cis/trans* isomerisation in a local region has to be significant enough in order to trigger the downstream process of regulation.

On the basis of such considerations, one speculation was proposed that the *cis/trans* isomerisation might act as a molecular switch for regulating protein structure and function. For example, molecular switches could regulate domain motions involving substrate binding sites or the regulatory sites in order to control downstream biological functions. Similarly, a conformational switch might occur far from an active site but involve protein-protein interactions necessary for protein oligomerisation or contacts.

Domain motions have been proposed as a mechanism for changing protein conformations in order to regulate protein functions. For example, it has been observed that slight changes in backbone torsion angles may result in interconversion of regular secondary structure elements and irregular loops (Ikura et al., 1992). Hinged domain movements near the catalytic subunit of cyclic adenosine-monophosphate dependent protein kinase has been identified to be involved in the interaction with a peptide inhibitor of this enzyme (Knighton et al.,

1991). Proline residues are commonly found in the putative transmembrane helices of many integral membrane proteins, which act as transporters, channels and receptors. The helical structures of these transmembrane segments can be altered by kinks induced by the proline residue, and it has been suggested that kinks in transmembrane helices involving prolines could be of functional importance (Sansom, 1992).

3.3.2 Speculation.

Can *cis/trans* isomerisation of peptide bonds, especially of peptidyl prolyl bonds, act as a molecular switch?

No special motifs were identified specific for *cis* imide bonds in different protein domains in the SCOP database. During the searching process, however, a number of *cis* imide bonds were detected closed to the active site loops in enzymes or proteins which bind to carbohydrates and carbohydrates derivatives by visual inspections.

Mannose-binding Protein.

The structural basis of carbohydrate recognition by rat liver mannose-binding protein (MBP-C) was determined at 1.9 Å resolution (Ng et al., 1996), PDB code 1RDM. This is a two-chain protein with two ligand-binding sites at both identical chains. A *cis* amide bond was detected between Glu190 and Pro191 for both chains. This *cis* peptide bond is located on a loop near the ligand-binding site, (Fig. 3.7).

The average B factor for this protein structure is estimated at 13.8 ± 4.3 Å^2. The B factors for Cα atoms of both residues are ca. 15.4 Å^2, which are relatively higher

than the average. The accessible areas for the Glu190 and Pro191 are estimated as 5.4 $Å^2$ and 11.3 $Å^2$ respectively, which indicate that this *cis* peptide bond is not very exposed to the outer environment.

The protein 1RDM is classified in the family of C-type lectin domain by SCOP definition. In this family, there are 4 protein domain types: (1) lithostathine, (2) E-selectin, (3) mannose-binding protein A, lectin domain, and (4) tetranectin. This Glu-*cis*Pro motif in the ligand-binding loop is conserved, and appears in all of the proteins in these protein domains.



Figure 3.7 The structure of protein 1RDM chain 2 with substrate a-methyl-D-mannopyranose (MMA) The distance between Glu190 and MMA is less than 3 Å.

Glucose 6-Phosphate Dehydrogenase.

Glucose 6-phosphate dehydrogenase (G6PD) is the first enzyme of the pentose phosphate pathway, which catalyses the dehydrogen reaction of glucose 6-phosphate into 6-phosphoglucono-δ-lactone. The structure of this enzyme from *Leuconostoc mesenteroides* was determined and refined to 2.0 Å resolution (Rowland et al., 1994), PDB code 1DPG. This enzyme is a dimer, each subunit consisting of two domains. One significant structural difference between two subunits is that the Pro149, on the coenzyme-binding domain, is *cis* in chain A and *trans* in chain B; however, Pro375, on a loop near the substrate binding pocket, in both chain A and chain B are both *cis*, (Fig. 3.8).

The average B factor for this protein structure is estimated at 33.5 ± 14.7 $Å^2$. The B factors for Cα atoms of both residues in the *cis* peptide bond between Lys148 and Pro149 in chain A are 42.8 $Å^2$ and 49.3 $Å^2$ respectively. The B factors for Cα atoms of Lys148 and Pro149 in chain B are 49.3 $Å^2$ and 48.0 $Å^2$ respectively. These four residues have evidently higher B values than other residues. The B factors of Cα atoms for *cis* peptide bonds franked by Asp374 and Pro375 in both chain A and B are comparatively lower than average and estimated at ca. 36 $Å^2$ for Asp374 and 15 $Å^2$ for Pro375.

The conserved Pro149 exhibits different conformations in different chains; *cis* for chain A and *trans* for chain B. A domain boundary differ by 1.73 Å in main chain coordinates were detected, which might arise in part from crystal contacts made by side chains of neighbouring residues (Rowland et al., 1994).

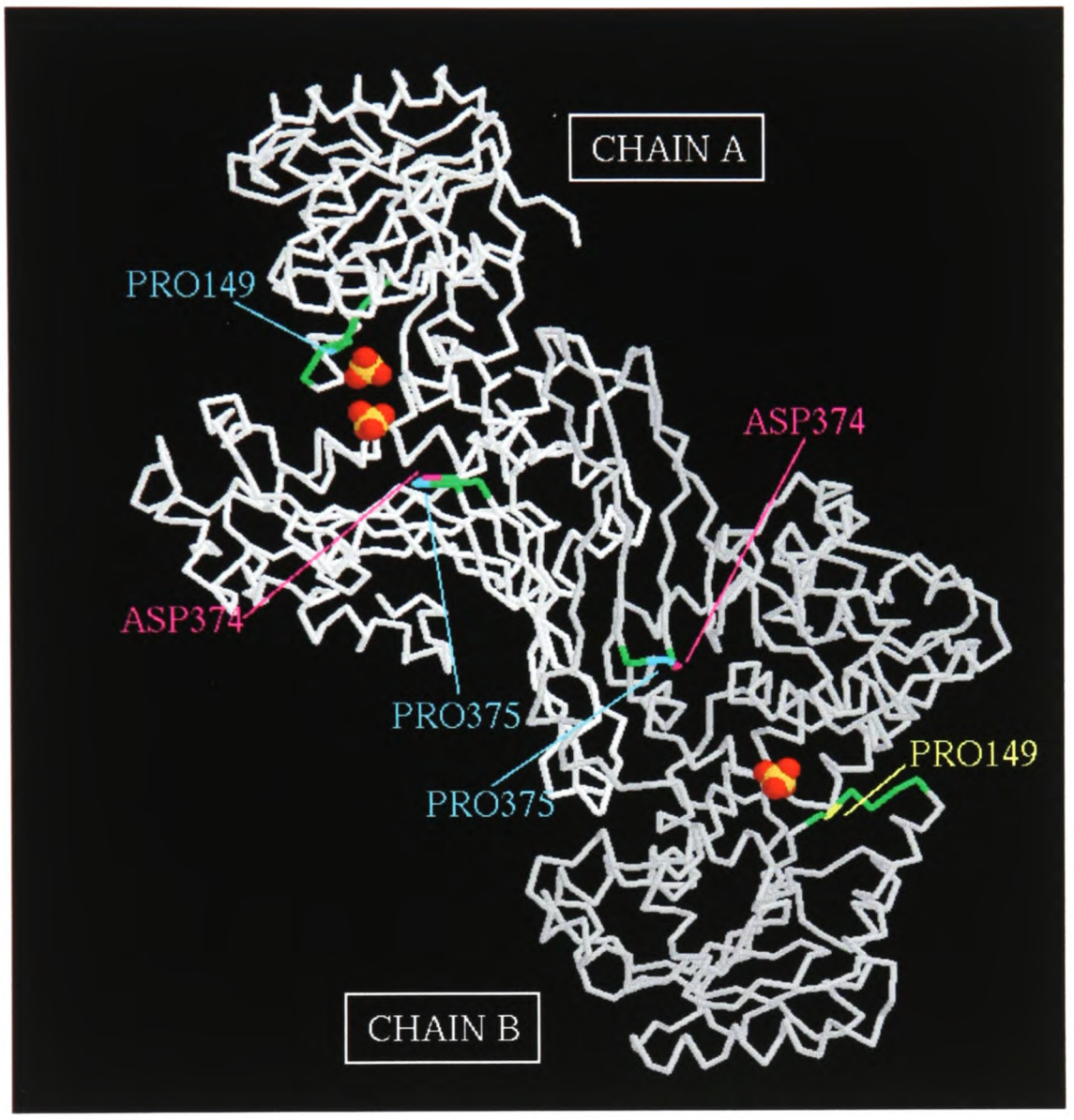


Figure 3.8 The structure of protein 1DPG. Residues denoted in cyan are *cis* proline residues. PO_4^{2-} molecules are within the catalytic pockets.

Superimposition of chain A and chain B by STAMP package (Russell and Barton, 1992) reveals that local structures near Pro375 at both chains are similar, (Figure 3.9). However, different *cis/trans* configurations for Pro149 at chain A and B cause a shift of the active site loop. The loop in chain B with *trans* Pro149 is more open than the loop in chain A with *cis* Pro149, (Figure 3.9). In addition, the conformations at the local regions near Pro149 are different in the two chains; the Cα atom displacement from chain A to chain B at the local region is estimated at an average of 2.14 Å for residues of Glu147 to Asp155.

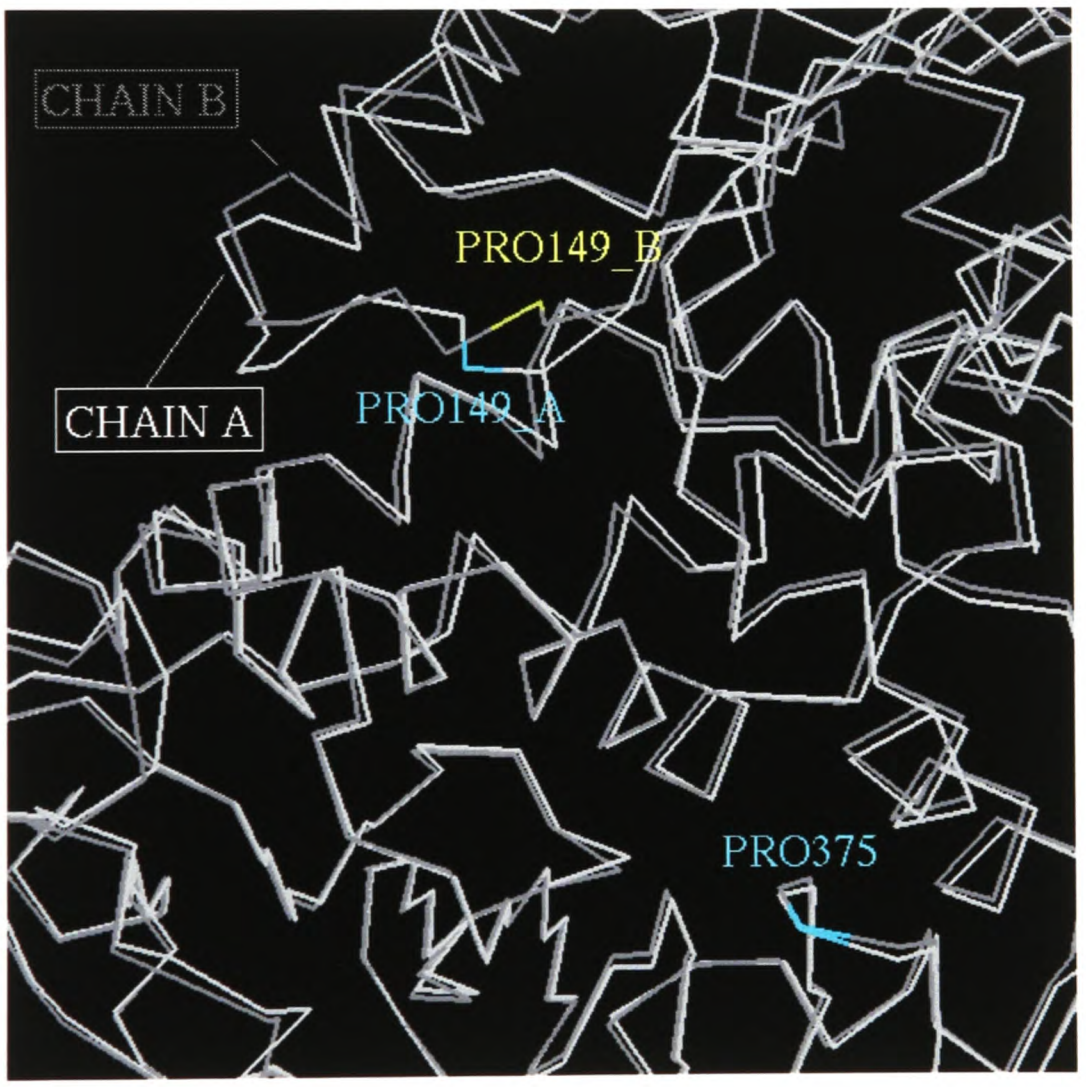


Figure 3.9 Superimposition of chain A and chain B of protein 1DPG. Chain A is denoted with white and chain B is denoted with gray.

TIM Protein Family.

An interesting motif of a negatively charged amino acid residue followed by a proline residue residing on the active site loop was also identified in the protein family of triosephosphate isomerase (TIM) through the SCOP database search. However, these peptide bonds are all in *trans* conformations.

Triosephosphate isomerase is a dimeric enzyme, with each subunit composed of 247 residues. This enzyme catalyses the interconversion of dihydroxyacetone phosphate (DHAP) into D-glyceraldehyde 3-phosphate (GAP) in the glycolytic pathway. In the catalytic mechanism of TIM, it is proposed that there is a hinge motion of an active site loop from residue 165 to 175 in order to control the uptake of DHAP into the substrate binding pocket (Joseph-McCarthy et al., 1994). The amino acid sequences of the hinge regions in different species, ranging from *E. coli* to human, are highly conserved through evolution, (Fig. 3.10). A proline residue was noted at the N-terminal end of this active site loop in all species; as shown in the structure of *Trypanosoma brucei* triosephosphate isomerase complexed with glycerol-3-phosphate, PDB code 6TIM (Noble et al., 1991), (Fig. 3.11). The Glu residue prior to the proline was proposed to participate directly in proton transfer in order to carry out the acid/base chemistry of this interconversion reaction (Bash et al., 1991).

A MD simulations study has been reported and performed in order to explore the active site loop's motion of the triosephosphate isomerase (Derreumaux and Schlick, 1998). The energy difference between the closed and opened forms of the active site loop was estimated at ca. 25 kcal/mol. This study suggests that even a local conformational change with high energy difference, for example hinge

movement or *cis/trans* isomerisation, could still be of functional importance in biology.

Species	*PDB*	*Active Site Loop*
Escherichia coli	1tre	AYEPVWAIGTGKS ATPA
Bacillus stearothermophilus	1btm	AYEPIWAIGTGKSSTPE
Plasmodium falciparum	1ydv	VYEPLWAIGTGKTATPE
Vibrio marinus	1aw1	ATEPIWAIGTGKATAED
Leishmania mexicana	1amk	ATEPVWAIGTGKVATPE
Trypanosoma brucei	6tim	AYEPVWAIGTGKVATPQ
S. cerevisiae	7tim	AYEPVWAIGTGLAATPE
Gallus gallus	1tph	AYEPVWAIGTGKTATPQ
Homo sapiens	1hti	AYEPVWAIGTGKTATPQ

Figure 3.10 Sequence conversions in the TIM protein family. Both Glu and Pro are essential for biological function in different species.

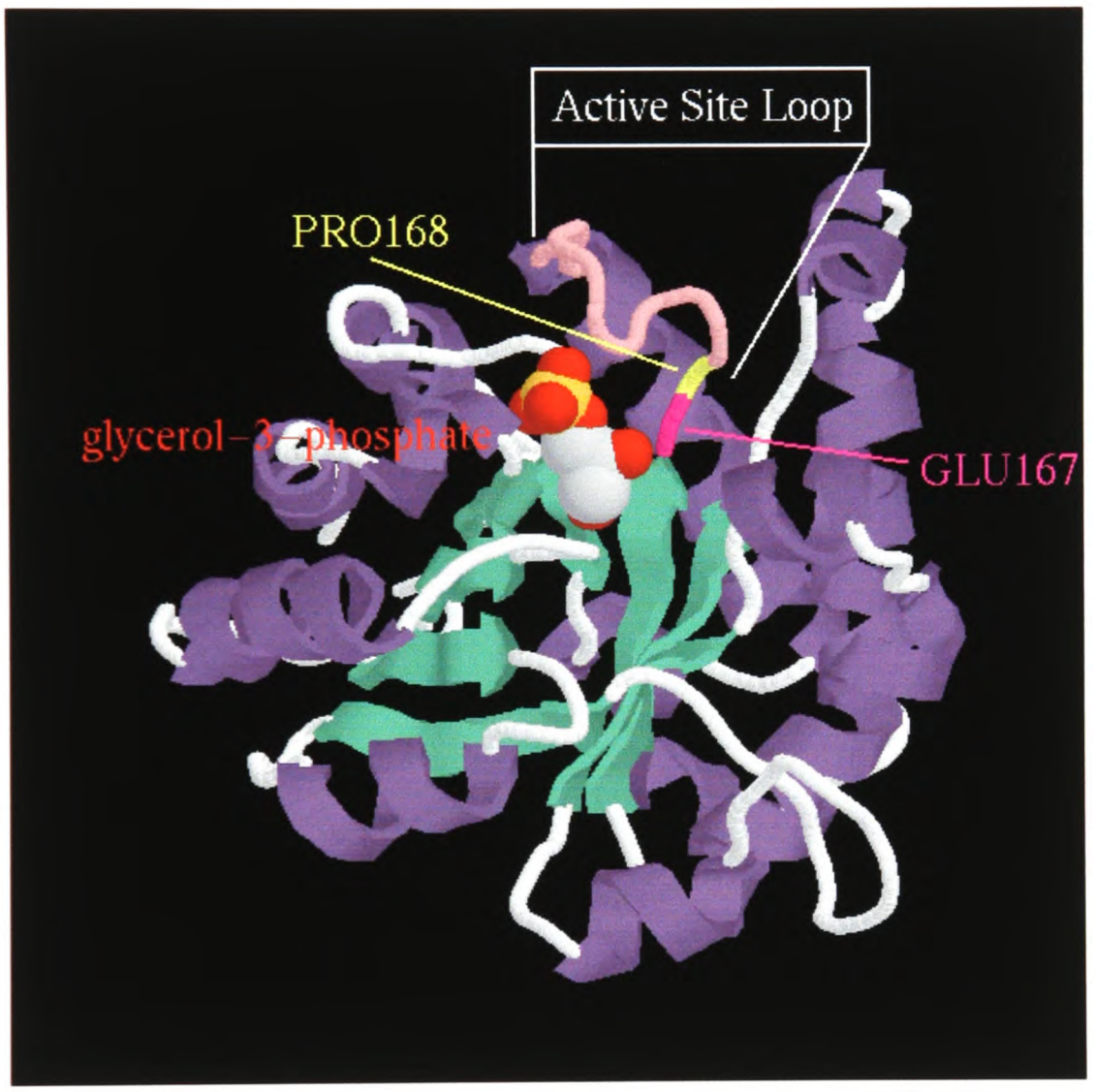


Figure 3.11 Structure of protein 6TIM. Two essential residues on the active loop are conserved through out evolution to control the open and close of the loop, which acts as a lid.

3.3.3 Observation.

After these interesting findings from the protein domain database search, it becomes clear that *cis/trans* isomerisation of peptidyl prolyl bond may play a role in regulating biological function. However, are there any experimental studies supporting an involvement of imide bond isomerisation in regulation of protein function?

Adenylate kinase.

Adenylate kinase catalyses the reversible phosphoryl transfer from ATP to ADP:

MgATP + AMP ⇔ MgADP + ADP.

The structure of porcine adenylate kinase was solved at 2.1 Å resolution (Dreusicke et al., 1988), PDB code 3ADK. There are 6 proline residues in this enzyme, and only Pro96 is *cis*, (Fig. 3.12). Three conserved proline residues were identified within the catalytic pocket by sequence comparison of adenylate kinases from different organisms (Muller and Schulz, 1992). They are Pro17, *cis*Pro96 and Pro159 in the porcine adenylate kinase.

An early study suggested that this enzyme undergoes a substrate-induced conformational change in order to surround the substrate with catalytic residues (Gerstein et al., 1993). In a recent study, different concentrations of peptidyl prolyl *cis/trans* isomerases (PPIase) were employed to accelerate the isomerisation rate controlling the conformational change of adenylate kinase. A fluorescent compound, 8-anilino-1-naphthalenesulfonic acid (ANS), was used to mimic the natural substrate ATP in order to measure the ligand binding kinetics of this enzyme. Comparing different reaction rates at different levels of PPIase

concentration, it has been suggested that the domain movement during the substrate binding to rabbit muscle adenylate kinase may involve proline isomerisation (Sheng et al., 1997). It is difficult to distinguish which proline residues are involved in the isomerisation. However, the activation energy was estimated to be 74.6 kJ/mol by kinetic studies.

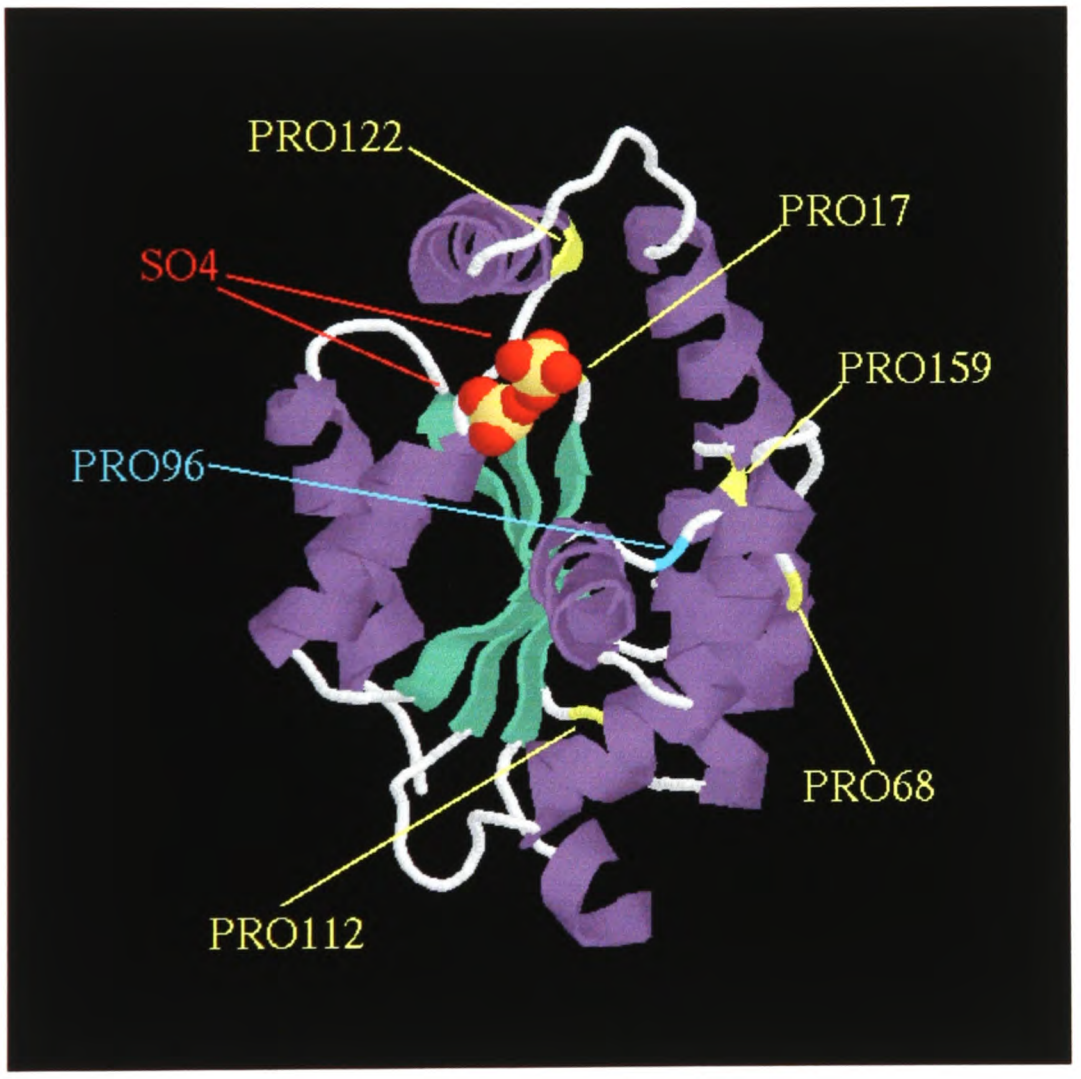


Figure 3.12 Structure of protein 3ADK. Pro17, Pro96 and Pro159 are within the catalytic pocket, and Pro96 is *cis* (cyan). Isomerisations of these proline residues are suspected to be involved in substrate binding and catalysis.

Mitochondrial Creatine Kinase.

Another reported case of a special role of isomerisation in biological function comes from another enzyme involved in energy metabolism. Creatine phosphate (CrP) is one of the high-energy phosphate compounds which store energy in cells. Creatine kinase is the key enzyme catalysing the phosphoryl transfer reaction from creatine phosphate to ADP (Wallimann et al., 1992),

$$Cr + ATP \Leftrightarrow CrP + ADP.$$

The structure of mitochondrial creatine kinase from chicken has been solved at 3.0 Å (Fritz-Wolf et al., 1996), PDB code 1CRK. This structure revealed that one of the 22 proline residues in this protein is in a *cis* conformation. This *cis* proline residue is positioned at residue 207, which is far from the active site but resides in a loop, which is likely to be responsible for the monomer-monomer contacts within the dimeric building blocks to form a functional octameric unit of this enzyme, (Fig. 3.13).

It has been proved that the overall conformations for both dimeric building component or octameric functional unit of this enzyme are changed when a mutation was introduced at residue 207 from proline to alanine (Forstner et al., 1998a). Furthermore, a decrease in the pH-optimum by 1.5 unit for the reaction of ATP synthesis and a reduction of 53% of the enzyme activity were observed. A theoretical study suggests that the changes in enzyme kinetics is likely to be induced by a structural change at the residue 207 peptide bond from a *cis* to *trans* configuration (Forstner et al., 1998b).

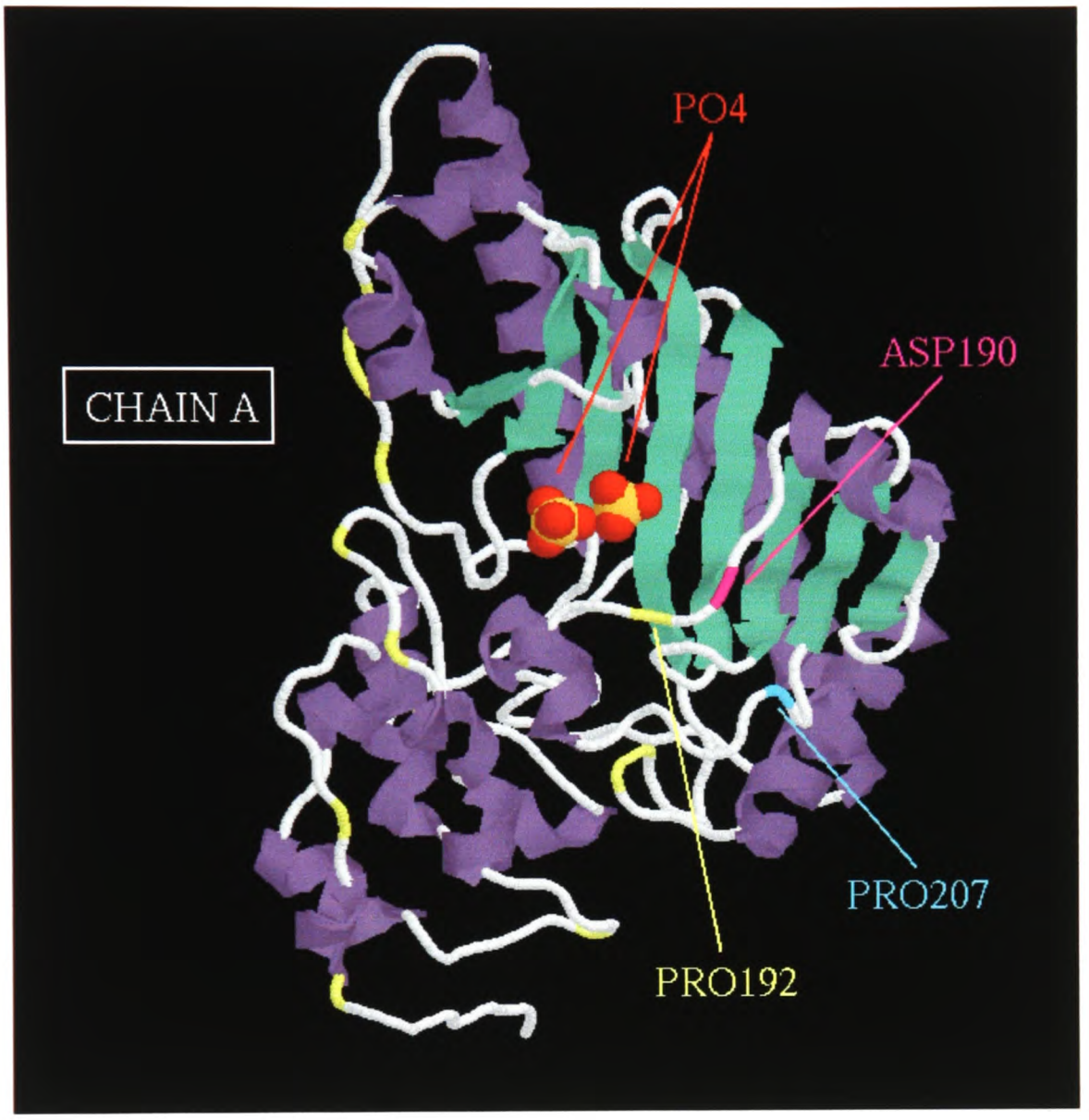


Figure 3.13 Structure of protein 1CRK. The Pro207 (cyan) is on a loop for protein-protein interaction to form the building block dimer, and is *cis*. The *cis*Pro207 is suspected to be of structural importance to maintain the function of this enzyme.

Mitochondrial Cytochrome$_{bc1}$ Complex.

Cytochrome bc_1 is a transmembrane protein complex composed of a [2Fe-2S] cluster-containing unit, a heme *c*-containing cytochrome c_1, a cytochrome *b* that contains two functionally inequivalent heme *b*'s, and other subunits dependent on different species. This huge protein complex, which mediates electron transfer in the photosynthetic reaction of purple photosynthetic bacteria or in the mitochondrial respiratory chain, is found in the plasma membrane of purple photosynthetic bacteria or in the inner mitochondrial membrane of eukaryotes (Brandts and Trumpower, 1994). The structure of bovine mitochondrial cytochrome bc_1, which contains 11 subunits, was solved and refined to 3.0 Å (Iwata et al., 1998), PDB code 1BE3.

From this structure, it was identified that the functional domain of the [2Fe-2S] cluster-containing unit has two separate folds; a metal cluster binding fold which is from residue 137 to 181 and a base fold which contains residue 74 to 136 and residue 182 to 196, (Fig. 3.14). It was also observed that the [2Fe-2S] cluster-containing unit interacted with cytochrome bc_1 and cytochrome *b* *via* contacts between their loops. During the contact process, small but significant positional displacements between metal cluster binding fold and base fold of the [2Fe-2S] cluster-containing unit were observed, which involved a *cis/trans* transition at Pro175 in the loop region of residue 104 to 166 (Iwata et al., 1998).

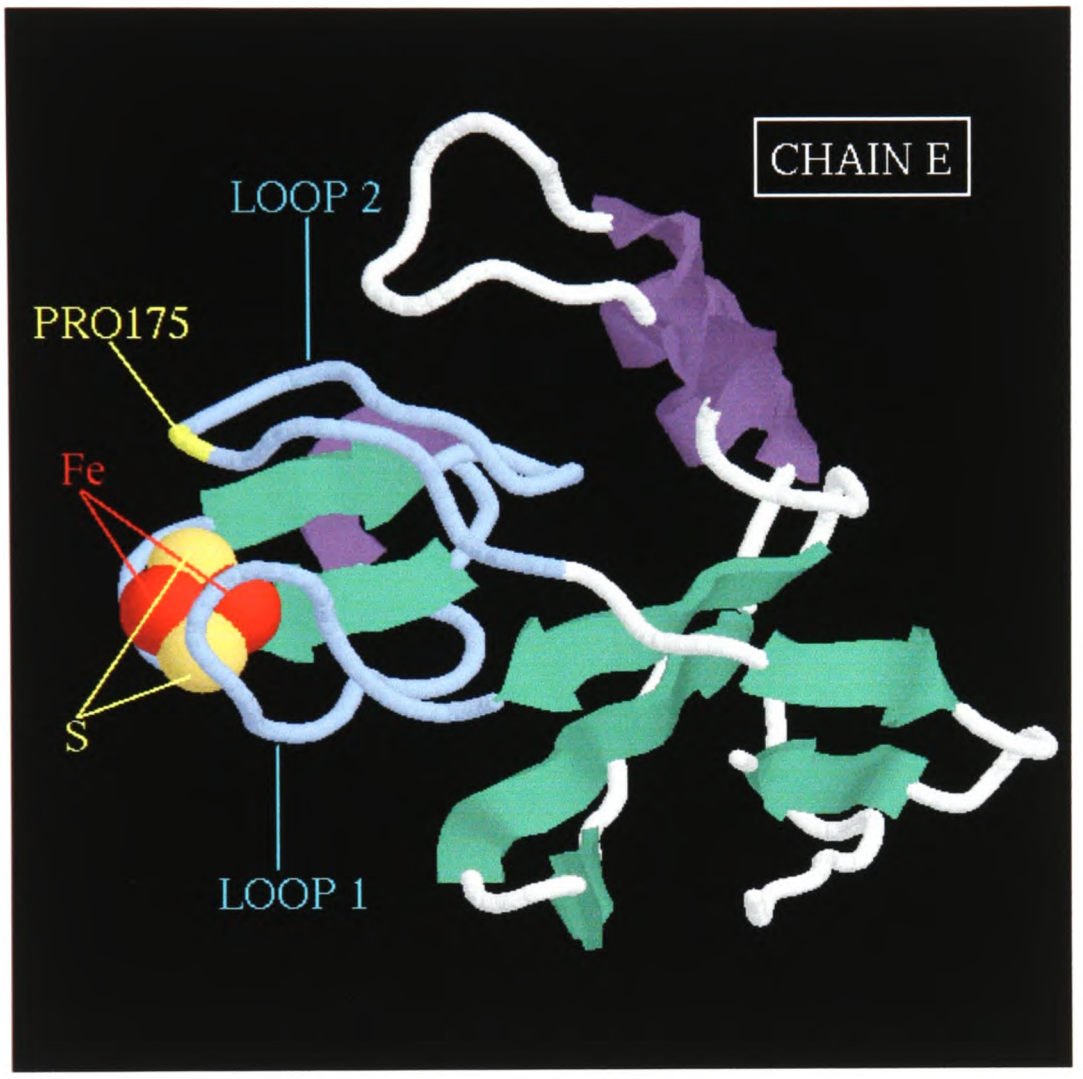


Figure 3.14 Structure of the (2Fe-2S) cluster-containing unit of protein 1BE3. There is one metal cluster binding fold (blue) and one base fold (white) in this domain. Loop1 and loop2 are involved in interactions with cytochrome *b* and cytochrome c_1. *Cis/trans* isomerisation of Pro175 (yellow) is suggested to be a part of the mechanism for the loop movement and protein-protein interactions.

TB Domain.

If *cis/trans* isomerisation does regulate biological function, the next question to be asked is whether these two conformers can coexist at the same time in a natural environment. This question has been answered in an elegant experiment by nuclear magnetic resonance (NMR) analysis.

The transforming growth factor β (TGF-β) binding protein-like (TB) domain is found in proteins localised to extracellular matrix fibrils, such as fibrillins and latent transforming growth factor β binding proteins (LTBPs) (Bork et al., 1996). Different conformers, due to *cis/trans* isomerisation of two proline residues, of the sixth TB domain (TB6) from human fibrillin-1 have been identified as its solution structure was solved (Yuan et al., 1997). There are 6 proline residues in the TB6 domain of 74 residues. Analysis of NMR data for the TB6 indicated that only Pro22 and Pro52 adopt both *cis* and *trans* conformations, which lead to the co-existence of one major conformer (*trans* Pro22 and *cis* Pro52) and one minor conformer (*cis* Pro22 and *trans* Pro52) in solution (Yuan et al., 1998).

Analysis of the secondary structure elements of these two conformers revealed that although these two conformers have similar overall structures, the secondary structure topologies were affected by the *cis/trans* isomerisation of Pro22 and Pro52 (Yuan et al., 1998). This finding demonstrates that not only the isomerisation of peptide bonds can result in changes of protein fold at the level of secondary structure locally, but also that the *cis* and *trans* isomers coexist in solution.

These findings reinforce the argument that the *cis/trans* isomerisation is of structural and biological significance.

3.3.4 Hypothesis.

As briefly mentioned in the section 3.1, an early statistical survey of the putative transmembrane helices of transporter proteins revealed that proline residues are frequently found at membrane-spanning helices (Brandl and Deber, 1986). The proline residue has relatively rigid backbone angle and the imide bond lacks a proton to form the hydrogen-bonding patterns which is required for helical structures, thus proline is known as a helical destabliser (Richardson and Richardson, 1989). The abundance of membrane-embedded proline residues in transmembrane helices remains arcane. One possible structural and functional role for these membrane-buried proline residues proposed that the *cis/trans* isomerisation of the Xaa-Pro bond could regulate the reversible conformational changes of transporter proteins (Brandl and Deber, 1986). One of the well-studied transmembrane proteins, bacteriorhodopsin, provides a good model to examine this theory.

There are three proline residues identified within the transmembrane helices of bacteriorhodopsin, (Fig. 3.15). Bacteriorhodopsins are found in the cell membrane of halobacteria and convert light energy into biological energy by acting as a proton pump (Stoeckenius, 1999). A structural role for regulation of these three membrane embedded proline residues, Pro50, Pro91 and Pro186, in proton pumping was suggested. In this model,

(1) the *cis/trans* isomerisation of the membrane embedded peptidyl propyl bond facilitates changes of the local structure,

(2) the overall electron resonance energy of an imide bond is affected by its local conformations, such as *cis* or *trans* configuration, and

(3) the proton affinity of the carbonyl oxygen of a peptidyl propyl bond, which behaves as the cation liganding site, is modulated by changes in the imide electron resonance energy (Dunker, 1982).

However, the site-directed mutagenesis studies on these three proline residues proved that the membrane-embedded proline residues are not essential for proton pumping in bacteriorhodopsin (Mogi et al., 1989). Even though, the proposal of proline residues as molecular switches to regulate conformational changes at the local region near proline residues on the transmembrane helices by *cis/trans* isomerisation of imide bonds remains possible. Since mutations of proline residue to other amino acid residues may disturb the local environment, unidentified proton liganding sites could emerge. In other word, the site-directed mutagenesis may not be the best way to test this proposal. This proposal has been further tested with X-ray crystal structure of bacteriorhodopsin at 1.9 Å resolution; however, no significant functions or peptide bond isomerisation are detected for these proline residues (Belrhali et al., 1999).

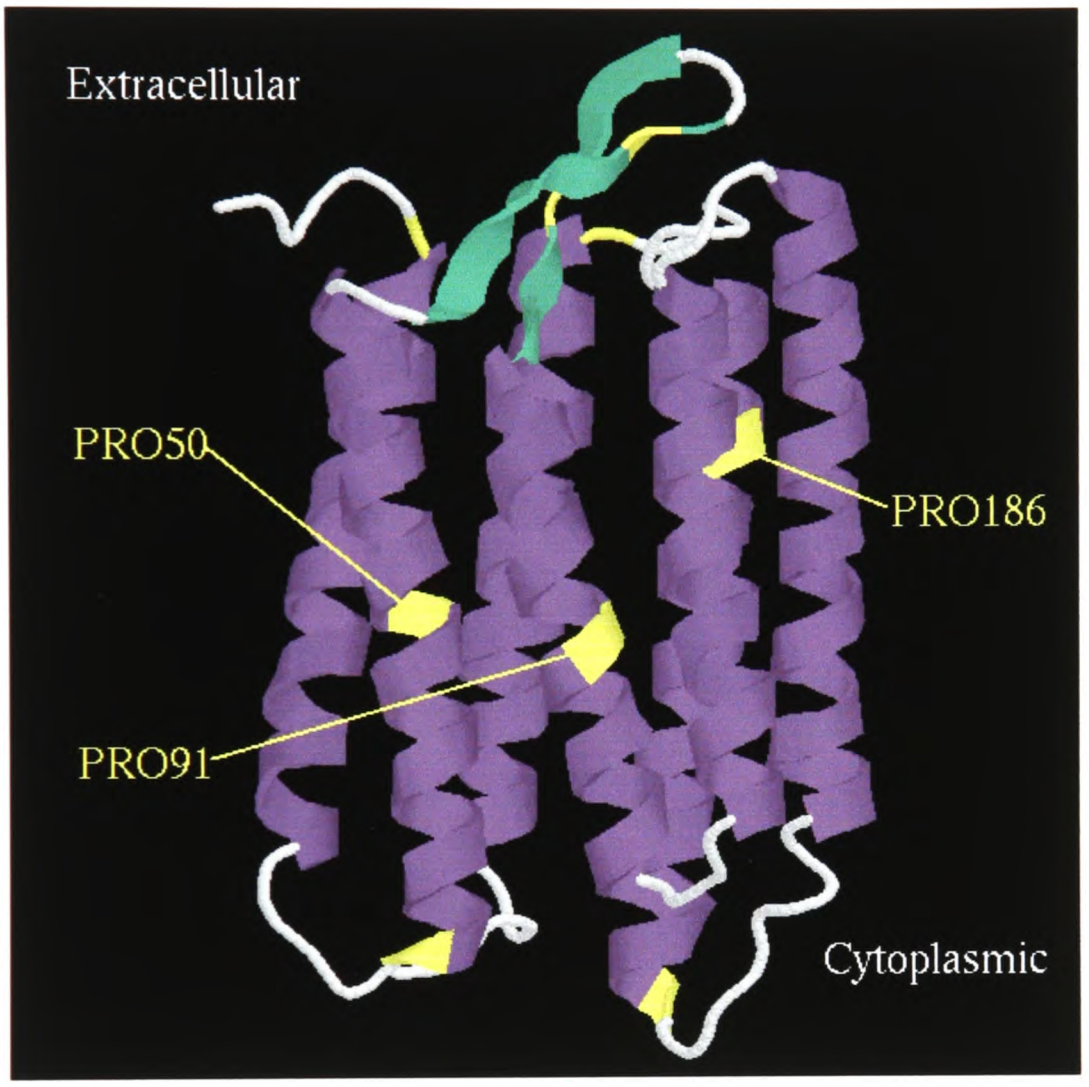


Figure 3.15 Structure of protein 1AT9. Three membrane-embedded proline residues, Pro50, Pro91 and Pro186, are denoted in yellow. Their *cis/trans* isomerisations are proposed to facilitate proton transportation

3.4 Discussion.

The predominantly frequent occurrence of the *trans* peptide bond in protein structures has been long observed and taken for granted. The rare occurrence of *cis* peptide units are explained by the higher intrinsic energy of the cis over the *trans* form, and the conformational restriction imposed by the occurrence of a *cis* peptide bond in the sequence of *trans* peptide units (Ramachandran and Sasisekharan, 1968). However, the proline residue, due to its lower intrinsic energy for isomerisation and its unique cyclic structure, has higher frequency of being observed in the cis conformation (Ramachandran and Mitra, 1976).

There have been some speculations about the possible roles of *cis/trans* isomerisation of peptide units in protein biological function. However, isomerisation of proline residues has only been shown to play an important role in the protein folding process (Brandts et al., 1975). No other special role has been found for *cis* peptide bonds, which are sometimes regarded as disordered structure.

In a recent study, a significant correlation between the resolution of the structure solved and the number of *cis* peptide units detected has been found (Weiss et al., 1998). It is also indicated, in the same study, that most commonly used protein structure refinement programs have high energy penalities for *cis* conformations and only allow *cis* conformation for the imide bond, unless specified explicitly for other amide bonds. It is clear that the occurrences of the *cis* peptide bond is likely to be underestimated due to the limited amount of high-resolution protein structures and refinement methods employed.

From an overall survey of the Protein Data Bank, it is found that there are still numerous *cis* peptide units existing in protein structures. This finding suggests that

peptide units in *cis* conformation may not be insignificant. For *cis* propyl peptidyl bonds, there are no significant preferences for specific amino acid residues preceding proline residue. This result is similar to a study indicating that the imide bond conformation is mainly determined by local effects in proteins (Reimer et al., 1998).

Searching for specific motifs containing *cis* peptide units did not reveal any useful information. This could be explained due to:

(1) underestimation of *cis* peptide bond's occurrence in protein structures, and

(2) formation of *cis* peptide bonds being largely dependent on the local environment within a protein molecule.

However, some interesting patterns of *cis* imide bonds are located in a series of enzymes or proteins, which have similar reaction mechanism.

Since *cis* peptide bonds are thermodynamically unfavourable over the *trans* peptide bonds, their existence has to provide some functional advantages for the proteins that bear them. This could be achieved either by

(1) direct organisation of special local structure for the active site or

(2) indirect provision of a special fold for the structural domain.

In the protein domain structure survey, several proteins binding to carbohydrates serve as examples of the first case. Proline residues, some of which are *cis*, are detected at the active loop regions which may control the entry of the substrate and provide an ideal environment for acid catalysis reactions. And the local environment of acid catalysis has been proposed to be essential for *cis/trans* isomerisation of peptide bonds (Forstner et al., 1998b).

The biological and structural importance of *cis/trans* isomerisation of peptide bond is worth more attention, due to the fact that many *cis* peptide bonds may have been misinterpreted because of the limited structure resolutions and the structure refinement protocols. Furthermore, how the peptide bond isomerisation might play a role as a molecular switch to regulate biological functions of protein remains to be proven by more protein structures to be solved at either high resolution or in solution.

Chapter 4

ABC Transporters:
From Genome Sequences to Transmembrane Helical Structures

4 Preface to Chapter 4.

During the studies of isomerisation of peptide bonds, unexpected high occurrences of membrane-embedded proline residues in transmembrane helices were noted. Since the proline residue is known to be a helical breaker (Richardson and Richardson, 1989), it's recurring existence in transmembrane helices is intriguing (Brandl and Deber, 1986). It seemed appropriate for concentrating my research on integral membrane protein structures.

However, transmembrane protein structures are notorious known for presenting great difficulties to either crystallography or NMR spectroscopy (Martonosi, 1998; Preusch et al., 1998), details see section 1.2. One alternative way is to use different computational methods to analyse amino acid sequences, to predict their possible topologies within lipid bilayers and to build transmembrane helical models by molecular dynamics simulations (Hong and Sansom, 1999; Biggin and Sansom, 1999).

Presented here are the sequence and structure prediction studies of transmembrane helices of a special protein family, ABC transporters (details see section 4.1.2), in different genomes. This chapter is divided into 5 parts and uses ABC transporters as a model system. Section 4.1 begins with introductions to membrane protein structures and biology of ABC transporters. Section 4.2 covers different methods used in this study, ranging from sequence databases, structure prediction to simulated annealing/molecular dynamics simulations. Section 4.3 discusses results from different genomes. Finally, Section 4.4 indicates the biological relevance of the analysis, and has an overall discussion about applications of this analysis.

4.1 Introduction.

4.1.1 Integral Membrane Proteins.

Biology of Integral Membrane Proteins

Transmembrane proteins comprise a large group of proteins, which span the cell membrane and represent an important class of proteins. They are employed in a wide range of essential cellular functions. For example, ion channels allow passive flux of inorganic ions through a central pore (Hille, 1992); ABC transporters mediate selective movements of various molecules across cell membranes in different species (Nikaido and Hall, 1998); G-protein-coupled receptors catalyse biochemical reactions, receive inward messages, and then transfer biochemical signals from outside environment into the cell (Freedman, 1996); bacteriorhodopsin absorbs light energy by its chromophore retinal and then converts light energy into biological energy in order to pump protons from the cytoplasm of purple bacteria into extra-cellular space (Oesterhelt and Stoeckenius, 1973).

Membrane proteins are no less important biologically than water-soluble proteins; however, they have been less thoroughly studied due to technical difficulties in expression, purification and crystallisation. However, they often are of pharmaceutical interest and have great potential as targets for drugs (Mingarro et al., 1997). One of the best examples is the ABC proteins, a large family of α-helical transmembrane proteins, which mediate varieties of solute across cell membranes. The ABC transporters are of great clinical and medical implications, since they are responsible for the antibiotic resistance in bacteria and also responsible for some human diseases, details see section 4.1.2.

Genomic Sequence of Transmembrane Protein

One of the resources which may be exploited in studying membrane proteins is the enormous and still increasing body of genome sequence data. At the time of writing, complete genomic sequences are known for 21 prokaryotes (Mewes et al., 1998), for yeast (Mewes et al., 1997) and for *C. elegans* (Consortium, 1998).

``How many membrane proteins are there?" (Boyd et al., 1998) is a frequently asked question of membrane protein research in the era of genome analysis. Attempts have been made to determine the portions of different genome that correspond to membrane proteins on the basis of hydropathy analysis. From these studies, it is generally agreed that integral membrane proteins comprise of at least 20% to 30% of encoded protein sequences in different genomes (Boyd et al., 1998; Wallin and von Heijne, 1998). This provides a strong motivation to exploit this information in terms of trying to understand the relationship between membrane protein sequence and structure and eventually to a better prediction of their function.

Efforts have also been spent on detecting the possible preferences of transmembrane topology of membrane-spanning proteins. It is reported that no dominant membrane protein families were found in the genomes of *M. genitalium*, *H. influenza*, *M. jannaschii* and *S. cerevisiae* (Arkin et al., 1997). However, similar studies suggests that unicellular organisms appeared to have a preference for 6 and 12 transmembrane helix membrane proteins, whereas *C. elegans* and *H. sapiens* have a preference for 7 TM helix membrane proteins (Wallin and von Heijne, 1998). This discrepancy simply reveals the difficulties of studying

membrane protein structures due to lack of reliable methods to locate the membrane-spanning regions either computationally or experimentally.

Structure and Topology of Transmembrane Protein

Another frequently asked question about membrane protein research is ``Do transmembrane protein superfolds exist?" (Jones, 1998). To answer this question, we first have to look into the structures of membrane proteins. However, as mentioned in section 1.2.2, the membrane proteins present difficulties for analysis by X-ray crystallgraphy or NMR spectroscopy, due to their hydrophobic nature, low solubility and requirement of co-crystallisation with detergent molecules to form protein-detergent complex (von Heijne, 1994; von Heijne, 1996).

Many attempts and different biophysical techniques have been drawn into the studies of membrane protein structures. For example, neutron diffraction has been applied to investigate the rotational orientation of the seven transmembrane α-helices in bacteriorhodopsin (Samatey et al., 1994); solid state NMR is progressively used to determine membrane protein structures (Watts et al., 1995); and the structure of the KcsA potassium channel from *Streptomyces lividans* has been determined at 3.2 Å resolution (Doyle et al., 1998). However, despite these advances, the number of membrane proteins for which high resolution structures are known remain small, (Tab. 1.1).

From those membrane proteins with known structures, it has been identified that membrane-spanning structures can be built either from transmembrane helices (von Heijne, 1996), *e.g.* bacteriorhodopsin (Pebay-Peyroula et al., 1997), or from β-barrels, *e.g. E. coli* outer-membrane protein OmpA (Pautsch and Schulz, 1998). This study will mainly focus on α-helical transmembrane proteins, since the

majority of structure-solved membrane proteins are made of transmembrane α-helices (Smith, 1999). The details of β-barrel membrane-spanning protein will not be discussed and can be found elsewhere (Welte et al., 1995).

4.1.2 ABC Transporters.

Biology of ABC Transporters

The ABC proteins were first identified by two ATP-binding subunits, which have hydrophilic sequences and contain no signal sequences, and are a class of protein sharing related structures and functions (Higgins et al., 1986). Subsequently, a large group of proteins, which transport a wide range of solutes across cell membrane, were found to have one or two of the ATP-binding domains or ATP-binding component proteins of the systems. They were collectively sorted into the family of **A**TP-**B**inding **C**assette protein and were named as ABC transporters (Nikaido and Hall, 1998).

The ABC transporters make up one of the larger superfamilies of membrane transport proteins, and members of the ABC transporter family are widely found in prokaryotes and eukaryotes. They are responsible for influx or efflux of a wide range of solutes across cell membrane, from inorganic ions to carbohydrate metabolites and oligo-peptides (Higgins, 1995; Pao et al., 1998), (Tab. 4.1).

Species	Transporter	Solute	Function
E. coli	MalF, G	maltose	oligosaccharide importing
E. coli	RbsC	ribose	oligosaccharide importing
S. typhimurium	HisM, Q	histidine	amino acid importing
C. glutamicum	GluC, D	glutamate	amino acid importing
B. subtilis	DppB, C	dipeptide	dipeptide importing
E. coli	NikB, C	nickel	metal importing
S. marcescens	SfuB	iron	metal importing
E. coli	BtuC	vitamin B_{12}	vitamin importing
E. coli	HlyB	hemolysin	protein exporting
B. pertussi	CyaB	cyclolysin	protein exporting
S. fradiae	TlrC	tylosin	protein exporting
S. cerevisiae	Ycf1	Cd-glutathione	MRP[1]
S. cerevisiae	Pdr5	toxics	PDR[2]
H. sapiens	CFTR	Cl^- ion	ion transporting
H. sapiens	Tap1, 2	peptide	TAP[3]
H. sapiens	P-gp	toxins/drugs	MDR[4]
H. sapiens	Mrp1, 2	toxins/drugs	MRP[1]

[1]MRP: multi-drug resistance-associated protein

[2]PDR: pleiotropic drug resistance

[3]TAP: transporter associated with antigen processing

[4]MDR: multi-drug resistance

Table 4.1 Examples of ABC transporters from different species and their functions.

In prokaryotes, some ABC transporters are of great clinical implication because of their role in antibiotic resistance in micro-organisms (Saier Jr, 1998). In eukaryotes, several ABC transporters are of particular biomedical importance. These include the transporter associated with the genetic defect disease cystic fibrosis, the cystic fibrosis transmembrane conductance regulator (CFTR), details see Chapter 5; transporters associated with resistance of tumour cells to chemotherapeutic agents, the P-glycoprotein (P-gp); and the transporters associated with antigen processing and presentation, Tap1 and Tap2 (Higgins, 1995).

General Structures of ABC Transporters

Most ABC proteins share a common transmembrane topology, which is made up of two transmembrane helical domains and two cytoplasmic ATP-binding domains, (Fig. 5.1). In general, each transmembrane helical domain contains 6 transmembrane helices; however, there are still some exceptions which have integral-membrane domains of 5 membrane-spanning helices or 7 membrane-spanning helices (Hediger, 1994). These domains may all be part of a single polypeptide chain, *e.g.* the cystic fibrosis transport regulator (CFTR) protein; or the two groups of six TM helices may occur on distinct polypeptide components, as is the case in many bacterial transporters (Sarkadi, 1997).

Some prokaryotic ABC systems for influx may have additional extra-cytoplasmic solute-binding receptors which in Gram-negative bacteria is found in the periplasm, and in Gram-positive bacteria is present either as a lipoprotein, tethered to the external surface of the cytoplasmic membrane, or as a cell surface-

associated protein, bound to the external membrane surface *via* electrostatic interactions (Saier Jr, 1994).

A crystal structure at 1.5 Å resolution of the ATP-binding subunit of HisP, an ABC protein which transports histidine from *Samonella typhimurium* has been reported (Hung et al., 1998). Correlations of biochemical, genetic, and biophysical properties of this wild type protein with numerous mutant HisP proteins, mutant MDR proteins and mutant CFTR proteins were performed on the basis of this structure, which provides an insight for understanding properties of ABC transporters. However, the structure of the membrane-spanning region of an ABC transporter is yet to be determined. Thus, computational methods are employed in this chapter to study the transmembrane domains of ABC transporters.

4.1.3 Aims of This Study

In this study, different computational methods are used in order to:

- identify fold arrangements for TM helices of ABC transporters, and
- explore evolutionary and functional relationships between ABC transporters on the basis of their structural similarities.

A long-term goal of this study is to develop structure-based functional genomics methods for membrane proteins based on the strategy proposed here to investigate integral membrane protein sequences and structures computationally at the genomic level.

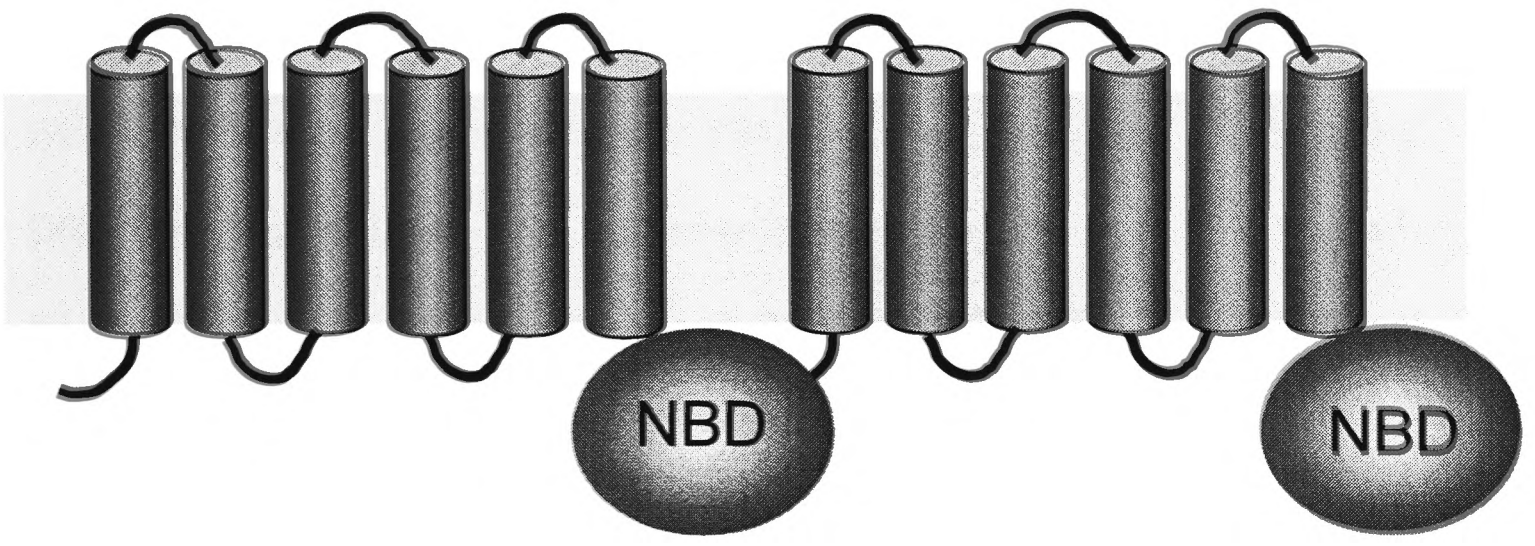


Figure 4.1 A putative structure of the ABC protein. In basic, it consists of two membrane-spanning domains, denoted by cylinders; and two ATP-binding domains, labelled with "NBD".

4.2 Methodology.

4.2.1 Sequence Database.

Introductions to Different Sequence Database

The first step to study a protein molecule by computational methods is to obtain the protein sequences of interest from sequence databases. There are a number of sequence databases which can be explored. Major nucleic acid sequence databases are collected and maintained at GenBank (NCBI) in USA (Benson et al., 1998), EMBL Nucleotide Sequence Database (EBI) in Europe (Stoesser et al., 1998) and DNA Data Bank of Japan (DDBJ) in Japan (Tateno et al., 1998), (Tab. 4.2). There are also two major protein sequence databases: SWISS-PROT (Bairoch and Apweiler, 1998); and PIR (Protein Information Resource) (Barker et al., 1998). SWISS-PROT, and it's supplement TrEMBL which contains all the translations of EMBL nucleotide sequence entries not yet integrated in SWISS-PROT, are designed to provide a high level of annotations for protein functions, domains and other information. The PIR-International Protein Sequence Database is a new joint venture between the Munich Information Center for Protein Sequences (MIPS) (Mewes et al., 1998), and the Japanese International Protein Sequence Database (JIPID) to create a comprehensive and well-verified database, which is also organised according to biological principles including structural, functional, and evolutionary relationships. In addition to these two major protein sequence databases, there is also a non-redundant composite of SWISS-PROT, PIR, GenBank (translation) and NRL-3D, called OWL, which is designed in accordance with a philosophy of providing a database which is "small and hence efficient in similarity searches" (Bleasby et al., 1994), (Tab. 4.2).

Nucleic Acid Sequence Database	
DDBJ	http://www.ddbj.nig.ac.jp/
EBI	http://www.ebi.ac.uk/Databases/index.html
GenBank	http://www.ncbi.nlm.nih.gov/
Protein Sequence Database	
SWISS-PROT	http://expasy.hcuge.ch/sprot/
PIR	http://www-nbrf.georgetown.edu/pir/
MIPS	http://www.mips.biochem.mpg.de/
OWL	http://www.biochem.ucl.ac.uk/bsm/dbbrowser/OWL/OWL.html

Table4.2 Sequence databases and their URLs.

Extractions of ABC Transporters from Different Genomic Databases

The simplest way to gather protein sequences of interest from various databases is probably by keyword searching. In the case of finding ABC transporter proteins in different organisms, one could submit the string "ABC transporter" to the search engine of a genome database in order to extract sequences of known ABC transporters from the given genome. However, the aim of this study is to analyse ABC transporters from different genomes; key word searching at a remote server is not applicable and not precise enough, which could cause ambiguities in further analysis.

Thus, entire translated sequences of different genomes were obtained from the public domains and were maintained locally. By October 1998, there were 17

organisms for which completed genomic sequences were available, (Tab. 4.3). These 17 genomes were employed in this study, which contains 3 archaea, 13 bacteria, and 1 eukaryote.

The ABC transporters were then extracted directly from translated sequences of different genomes on the basis of annotations listed in the gene function tables of each genome displayed in original publications, (Tab. 4.3). The extracted candidates were inspected manually in order to minimise possible errors. Yet there are still some unavoidable errors due to mistakes in the gene sequence annotations in different databases, see section 4.4.3.

Archaeoglobus fulgidus; **A; (Klenk and al., 1997)**
http://www.tigr.org/tdb/mdb/afdb/afdb.html

Methanobacterium thermoautotrophicum; **A; (Smith, 1997)**
http://www.biosci.ohio-state.edu/~genomes/mthermo/

Methanococcus jannaschii; **A; (Bult and al., 1996)**
http://www.tigr.org/tdb/mdb/mjdb/mjdb.html

Aquifex aeolicus; **P; (Deckert and al., 1998)**
http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/framik?db=Genome&gi=133

Bacillus subtilis; **P; (Kunst and al., 1997)**
http://www.pasteur.fr/Bio/SubtiList.html

Borrelia burgdorferi; **P; (Fraser and al., 1997)**
http://www.tigr.org/tdb/mdb/bbdb/bbdb.html

Escherichia coli; **P; (Blattner and al., 1997)**
http://www.genetics.wisc.edu/

Haemophilus influenzae; **P; (Fleischmann and al., 1995)**
http://www.tigr.org/tdb/mdb/hidb/hidb.html

Helicobacter pylori; **P; (Tomb and al., 1997)**
http://www.tigr.org/tdb/mdb/hpdb/hpdb.html

Mycobacterium tuberculosis; **P; (Cole and al., 1998)**
http://www.sanger.ac.uk/Projects/M_tuberculosis/

Mycoplasma genitalium; **P; (Fraser and al., 1995)**
http://www.tigr.org/tdb/mdb/mgdb/mgdb.html

Mycoplasma pneumoniae; **P; (Hilbert and al., 1996)**
http://www.zmbh.uni-heidelberg.de/M_pneumoniae/MP_Home.html

Pyrococcus horikoshii; **P; (Kawarabayasi and al., 1998)**
http://www.bio.nite.go.jp/ot3db_index.html

Rhizobium sp; **P; (Freiberg and al., 1997)**
http://genome.imb-jena.de/~cfreiber/pNGR234a2.html

Synechocystis PCC6803; **P; (Kaneko, 1996)**
http://www.kazusa.or.jp/cyano/cyano.orig.html

Treponema pallidum; **P; (Fraser and al., 1998)**
http://utmmg.med.uth.tmc.edu/treponema/tpall.html#tpgenome

Saccharomyces cerevisiae; **E; (Mewes et al., 1997)**
http://genome-www.stanford.edu/Saccharomyces/

Table 4.3 Genomic databases used in this study, their URLs and the original references. **P** denotes prokaryote, **E** denotes eukaryote and **A** denotes archaea.

4.2.2 Searching for Similar Sequences.

Introduction to Sequence Searching

Although, ABC sequences were extracted manually from genome databases in this study, the most reliable and efficient way to identify proteins that are related to a "target" protein is to search for those proteins that exhibit similarity of sequence with the target. These similar sequences are of great use for further analysis such as structure predictions and periodicity analysis. Several different algorithms for sequence database searching have been developed to achieve this goal. Among them, BLAST (Basic Local Alignment Search Tool) (Altschul et al., 1990) and FASTA (Pearson and Lipman, 1988) are probably the most widely used programs for exploring all of the available sequence databases. These two algorithms were developed on the basis of different concepts. However, they were both designed with the assumption that the best overall sequence similarities can be detected by local alignments (Pearson, 1996). Both BLAST and FASTA have been constantly improved and refined since their first releases. Both programs are widely used in different sequence databases, and available at public domains.

Some new capabilities have been implemented into the latest version of BLAST (Altschul et al., 1997). The newer Gapped BLAST allows sequence database search with gaps, i.e. deletions and insertions, which is not available in the original BLAST. Furthermore, the new PSI-BLAST iteratively searches sequence databases using a position-specific score matrix. This gives higher sensitivity, enabling one to detect weak but biologically relevant sequence similarities that may be missed by conventional pair-wise comparison. However, it is important to have a statistical measure of the significance of a weak similarity. Statistical measures of sequence similarity, like the *E* value of BLAST2, provide good

indicators of distantly homologous sequences. The newer version of FASTA package has been enhanced to provide statistical estimates for length-corrected sequence similarity scores, which is reported to be more effective than raw scores at identifying distantly related homologues (Pearson, 1998).

In addition to comparison algorithms, scoring matrices play an important role in determining the sensitivity of pair-wise comparison of two protein sequences. Two major series of scoring matrix are widely used: PAM (Point Acceptable Mutations per unit time); and BLOSUM (Best Local SUMmed alignment percentages).

The PAM matrix series were developed from observed amino acid substitutions frequencies by Dayhoff and co-workers (Dayhoff et al., 1978). Thus, the higher number denoted for each matrix, eg. PAM250, implies a higher degree of substitution. The BLOSUM matrix was developed by Henikoff and Henikoff (Henikoff and Henikoff, 1992) based on analysis of conserved regions in aligned protein sequence families. Therefore, a high denoted BLOSUM number, eg. BLOSUM62, implies a high degree of conservation.

Vogt *et al* (Vogt et al., 1995) analysed alignment accuracies as a function of matrix. In their study, PAM, BLOSUM and other matrices were used to align protein sequences which were also matched by superposition of known three dimensional structures. The comparisons showed relatively similar results for the PAM and BLOSUM series. This result suggests that no single matrix is optimal for all protein sequence comparisons. The default scoring matrices are BLOSUM62 (in BLAST) and BLOSUM50 (in FASTA).

Homologous Sequence Searching for ABC Transporters

After extraction of ABC transporters from different genomes, the next step is to search for homologous sequences. First, it is necessary to search a protein database in order identify related sequences of the target sequence. Then, evolutionary information of this particular protein can be obtained from those similar sequences and from their alignments.

SWISS-PROT version 35 is used as a template for sequence searching in this study. It was released in November 1997 and contained 69,113 protein sequence entries with 25,083,768 amino acid residues.

BLAST2 is employed in this study because of its sensitivity in finding remote homologous protein sequences. For example, BLAST2 has been reported to successfully detect distant relatives of mammalian LPS-binding and lipid transporter proteins (Beamer et al., 1998). The E-value cut-off was set at 0.001 in order to get a balance between searching sensitivity and accuracy. The default scoring matrix for BLAST2, BLOSUM62, is applied thoroughly in this study.

4.2.3 Multiple Sequence Alignment.

Introductions to Protein Multiple Sequence Alignments

Aligning a large set of N protein sequences is in principle the extension of pair-wise alignment to the alignment of N sequences simultaneously. In theory, this task can be achieved by generalisation of pair-wise dynamic programming to N-dimensional dynamic programming. However, due to finite CPU time and computer memory, such a simultaneous approach limits the maximum number of sequences in a set and the number of residues in each sequence to be compared. A

similar idea was implemented to develop the program, MSA, which used a rather sophisticated approach to reduce the volume of the N-dimensional dynamic programming matrix (Lipman et al., 1989). Even with this refinement, MSA is only suitable for up to eight protein sequences with average lengths of 200 to 300 residues. This limits the practical use of the truly simultaneous approach.

An alternative approach is to align multiple sequences progressively. The first step is to construct a succession of optimal pair-wise alignments by dynamic programming. The two most similar sequences are aligned first. Then, the next most similar sequence is added to the existing alignment. This process is iterated until all sequences have been aligned. This approach has been widely implemented in a number of multiple sequence alignment programs: PILEUP (in the GCG package); AMPS (Barton and Sternberg, 1987); MAXHOM (Sander and Schneider, 1991); and CLUSTALW (Thompson et al., 1994).

In addition to dynamic programming algorithms, other mathematical methods have been used to solve this problem of multiple sequence alignment. The genetic algorithm (GA) has been implemented in developing the program, SAGA (Notredame and Higgins, 1996). Hidden Markov Models have been implemented in the HMMER package for multiple sequence alignments (Eddy, 1995). By extension of HMMER, a database, Pfam, which contains multiple alignment and Hidden Markov Model based profiles to detect remote homologous sequences was constructed (Sonnhammer et al., 1997).

A recent study has been performed to compare different multiple sequence algorithms. In this comparison, different algorithms were used to align protein sequences which were also matched by superposition of known three dimensional

structures. The comparison results suggest that programs based on dynamic programming in general perform better than other methods (Raghava and Barton, 1998).

A list of programs for protein multiple sequence alignments is summarised at Table 4.4.

CLUSTALW	http://www2.ebi.ac.uk/clustalw/ http://www.clustalw.genome.ad.jp/ http://helix.nih.gov/science/clustalw.html
MAXHOM	http://dodo.cpmc.columbia.edu/pp/doc/ help_08.html#P8_method_maxhom
AMPS	http://barton.ebi.ac.uk/manuals/amps/amps.html
HMMER	http://hmmer.wustl.edu/
SAGA	http://mathbio.nimr.mrc.ac.uk/~cnotred/Projects_home_page/ saga_home_page.html

Table 4.4 Programs of protein multiple sequence alignments and their URLs.

Multiple Sequence Alignments of ABC Transporter

On the basis of this comparison, MAXHOM was used in this study to align protein sequences for further analysis in structure prediction and periodicity analysis; and ClustalW was used for phylogene analysis.

Sequence Analysis by MAXHOM.

After generating a long list of hit sequences *via* similarity searching for each ABC transporter, by measuring the statistical significance with an E value cut-off of

0.001, homologous sequences are extracted from the SWISS-PROT 35 automatically. Extracted protein sequences are gathered together, and then aligned by MAXHOM with the default parameter in order to detect the amino acid conservation patterns for further analysis of structure predictions and periodicity analysis.

Sequence Analysis by ClustalW.

Sequence comparisons were carried out using ClustalW with default parameters and BLOSUM protein weight matrix in order to detect phylogenetic relationships between different ABC transporters within each individual genome. Unrooted phylogenetic trees were plotted on the basis of the sequence comparison results for every genome using "DRAWTREE" in PHYLIP package (Baum, 1898).

4.2.4 Prediction of Transmembrane Helices.

Introduction to Transmembrane Helix Prediction

Prediction of TM helices is a crucial step in computational analysis of membrane protein structure. Knowledge gained from different membrane protein structures and the advances in membrane protein structure determinations facilitate and enhance many computational methods available to predict membrane protein structures and topologies directly from their sequences. It is generally believed that prediction of TM helices is easier than prediction of secondary structures in globular proteins, since the lipid-bilayer provides extra restraints upon the sequences of individual TM helices and on the topology of a membrane protein as a whole (Taylor et al., 1994). Many attempts have been made to develop

algorithms for locating putative TM helices either from single sequences or from multiple alignments of protein sequences.

A number of algorithms have been developed to predict the location of TM helices within a protein sequence. Early methods simply searched for runs of about 21 predominantly hydrophobic amino acids in a sequence, *i.e.* a sequence element of sufficient length to form a hydrophobic α-helix able to span about 30 Å thick hydrophobic core of a lipid bilayer (Kyte and Doolittle, 1982; Engelman et al., 1986). This resulted in much discussion of the optimum hydrophobicity scale for use in prediction of transmembrane proteins based on the sequence hydrophobicity pattern (Cornette et al., 1987).

More recent work has analysed the sequences of known transmembrane helices in order to develop more sophisticated algorithms based on pattern recognition of hydrophobicity or known structures for detecting transmembrane helix. A number of such methods have been developed, all of which require a single query sequence as input.

Based on the observations of higher residue hydrophobicity exhibited by transmembrane helices (Kyte and Doolittle, 1982) and the more recurrent distributions of positively charged amino acid residues within cytoplasmic loops (von Heijne, 1992), it becomes possible to deduce likely transmembrane topologies of a membrane-spanning protein using amino acid analysis of hydrophobicity and charge distribution along a protein sequence. These ideas have been implemented to develop a prediction program, TOPPRED2, in order to locate putative helices from a single sequence (von Heijne, 1992). On the basis of data derived from well-studied transmembrane proteins and application of constrained

dynamic programming, MEMSAT was designed to predict both the location and orientation, in-out *vs.* out-in, of transmembrane helices (Jones et al., 1994). DAS was developed by using a dense alignment surface method, which employed low-stridency dot-plots of the query sequence against a collection of non-homologous membrane proteins to locate putative helices (Cserzo et al., 1997). An amphiphilicity index and two sets of physical-chemical indices were applied in designing SOSUI to detect putative helices (Hirokawa et al., 1998). The Hidden Markov Model was also implemented to predict location and orientation of putative transmembrane helices. This method, TmHMM, avoids using a fixed hydrophobicity cutoff. Thus it avoids a procedure which has the disadvantage of underestimation of potential TM helices (Sonnhammer et al., 1998).

Predictions of protein secondary structure are more accurate if evolutionary information is considered using multiply aligned related sequences (Russell and Barton, 1993). This also seems to apply to prediction of transmembrane helices. TMAP (Persson and Argos, 1994) uses calculations of hydrophobicity propensity and terminal polar propensity for each aligned residue, allowing prediction of TM helices from multiply sequence alignments. PHD is probably the most used program for protein secondary structure prediction. Two sub-routines of PHD program, PHDhtm and PHDtop, take multiple alignments of protein sequences as input to a neural network and then apply dynamic programming to refine the raw prediction results from the neural network in order to have a more accurate prediction of transmembrane protein structure and topology (Rost et al., 1995; Rost et al., 1996).

A list of TM helix prediction methods and associated websites is given in Table 4.5. The overall accurate rate of most TM helix prediction methods is about 80% (von Heijne, 1996). However, if the same protein sequence is analysed using a range of different algorithms, subtle differences in the resultant predictions are obtained, especially in terms of identification of the precise start and end residues of each TM helix. This possible source of error may be important in subsequent modelling and simulation studies. A comparison of some TM helix prediction programs suggested that in general PHDhtm yielded better prediction results (Sonnhammer et al., 1998). However, it was noticed that PHDhtm tends to predict the TM helices to be shorter than they are shown to be when structural data are available, which is an artifact of shorter helical length cut-off set by the author (Rost et al., 1996), (Rost, personal communication). Even so, this program is reasonably good at identifying the location and orientation of TM helices.

DAS	http://www.biokemi.su.se/~server/DAS/
TOPPRED2	http://www.biokemi.su.se/~server/toppred2/
MEMSAT	http://globin.bio.warwick.ac.uk/~jones/memsat.html
SOSUI	http:// www.tuat.ac.jp/~mitaku/ady_sosui/
TmHMM	http://www.cbs.dtu.dk/services/TMHMM-1.0/
TMAP	http://www.embl-heidelberg.de/tmap/tmap_info.html
PHD	http://www.embl-heidelberg.de/predictprotein/predictprotein.html

Table 4.5 Summary of different TM helix prediction programs and their URLs.

To introduce a note of caution, examine Figure 4.2 in which the results of seven different prediction methods applied to bacteriorhodopsin are compared with one another and with the secondary structure assignment obtained by DSSP analysis (Kabsch and Sander, 1983) of the 2.5 Å resolution X-ray structure of this protein (Pebay-Peyroula et al., 1997). It is clear that no single method outperforms the others for all seven TM helices. Whilst not pretending that this is a systematic comparison of available methods, it does reinforce the message that TM helix predictions should be treated in the same critical fashion as the results of any other analysis, computational or experimental. It should be remembered that current prediction methods are based on a relatively small database of experimental three dimensional structure for membrane proteins. It is likely that improvements in prediction accuracy may occur as this database continues to expand.

Predictions of TM Helices for ABC Transporters

The ABC transporter sequences were submitted, either as single sequences (to DAS, MEMSAT and TmHMM) or as multiple sequence alignments against homologous sequences from MAXHOM analysis (to TMPPRED and PHD). The prediction results of each program were analysed. A protein sequence is rejected if it is predicted as a non-transmembrane protein by any one of the methods.

Prediction results used in this study were mainly from PHDhtm. However, prediction results for some transmembrane proteins were missed because of failures in generating multiple sequence alignments. Under this condition, prediction results from TmHMM would be employed.

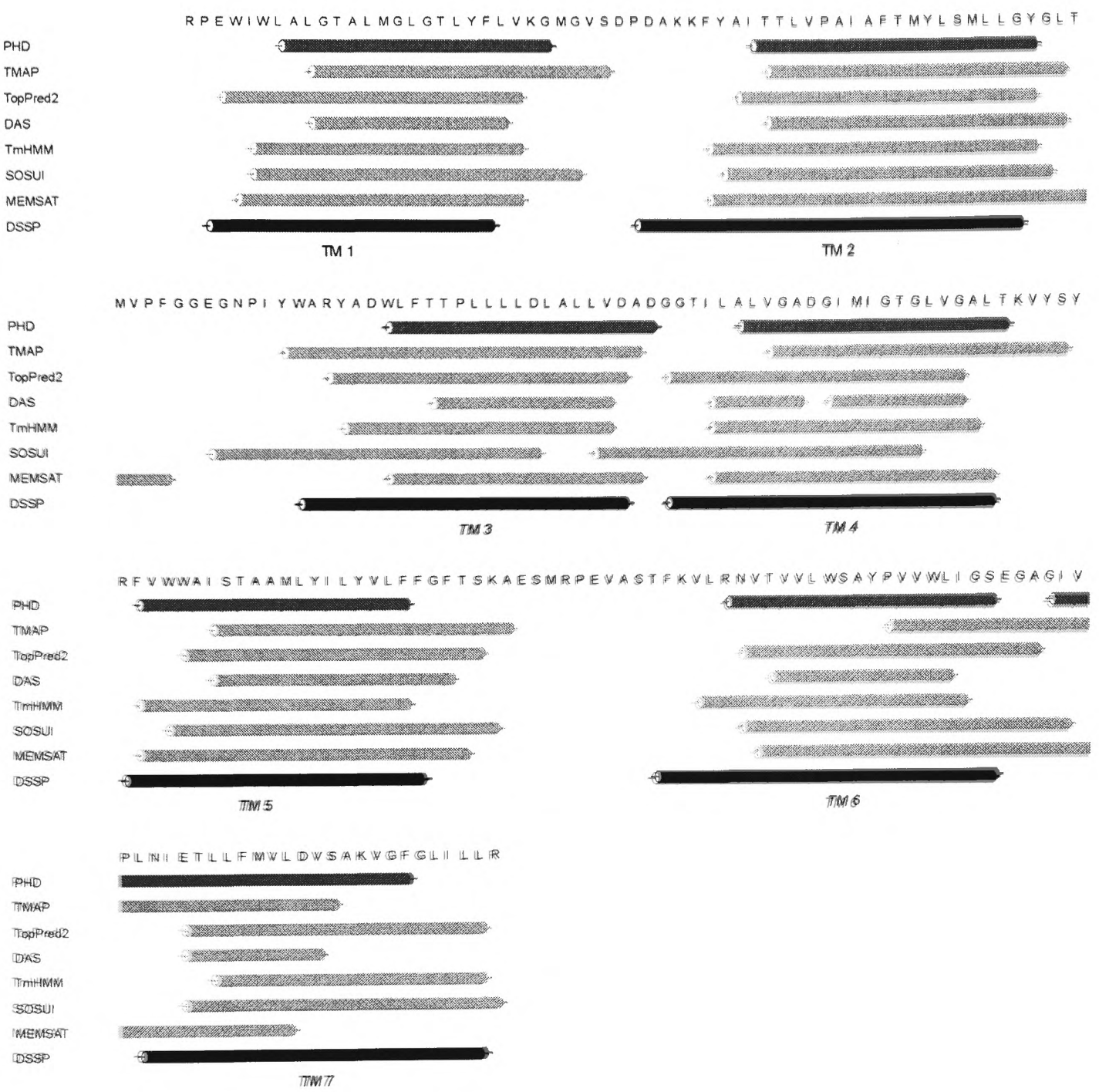


Figure 4.2 Predictions of transmembrane helices by different algorithms. Different programs tend to give different predictions. Comparing with the structure-solved bacteriorhodopsin, PHDhtm has an advantage in locating the cores of TM helices.

4.2.5 Periodicity Analysis.

Introductions to Periodicity Analysis

Having defined the locations within the sequence of putative transmembrane helices, and their orientations relative to the cell membrane, the next step is to predict which residues of the helix face the surrounding lipid and which point inwards to the centre of the helical bundle.

Early work on the photosynthetic reaction centre from *Rhodobacter sphaeroides* indicated that interior-facing residues of transmembrane regions are more hydrophilic than membrane-exposed residues (Yeates et al., 1987; Komiya et al., 1987). Similar results were found in the putative transmembrane regions of related species of *Rhodobacter sphaeroides* (Rees et al., 1989). This work was extended to G-protein coupled receptors and revealed that membrane-exposed residues are more variable and less conserved than interior-facing residues of transmembrane regions (Donnelly et al., 1989).

Based on these observations, tables of hydrophobicity and of variability/conservation of amino acid residues along a transmembrane helix were constructed. These parameters of conservation/variability and hydrophobicity provide a better estimation of which residues are facing the lipid-bilayer and which are facing the membrane interior for predicted transmembrane helices, and thus serve as better indications of the spatial arrangements of putative helices within lipid-bilayer. By using Fourier transformation analysis to filter out noise and to discern sequence patterns, a package, PERSCAN, enables one to detect the periodicities of residue conservation and hydrophobicity of a transmembrane helix by sequence alignments (Donnelly et al., 1993).

Periodicity Analysis of ABC Transporters

Periodic patterns of residue properties along predicted TM helices were analysed using the PERSCAN package (Donnelly et al., 1993). Considering the evolutionary information by using multiple protein sequence alignments from MAXHOM analysis as input, section 4.2.3, and then employing Fourier transform power spectrum to analyse this information, the PERCON module detects the periodicity of residue conservation whereas PERHYD detects periodicity of hydrophobicity. The analysis results of hydrophobicity and amino acid residue conservation for each helix were recorded and sorted for comparisons with helices at similar positions from homologous ABC transporters.

In general, interior residues of a TM helix tend to be less hydrophobic and more conserved; while lipid-facing residues of a TM helix tend to be more hydrophobic and less conserved. Thus, transporters sharing similar spatial locations and arrangements for their TM helices will show similar patterns of hydrophobic periodicity and conservation periodicity.

PERCON: periodicity of residue conservation.

PERHYD: periodicity of residue hydrophobicity

PERMUT: periodicity of lipid substitution parameter

PERVAR: periodicity of residue variability

Table 4.6 Modules in PERSCAN package for detecting different patterns of periodicity.

4.2.6 α-Helical Structure Model Building.

Introduction to Molecular Modelling

Helical models are built by the molecular dynamics simulations incorporating the general procedure of simulated annealing (Kirkpatrick et al., 1983) in order to explore possible side chain conformations by performing high temperature molecular dynamics, simulated annealing/molecular dynamics (SA/MD). This technique is used extensively in Chapter 4 and Chapter 5, thus details of molecular dynamics simulations will be given in Appendix.

Building Idealised α-helical Model for ABC Transporter

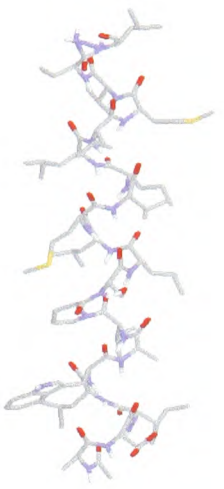
The ideal α-helical structure is constructed on the basis of the putative α-helix located by PHDhtm. Models are build by a two-stage SA/MD protocol (Kerr et al., 1994) using Xplor (Brünger, 1992), details see Appendix.

The α-helices generated by the SA/MD method are all in the N-acetylated and N-amidated form, Ac-HELIX-NH_2, in order to mimic more closely the backbone of a protein structure.

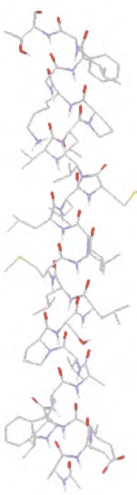
The structures of the TM helix 3 of ABC transporter OppC in *B. subtilis* generated from different stage of the SA/MD protocol are shown to illustrate the progress of model building, (Fig. 4.3).



Stage 0



Stage 1



Stage 2

Figure 4.3 The progress of helical structure building for the TM helix 3 of *B. subtilis* OppC transporter. Structures were obtained after each stage of SA/MD protocol, for details see section A.1.3 and Fig. A.1.

4.2.7 Methods Integrated in This Study

To obtain a clear view of methods used in this study, A flow diagram of methodology is summarised, (Fig. 4.4).

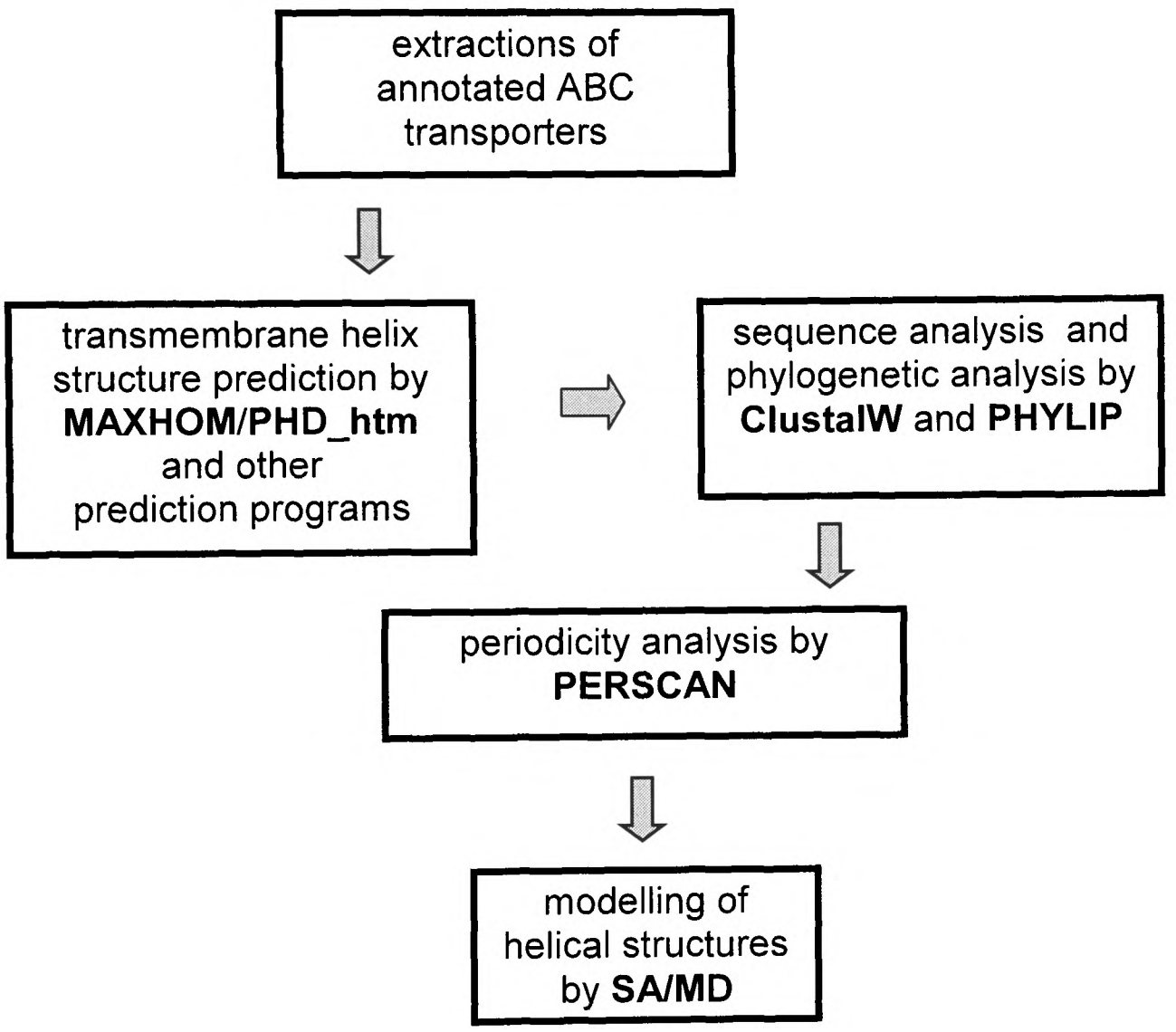


Figure 4.4 Flow diagram summarising methods used in this study.

4.3 Results

4.3.1 Extraction and Prediction of ABC Transporters.

The ABC proteins were extracted from translated open reading frames (ORFs) of different genome databases maintained locally on the basis of their published functional tables. This is to ensure a method of extracting proteins in an accurate and objective fashion.

Not all of these proteins were integral membrane proteins. In order to identify the transmembrane proteins, different algorithms for membrane protein structure prediction were applied. Since some of these TM helix prediction programs require multiple sequence alignments, for each extracted sequence, homologous sequences were identified using BLAST2 to search against the SWISS-PROT 35 sequence database. Results of sequence similarity searching by BLAST2 were further refined by MaxHom, with the E value cut-off set at 0.001. Multiple sequence alignments were generated for all potential ABC transporters. The multiple sequence alignment for each sequence was submitted to TMAP and PHDhtm; each single sequence was submitted to MEMSAT, DAS and TmHMM. Only those sequences predicted by all methods to contain multiple TM helices were retained for further analysis.

This protocol of TM protein identification reduced sequence sizes on each genome set. For example, the sequence size of *B. subtilis* was reduced from 144 sequences to 47. Many of the excluded sequences corresponded to soluble periplasmic binding proteins that are associated with ABC transporters. Details of the quantity of ABC transporter sequences are shown, (Tab. 4.7).

Species	Sequence Size		
	ORF[1]	ABC protein	TM ABC Transporter
Archaeoglobus fulgidus	2409	79	30
Methanobacterium thermoautotrophicum	1869	28	17
Methanococcus jannaschii	1715	26	17
Aquifex aeolicus	1522	24	12
Borrelia burgdorferi	850	35	16
Bacillus subtilis	4100	144	47
Escherichia coli	4290	80	75
Haemophilus influenzae	1709	74	29
Helicobacter pylori	1577	34	15
Mycoplasma genitalium	467	26	13
Mycoplasma pneumoniae	677	34	17
Mycobacterium tuberculosis	3924	35	10
Pyrococcus horikoshii	1979	53	21
Rhizobium sp	417	24	13
Synechocystis PCC6803	3168	41	40
Treponema pallidum	1031	40	14
Saccharomyces cerevisiae	6296	26	21

[1]ORF: open reading frame

Table 4.7 Numbers of protein sequences at genomic levels, extracted ABC proteins and predicted ABC transporters with transmembrane domains from different genomes.

4.3.2 Characterisations of Each Genome.

Those ABC protein sequences which were predicted to contain membrane-spanning domains, (Tab. 4.7), were gathered together on the basis of individual genomes. Sequence comparisons were carried out using ClustalW. A phylogenetic tree was constructed on the basis of the sequence comparison results for each genome using "PHYLIP". In general, patterns of either two-branched or three-branched trees were observed.

Some possible correlations between sequence-generated clusters and putative transmembrane helices were detected, which include numbers of TM helices and locations of the putative membrane-spanning helical segments. It is generally observed that proteins within the same clusters are likely to have same numbers of putative integral membrane helices, or at least similar amounts of TM helices. In addition, it is also noted that proteins within the same clusters tend to have similar arrangements of location for predicted TM helices along the primary protein sequences

It is also noted that transporters within the same clusters are more likely to mediate transportation of solutes with similar chemical and biological properties across the cell membrane. In other words, there are possible correlations between sequence-generated clusters and transport functions.

Bacillus subtilis

The 47 extracted ABC transporters in the *B. subtilis* were clustered into a three-branched phylogenetic tree, (Fig. 4.8).

In Cluster 1, most of the proteins transport small peptides across cell membrane. For example, OppC, AppC and AppB are oligopeptide transporters; DppC, and DppB are dipeptide transporters. OpuAB is a transporter of glycine and betaine. All proteins in cluster 1 have 6 putative TM helices; except for OpuAB, which was predicted to have 7 TM helices.

In Cluster 2, CydC and CydD are the only two proteins whose functions have been identified; they are required for expression of cytochrome *bd*. However, the majority of proteins in this cluster were predicted to have 4 TM helices; which include YwjA, YfiB, YvcC, YheI, YknV, and YfiC. Three proteins, CydD, CydC and YgaD, were predicted to have three TM helices. YknU only has 5 putative TM helices. In addition, nucleotide-binding domains were detected in proteins in this cluster.

Functions of most of the proteins in Cluster 3 remain to be identified. However, four proteins in Cluster 3 of known function are transporters for basic solutes; OpuBB and OpuBD for choline; OpuCB and OpuCD for glycine, beatine, carnitine and choline. The numbers of predicted TM helices for proteins in this cluster were largely five.

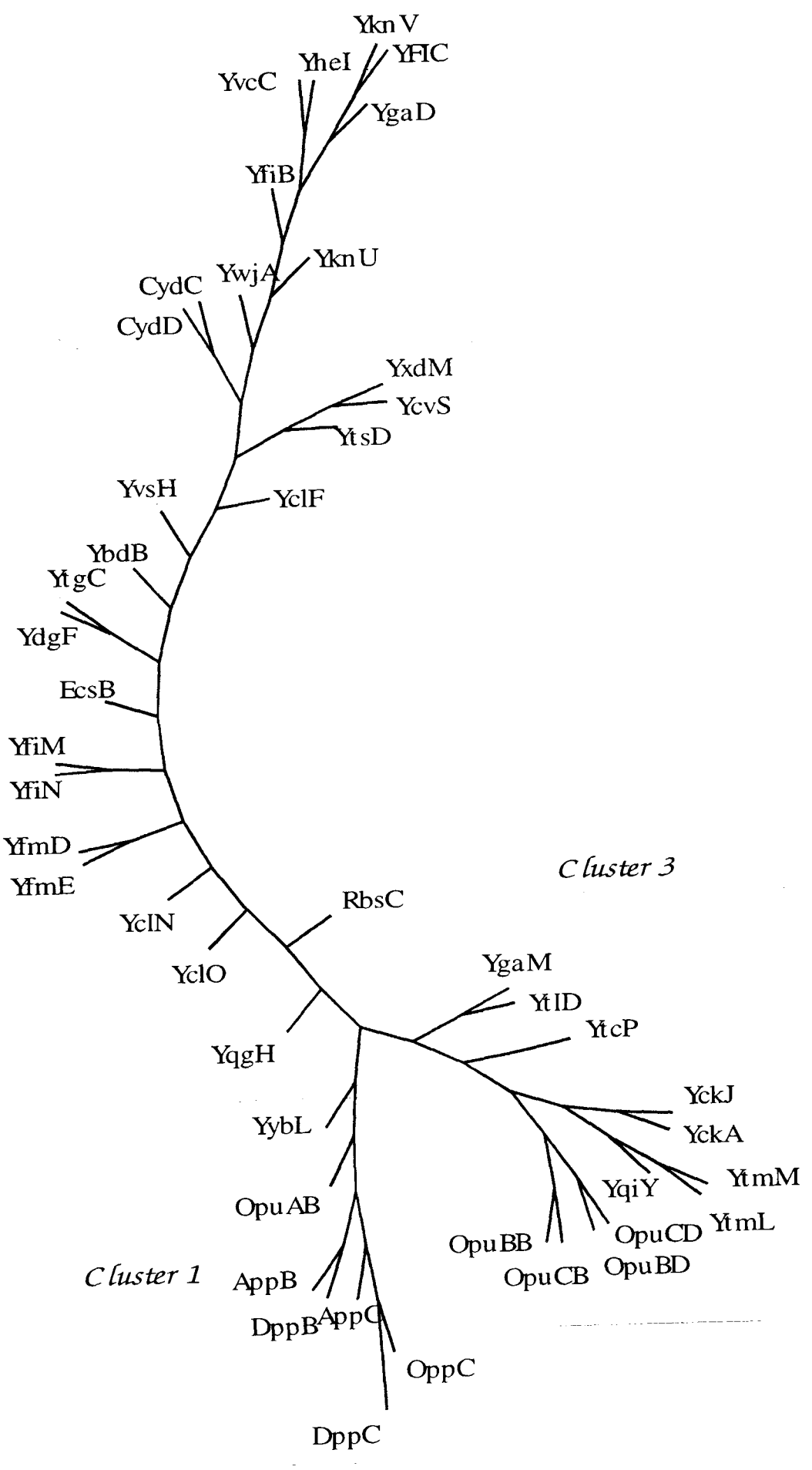


Figure 4.8 *Bacillus subtilis.*

Similarities in locations of putative TM helices along the protein sequences were also detected in ABC transporters of *B. subtilis*. For example, proteins in cluster 1 can be further categorised into two sub-clusters; DppC, OppC and AppC for cluster 1.1 and DppB and AppB for cluster 1.2. It was observed that proteins in both sub-cluster have similar arrangements of putative TM helices along their primary sequences. In addition, similar arrangements of TM helices were detected between the two sub-clusters.

The similarities on linear arrangements of putative TM helices can be easily observed by viewing their multiple sequence alignments. For proteins DppC, OppC and AppC, it can be seen that their 6 putative TM helices share highly conserved locations along their primary sequences, (green segments in Fig. 4.9). For proteins DppB and AppB in sub-cluster 1.2, similar arrangements of TM helices were observed as well, (orange segments in Fig. 4.9).

Similarities were also detected between sub-cluster 1.1 and sub-cluster 1.2, which included TM helices 2 for position 120 to 140, TM helices 5 for position 260 to 280, and TM helices 6 for position 315 to 330, (Fig. 4.9). These three TH helical regions could be regarded as cores of TM helix for proteins within Cluster 1 in *B. subtilis*. This implies that proteins within similar sequence-generated clusters may share spatial arrangements of TM helices in the lipid bilayer.

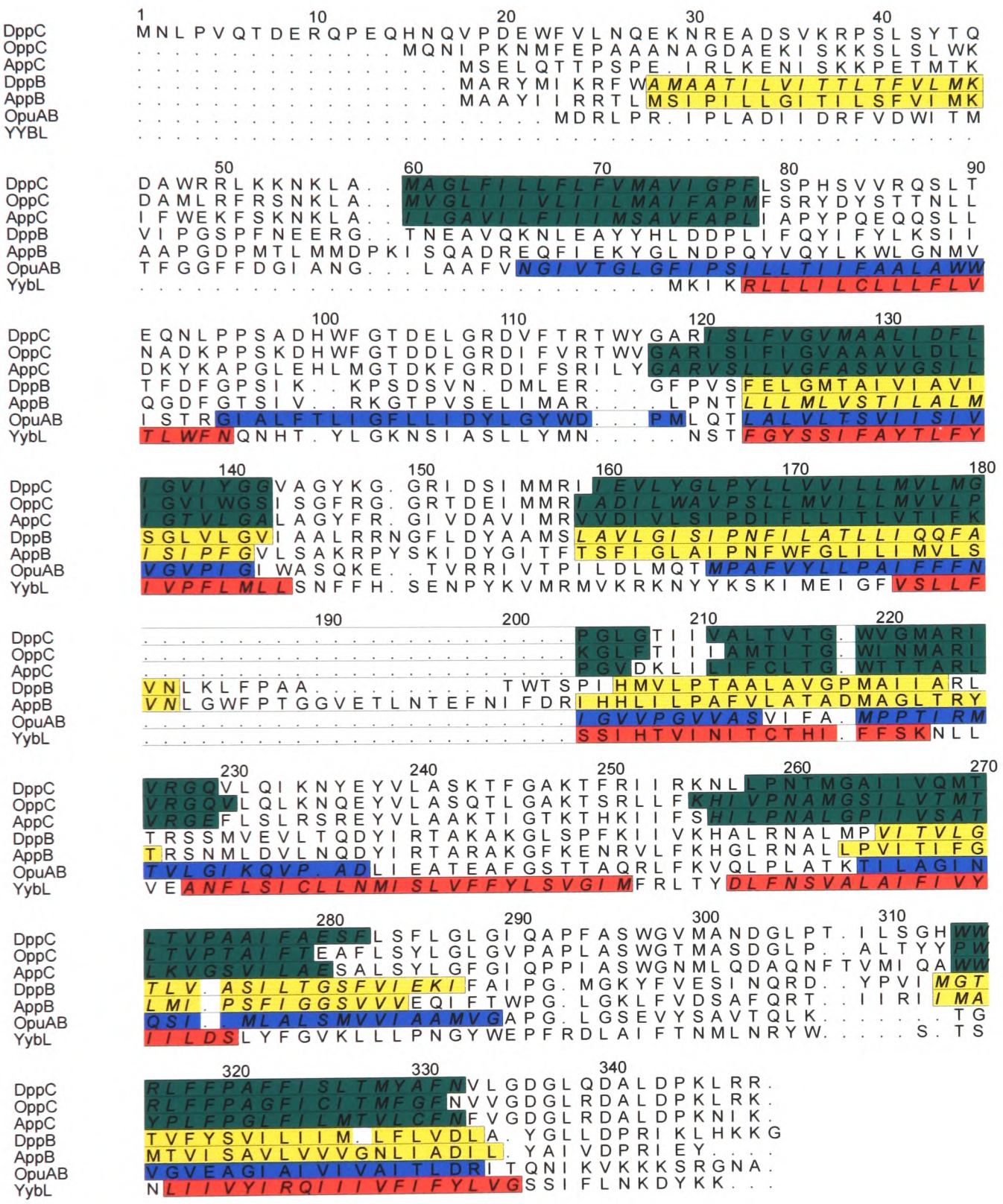


Figure 4.9 Putative helical structure arrangements of transporters in cluster 1 of *B. subtilis*. Two sets of proteins were observed with similar locations of TM helices along their primary sequences; green for DppC, OppC and AppC, and orange for DppB and AppB. In addition, conserved locations of TM helices were detected among these 7 proteins at TM helices 2 (120-140), 5 (260-280), and 6 (315-330).

Aquifex aeolicus

A two-branched tree is observed for the phylogentic analysis for *A. aeolicus*, (Fig. 4.10).

In cluster 1, gi2983693 and gi2983712 were predicted to have 8 and 11 TM helices respectively. For the biological function, gi2983693 transports ferrous iron and gi2983712 transports sulfate.

The numbers of putative TM helices for the cluster 2 of *A. aeolicus* are 6 for gi2983575 with unknown function, 5 for gi2982839 which is a molybdenum transporter, 6 for both gi2984255 and gi2984254 which are phosphate transporters. In addition, gi2984255 and gi2984254 do not only share similarities in the numbers of their TM helices and their functions, their first three putative TM helices have similar locations along the primary sequences as well.

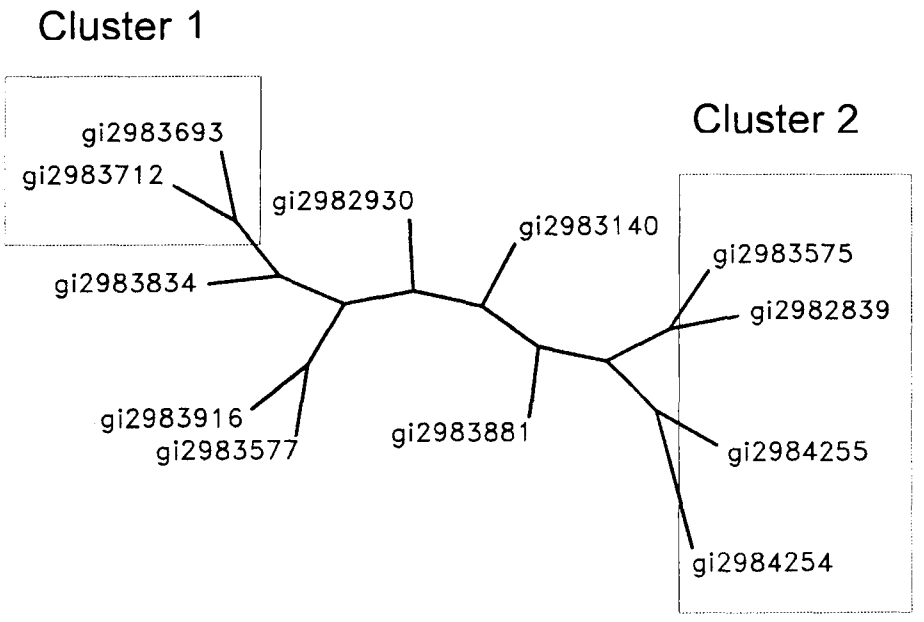


Figure 4.10 *Aquifex aeolicus*.

Borrelia burgdorferi

A curved two-end tree was identified in the phylogenetic analysis of *B. burgdorferi*, (Fig. 4.11).

Two proteins in cluster 1 mediate transportation of solutes, which are related to nitrogen compound metabolism; both BB0640 and BB0641 are transporters of spermidine and putrescine. BB0640 was predicted with 8 TM helices, and BB0641 was predicted with 7 TM helices. Another protein in this cluster, BB0746, was predicted to have 5 TM helices and this protein is likely to be a transporter of oligopeptide.

Proteins in cluster 2 of *B. burgdorferi* are all responsible for transportation of sugars. Both BB0408 and BB0629 are fructose transporters. Both BB0116 and BBB29 are transporters of maltose and glucose. BB0645 is glucose transporter.

The two fructose transporters, BB0408 and BB0629, were predicted to have 8 TM helices. The maltose/glucose transporters, BB0116 and BBB29, were predicted to have 11 and 10 TM helices respectively. BB0645 has 10 putative TM helices. The numbers of putative helices are similar for proteins in this cluster. For the two fructose transporters, they not only have the same putative TM helices, the linear arrangement of the putative TM helices along their primary sequences are similar as well.

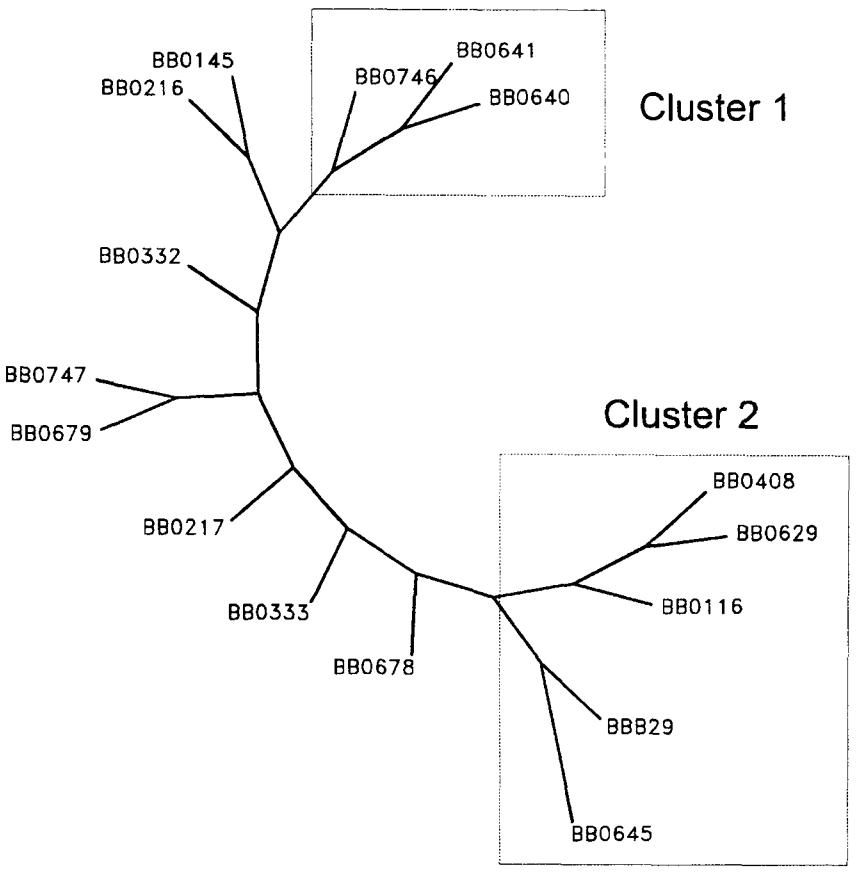


Figure 4.11 *Borrelia burgdorferi*

Escherichia coli

The phylogenetic analysis of ABC transporters of *E. coli* revealed a complicated pattern, (Fig. 4.12). In general, three major clusters were identified. However, proteins in both cluster 2 and 3 can be further subdivided into two parts. Thus, ABC transporters in this genome were divided into five parts.

Proteins in cluster 1 are mainly responsible for transportation of sugar. P23200 is a galactoside transporter. P08532 and gi11788896 are ribose transporters. P04984 and gi1742491 are arabinose transporters. P37389 is a xylose transporter. Only P45767 is a transporter of amino acids. The numbers of predicted TM helices

range from 6 to 10. Structural similarities on putative TM helices in the context of number or locations on primary sequence among proteins in cluster 1 were not observed.

Most proteins in cluster 2.1 were predicted to have 6 TM helices, except for P10906. Proteins P77315 and Q47824 are dipeptide transporters. They share strong similarities on the arrangement of TM helices along protein sequences. These similarities were not detected in protein P33592, a nickel transporter, P07622, a maltose transporter and P10906, a *sn*-glycerol-3-phosphate transporter.

Proteins in cluster 2.2 are all transporters of amino acids. Both proteins P30861 and P30862 are arginine transporters with 6 putative TM helices. Both proteins P20991 and P52094 are transporters of histidine, and have 6 putative TM helices. Both proteins P41074 and P41075 are transporters of glutamate/aspartate, and have 6 and 5 putative TM helices respectively. Protein P10345 is a glutamine transporter and has 6 predicted TM helices. These proteins show similarities on their solutes, numbers of predicted TM helices, and the arrangements of predicted TM helices along the primary sequences.

Physiological functions of proteins in cluster 3.1 are largely unknown. These proteins were predicted to have 5 TM helices for P37624, gi1773133 and P27624.

The solutes for proteins in cluster 3.2 are diverse. The numbers of predicted TM helices for proteins in this cluster are different as well. It is noted that protein P06972, which is a ferrichrome transporter, has 18 putative helices. This unusual abundance of TM helices for an ABC transporter is likely to reflect the large and rigid solute of ferrichrome.

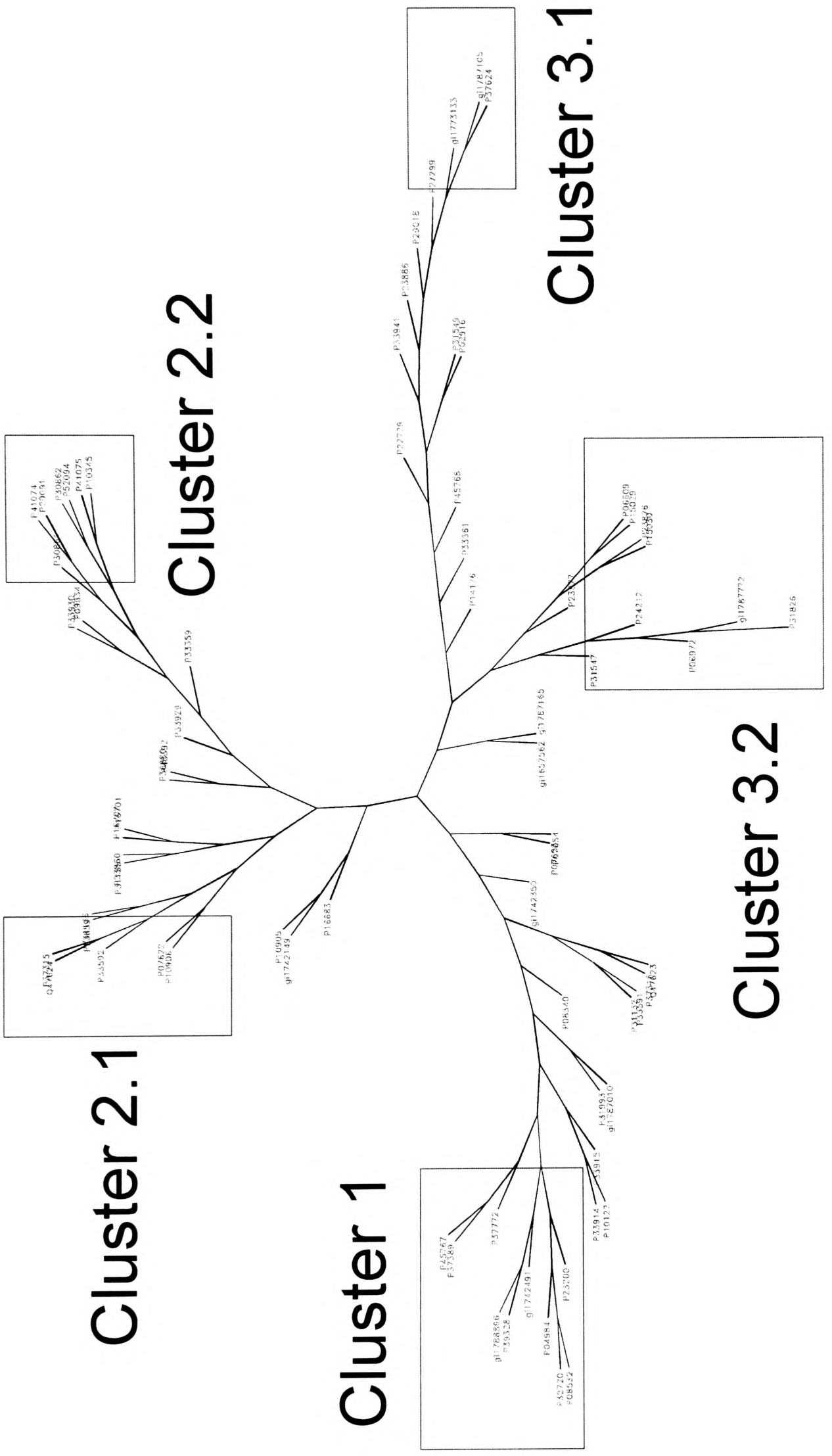


Figure 4.12 *Escherichia coli.*

Haemophilus influenzae

The phylogenetic analysis of ABC transporters in the *H. influenza*, revealed a curved two-end tree, (Fig. 4.13).

The solutes of transporters in cluster 1 are mostly unidentified. Structurally, these five proteins all contain the nucleotide-binding domains, and their membrane-spanning domains are composed with either 5 putative TM helices (HI1051, HI0060, and HI0036), or 6 putative TM helices (HI0664 and HI1467).

The four proteins in cluster 2 were predicted to have 6 TM helices. Most of them are transporters of positive amino acids, HI1178, HI1177 and HI1079; except for HI1692 is a molybdenum transporter. The linear arrangements of putative helices for the transporters HI1177 and HI 1178 are similar.

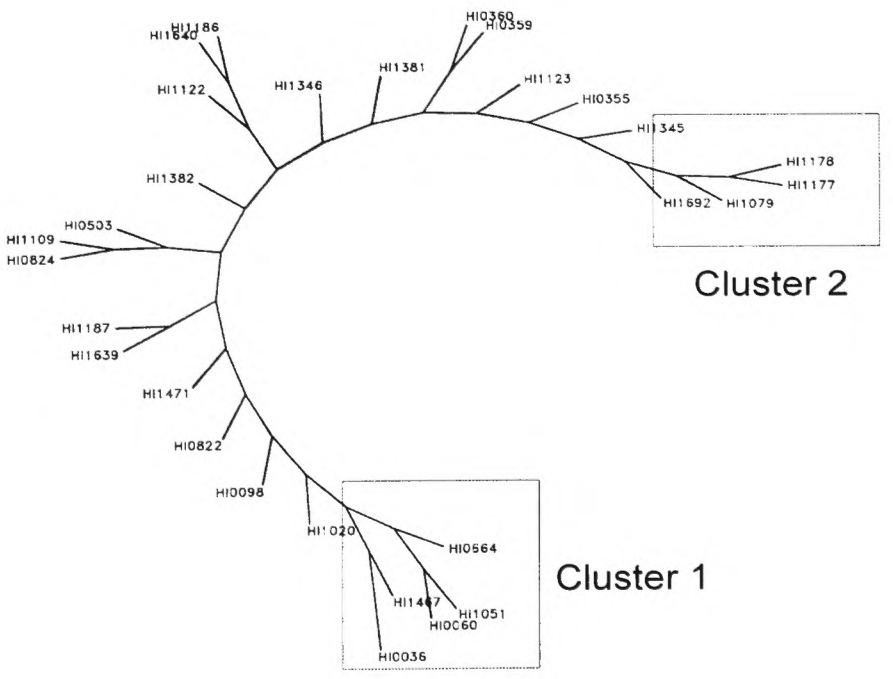


Figure 4.13 *Haemophilus influenza*

Helicobacter pylori

The evolutionary relationships between the ABC transporters of the *H. pylori* revealed a two-end tree, (Fig. 4.14),

The three proteins in cluster 1 were predicted to have 6 TM helices. They have similar solutes as well; transporters HP1170 and HP1169 for glutamine, and transporter HP0939 for amino acids.

The four proteins in cluster 2 are all related to multi-drug resistance. Even though they have similar functions, the numbers of their predicted TM helices diversify; HP1082 with 5, HP1206 with 6, HP0600 with 4 and HP1181 with 12.

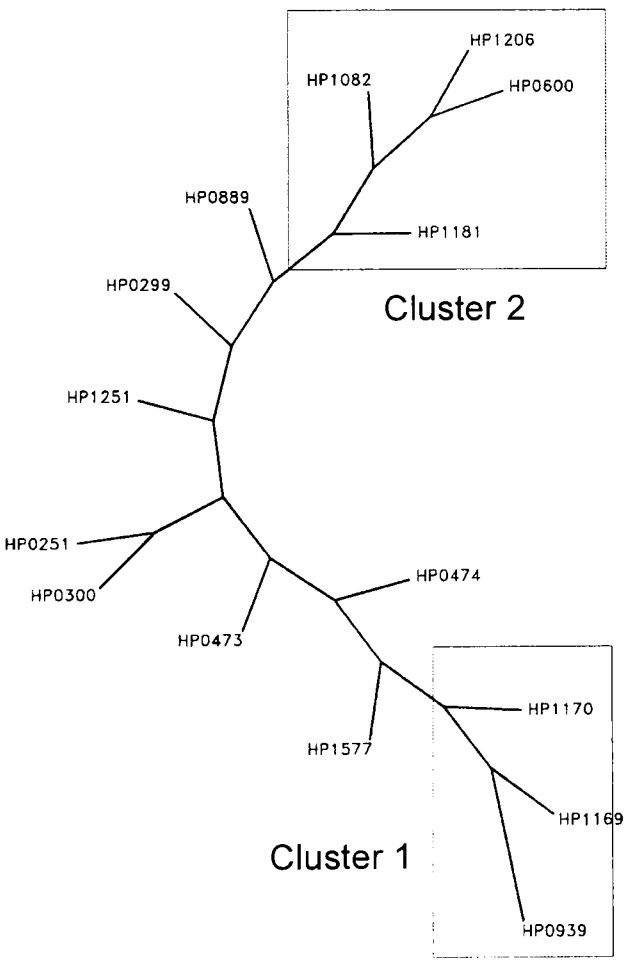


Figure 4.14 *Helicobacter pylori.*

Mycoplasma genitalium

A curved two-branched phylogenetic tree was detected in ABC transporters of *Mycoplasma genitalium*, (Fig. 4.15).

Three proteins in cluster 1 were predicted to have 6 TM helices. However, only MG043 and MG044 have known functions, which transport spermidine and putrescine.

Neither functional similarities nor structural similarities were detected for proteins in cluster 2. MG411 is a phosphate transporter, with 13 putative TM helices; MG069 is a glucose transporter, with 10 putative TM helices; MG468 is functionally unknown, with 5 putative TM helices.

Two proteins, MG014 and MG015, were located at the curve-transition position of the phylogenetic tree. Their functions remain to be identified; however, both of these two proteins have 6 putative TM helices. In addition, their predicted helical arrangements are similar along their primary sequences.

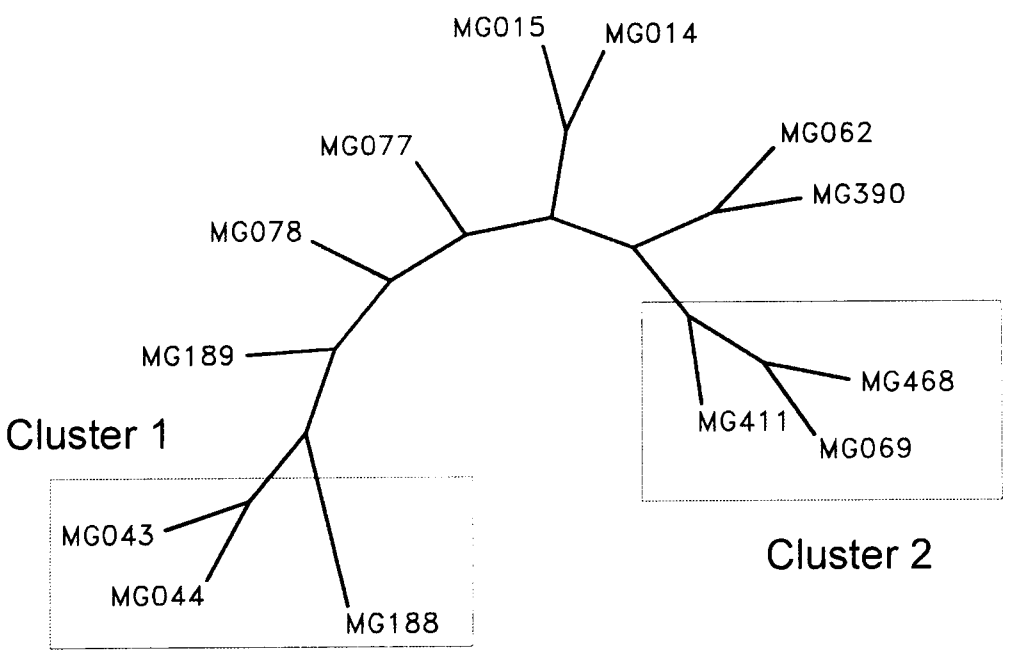


Figure 4.15 *Mycoplasma genitalium*.

Mycoplasma pneumoniae

For ABC transporters in the *M. pneumoniae*, their phylogenetic analysis revealed a three-branched tree, (Fig. 4.16).

The four proteins in cluster 1 were all predicted to have 6 TM helices. Two proteins are transporters of *sn*-glycerol-3-phosphate, E07_orf329 and E07_orf319; other two proteins are transporters of spermidine and putrescine, D09_orf286A and D09_orf286B. High degrees of similarities in their predicted TM helix locations were detected between proteins at each pair.

Proteins in cluster 2 showed no similarities in either their functions or structures.

Three proteins were identified in cluster 3, F10_orf491, F10_0rf503 and A05_orf542. The first two proteins are likely to be transporters of amino acids, and have 12 putative helices. Their putative helices located on similar positions along their primary sequences as well. Protein A05_orf542 has 11 putative TM helices; however, its function is still unknown.

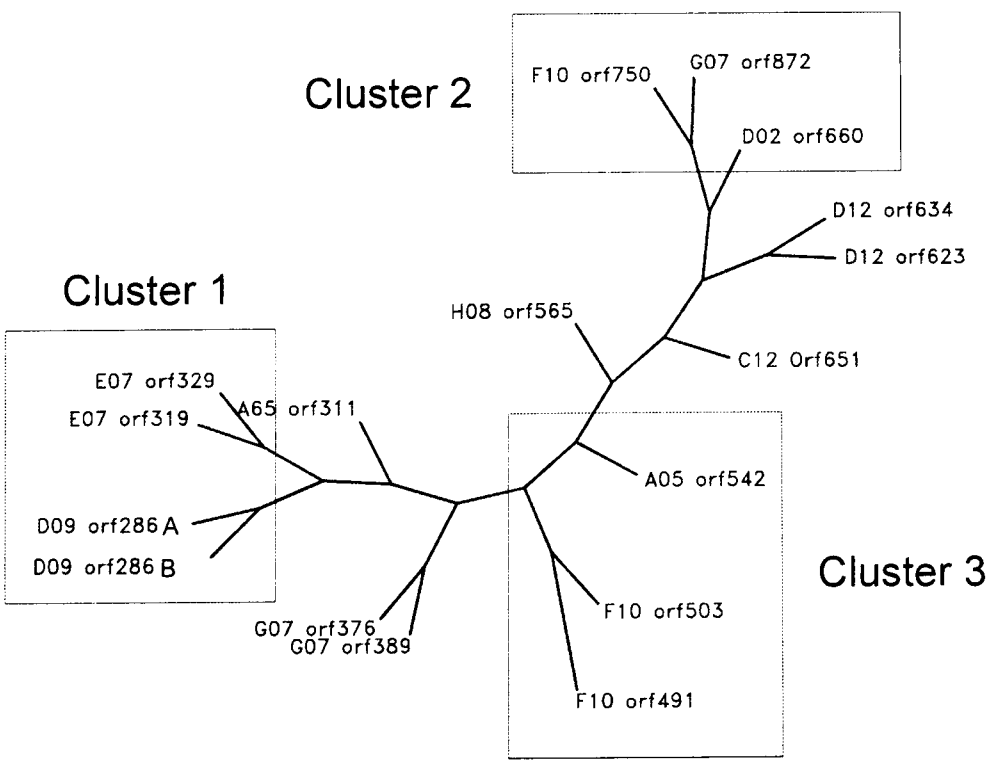


Figure 4.16 *Mycoplasma pneumoniae.*

Mycobacterium tuberculosis

Compared with other genomes, *M. tuberculosis* showed a relatively simple pattern on its ABC transporter phylogenetic analysis, (Fig. 4.17).

The putative helices for these proteins range from 4, Rv0072 and Rv2326c, to 12, Rv1672c. Most of their functions are unidentified. Only Rv1620c and Rv1621c are homologous proteins of CydC and CydD in *B. subtilis*, which are essential for expression of cytochrome *bd*. These two proteins were predicted to have 5 putative TM helices. However, the CydC and CydD in *B. subtilis* have 6 putative helices.

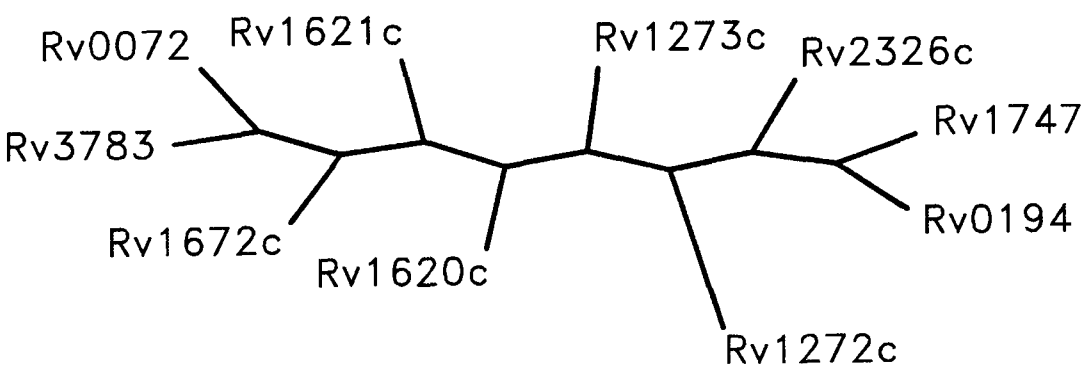


Figure 4.17 *Mycobacterium tuberculosis*

Pyrococcus horikoshii

A pattern between two-branched tree and three-branched tree was identified for the ABC transporters phylogenetic analysis of *P. horikoshii*, (Fig. 4.18).

The functions of only three proteins at the cluster 1 are identified; PH0809 and PH0504 for oligopeptide transporters and PH0204 for maltose transporter. These three proteins all have 6 putative TM helices. For the protein with unknown functions, the numbers of their predicted TM helices are diverse; PH0755 and

PH0911 for 6, PH0943 for 5 and PH0154 for 7. Proteins PH0204 and PH0755 show similar patterns on the locations of putative TM helices.

The cluster 2, including PH0808, PH0503 and PH1410, could be seen as an independent cluster even though it is not distinct on the phylogenetic tree. They were all predicted to have 6 putative TM helices and are all oligopeptide transporters. In addition, strong similarities were seen between patterns on the locations for putative TM helices of these three proteins.

Proteins in cluster 3 show a diverse pattern on the numbers of predicted TM helices; PH1736 for 7, PH1735 for 6, PH1271 for 13, PH1361 for 12 and PH1960 for 6. Only proteins PH1361 and PH1960 are with identified functions, which are likely to be proline transporter and oligopeptide transporter respectively.

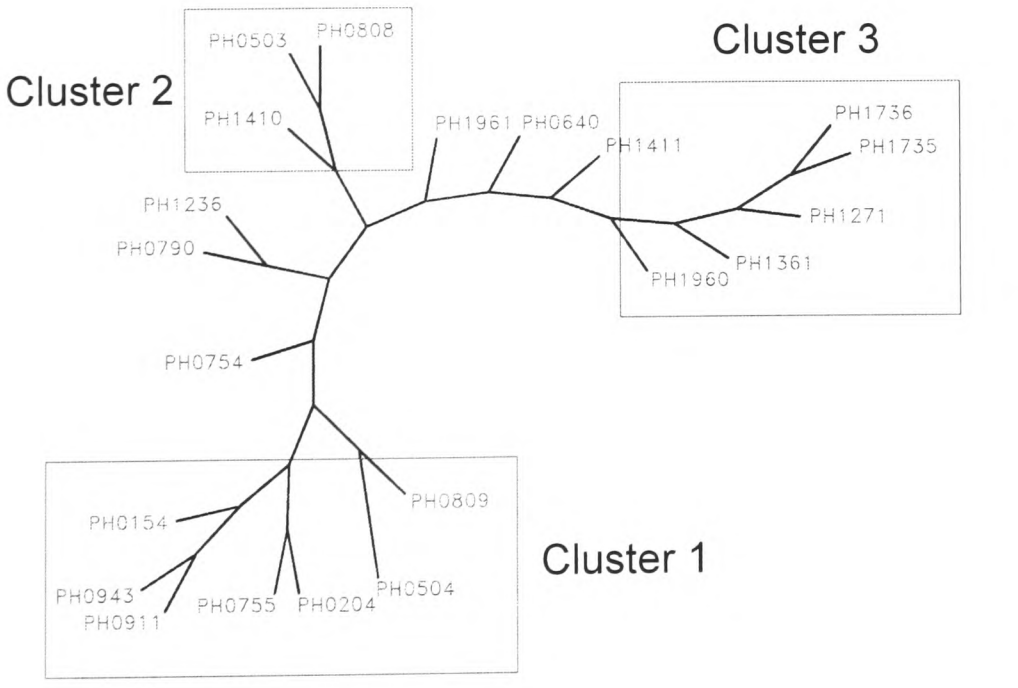


Figure 4.18 *Pyrococcus horikoshii.*

Rhizobium sp

A simple pattern was found for the ABC transporters phylogenetic analysis of *Rhizobium sp*, (Fig. 4.19).

Most of the proteins' functions remain to be verified. The only exception is the Q5391, which is likely to be a dipeptide transporter, and was predicted to have 6 TM helices. For other proteins, the numbers of putative helices are not unified, ranging from 4, P55660 and P55661 on one side of the phylogenetic tree, to 14, P55452 on the other side of the phylogenetic tree. No clear pattern was identified in this genome.

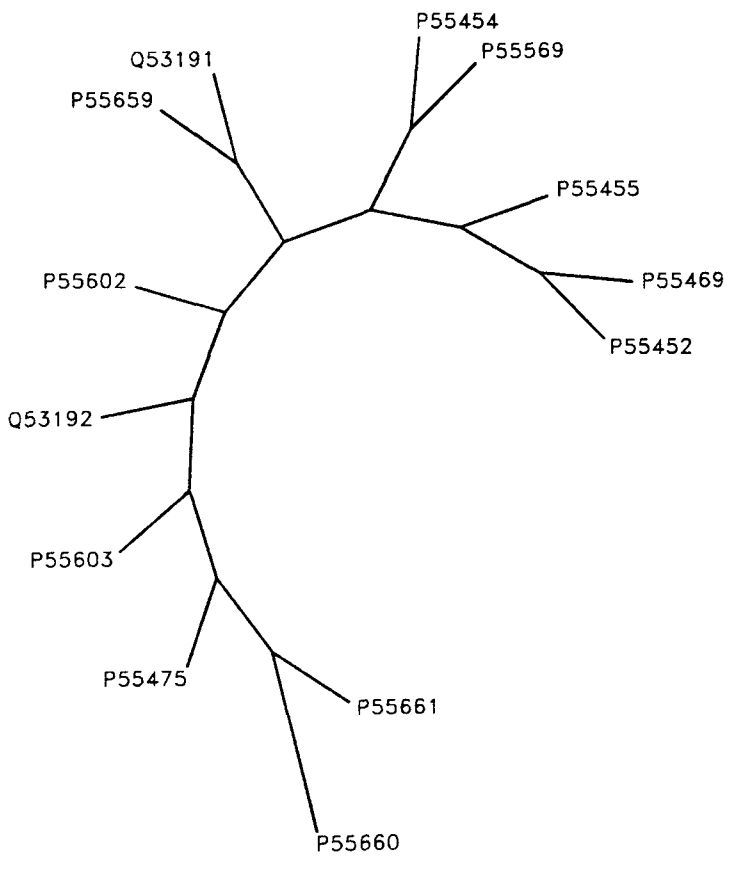


Figure 4.19 *Rhizobium sp*.

Synechocystis PCC6803

The phylogenetic analysis of ABC transporters in *Synechocystis PCC6803* revealed a two-branched tree, (Fig. 4.20). The functions of ABC proteins in the genome mostly remain unidentified.

Proteins in cluster 1 were largely predicted to have either 4 or 5 TM helices. The only exception is the protein slr0527, which is an ion transporter.

For proteins in cluster 2, three of them are transporters of sugars, slr1202 and slr0574 for lactose and sll0530 for sugar. Their putative TM helices range from 4 (sll06150), 6 (slr0977 and sll0574), 7 (sll1600, slr1202 and slr0530), and 8 (slr2045).

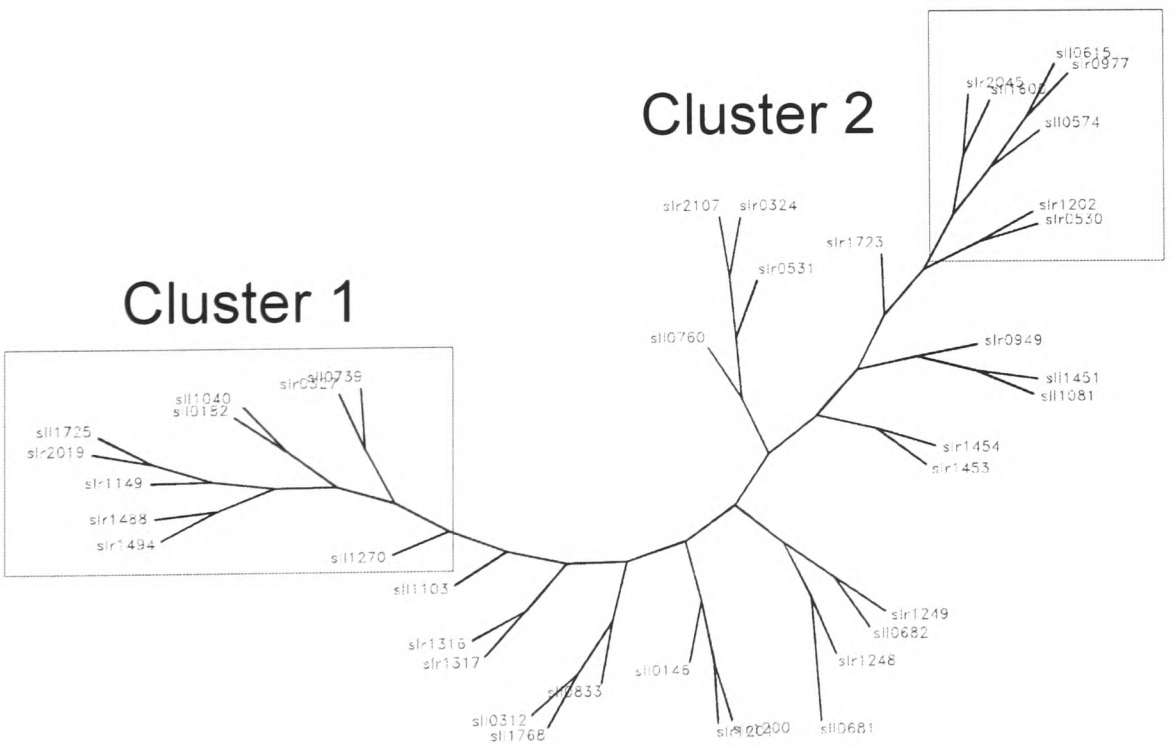


Figure 4.20 *Synechocystic PCC6803.*

Treponema pallidum

The phylogenetic analysis of ABC transporters in *T. pallidum* revealed a two-end tree, (Fig. 4.21).

Proteins in cluster 1 transport different solutes. Both TP0653 and TP0654 are transporters of spermidine and putrescine, and they were predicted to have 7 and 6 TM helices respectively. Protein TP0076 is a sugar transporter and has 6 putative TM helices. Protein TP0119 is an amino acid transporter and has 4 putative TM helices. The function of protein TP0036 is unknown, and this protein was predicted to have 7 TM helices.

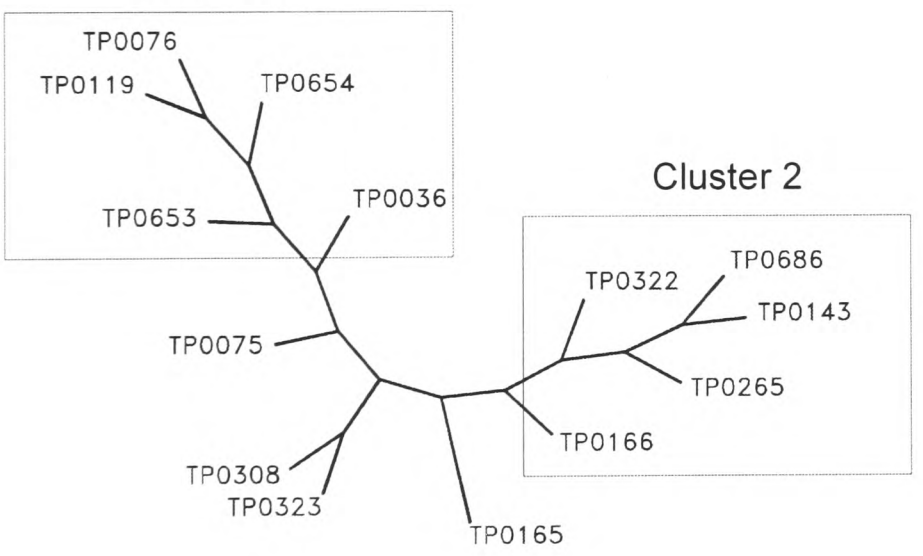


Figure 4.21 *Treponema pallidum.*

Proteins in cluster 2 show no similarities either. TP0322 is a transporter of ribose and galactose, and has 8 putative TM helices. TP0686 is a methylgalactoside

transporter, and has 9 putative TM helices. TP0143 is a thiamine transporter, and has 14 putative TM helices. TP0322 is a transporter of amino acids, has 12 putative TM helices. TP0166 is has no identified function, and has 8 putative TM helices.

Archaeoglobus fulgidus

The phylogenetic tree of *Archaeoglobus fulgidus* shows a two-branched pattern, (Fig. 4.22).

For the cluster 1, three out five transporters have 6 putative TM helices; AF1980, AF0289 and AF0042. Other proteins were predicted to have 5 TM helices for AF0232 and 7 TM helices for AF1981. Even though, they have similar numbers of TM helices, these transporters do not share a strong similarity in their solutes. However, two of them, AF0289 and AF0042, transport similar solutes, polysaccharide, do share similar linear locations of predicted TM helices.

For the cluster 2 of *A. fulgidus*, three transporters, AF1325, AF2258 and AF1144, are responsible for multi-drug resistance and they were predicted to have 11, 14 and 14 TM helices respectively. AF1393 is an amino acid transporter and was predicted to have 8 TM helices; AF1768 is a dipeptide transporter and was predicted to have 6 TM helices.

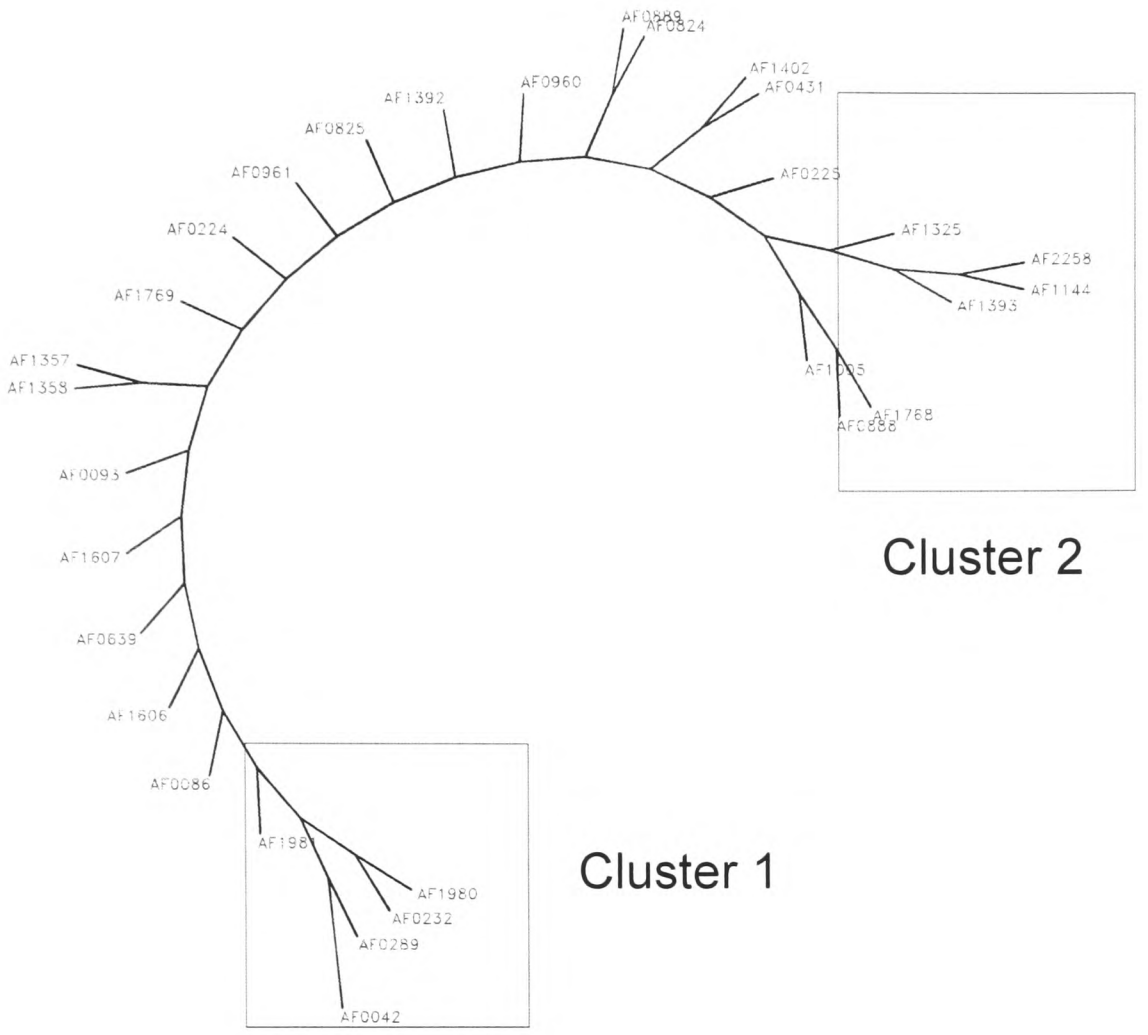


Figure 4.22 *Archaeoglobus fulgidus.*

Methanobacterium thermoautotrophicum

The phylogenetic tree of *M. thermoautotrophicum* shows a two-branched pattern, (Fig. 4.23). Proteins at two edges, cluster 1 and 2, were examined.

In general, proteins in cluster 1 have approximate 12 putative TM helices. MTH104 was predicted to have 12 predicted TM helices, and is likely to be responsible for multi-drug resistance. MTH1856 was predicted to have 13 TM helices, and is a proline permease. MTH546 was predicted to have 12 TM helices and is likely a transporter of positive-charged amino acids.

The numbers of putative TM helices for proteins in cluster 2 are around 4 to 6. Both MTH1705 and MTH132 are cobalt transporters; they were predicted to have 4 and 5 TM helices respectively. MTH1092 has 6 putative helices and is likely responsible for detoxification. MTH478 has 5 putative helices and is responsible for sulfate transportation.

Similar numbers of predicted TM helices for proteins in cluster 1 were found; however, these proteins show weak similarities in their functions. For cluster 2, proteins do not only have similar numbers of putative TM helices, their solutes are likely to be charged molecules like cobalt and sulfate.

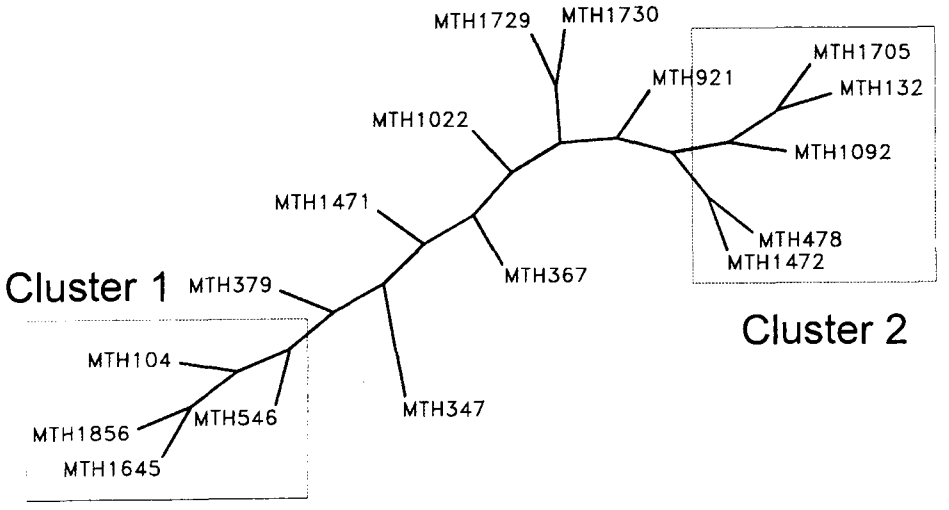


Figure 4.23 *Methanobacterium thermoautotrophicum*

Methanococcus jannaschii

The phylogenetic tree of *M. jannaschii* shows a two-branched tree, (Fig 4.24). Proteins in cluster 1 have 5, MJ1089, to 6, MJ0413 and MJ1013, putative helices. No similar solutes were detected for these three proteins. For proteins in cluster 2, 3 out of 4 transport positive-charged molecules; MJ0566 for ferrous iron, MJ0058 and MJ1343 for ammonium, and MJ0609 to be identified. Their predicted TM helices do not exhibit strong similarities; MJ0566 for 8, MJ0058 for 10, MJ1343 for 11 and MJ0609 for 12. The arrangements of predicted TM helices along their protein sequences are very different as well.

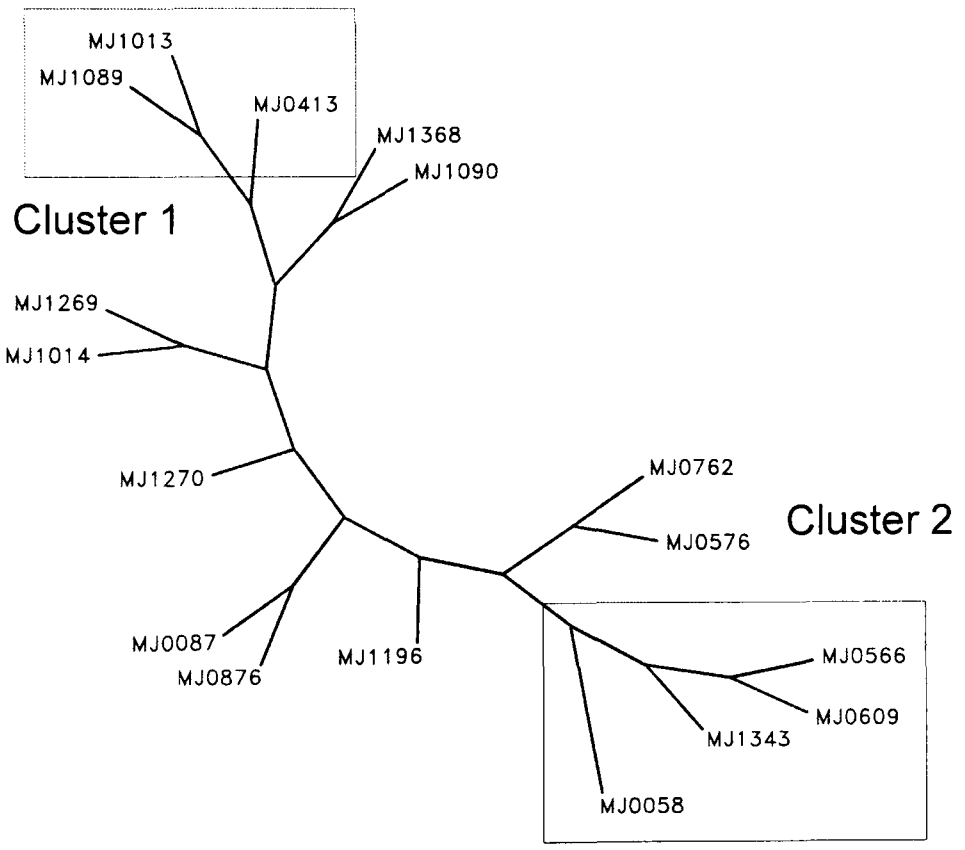


Figure 4.24 *Methanococcus jannaschii.*

Saccharomyces cerevisiae

A three-branched tree was identified on the phylogenetic analysis of ABC transporters in *S. cerevisiae*, (Fig. 2.25).

The functions of proteins in cluster 1 are largely unknown; except for YPL147W, which is a transporter of long chain fatty acid. Compared with other proteins in cluster 2 and 3, these proteins have smaller numbers of putative TM helices; YPL147W, YPL270W and YMR301C for 4, and YLR188W for 5.

The four proteins in cluster 2 are all composed of two TM domains occurring within a single peptide. Thirteen putative TM helices are predicted for YOR328W, YOR153W and YDR406W, and twelve TM helices are predicted for YPL058C. Only two of them are with identified function. YOR153W is a transporter of suporidesmin, which is a toxin. YPL058C is a likely a multidrug resistance transporter.

In general, proteins in cluster 3 have more predicted TM helices in quantities than proteins in other clusters. YHL035C, YLL048C and YDR135C have 14 putative TM helices. YLL015W has 13 putative TM helices. The only exception is YGR281W, which has only 11 putative TM helices. These proteins are likely to be transporters of larger organic molecules. YGR281W is a transporter of oligomycin, YDR135C is a transporter of glutathione S-conjugated metal. YLL015W is likely a transporter of metal. YLL048C is a bile transporter. YHL035C is likely to be a multidrug resistance transporter.

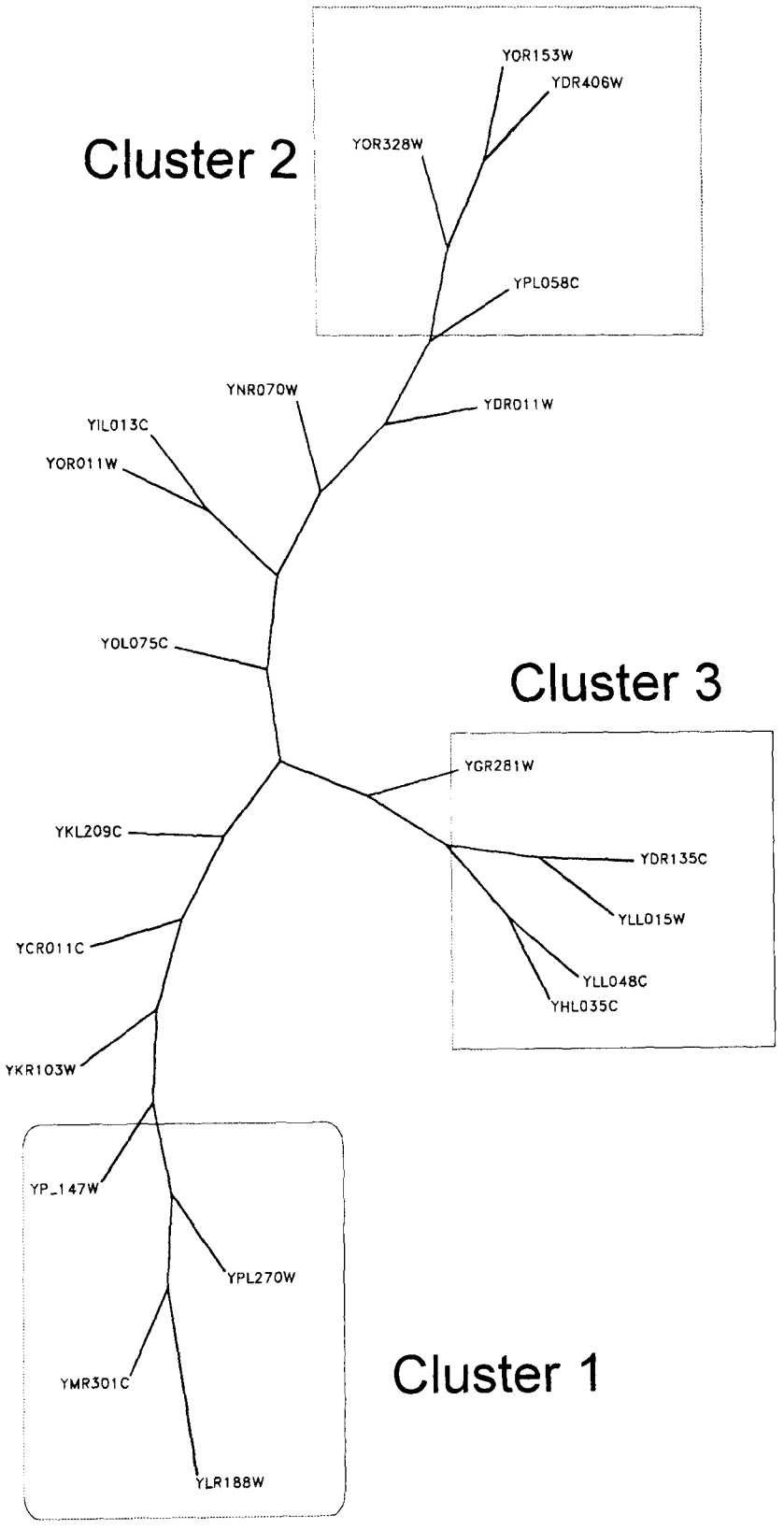


Figure 4.25 *Saccharomyces cerevisiae*

4.3.3 Periodicity Analysis

It has been observed that proteins within similar sequence-generated clusters may share similar locations of transmembrane helices on their primary protein sequences. Does this imply that they share similar helical topology in the lipid bilayer, and adopt similar spatial arrangements and orientations of helices within transmembrane bundles? By analysing the periodic patterns of residue properties along putative TM helices, power spectrum of sequence conservation and residue hydrophobicity could provide an indication. In general, a peak at 100° corresponds to a sequence pattern that repeats once every 3.6 residues, *i.e.* with the periodicity expected for an α-helix. In addition, a helix identified with a peak at 100° on its periodic spectra also suggests this helix has a clear discrimination on residues of lipid-facing phase and of interior-facing phase.

Three proteins, AppC, DppC and OppC, in the cluster 1 of *B. subtilis* showed similar linear arrangements of putative TM helices, (Fig. 4.8 and 4.9). Spectra for each helix were derived by calculating its periodicities of conservation and hydrophobicity using PERSCAN, for details see section 4.2.5. Periodic spectra of helices at the same positions were superimposed in order to have a better observation for their similarity.

It was easily observed that the TM helices 1 (Fig. 4.26 A), 2 (Fig. 4.26 B), 3 (Fig. 4.26 C) and 5 (Fig. 4.26 E) of these three transporters exhibiting ideal patterns of α-helix on sequence conservation. Further, the three transporters also shared strong similarities on the spectra of conservation periodicity at these four TM helices. However, the clear spectrum patterns for α-helices were not observed in TM helix 4 (Fig. 4.26 D) and TM helix 6 (Fig. 4.26 F). No similarities were

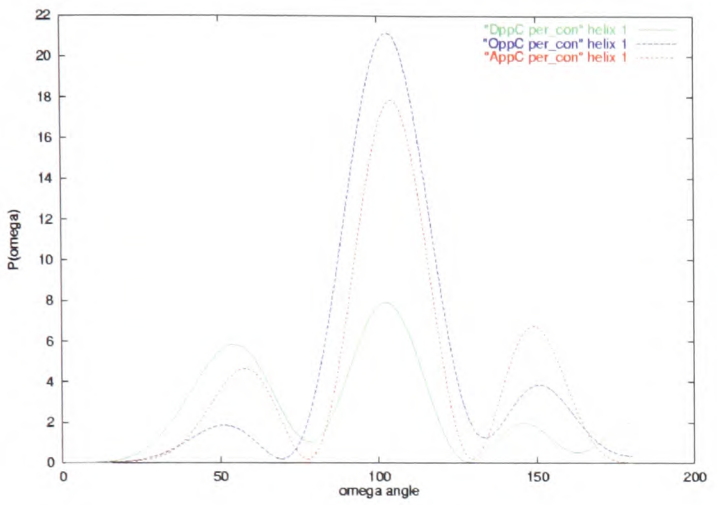
observed for sequence conservation or residue hydrophobicity in TM helices 4 and 6 of these three transporters either.

For the sequence periodicity of hydrophobicity, spectra of ideal α-helices were observed for TM helix 1 (Fig. 4.27 A), TM helix 2 (Fig. 4.27 B), TM helix 3 (Fig. 4.27 C), TM helix 4 (Fig. 4.27 D) and TM helix 5 (Fig. 4.27 E). In addition, similarities of hydrophobic periodicity between these three transporters were detected on these 5 TM helices as well. For the TM helices 6, the similarity of hydrophobic periodicity was not detected between TM helices of different transporters, (Fig. 4.27 F).

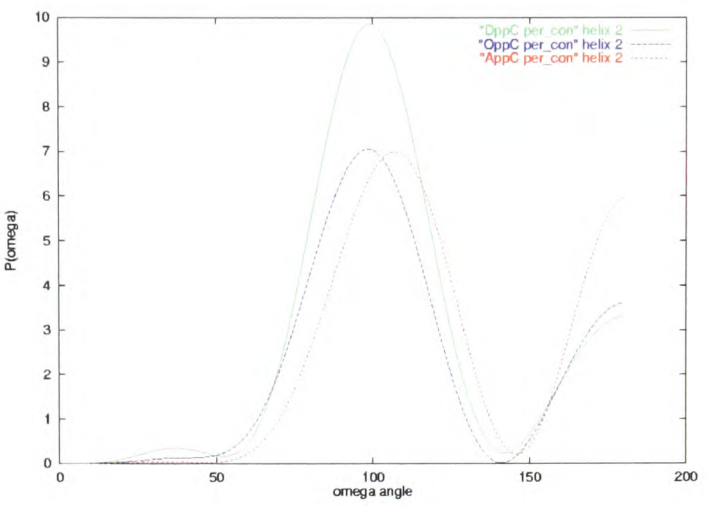
It has been observed that transporters DppC, OppC and AppC have similar periodic patterns in both sequence conservation and residue hydrophobicity for the majority of their TM helices; TM helices 1, 2, 3, and 5. This reveals that the overall arrangement and orientations for the TM helices within the intact helical bundles of these three proteins are likely to be similar. TM helices 4 and 6 display less ideal patterns of conservation and hydrophobicity for membrane-spanning helices which reveal that these two helices do not have a clear interface between interior-facing residues and lipid-facing residues. This suggests that TM helices 4 and 6 might be surrounded by other helices, which have clear periodic patterns, in the intact transmembrane helical bundle.

The trend of transporters sharing similar periodic patterns was also detected in proteins with similar one-dimensional helical arrangement in cluster 2 (YvcC, YheI and YfiC), and cluster 3 (OpuBB, OpuCD, OpuBD and OpuCD) of *B. subtilis*.

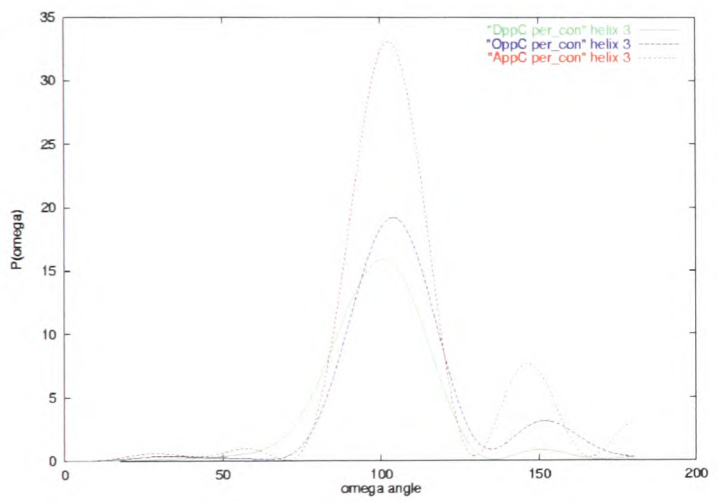
A. TM helix 1



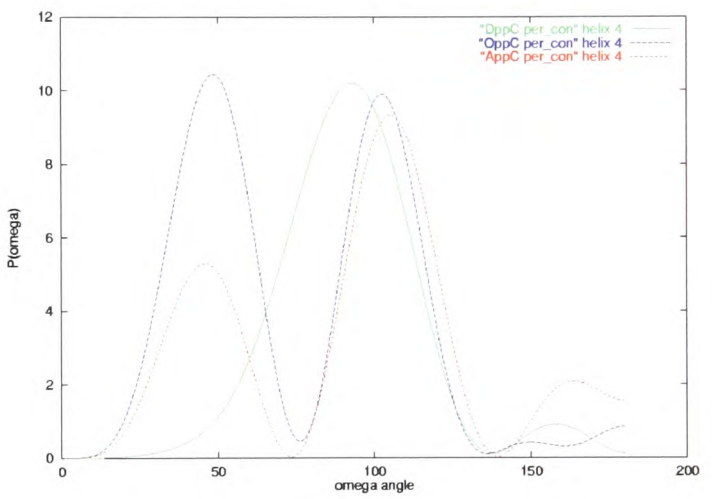
B. TM helix 2



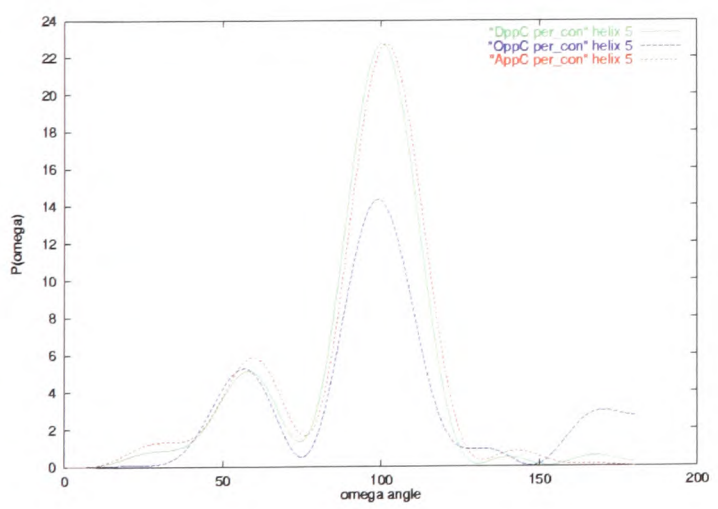
C. TM helix 3



D. TM helix 4



E. TM helix 5



F. TM helix 6

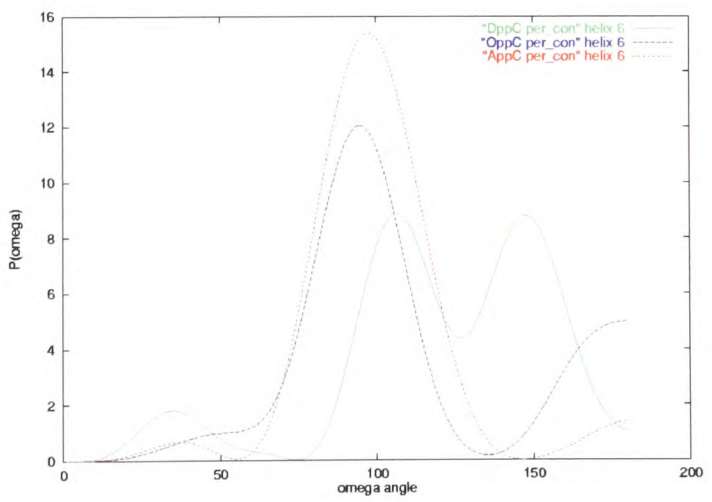
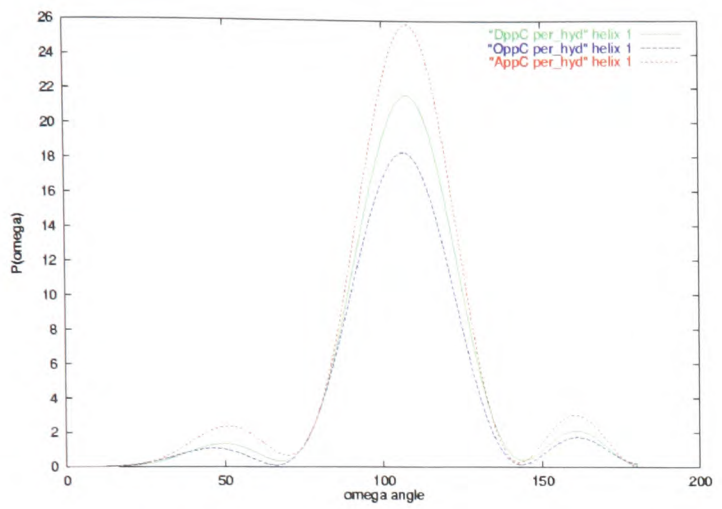
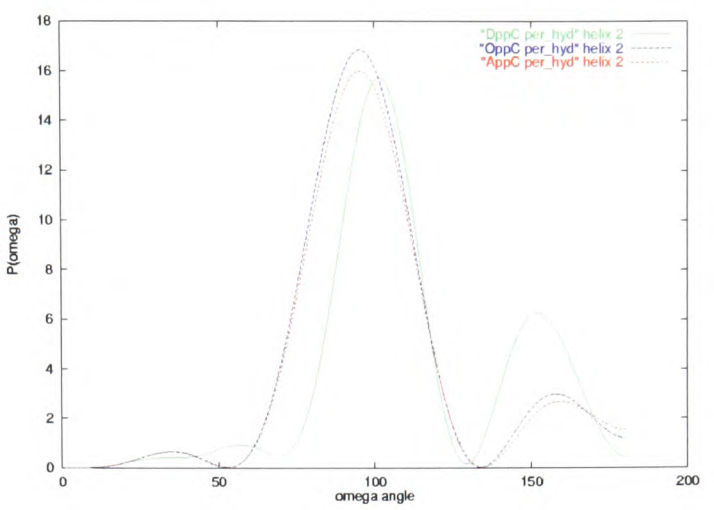


Figure 4.26 Spectra of sequence conservation periodicity of transporters AppC, DppC and OppC in *B. subtilis*, for details see section 4.2.5.

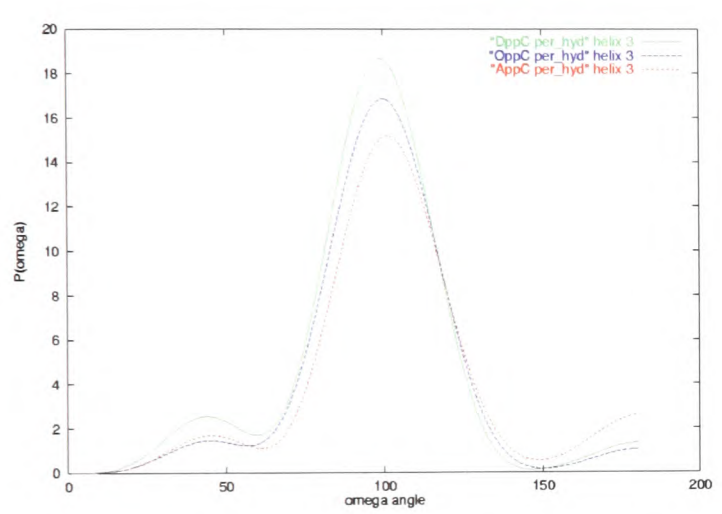
A. TM helix 1



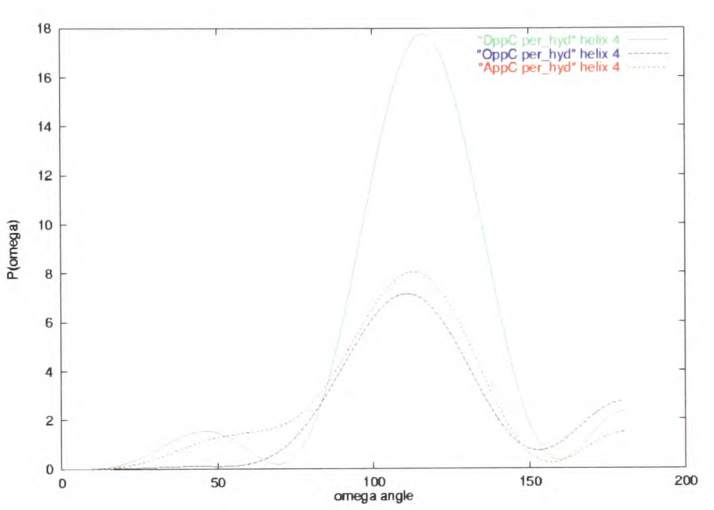
B. TM helix 2



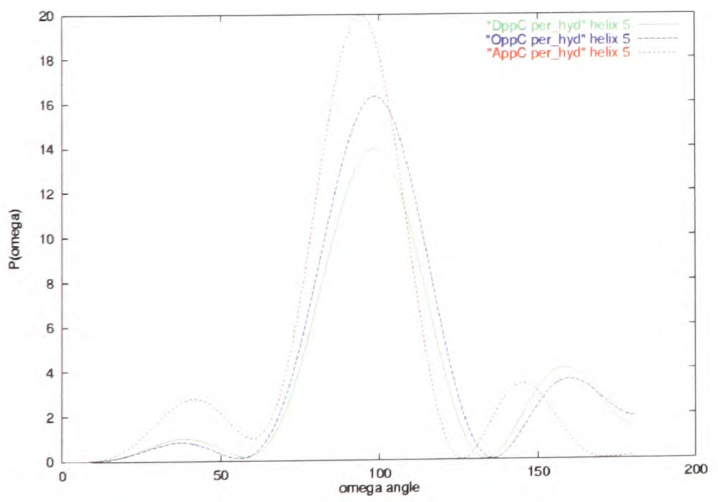
C. TM helix 3



D. TM helix 4



E. TM helix 5



F. TM helix 6

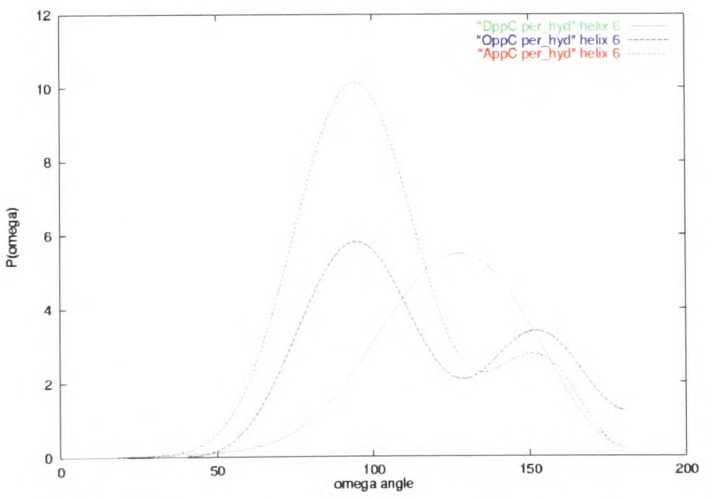


Figure 4.27 Spectra of amino acid residue hydrophobicity periodicity of transporters AppC, DppC and OppC in *B. subtilis*, for details see section 4.2.5.

The trend of transporters in the same sequence-generated cluster sharing similarities on the linear arrangements of their putative helices will lead to similar periodic patterns has been detected in other genomes too.

Three out of four proteins in cluster 2 of *S. cerevisiae* were predicted to have 13 TM helices; YOR328W, YDR406W and YOR153W. These proteins shared strong similarities on their putative helical locations, section 4.3.2. From the periodic analysis of sequence conservation, conservation spectra of these proteins were revealed, (Fig. 4.28). Many TM helices were identified with ideal conservation patterns of membrane-spanning α-helix; for example, TM helices 6 of these proteins (Fig. 4.28 F). The TM helices showing spectra of conservation for ideal integral membrane helix were summarised, (Tab. 4.8). The spectra of amino acid hydrophobicity for these 13 TM helices were acquired with similar periodic analysis, (Fig. 4.29). The TM helices with spectra of hydrophobicity for ideal membrane-spanning helix, *e.g.* TM helices 5 (Fig. 4.29 E), were summarised, (Tab. 4.8).

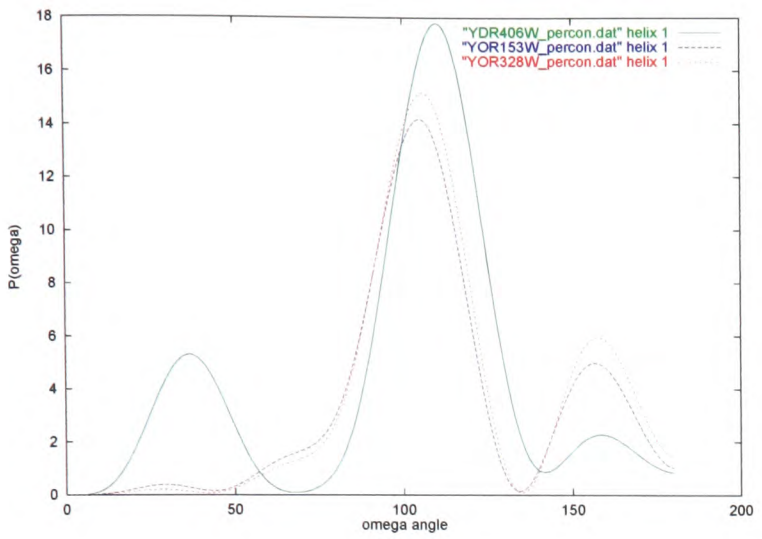
Combining information from the periodic analysis of conservation and hydrophobicity, possible spatial arrangements of these helices could be deduced. For example, TM helices 1, 2, 4, 5, 7, 8, 9, and 10 were proposed to form a circle on the outer layer of the transmembrane bundle and to face the lipid bilayer, due to the clear spectra of α-helix for conservation and hydrophobicity they have shown. In contrast, the other helices were likely to be on the interior part of the transmembrane bundle, surrounded by helices facing lipid bilayer.

Analysis	TM helices
Conservation	**1, 2, 4, 5, 6, 7, 8, 9, 10, 11**
Hydrophobicity	**1, 2, 3, 4, 5, 7, 8, 9, 10, 12**

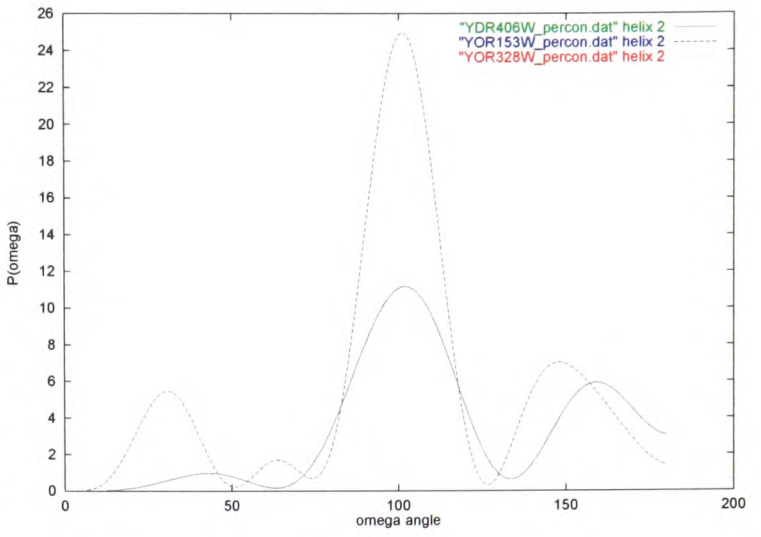
Table 4.8 Helices sharing similar periodic patterns of ideal α-helix in protein YDR406W, YOR153W and YOR328W of *S. cerevisiae*.

Four proteins were sorted into the cluster 1 of *S. cerevisiae*; YPL147W, YMR301C, YLR188W and YPL270W. Except YLR188W, which was predicted to have 5 TM helices, other transporters have 4 putative TM helices. The four putative TM helices of YPL147W, YMR301C and YPL270W and the first four TM helices of YLR188W were predicted at similar locations along their primary protein sequences. Periodic analysis of sequence conservation and amino residue hydrophobicity on these TM helices of the four proteins has shown similarities on patterns of periodicity for TM helices 2 and 4 of these four transporters. This provided an indication that these four transporters may share similar fold arrangements of TM helices 2 and 4 within the cell membrane.

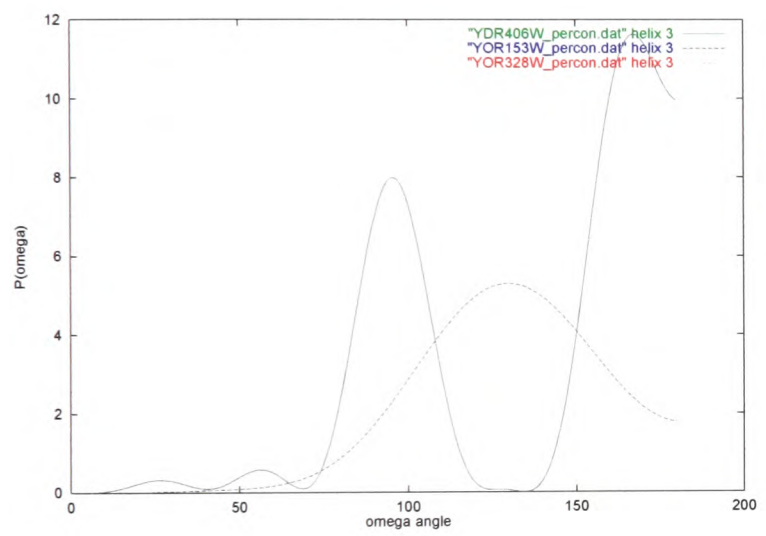
A. TM helix 1



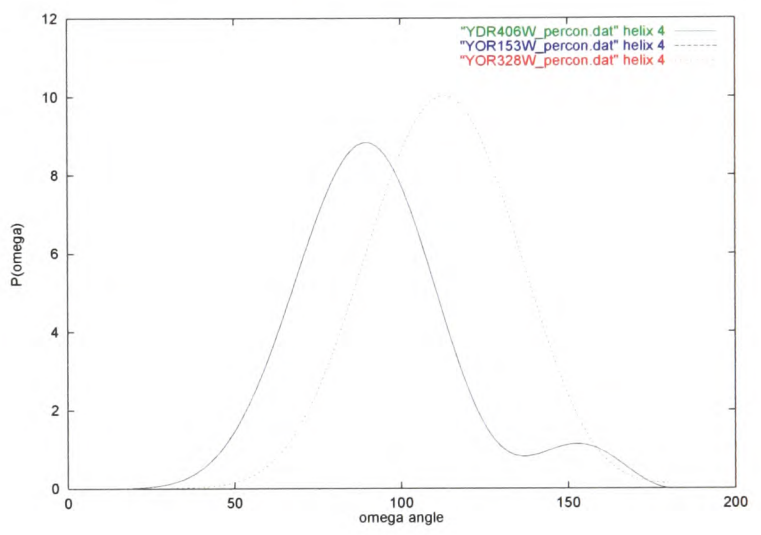
B. TM helix 2



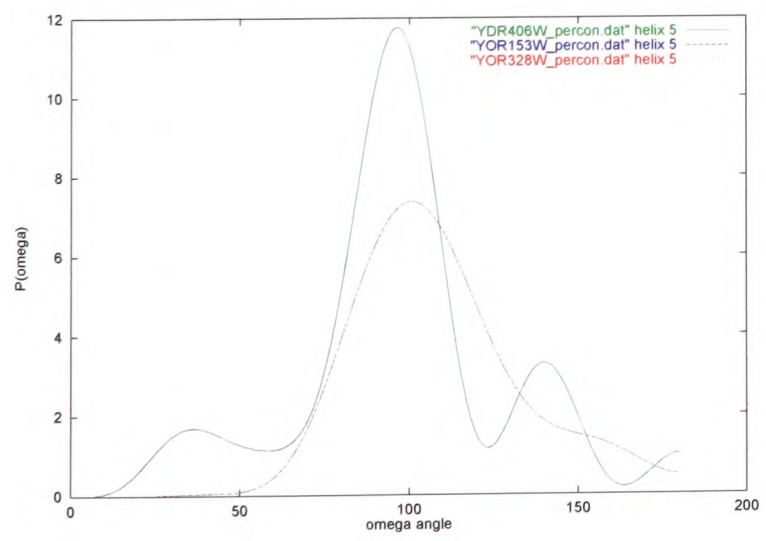
C. TM helix 3



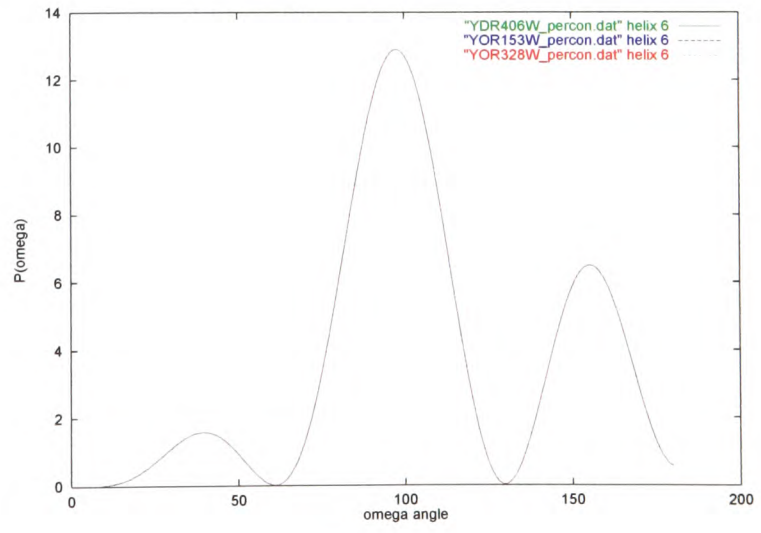
D. TM helix 4



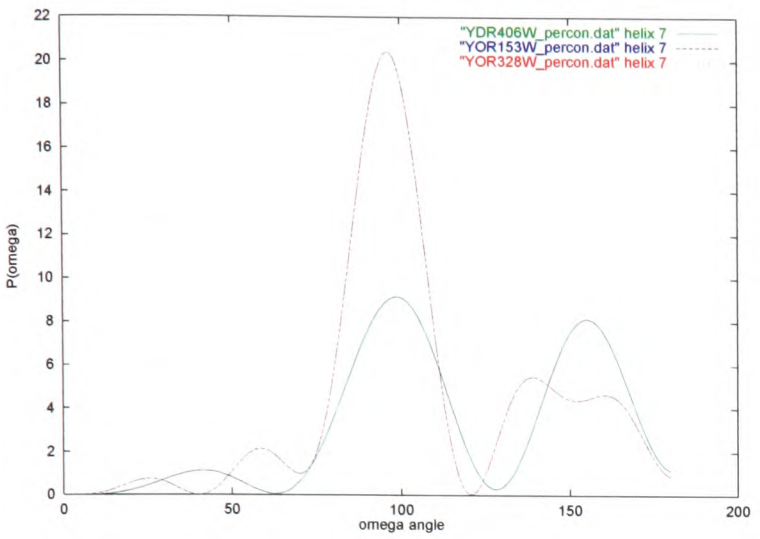
E. TM helix 5



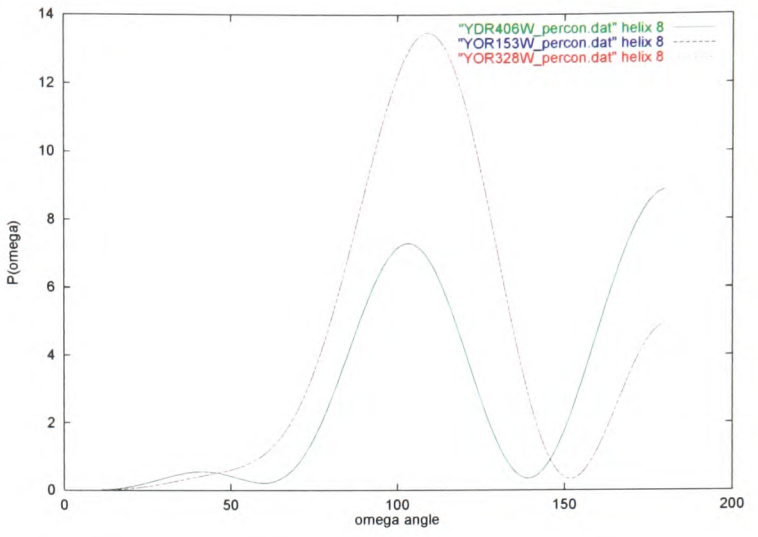
F. TM helix 6



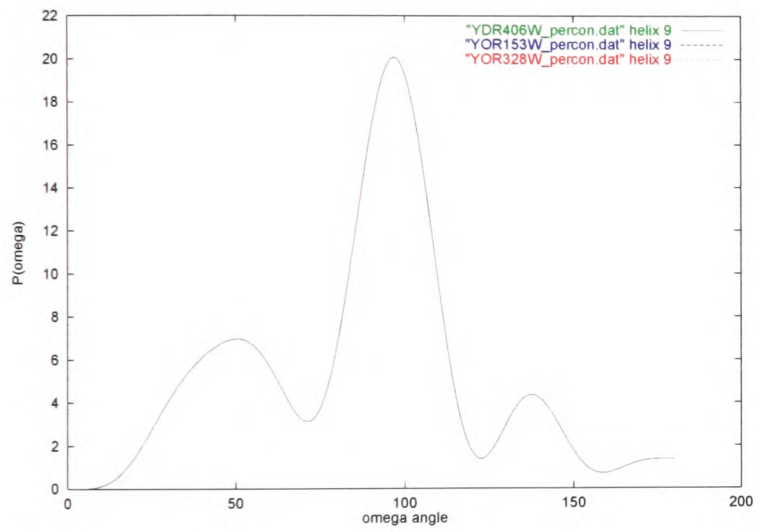
G. TM helix 7



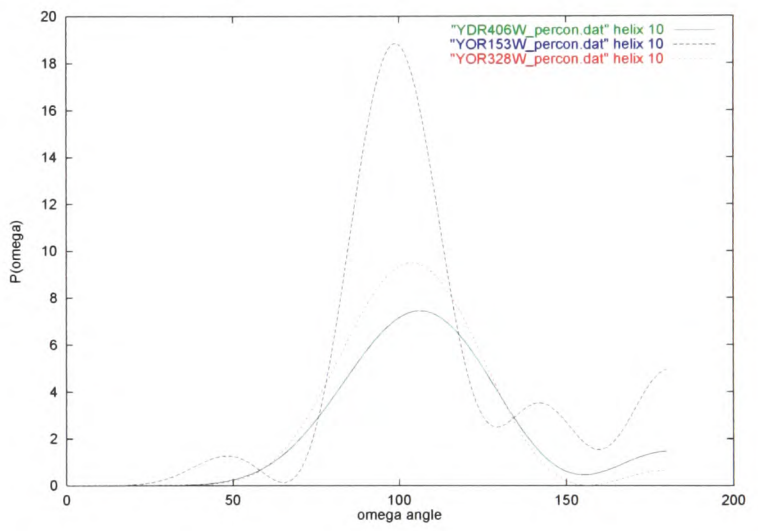
H. TM helix 8



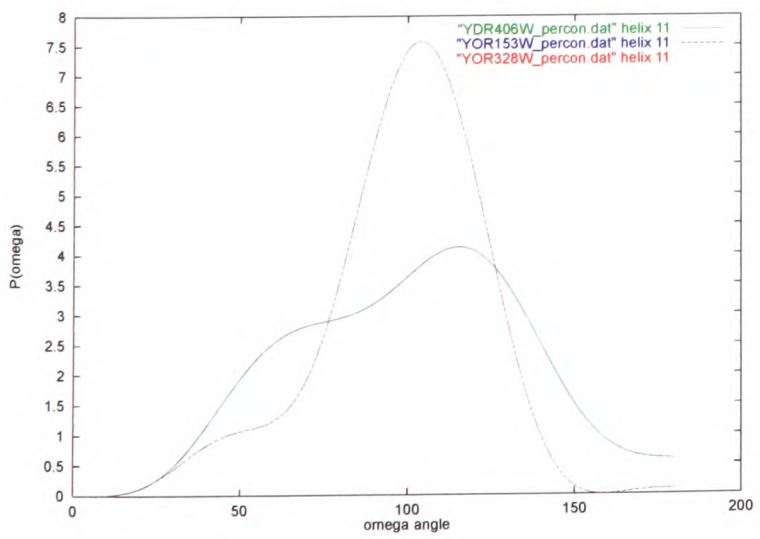
I. TM helix 9



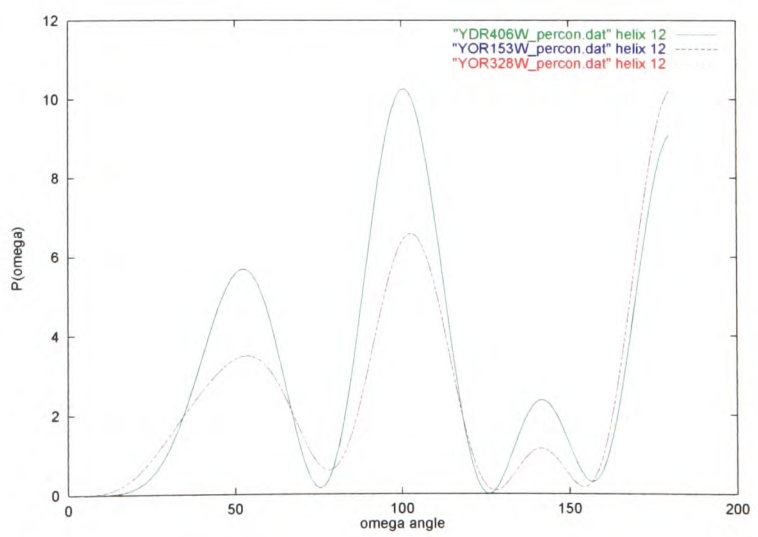
J. TM helix 10



K. TM helix 11



L. TM helix 12



M. TM helix 13

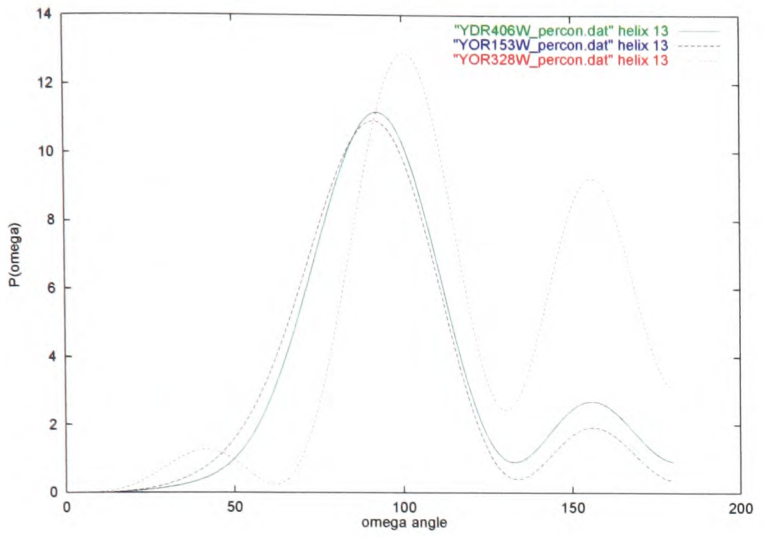
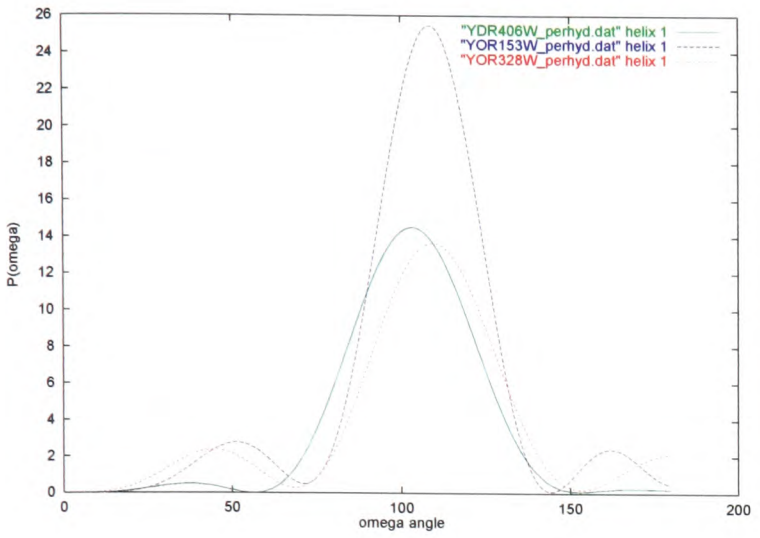
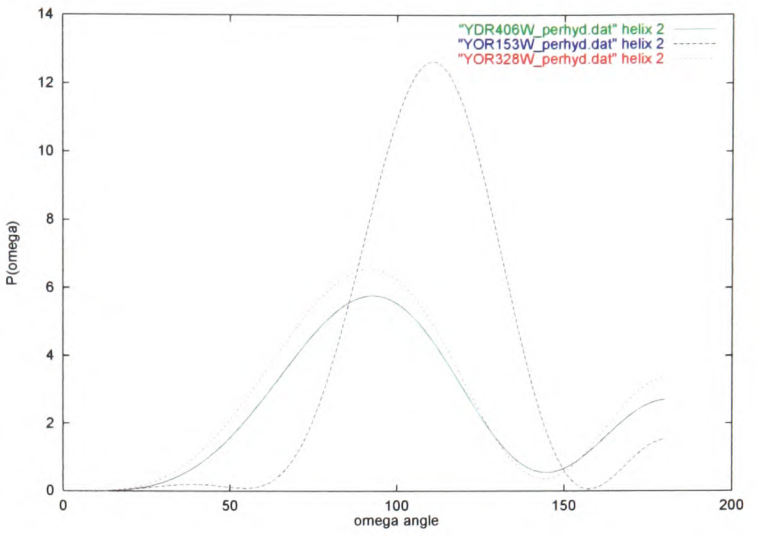


Figure 4.28 Spectra of sequence conservation periodicity of transporters YOR328W, YDR406W and YOR153W in *S. crevisiae*, for details see section 4.2.5.

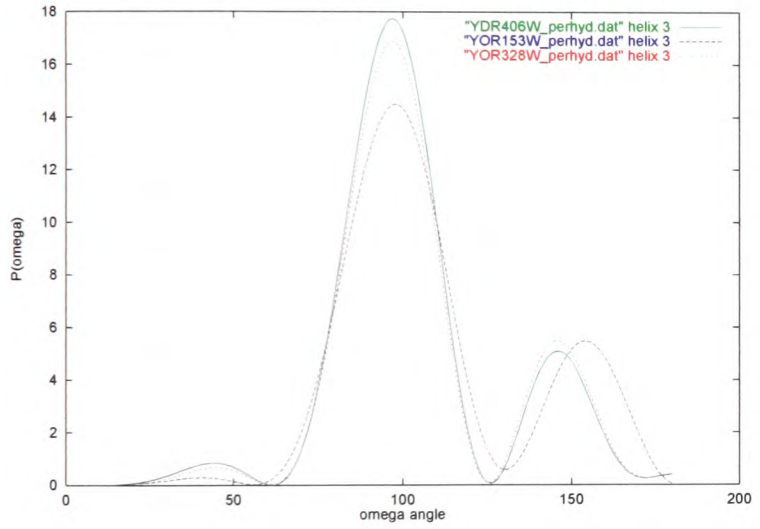
A. TM helix 1



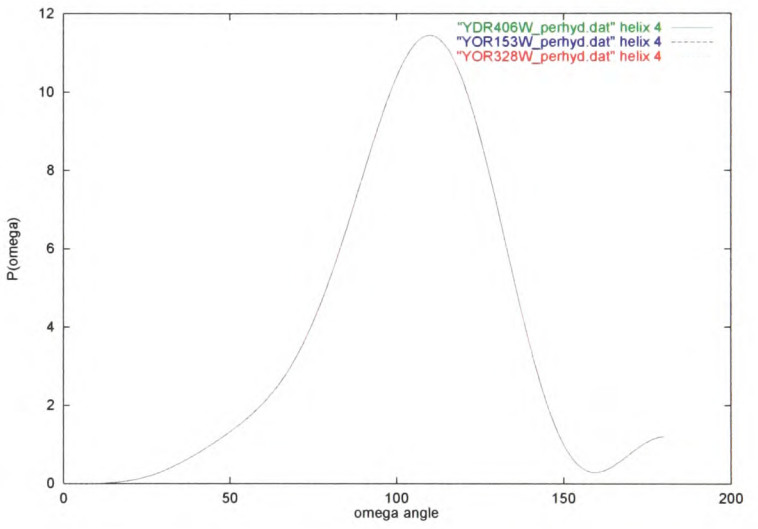
B. TM helix 2



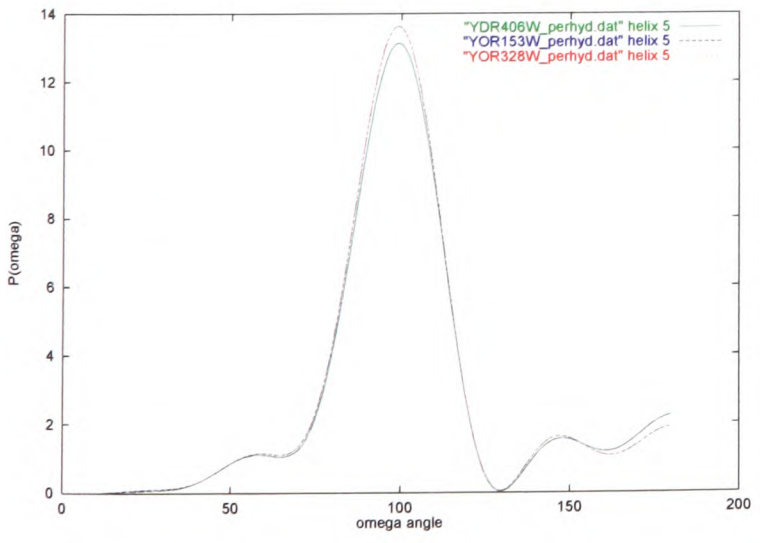
C. TM helix 3



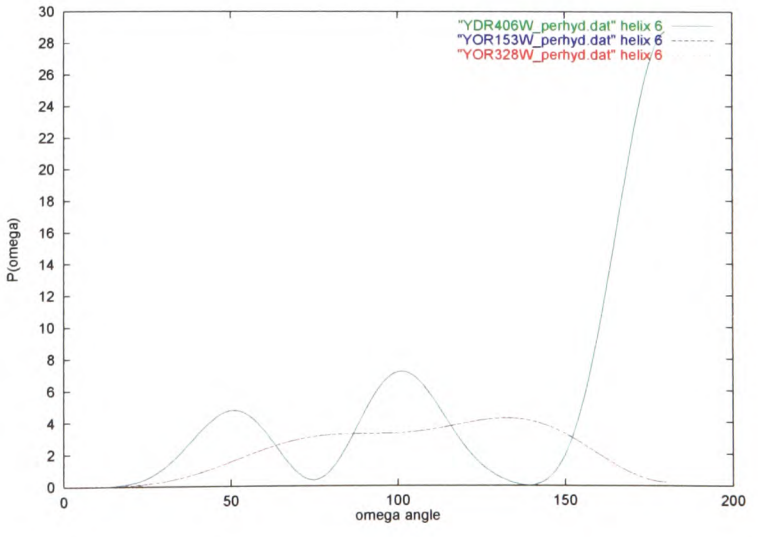
D. TM helix 4



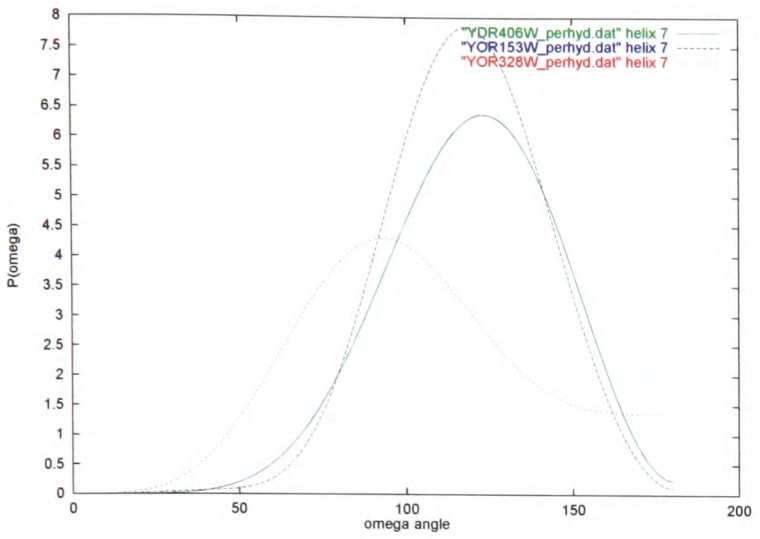
E. TM helix 5



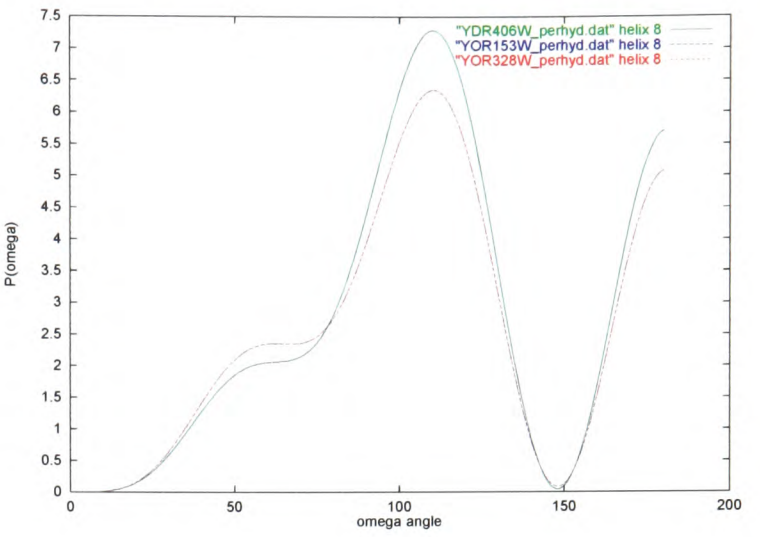
F. TM helix 6



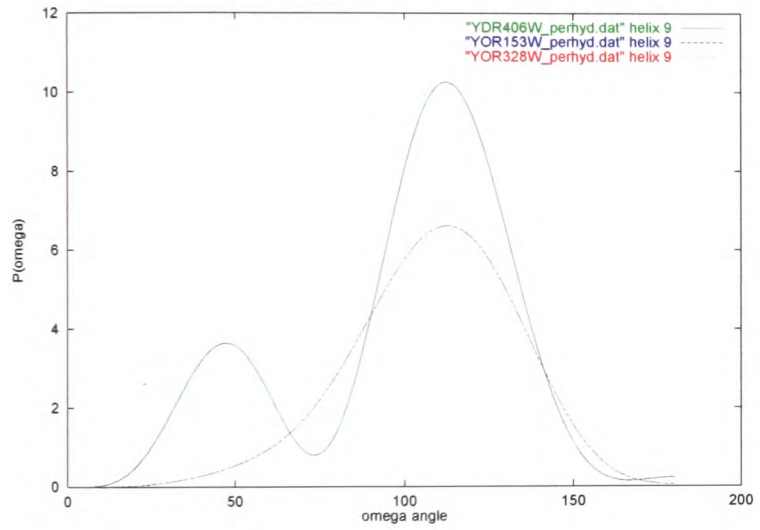
G. TM helix 7



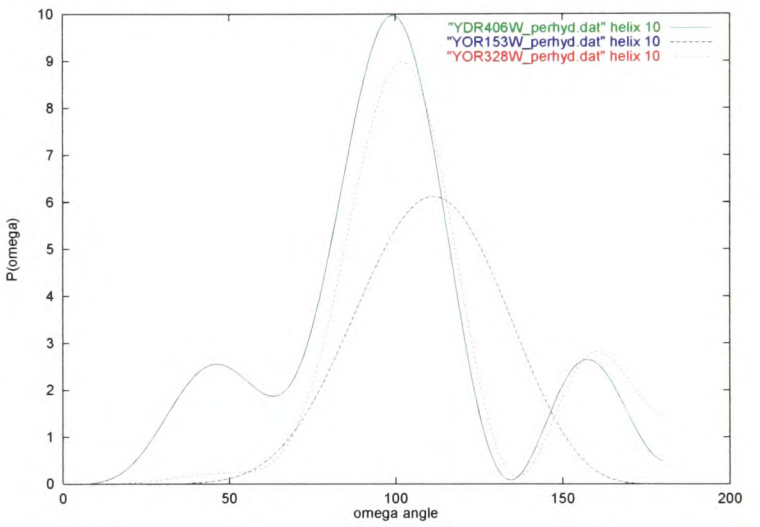
H. TM helix 8



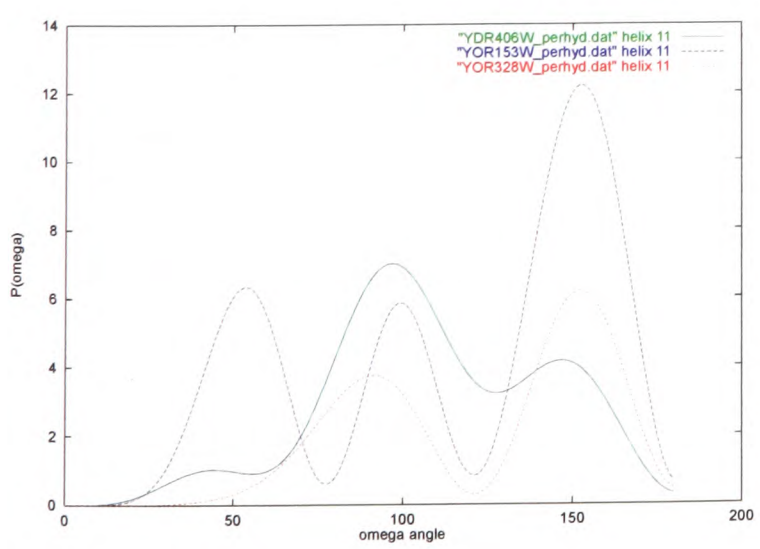
I. TM helix 9



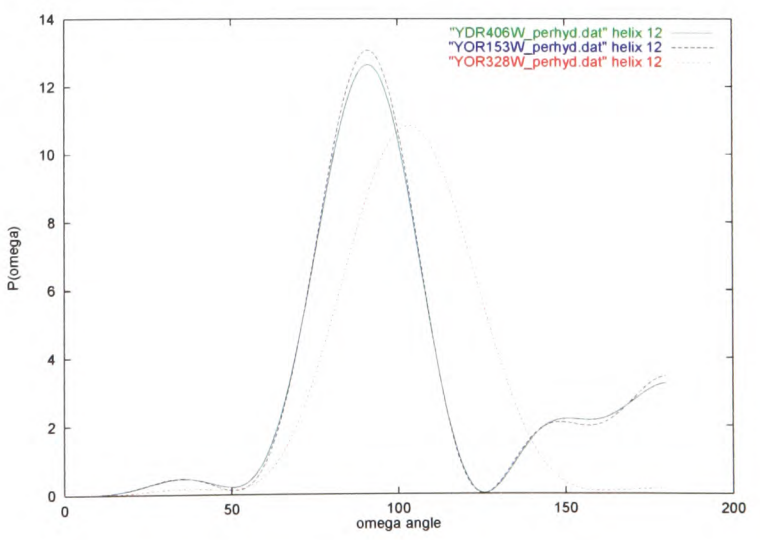
J. TM helix 10



K. TM helix 11



L. TM helix 12



M. TM helix 13

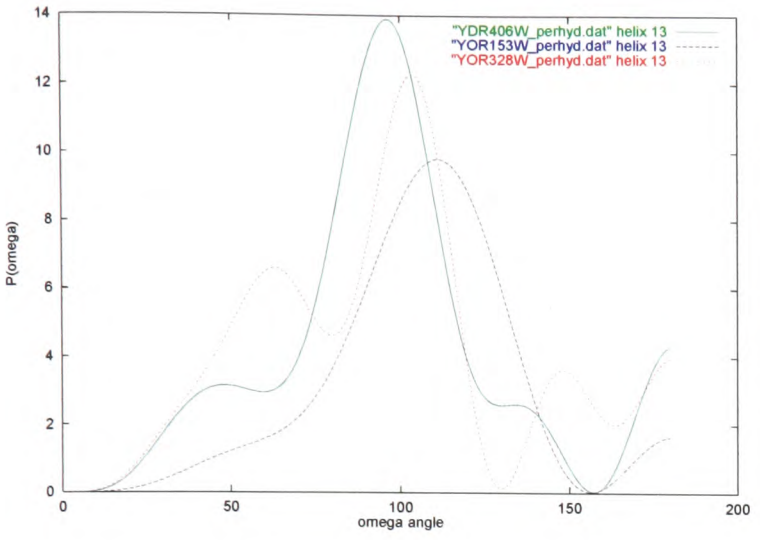


Figure 4.29 Spectra of amino acid residue hydrophobicity periodicity of transporters YOR328W, YDR406W and YOR153W in *S. crevisiae,* for details see section 4.2.5.

Two out of three proteins in cluster 1 of *M. jannaschii* were predicted to have 6 TM helices. Even though, these two transporters, MJ1013 and MJ0413, have the same amount of predicted TM helices, the linear arrangements of their putative helices are different. Spectra derived from periodic analysis of conservation and hydrophobicity were noisy.

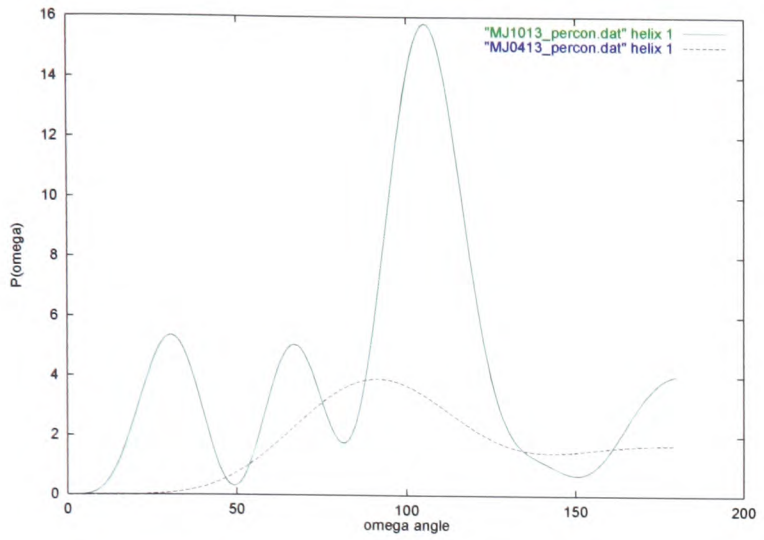
For the transporter MJ1013, TM helices 1, 2, 4, and 5 showed an ideal pattern for transmembrane α-helix on conservation, (green lines in Fig. 4.30 A, B, D, and E). TM helices 1, 3, 4, 5 and 6 showed an ideal pattern for transmembrane α-helix on hydrophobicity, (green lines in Fig. 4.31 A, C, D, E and F).

For the transporter MJ0413, TM helices 2, 4 and 5 showed an ideal pattern for transmembrane α-helix on conservation, (blue lines in Fig. 4.30 B, D, and E). TM helices 1, 2, 3, 4 and 5 showed an ideal pattern for transmembrane α-helix on hydrophobicity, (blue lines in Fig. 4.31 A, B, C, D and E).

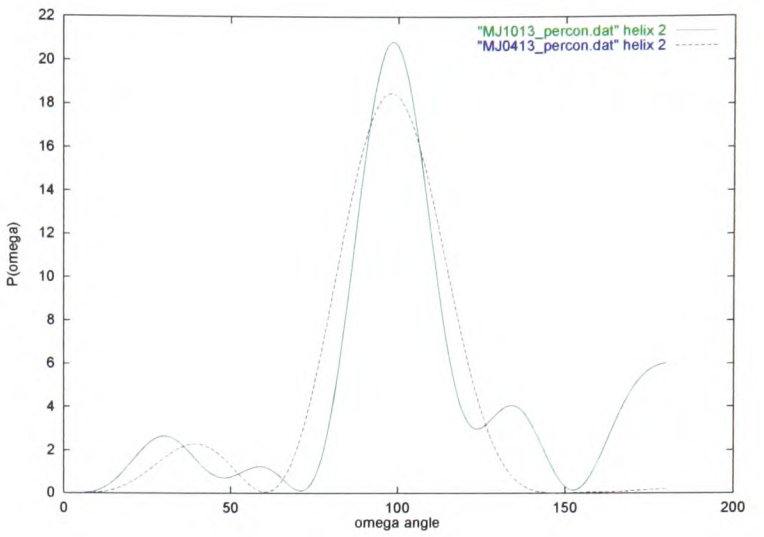
However, combining information of both conservation and hydrophobicity periodicities for these two transporters, similarities were not observed between these two proteins. This suggests that these two proteins may adopt different spatial arrangements for the helices within the lipid bilayer.

In general, the correlation between sequence-generated clusters and similarities of helical arrangement was identified. This suggests that transporters, which are evolutionary related, may have similar locations for the TM helices along the primary protein sequences, and then share similar spatial locations of TM helices within the helix bundles of the intact transmembrane proteins.

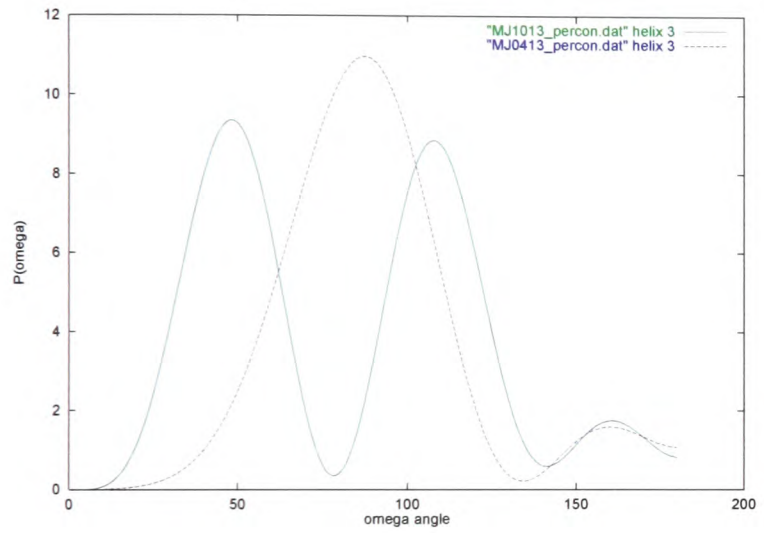
A. TM helix 1



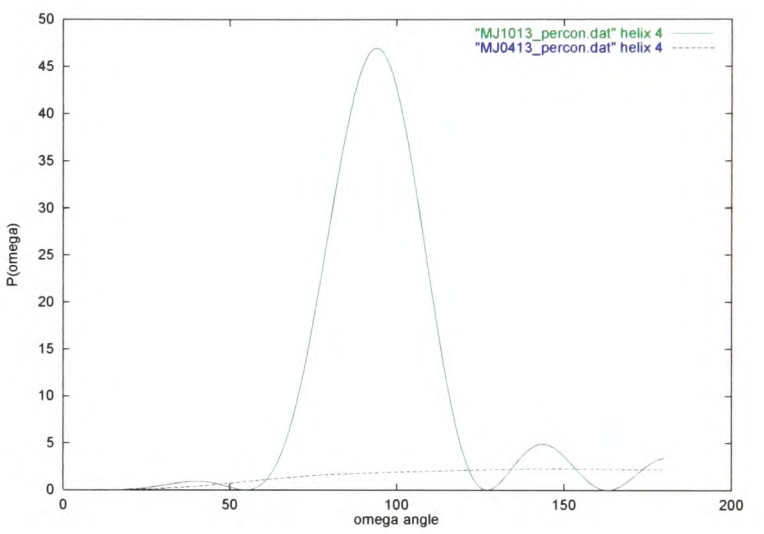
B. TM helix 2



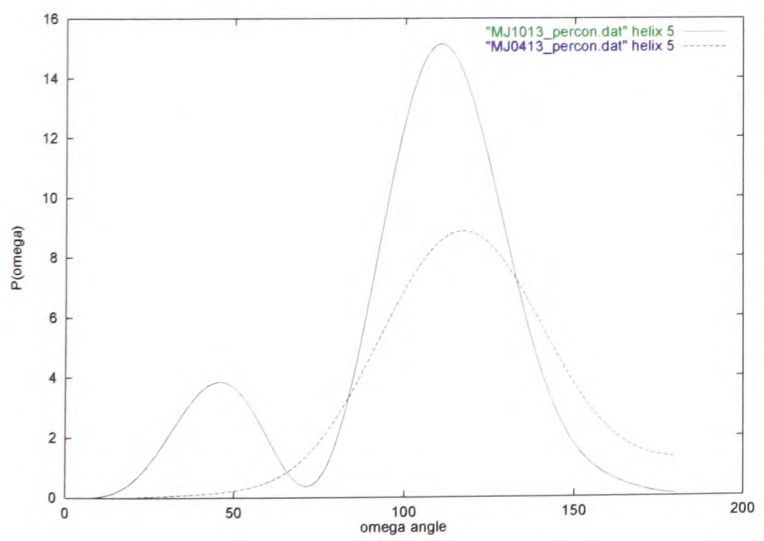
C. TM helix 3



D. TM helix 4



E. TM helix 5



F. TM helix 6

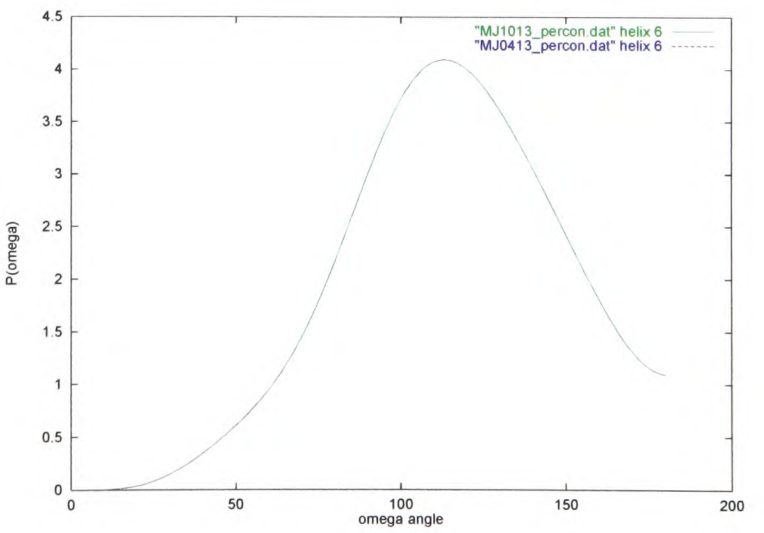
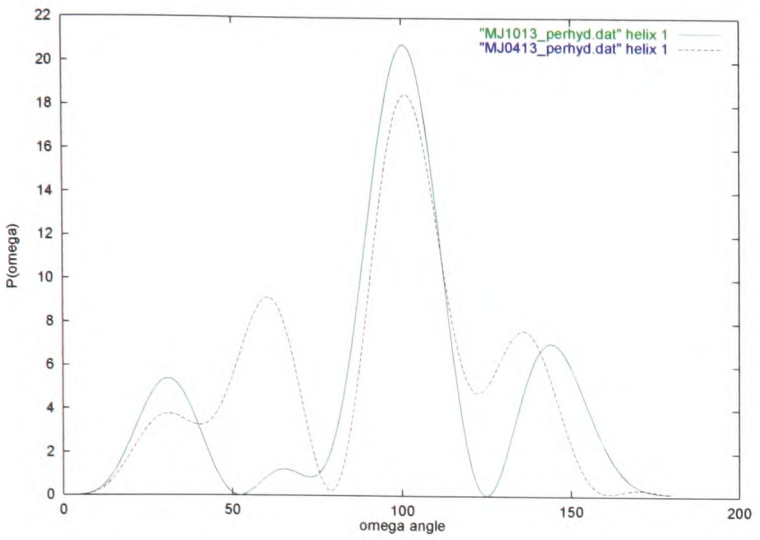
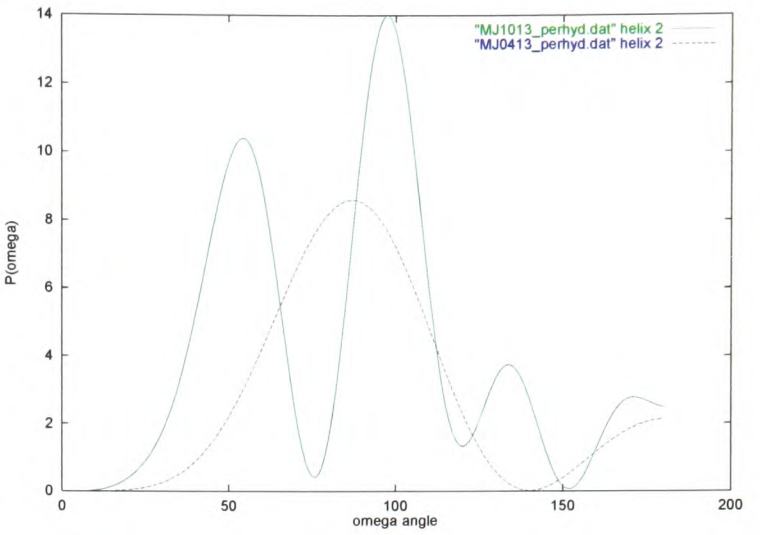


Figure 4.30 Spectra of sequence conservation periodicity of transporters MJ1013 and MJ0413 in *M. jannaschii*, for details see section 4.2.5.

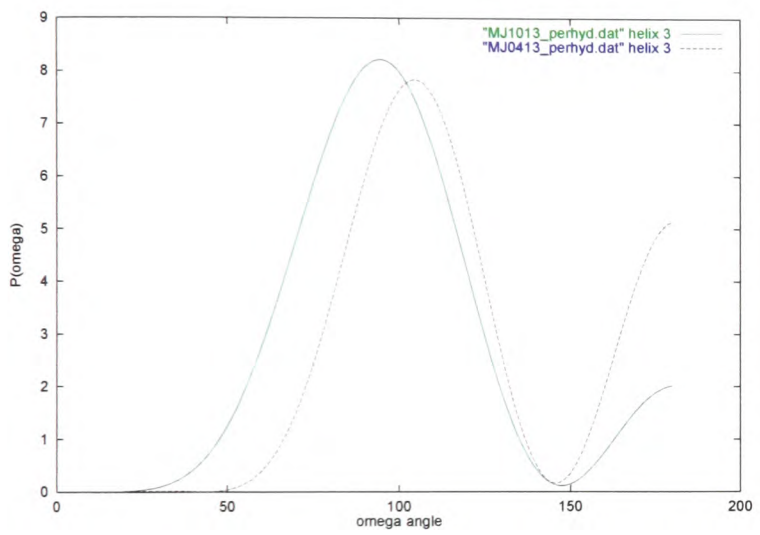
A. TM helix 1



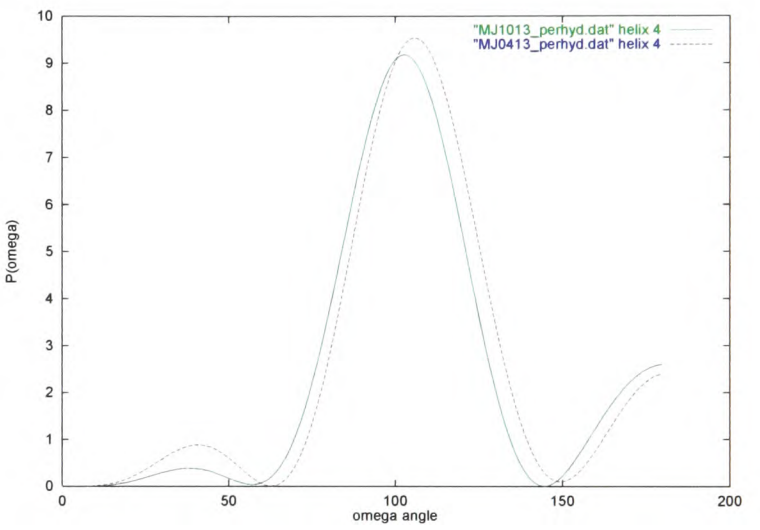
B. TM helix 2



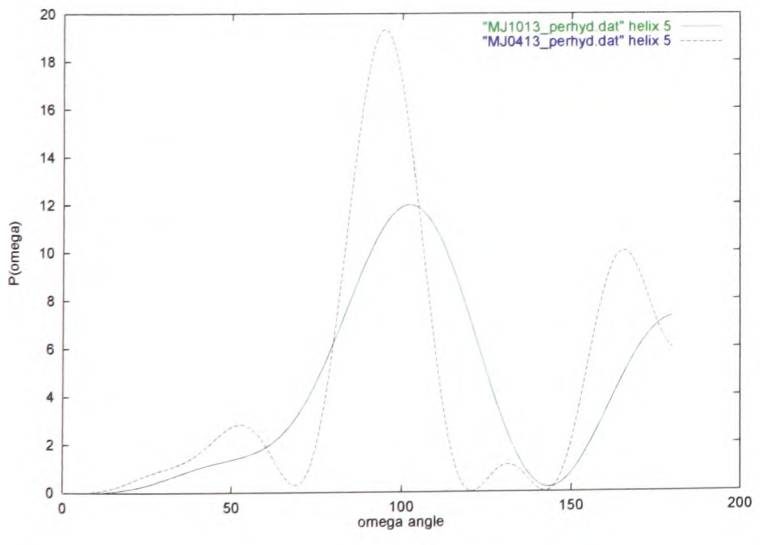
C. TM helix 3



D. TM helix 4



E. TM helix 5



F. TM helix 6

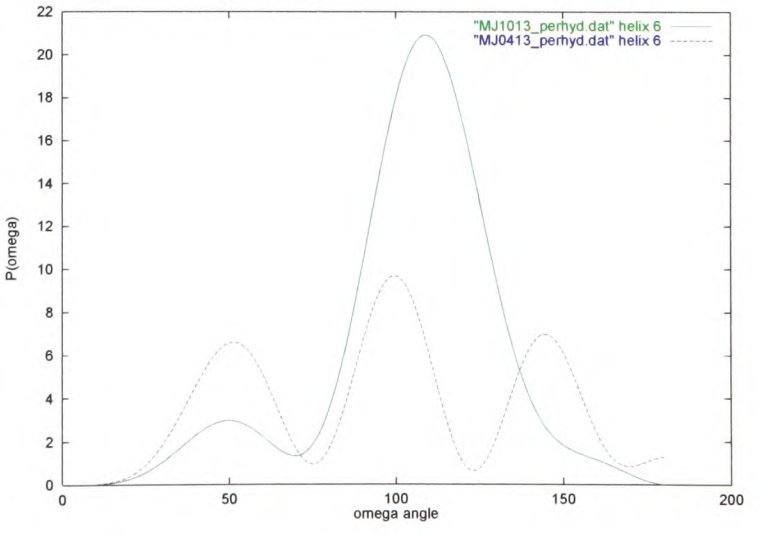


Figure 4.31 Spectra of amino acid residue hydrophobicity periodicity of transporters MJ1013 and MJ0413 in *M. jannaschii*.

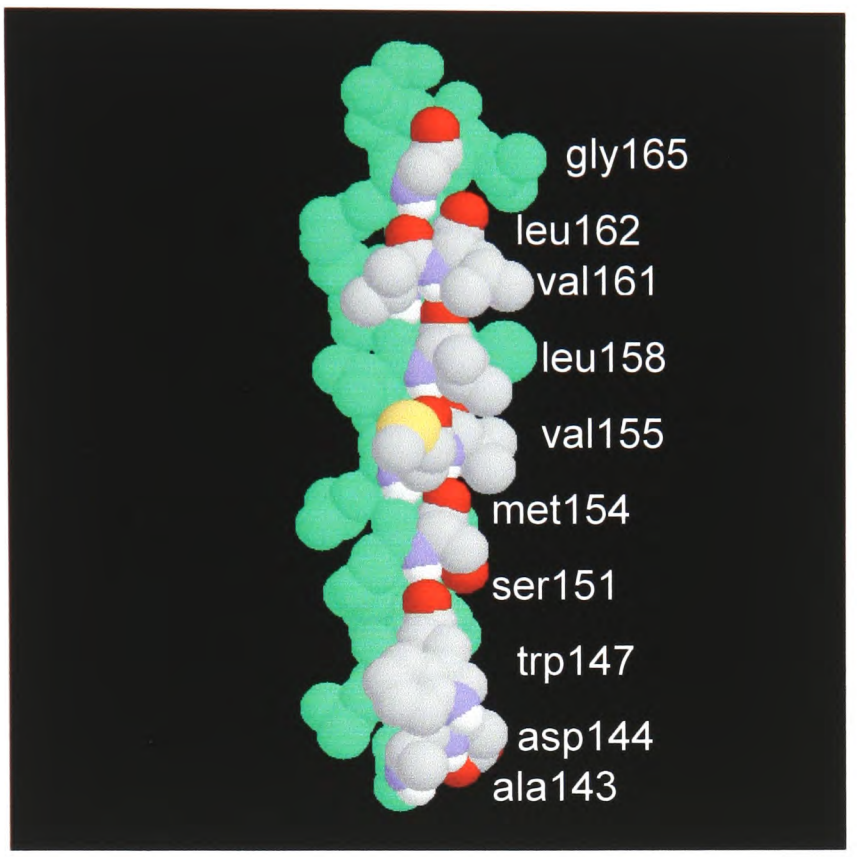
4.3.4 Models.

In order to acquire a better understanding of the predicted TM helices in the context of tertiary structure, ideal α-helical models were generated by restrained *in vacuo* molecular dynamics.

For example, the structure of the TM helix 3 of OppC in *B. subtilis* is shown, (Fig. 4.32 A). The residues predicted on the lipid bilayer-facing surface by periodic analysis are labelled. The clear interface between hydrophilic and hydrophobic residues, identified by the periodicity analysis, is clearly observed as displayed on a helical wheel, (Fig. 4.32 B). Residues predicted to be within the lipid-facing phase were franked with two arrows, and those hydrophobic residues within were denoted in shade.

Of course, this model was generated *in vacuo*. In order to get a more realistic model of the structure of this helix in a bilayer environment it is possible to refine such models via multi-nanosecond molecular dynamics simulations which include helix, lipid and water atoms. This is going to be discussed in Chapter 5.

A



B.



Figure 4.32 (A) Helical model of TM helix 3 of OppC; (B) helical wheel representation of residues on the interior face.

4.4 Discussion.

In this study, attempts have been made to study the ABC transporters from different organisms on the basis of sequence similarities and the particularities of predicted TM helices.

To achieve this goal, ABC transporters were picked up from the translated open reading frames of different genomes objectively, based on the deposed function tables, in order to eliminate possible mistaking of other proteins or transporters. After gathering reliable sets of ABC proteins, sequence analysis were carried out to detect basic evolutionary information of ABC transporters within the same genomes, and to pull out homologous sequences of each query sequence for enhancing accuracies of secondary structure prediction, as discussed at section 1.3.2 and 4.2.4.

Detailed structures of transmembrane domains of ABC transporters at atomic resolutions are not yet available. Structural information of each transporter could only be collected from prediction results. It became crucial to have reliable prediction methods. Different prediction programs have been tested. A program, PHDhtm, based on neural network was used mainly, with reinforcements from a program, TmHMM, using hidden Markov models.

Three traits were of interest to be examined; the number of predicted TM helices, the linear positions of predicted TM helices along the primary protein sequences and the possible fold arrangement within lipid bilayer of putative helices. The numbers of predicted TM helices for ABC transporters differ. Proteins with unexpected numbers of TM helices were observed frequently, for example 5, 7, 8 and 9. This implies two things. Firstly, the numbers of TM helices for ABC

transporters may not be always 6 or 12 as traditionally suggested. Actually, this has been observed in a previous study on *E. coli* genome (Linton and Higgins, 1998). Secondly, there are too many potential mistakes leading to these confusing figures; for example, errors in the sequence data, errors from extractions of ABC transporters or errors from secondary structure predictions. Uncertainties of prediction might be the major reason. For the same reason, detailed statistical analysis has not been performed, which had been used extensively in other studies of membrane proteins (Arkin et al., 1997; Arkin and Brunger, 1998; Wallin and von Heijne, 1998).

However, a crude survey of ABC transporters within identified clusters reveals two peaks at 6 and 12 for the distribution of TM helical numbers, (Fig. 4.33). This may indicate that most ABC transporters adapt a TM helical arrangement of six or twelve. The putative TM helices of 5, 7, 10, 11 and 13 are likely to be a reflection of inaccuracies for predictions. It is noted that a protein from *E. coli* is predicted to have 18 TM helices. Whether this protein is really composed of 3 TM domains and each TM domain contains 6 integral membrane helices remains to be examined experimentally.

Alternatively, attentions have been more focussed on the locations of predicted TM helices relative to their primary sequences and the likely TM helical packing patterns within the lipid bilayer deduced from sequence analysis. In general, it is found that proteins within the same sequence-generated clusters are likely to have similar numbers of predicted TM helices. Frequently but not always, proteins within the same sequence-generated clusters tend to transport similar solutes. In addition, they share similar linear pattern of locations for putative TM helices on

the one-dimensional sequences. Furthermore, with detailed periodic analysis of conservation and hydrophobicity, similarities in likely packing patterns of TM helices and arrangements of TM helices in a transmembrane bundle within cell membrane are detected for proteins at similar sequence-generated clusters.

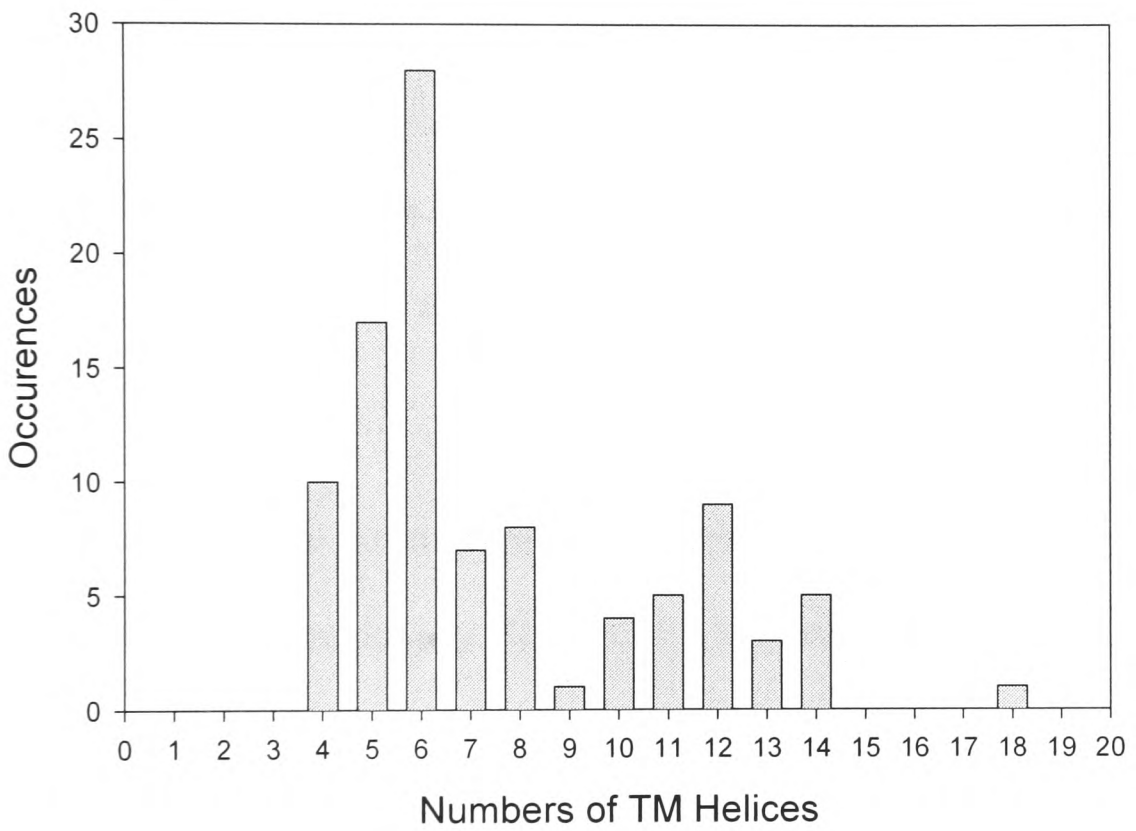


Figure 4.33 A crude survey of TM helical numbers in ABC transporters identified within different clusters. The predicted TM helices of 4, 5, 7 and 8 are likely to be inaccurate predictions for TM helices of 6. The predicted TM helices of 10, 11, 13 and 14 are likely to be inaccurate predictions for TM helices of 12.

In short, correlations between arrangements, numbers and locations, of putative TM helices and phylogenetic relationships are detected. In addition, the correlations between spatial arrangements and orientations of TM helices with lipid bilayer between phylogenetic relationships are seen by sequence periodic analysis. These findings suggest that ABC transporters with similar function are likely to develop similar structural complex of TM helices, and to evolve in a correlated fashion.

However, when proteins with similar solutes were mentioned, their transporting mechanism, influx or exflux, are not known. Since, this information is not annotated in the sequence data. And this could cost vast amount of human work and interventions to categorise these transporters on the basis of their transporting mechanisms.

Transporters with similar positions along the primary protein sequences for their putative TM helices were largely detected by human inspections. Locations of predicted TM helices were superimposed on the basis of their amino acid sequence alignments. Developments of a automated devise to match potential proteins of interest, and scoring schemes to evaluate similarities in the context of relative arrangement TM helices on one-dimension will be useful for more extensive survey.

In a similar manner, detection of similarities was done with combinations of semi-automated mechanisms and human inspections. Developments of a mechanism to recognise similar patterns of periodic spectra for proteins of interests will make this strategy of studying evolutionary and functional relationships between

membrane proteins by their structural similarities of predicted transmembrane α-helical domains more comprehensive and applicable.

As the genome data are constantly changed and updated, data used in this study simply reveal information at a certain time frame of sampling in the ongoing process, and could be subject to possible errors in sequencing and annotating. In addition, an alternative approach to explore phylogenetic correlations of ABC transporters, which focus on identified periplasmic binding proteins, was claimed to be more accurate than analysis including membrane-spanning (Saurin and Dassa, 1994). Even though, some interesting trends are identified in this study.

Recent progresses in completions of genome sequencing projects have provided a vast amount of information, which were not available before, and thus led to a new direction about how biological research is performed (Montelione and Anderson, 1999). Hugh amount of effort has been focussed on characterisation of these completed genomes by different approaches. Analysis based on predicted protein secondary structures and sequence comparisons with structure-solved molecules were performed in order to identify the structural census of different genomes (Gerstein, 1997; Gerstein, 1998). Homology searching and predicted fold recognition were applied to assign folds to all proteins encoded in *M. genitalium* (Huynen et al., 1998) and to detect possible protein folds in *M. pneumoniae* (Grandori, 1998). Combinations of fold assignment, comparative protein structure modelling and model evaluation were automated and used in the *S. cerevisiae* genome to detect likely structures of yeast protein sequences (Sanchez and Sali, 1998). Attempts of using sequence analysis and protein phylogenetic profiles had been tested to predicted the function of uncharacterised proteins (Pellegrini et al.,

1999). For membrane proteins, a systematic analysis on the basis of putative membrane topology, protein family, bioenergetics and substrate specificities was performed to estimate the overall transport capabilities of different genomes and to assigned probable functions to transporters identified (Paulsen et al., 1998).

Studies of ABC transporters have been also benefited from these completed genomes. The complete inventory of ABC proteins in the *S. cerevisiae* genome had been reported (Decottignies and Goffeau, 1997). Details about evolutionary relationships, functions and substrate specificities of ABC proteins in *E. coli* genome had been revealed (Linton and Higgins, 1998). With identifications of nucleotide-binding domain, membrane spanning domain and solute-binding protein, an assembly of ABC proteins had been compiled and studied (Quentin et al., 1999). By collecting similar proteins to a representative ABC protein, MalK, a list of ABC proteins were found; phylogenetic analysis of this ABC protein list suggested that ABC transporters for uptaking and exporting adapted different evolving routes (Saurin et al., 1999).

Chapter 5

MD Simulations of the CFTR Transmembarne Helices

5 Preface to Chapter 5

Studies of the ABC transporter in different genomes reveal the limitations the computational studies of membrane proteins only based on sequence analysis. Some more sophisticated methods are available to refine the analysis of putative structures. However, for these methods, more powerful computers, longer computational time and more efficient algorithms are needed. In this chapter, molecular dynamics simulations were applied to studies of the transmembrane helices of cystic fibrosis transmembrane conductance regulator (CFTR).

This chapter is divided into 4 sections. Section 5.1 starts with introduction to cystic fibrosis transmembrane conductance regulator (CFTR) and MD simulations. Detailed methods employed in this study are presented in section 5.2. Simulations results are analysed in section 5.3. Biological implications of these MD simulations and future directions are covered in section 5.4.

5.1 Introduction.

5.1.1 Cystic Fibrosis Transmembrane Conductance Regulator.

Cystic Fibrosis

Cystic fibrosis is an inherited disorder, which affects the cells and glands that produce mucus, tears, sweat, saliva and digestive juices. For normal people, these secretions are thin and slippery; however, for people with cystic fibrosis, these secretions are thick and sticky. These abnormal secretions tend to block tubes, ducts and passageways, especially in the pancreas and lungs. Lung disease and lung damage may occur and often lead to respiratory failure, which are the most

dangerous consequences of cystic fibrosis. The sweat glands and reproductive system are affected as well. Clinically, cystic fibrosis is a chronic disease and generally worsens with time. The result of this disease is almost always fatal. (Riordan, 1993). In addition, overstimulation of CFTR in intestinal crypt cell by various enterotoxins results in secretory diarrhea, which is one of the major causes of infant mortality (Field and Semard, 1993). These abnormalities of unusual thick, sticky mucus are caused with a defective gene, which leads to productions of a mutant protein that alters the normal movement of chloride in and out of epithelia cells. The product of this defective gene is thus named as cystic fibrosis transmembrane conductance regulator (CFTR) (Collins, 1992; Riordan, 1993).

Biology of CFTR

CFTR is a member of the ABC transporter family (Riordan et al., 1989). Electrophysiologic analysis of CFTR has shown that this protein is a chloride channel (Anderson et al., 1991b; Bear et al., 1992). It has been observed that CFTR forms an ion channel involved in the processes of chloride secretion and reabsorption in a variety of epithelia and other cells (Gadsby et al., 1995). The Cl^- transportation activity of CFTR is cAMP-dependent and under complex regulation by phosphorylation and cytoplasmic ATP (Cheng et al., 1991; Anderson et al., 1991a).

In addition to the well-defined cAMP-dependent Cl^- channel activity, CFTR has been suggested to be a regulator of other epithelial ion channels; such as an outward-rectifying Cl^- channel in airway epithelia (Schwiebert et al., 1995; Schwiebert et al., 1998), an epithelial amiloride-sensitive Na^+ channel (Stutts et al., 1995; Stutts et al., 1997; Kunzelmann et al., 1997), and a kidney K^+ channel

(McNicholas et al., 1996; Ho, 1998). However, the mechanisms by which these regulatory functions contribute to the maintenance of airway electrolyte and fluid balance remain to be identified.

CFTR has also been suggested to play a role in ATP release, which is regulated by the phosphorylation- and nucleotide-hydrolysis-dependent gating (Sugita et al., 1998; Cantiello et al., 1998). However, these suggestions are not conclusive and remain controversial, since contradictory evidences have been provided to show that there is no association of CFTR with ATP permeability (Reddy et al., 1996; Li et al., 1996). Recently, a study has shown that the CFTR-modulated ATP release is under the regulation of a Cl^- sensory mechanism, which is yet to be clarified (Jiang et al., 1998).

General Structure of CFTR

CFTR is a glycoprotein with a molecular weight of 140 kD approximately and a member of the ABC transporter family (Riordan et al., 1989). Like other ABC transporters, CFTR is predicted to be comprised with two major parts, transmembrane helical domains and cytoplasmic domains, on the basis of hydrophobicity analysis. The overall structure proposed consists of 5 functional domains: 2 hydrophobic membrane-spanning domains, each including 6 transmembrane helices; 2 hydrophilic ATP-binding domains; and a regulatory domain which carries consensus sequence for phosphorylation by protein kinase A and C, (Fig. 5.1), (Riordan et al., 1989).

Phosphorylation of the regulatory domain controls the channel activity of CFTR *via* a mechanism, which involves modulation of the ATP-binding domains' affinity for ATP (Winter and Welsh, 1997). The cAMP-dependent

phosphorylation of the regulatory domain by protein kinase A is likely to induce conformational changes or to modulate electrostatic interactions of the regulatory domain with other parts of the CFTR (Rich et al., 1993). Furthermore, it has been suggested that the CFTR gating is under the regulation of MgATP through a stabilisation of the opened channel conformation in a complex form of CFTR-MgATP (Aleksandrov and Riordan, 1998).

The structures and topologies of the two putative transmembrane domains are generally supported by experimental data derived from epitope detection by antibody binding to the loop regions (Schultz et al., 1997), and glycoslation site insertion to map the loops (Chang et al., 1994). A hypothesis has been proposed that the second transmembrane domain may contribute to the overall channel activity of the pore, and the first transmembrane domain may confer the ion selectivity of CFTR (Devidas et al., 1998). The membrane topology of the human multidrug resistance-associated protein, a protein related to CFTR, has been examined and found to have similar topology of CFTR (Bakos et al., 1996; Rosenberg et al., 1997).

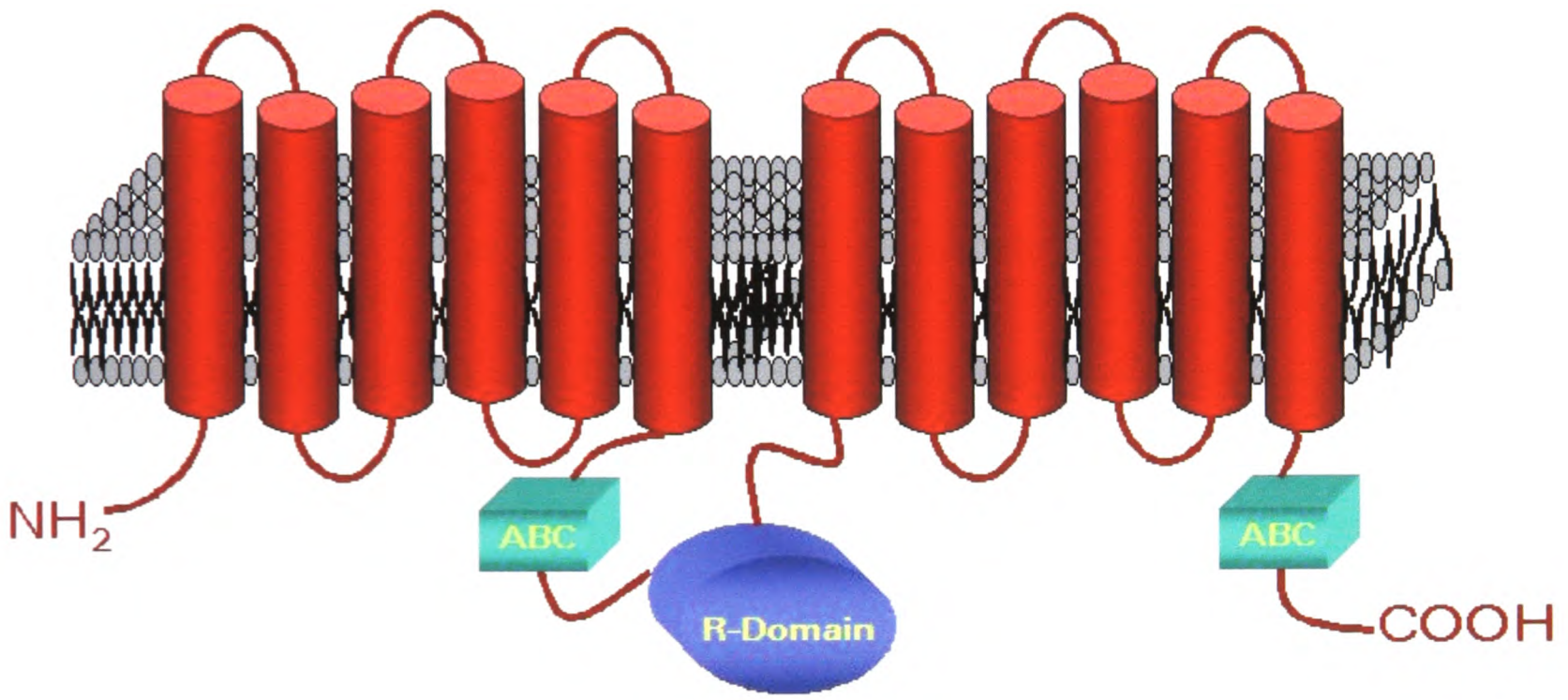


Figure 5.1 Schematic model of the predicted CFTR protein. The six transmembrane helices in each half of the molecule are denoted as red cylinders. The ATP-binding domains are shown as green cassette. The two separated halves are connected with the cytoplasmic regulatory domain, which is denoted as blue sphere.

Folding of CFTR

Various protein factors have been reported to be essential for facilitating folding process of cytoplasmic domains of CFTR. For example, a molecular chaperone, Hsp90, has been proposed to contribute to folding and maturation of the newly synthesised CFTR by interaction with the cytoplasmic domains of CFTR on the endoplasmic reticulum (Loo et al., 1998). In addition, other chaperonic proteins, Hdj-2 and Hsc70, have been suggested to facilitate early steps in CFTR assembly by catalysing the formation of an intramolecular complex of first ATP-binding domain and the regulatory domain (Meacham et al., 1999).

Detailed folding mechanisms for the 12 CFTR TM helices in the lipid bilayer are yet to be understood. However, the folding of α-helical integral membrane proteins has been proposed to be a two-stage mechanism (Popot and Engelman, 1990), which involves (1) formation of the integral helical structures in the non-aqueous environment of membrane, and then (2) spatial arrangement of TM helices to build up the tertiary fold within the lipid bilayer (Popot, 1993), (Fig. 5.2).

Technical difficulties in determining three-dimensional structures of CFTR transporter remain the greatest limitation to undergo its functional and structural analysis. However, considerable insight can be gained from their putative two-dimensional topological arrangement in the lipid bilayer. Efficient computational methods provide a way to understand better the biophysical behaviours of each helix.

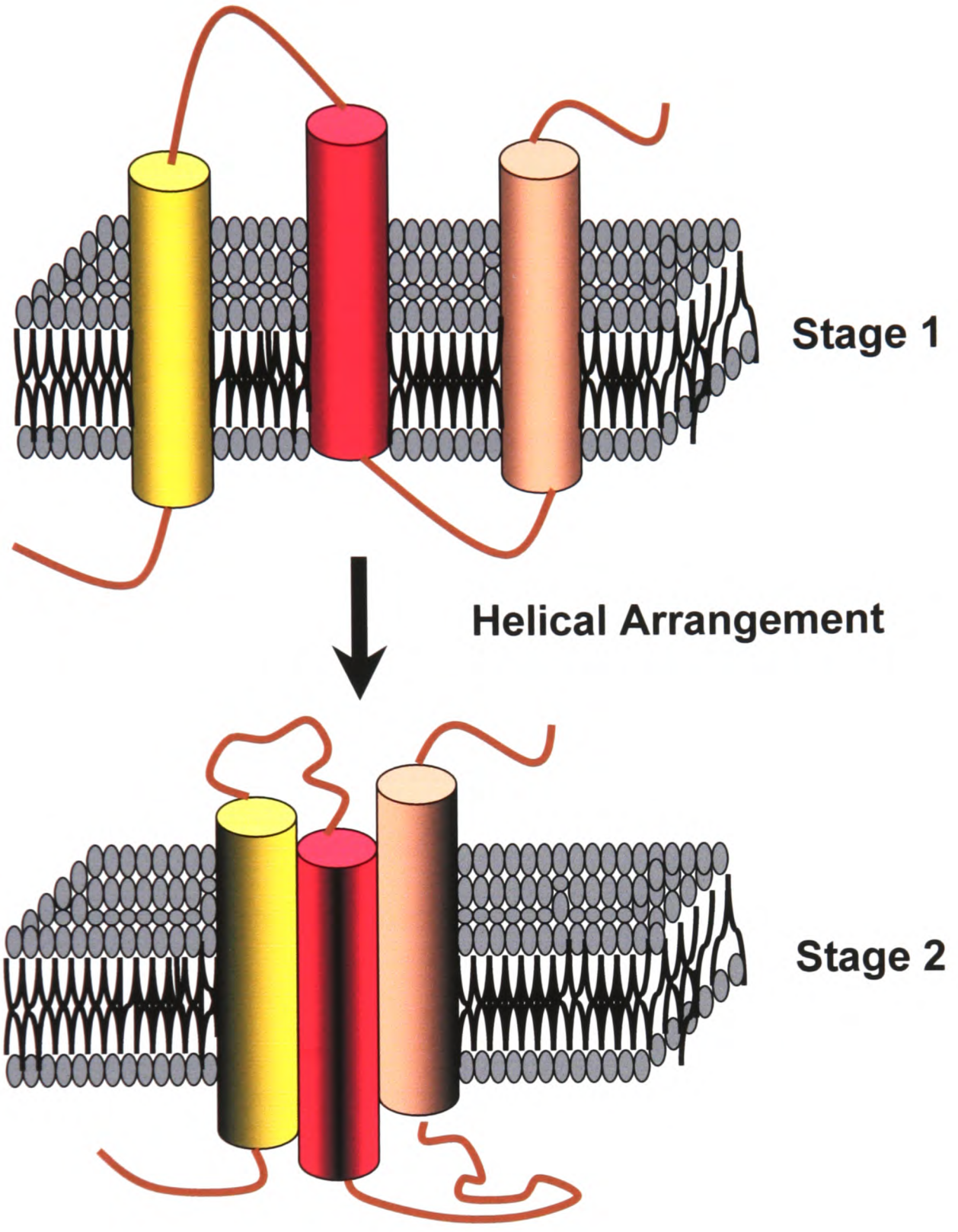


Figure 5.2 Two-stage model for the folding of α-helical integral membrane protein. The first stage is the formation of independently stable TM helices, which is in response to the hydrophobic effect and the formation of main-chain hydrogen bonds in the non-aqueous environment. The second stage is the interaction of the helices to form the tertiary fold of the polypeptide. Factors that could contribute to the energies of stage 2 are the links between helices, packing of helices and lipid molecules, polar interactions between helices, and, when applicable, association with prosthetic groups or with other proteins.

5.1.2 Applications of Molecular Dynamics to Integral Membrane Protein.

As mentioned in section 1.3.4, molecular dynamics simulations have advantages over examining interactions between molecules at atomic resolution for dynamical systems. Due to the fluid nature of membrane and limited membrane protein structures available, this technique has been applied extensively in studies of lipid bilayers and transmembrane proteins (Tieleman et al., 1997; Biggin and Sansom, 1999).

Simulations of Lipids

The liquid and mobile characters of lipid present problems for studies of lipid bilayers with traditional experimental techniques, such as X-ray diffraction and NMR. Molecular simulations contribute an alternative method to the studies of lipid bilayer on an atomic scale with the limited information gained from experiments.

For example, the MD simulations were applied in a system containing both dipalmitoyl phosphocholine (DPPC) and water molecules to study the process of water permeation through the lipid membrane (Marrink et al., 1996). The adhesion forces of lipids in the phospholipid bilayer were evaluated with the applications of MD Simulations in a DPPC system (Marrink et al., 1998).

In general, the MD simulations of nanosecond (*ns*) scale are capable of discerning the dynamics of individual lipid molecules and the different physical behaviours of different types of lipids. However, the slow motions like rotational or translational motions are still beyond the capacities of current algorithm designs and computational power (Tieleman et al., 1997).

Simulations of Peptide and Lipid Bilayer Interactions

The insights of interactions between membrane proteins and their environments can be obtained from studying interactions of α-helical peptides with lipid bilayers by MD simulations (Sansom, 1998).

More research interests have been focussed on the interactions between integral membrane α-helices and dimyristoyl phosphocholine (DMPC) lipid bilayers. An earlier molecular dynamics study between a TM helix and membrane was performed on a simple designed transbilayer helix (Belohorcova et al., 1997).

Detailed structures and dynamics of a TM helix in membranes were revealed from a MD simulations study which looked into interactions between the DMPC lipids and each helix derived from the bacteriorhodopsin structure, PDB code 1BRD, (Woolf, 1997). The concepts of hydrophobic moments and the correlation between amino acid types, their positions along the TM helix and details of interaction were clarified in a subsequent study (Woolf, 1998).

In addition, molecular dynamics are of great help in understanding the roles of lipid bilayers in membrane protein structure formations. This knowledge is especially useful in determining or predicting tertiary structures of a membrane protein. For example, it has been demonstrated that molecular dynamics simulations provides a more reliable method to refine the predictions of TM helical structures purely from hydrophobicity analysis of the primary amino acid sequences (Forrest et al., 1999).

Simulations of α-Helix Bundle and Lipid Interactions

With the advances of computer power and some recently solved membrane protein structures. It has become possible to extend simulations on a larger system to accommodate an α-helix bundle, consisting of more than one isolated TM helix, or even an intact TM protein (Tieleman and Berendsen, 1998). A recent MD simulations study has been reported on a system of a helix bundle containing 6 alamethicin TM helices in a solvated 1-palmitoyl-2-oledyl phosphocholine (POPC) lipid bilayer; ionisation states of certain amino acids, conformational dynamics of the channel pore and movements of water molecules were investigated (Tieleman et al., 1999b). Similar studies have been extended to the same bundle with different numbers of TM helices (Tieleman et al., 1999a).

Limitations of MD Simulations

There are limits for the applications of molecular dynamics simulations. Firstly, parameters are required to describe different physical interactions, *e.g.* charges, van der Waals interactions and force constants *etc.*, in the potential functions. These parameters are usually derived and simplified from experiments or quantum mechanics. How real do the simplified parameters reflect nature remain uncertain and are in need of further test on simple systems with supporting experimental data. In addition, the current computational capacity still limits the time scale of simulations at the *ns* range, during which biological and chemical interactions cannot be observed (Tieleman et al., 1997).

5.1.3 Aims of This Study.

In this study, molecular dynamics simulations were used in order to understand biophysical properties of individual CFTR TM helix in a solvated 1-palmitoyl-2-

oledyl phosphocholine (POPC) lipid bilayer system, to dissect interactions of specific amino acids and TM helices, POPC lipid and water, and explore the membrane-spanning regions of CFTR at the atomic resolution.

5.2 Methodology.

5.2.1 Overview of Methods.

The 12 putative TM helices were located with different secondary structure prediction algorithms for membrane proteins and in conjunction with experimental data, section 5.2.2. Idealised α-helical structures were constructed for the 12 predicted helical segments by the SA/MD protocol (Kerr et al., 1994). Each α-helical segment in the ensemble of constructed α-helices was examined by its stereochemistry, section 5.2.3. Each selected helix was then docked into a pre-existing POPC bilayer system (Tieleman et al., 1999b), section 5.2.4. After solvation and energy minimisation of the whole system, production MD runs for 1 ns were performed for each of the 12 TM helices. A flow diagram of methods used in this study is given in Figure 5.3.

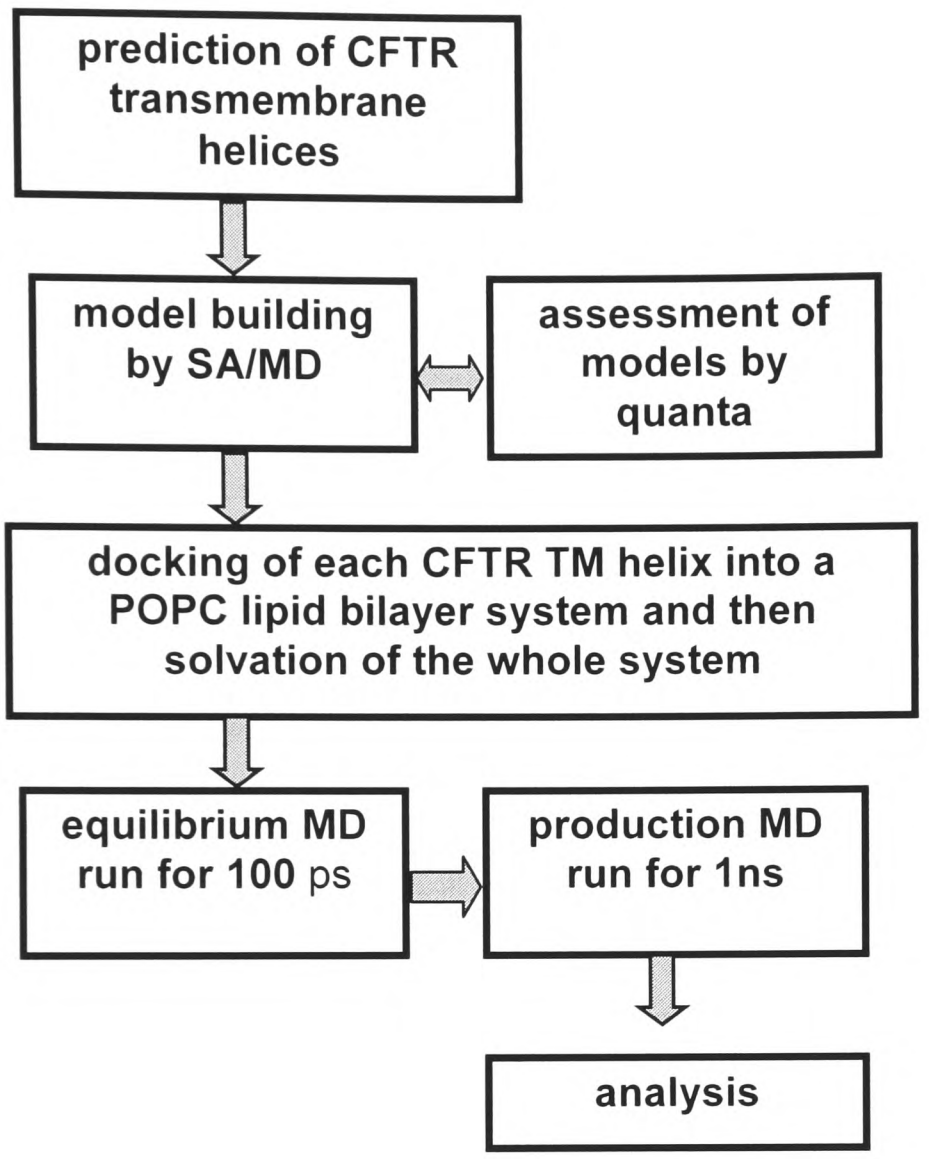


Figure 5.3 Flow diagram summarising methods used in this study.

5.2.2 Predictions of Transmembrane Helices.

The primary protein sequence of human cystic fibrosis transmembrane conductance regulator was downloaded from SWISS-PROT database (Bairoch and Apweiler, 1998), accession number P13569.

Different algorithms, Table 5.1, were employed to predict the TM helices of CFTR. Experimental data on loop detection either by amino acid glycoslation (Chang et al., 1994) or antibody mapping (Schultz et al., 1997) were used in conjunction with the prediction results in order to enhance the accuracy for locating the transmembrane regions.

DAS	http://www.biokemi.su.se/~server/DAS/
TOPPRED2	http://www.biokemi.su.se/~server/toppred2/
MEMSAT	http://globin.bio.warwick.ac.uk/~jones/memsat.html
SOSUI	http://www.tuat.ac.jp/~mitaku/ady_sosui/
TmHMM	http://www.cbs.dtu.dk/services/TMHMM-1.0/
TMAP	http://www.embl-heidelberg.de/tmap/tmap_info.html
PHD	http://www.embl-heidelberg.de/predictprotein/predictprotein.html

Table 5.1 Different TM helix prediction programs used in this study and their URLs.

A preliminary study on the structure-solved membrane proteins reveals that average length of the TM helices is ca. 25 residues (Hong, unpublished results). This is similar to a detailed statistical study which suggests that TM helix lengths range from 14 to 36 residues with an average length of 26.4 residues (Bowie,

1997). Combining information gathered from prediction results, experimental data and the length cut-off at 25 residues, the 12 CFTR TM helices were located. Their lengths range from 21 to 25 amino acid residues, (Tab. 5.2).

TMH	Amino Acid Sequence	Length	REGION
TM1	**FFWRFMFYGIFLYLGEVTKAVQPLL**	25	***77*** to ***101***
TM2	**IAIYLGIGLCLLFIVRTLLLHPAIF**	25	***119*** to ***143***
TM3	**GLALAHFVWIAPLQVALLMGLIW**	23	***194*** to ***216***
TM4	**SAFCGLGFLIVLALFQAGLGRMM**	23	***222*** to ***244***
TM5	**FNSSAFFFSGFFVVFLSVLPYAL**	23	***305*** to ***327***
TM6	**LRKIFTTISFCIVLRMAVTRQFPW**	24	***333*** to ***356***
TM7	**IFVLIWCLVIFLAEVAASLVVLW**	23	***860*** to ***882***
TM8	**YVGVADTLLAMGFFRGLPLVHTLIT**	25	***919*** to ***943***
TM9	**TIFDFIQLLLIVIGAIAVVAV**	21	***990*** to ***1100***
TM10	**FVATVPVIVAFIMLRAYFLQTSQQ**	24	***1016*** to ***1039***
TM11	**WFQMRIEMIFVIFFIAVTFISILTT**	25	***1098*** to ***1122***
TM12	**GIILTLAMNIMSTLQWAVNSSI**	22	***1130*** to ***1151***

Table 5.2 Predictions of CFTR transmembrane helices. "REGION" represents the location of each TM helix within the CFTR primary protein sequence.

5.2.3 Building Helical Models.

A restrained SA/MD protocol (Kerr et al., 1994) was used to obtain ensembles of helical structures by using Xplor (Brunger and Karplus, 1991), for details see Section 4.2.6. Twenty-five structures were generated for each CFTR TM helix.

Each TM helix was N-acetylated and C-amidated to mimic the preceding and following bonds within the intact protein.

Assess structural quality of each helix.

The stereochemistry of each helix was examined using Quanta (Biosym/MSI) to compare the side chain conformations with different rotamer libraries (Sutcliffe et al., 1987; Ponder and Richards, 1987; Dunbrack Jr and Karplus, 1993). After filtering through three libraries, the best structure was then picked up for further analysis. As an example, the structure for CFTR TM helix 3 is shown, (Fig. 5.4).

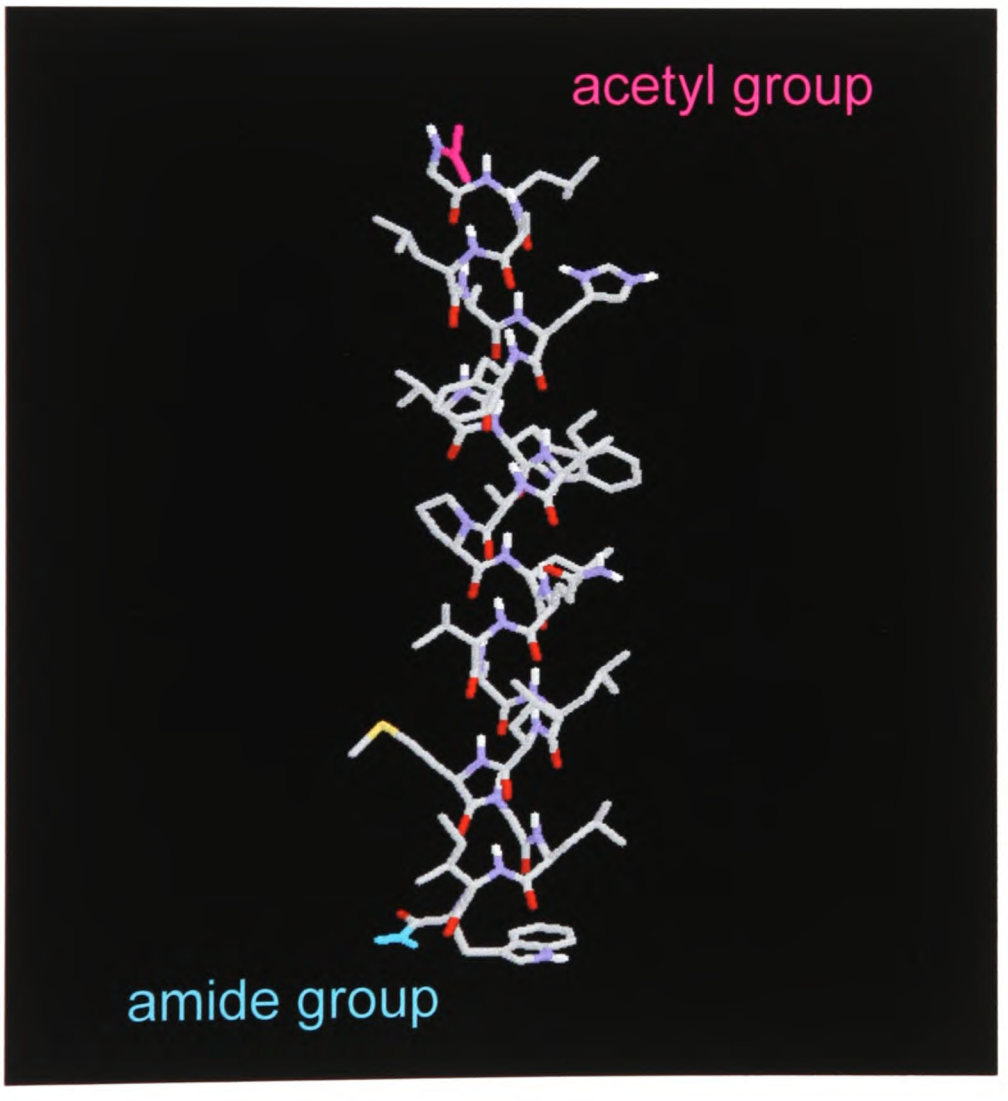


Figure 5.4 Idealised TM helix generated by SA/M for CFTR transmembrane helix 3.

5.2.4 Construction of Helix/Bilayer/Water Simulation Systems.

Each idealised TM helix of CFTR was docked into a pre-equilibrated bilayer system, which was of 127 molecules of 1-palmitoyl-2-oleoyl-*sn*-glycerol-3-phosphocholine (POPC). In this POPC system, a cylindrical hole of radius 0.7 nm was generated by a short MD simulations in order to accommodate a TM helix.

After docking, the system was energy minimised and then solvated with a minimum of 30 to 40 SPC water molecules per lipid molecule (Hermans et al., 1984; Fraternali and van Gunsteren, 1996). For those systems with a net charge on the TM helix, counterions, Na^+ or Cl^-, were added by a procedure which systematically positioned the ion instead of each water molecule by turn, eventually selecting the ion position with the lowest potential energy. After addition of ion(s), each system was once again energy minimised, followed by a short equilibration simulation of 100 ps, during which the Cα atoms of each helix were restrained to their initial coordinates. Twelve helix/bilayer/water simulation systems were constructed with this procedure for TM helices of CFTR. Details of each system are shown, (Table 5.3).

TMH	Helical Length	No. H_2O	No. Ion	Total Atom No.	CPU Time
TMH1	25	4475		20336	18635
TMH2	25	4257	1 Cl^-	19624	19301
TMH3	23	4262		19617	17453
TMH4	23	4509	1 Cl^-	20351	18961
TMH5	23	4508		20389	19383
TMH6	24	4514	3 Cl^-	20427	19100
TMH7	23	4272	1 Na^+	19658	18493
TMH8	25	4271		19661	20104
TMH9	21	4271	1 Na^+	19615	18258
TMH10	24	4268	1 Cl^-	19665	18224
TMH11	25	4255		19658	17828
TMH12	22	4270		19623	17591

Table 5.3 MD simulations systems for 12 CFTR TM helices. "No." denotes the number of each object. Molecular dynamics simulations were performed with a paralleled version of GROMACS on a SGI Power Challenge super computer. Each job was diverted on eight processors. The CPU time was estimated from the running time of the first CPU times 8, which could be slightly higher than the actual CPU time consumed. The CPU time is expressed in minute.

5.2.5 MD Simulations.

Protocols for molecular dynamics simulations were adopted from previous studies (Tieleman et al., 1999a; Tieleman et al., 1999b) using GROMACS (Berendsen et al., 1995) with periodic boundaries and under NTP conditions, *i.e.* constant number of particles, pressure and temperature. A production simulation of 1 *ns* was run for each helix. The timestep was 2 fs, with the LINCS algorithm used to constrain bond lengths (Hess et al., 1997). A constant pressure of 1 bar independently in all three directions was used, with a coupling constant of $\tau_P = 1.0$ ps (Berendsen et al., 1984). This allows the bilayer/peptide area to adjust to its optimum value for the forcefield employed. Water, lipid and protein were coupled separately to a temperature bath at 300 K, using a coupling constant $\tau_T = 0.1$ ps. For long-range interactions, a twin range cutoff is applied. For van der Waals interactions, 1.0 nm is used; and for electrostatic interactions, 1.7 nm was used.

The lipid parameters were derived from previous MD studies of lipid bilayers (Berger et al., 1997; Marrink et al., 1998), which give good reproduction of the experimental properties of a DPPC bilayer. The SPC water model has been shown to behave well in lipid bilayer/water simulations (Tieleman and Berendsen, 1996).

5.3 Analysis.

5.3.1 Progress of Simulations.

The progresses of simulations were monitored by potential energy, kinetic energy, total energy and density (kg/l) for each system in order to ensure the accuracy of the MD production run. High energy fluctuations are not expected throughout the 1 *ns* simulations. The density distributions were expected always to relax to ca. 1 kg/l, (Figure 5.5), a pattern shown for the TM helix 3 and also observed for the other 11 TM helices.

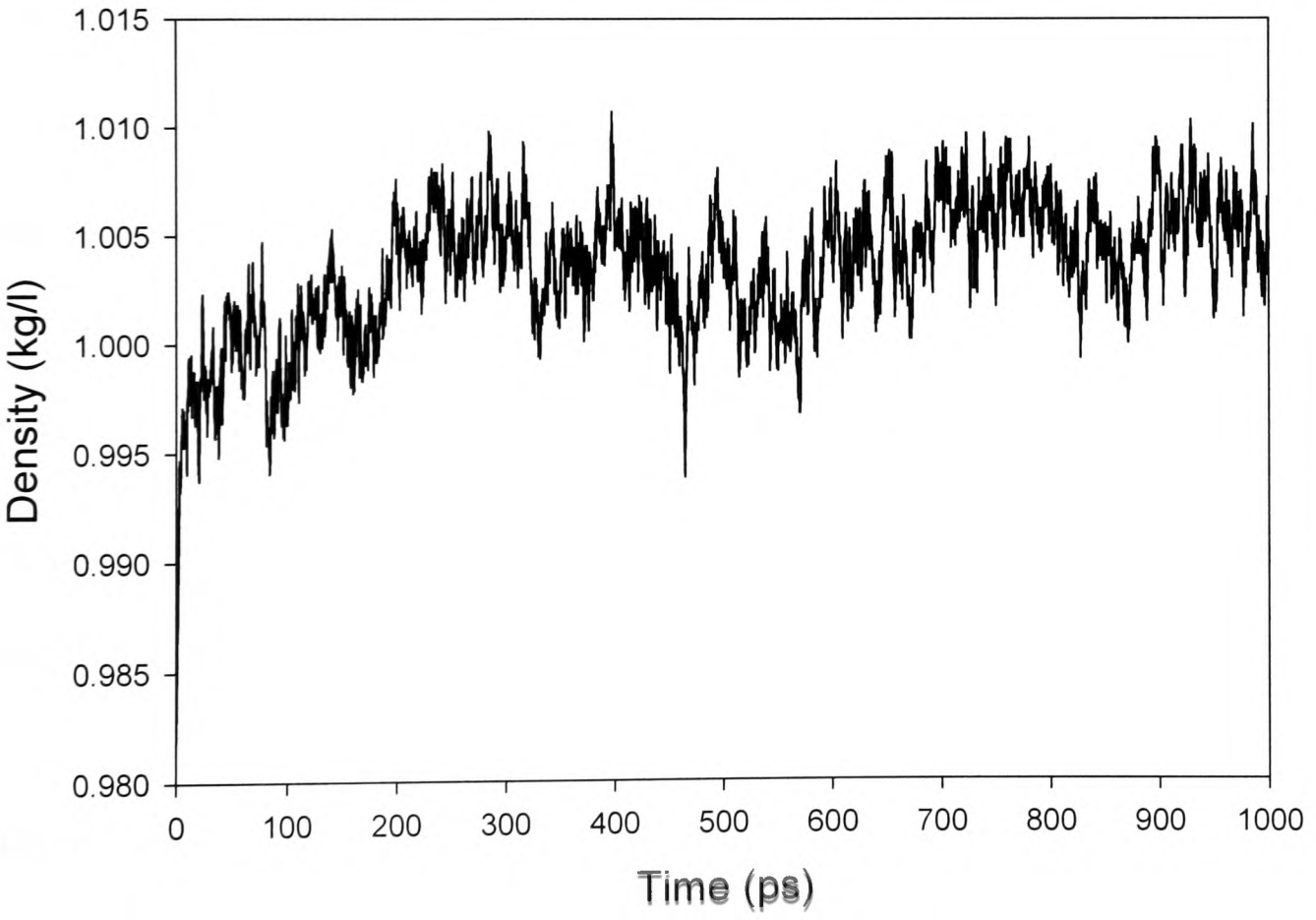


Figure 5.5 Density distributions of the TM helix 3 of CFTR. Similar patterns were seen for other TM helices of CFTR.

The macroscopic behaviour, conformations and positions, of the TM helix within the solvated POPC system could be observed from displaying the trajectory at different time points, (Fig. 5.6).

Considering the TM helix 3 as an example, at the starting point of the simulation, the TM helix 3 sat on the centre of the system after the equilibrium MD run, (0 ps at the Fig. 5.6). After the first 200 ps of production run, the TM helix was observed skewed and twisted to the left, (200 ps at the Fig. 5.6). A kink near the centre of the TM helix began to emerge after another 200 ps, (400 ps at the Fig. 5.6). At the timestep of 600 ps, the TM helix was observed to slide away from the centre toward the left, (600 ps at the Fig. 5.6). At the timestep of 800 ps, the TM helix was observed to drift further to the left and a kink angle became clear, which leads to the upper half of TM helix pointing toward to the right while the whole TM helix drifting to the left, (800 ps at the Fig. 5.6). At the end of 1 *ns* production run, a kink was clearly detected, see section 5.3.8, and possible interactions between C-terminus of the TM helix 3 and water molecules were seen due to the changes of the water molecule positions.

These snapshots for different TM helices at different time steps revealed possible intermolecular interactions, *e.g.* helix and lipid, or intramolecular interactions, *e.g.* amino acid residues within TM helix. With these interactions spotted, more detailed analysis on a microscopic level could be initiated efficiently.

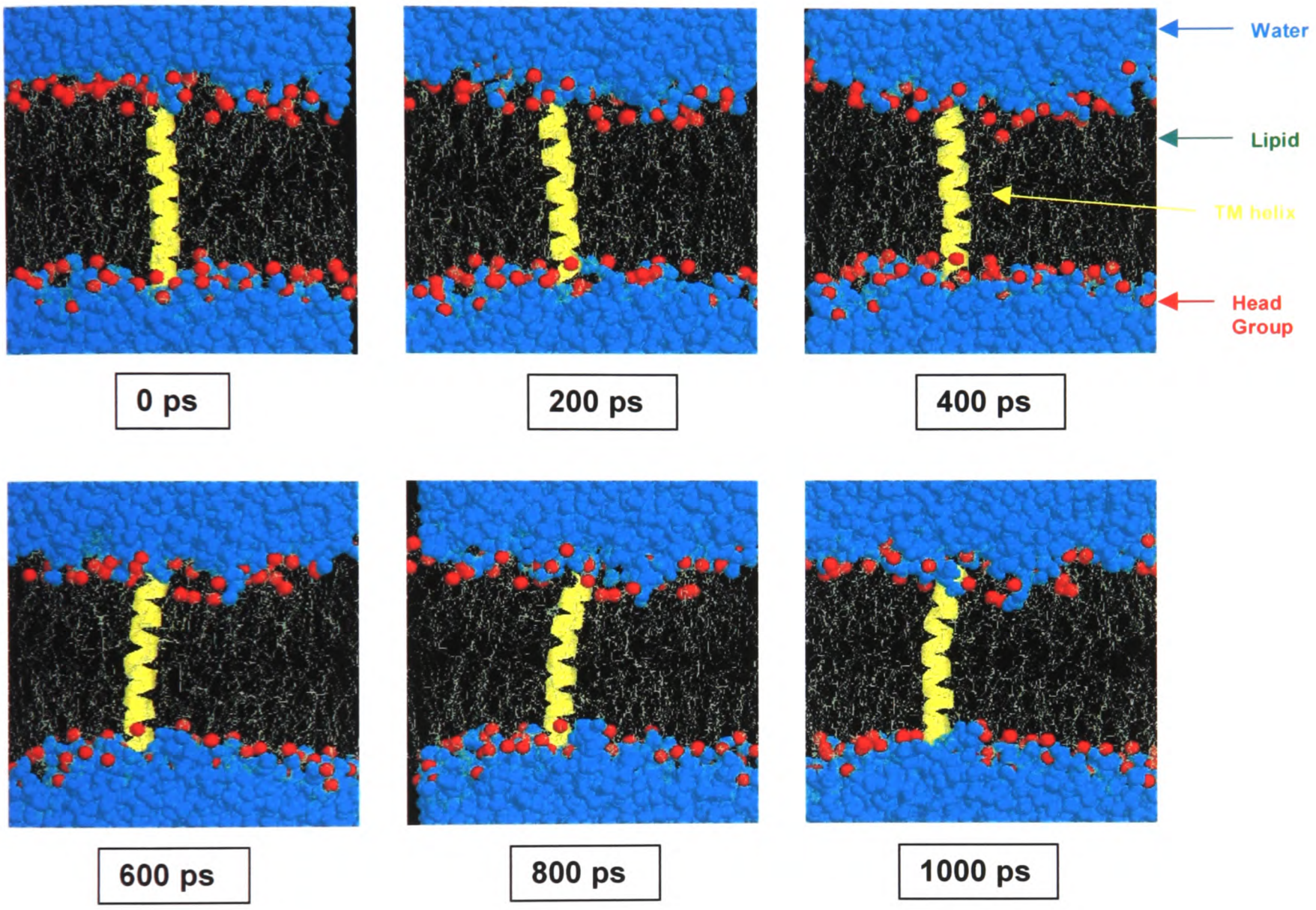


Figure 5.6 Monitoring the physical behaviour of the TM helix 3 within the simulation system at different time steps.

5.3.2 Dynamics of Transmembrane Helix Backbone.

The overall progress of simulations for each TM helix may be compared *via* their Cα RMSDs versus time. In general, the Cα RMSDs for these 12 TM helices range from 0.1 to 0.2 nm from their initial structures, (Fig. 5.7).

For TM helix 1, it can be seen that the RMSD increases steadily up to ~0.26 nm over the first 400 ps or so and then decreases steadily to ~0.15 nm at the end of the 1ns simulation, (Fig. 5.7 A). For TM helices 2 and 3, a steadily increasing pattern of the Cα RMSDs is observed; however, the final values of the Cα RMSDs at the 1 ns for both helices still range at ~0.15 nm, (Fig. 5.7 B and C). For TM helix 4, the RMSD stays approximately constant throughout the 1 ns simulation at a final RMSD less than 0.1 nm, (Fig. 5.7 D). For TM helix 5, the RMSD increases steadily up to ~0.16 nm over the first 220 ps or so and then stays constantly at ~0.15 nm to the end of simulation, (Fig. 5.7 E). For TM helix 6, a steadily increasing pattern is seen and has a relatively large RMSD, more than 0.2 nm, at the end of the simulation, (Fig. 5.7 F).

For TM helix 7, the RMSD rises sharply, from 0.1 nm to 0.2 nm, at ca. 200 ps of simulations, and then increases steadily to ~0.25 nm at the end of the 1ns simulations, (Fig. 5.7 G). For TM helix 8, the RMSD increases steadily during the first 250 ps to ~0.2 nm, then drops to ~0.1 nm at ca. 250 ps, and increases steadily to ~0.18 nm at the end of the simulation, (Fig. 5.7 H). For TM helix 9, the RMSD fluctuates during the first 200 ps of simulations, and then reaches a steady plateau of ~0.18 nm afterwards to the end of the simulation, (Fig. 5.7 I). For TM helix 10, the RMSD increases to ~0.2 nm during the first 250 ps of simulations, and then decreases steadily to ~0.1 nm at the end of the simulation, (Fig. 5.7 J). The TM

helix 11 shows a steady pattern of increase in its RMSD and reaches ~0.2 nm at the end of the simulation, (Fig. 5.7 K). For TM helix 12, the RMSD jumps up and down through the 1 ns simulation between 0.05 to 0.1 nm; the final RMSD of the simulation is less than 0.1 nm, (Fig. 5.7 L).

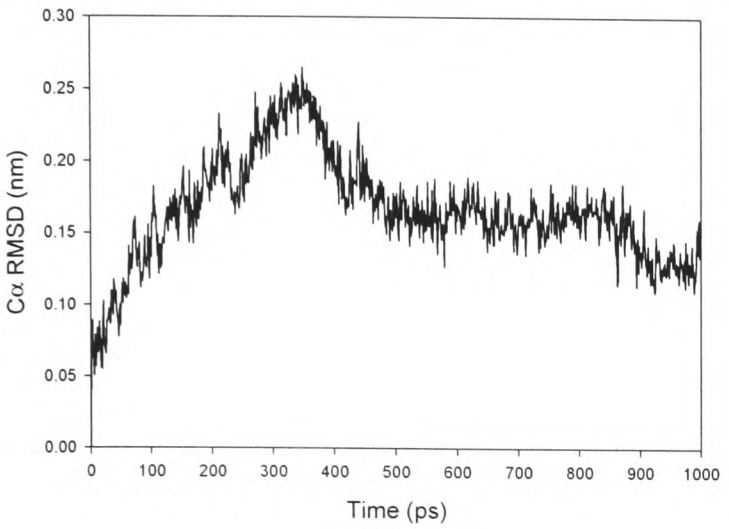
The low RMSDs, 0.1 to 0.2 nm, indicate relatively small motions for the helical structures over the course of the 1 ns simulations for the 12 CFTR TM helices. The overall RMSDs of this magnitude have been observed as well in simulations of other membrane proteins, which start from experimental X-ray structures of the TM helices (Tieleman and Berendsen, 1998; Shrivastava and Sansom, 1999).

No similarities of the RMSD patterns are detected between the first half and second half of CFTR transmembrane domains. For example, TM helices 2, 3, 5 and 6 in the first half of CFTR transmembrane domain show a pattern of steady increasing; however, this pattern is only observed For TM helix 11 in the second half of CFTR transmembrane domain. This might imply that the symmetrical arrangements of the TM helices between two transmembrane domains are not strictly maintained.

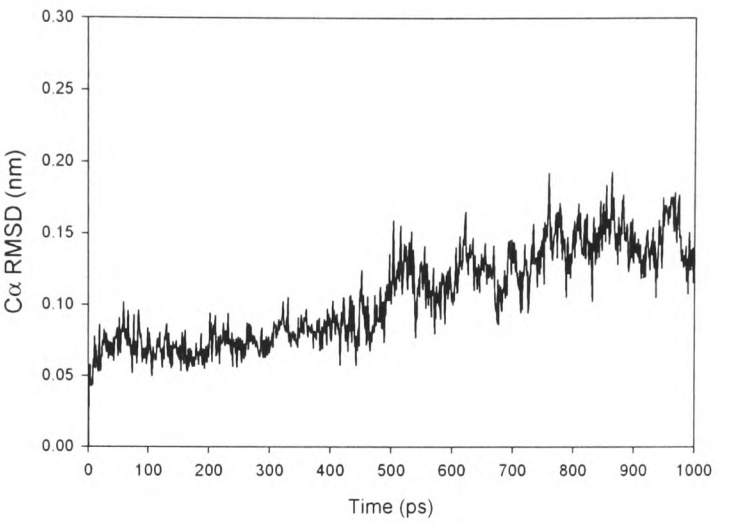
Some subtle differences in structural dynamics are revealed in the Cα RMSDs of different simulations. The drifting of a TM segment along the lipid bilayer is observed in TM helix 11, which shows a steadily increasing pattern for the Cα RMSD. In addition, the TM helix 7 has the largest Cα RMSD for 0.26 nm, and a sudden rise of Cα RMSD at ca. 200 ps. A splitting of the α-helix into two α-helical segments is detected by visual examination of the structures of TM helix 7 at these time frames. This strongly suggests some interesting interactions occurring at these time steps, and more detailed analysis is needed for TM helix 7.

In general, the longer helices or the shorter helices, *i.e.* those which do not match the bilayer thickness, might be expected to have larger Cα RMSDs in order to span across the hydrophobic lipid bilayer. For example, the longer helix could adapt itself within the limited thickness of lipid bilayer either by kinks or tilts, and the shorter helix could extend itself to stretch across the lipid bilayer. However, this tendency is not detected from the simulations of CFTR TM helices. For example, TM helix 4 and TM helix 12 are the two helices having RMSDs less than 0.1 nm and are short helices, with only 23 and 22 residues respectively. The shortest helix, TM helix 9, has 21 residues and has a Cα RMSD of 0.17 nm, which is similar to other helices of 25 residues, for example, TM helices 1, 2 and 8. This suggests the TM peptide motions observed reflect more truly for the interactions between helices and lipid bilayers rather than simple helix/bilayer mismatch. This also suggests that lengths of the 12 TM helices defined in this study are within a reasonable range.

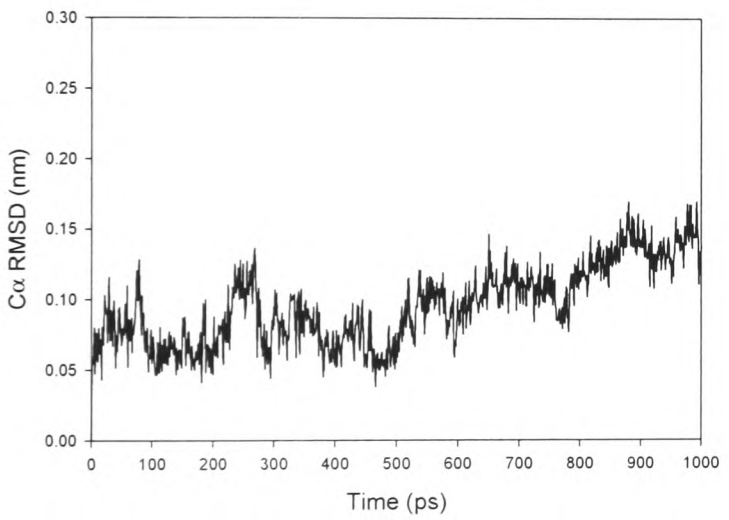
A. TM Helix 1



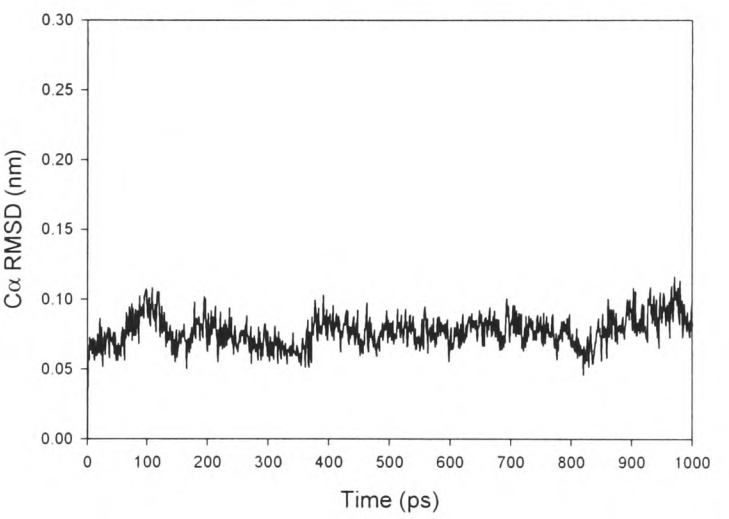
B. TM Helix 2



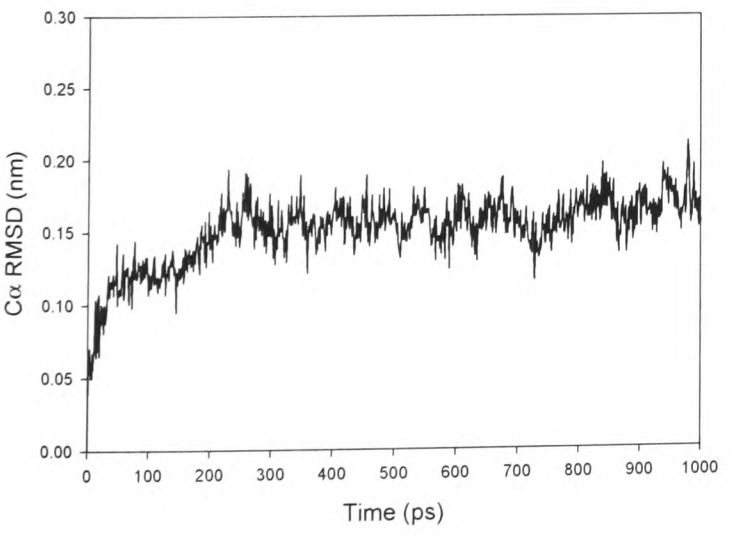
C. TM Helix 3



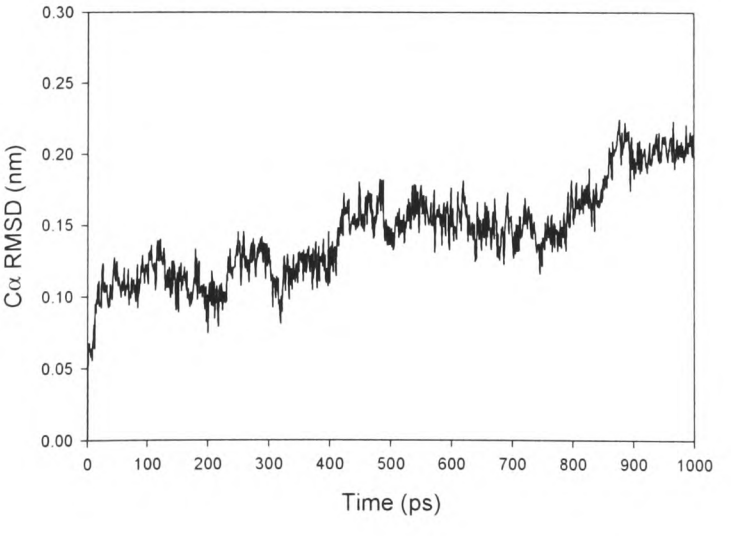
D. TM Helix 4



E. TM Helix 5



F. TM Helix 6



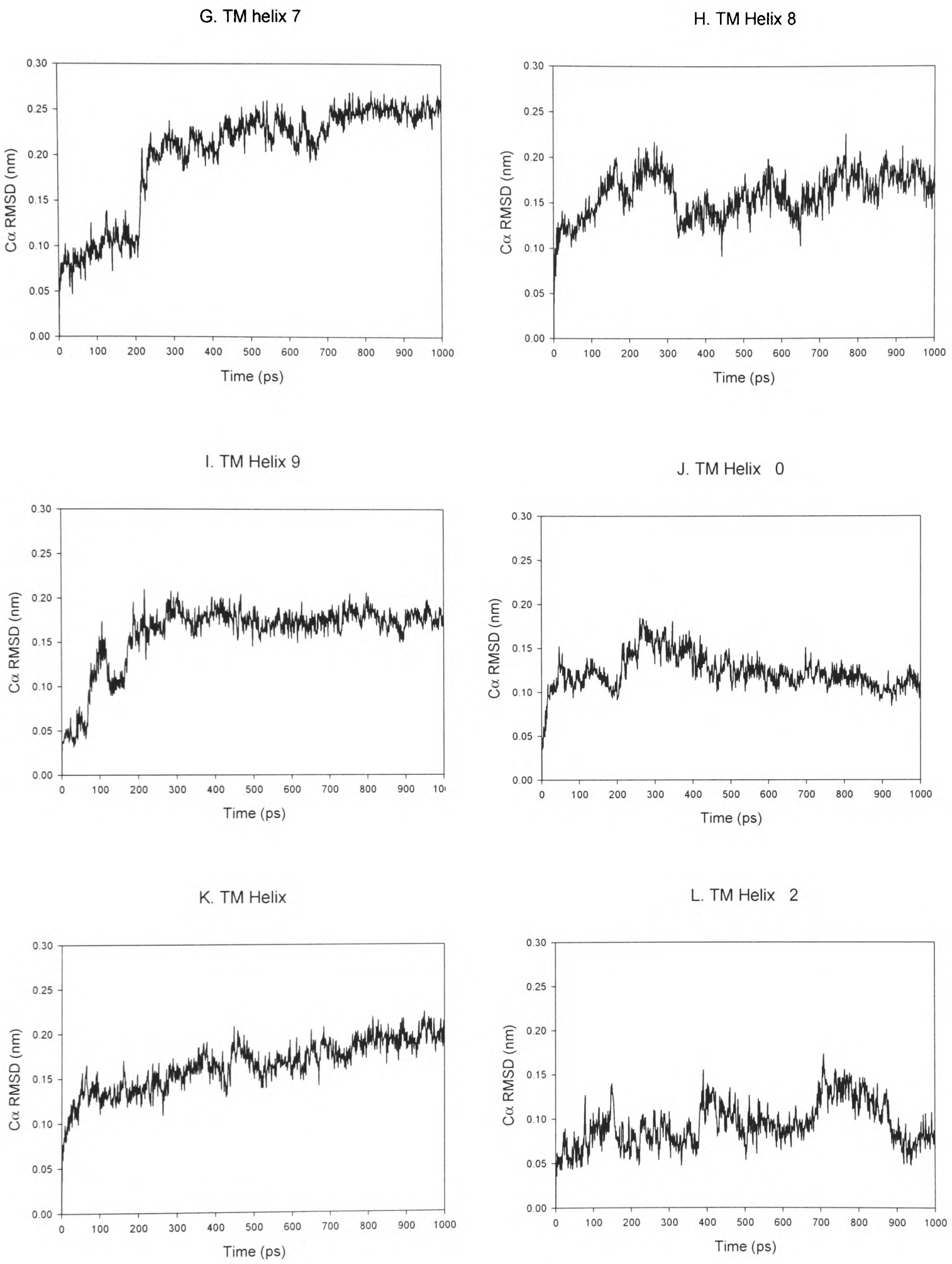


Figure 5.7 RMS deviations of Cα atoms. The RMS deviations are calculated from differences between the initial structure and structures at different time step for each TM helix.

5.3.3 Fluctuations of Transmembrane Helices.

While the Cα RMSD describes global peptide conformational drift, it is also of interest to examine the internal motions of a helix, which could be expressed in terms of the residue-by-residue Cα RMS fluctuations (RMSF) from the mean structure over the 1 ns simulation (Tieleman et al., 1999a).

For TM helix 1, the Cα RMSFs of the first 22 residues are all less than 0.1 nm, which indicates that these residues are not very mobile during the 1 ns simulation, and residues at the C-terminus have larger Cα RMSFs, more than 0.1 nm, (Fig. 5.8 A). This can be easily observed from the visual comparisons between the initial structure, (upper helix in Fig. 5.8 A), and the final structure, (lower helix in Fig. 5.8 A). For TM helix 2, the fifth residue, Leu123, is very stable through the 1 ns simulation, while other residues are relatively mobile, (Fig. 5.8 B). The sudden increase of the Cα RMSF after Leu123 can be explained by the following two glycine residues at positions 124 and 126. Clear fluctuations between the 6^{th} to 11^{th} residues and the last five residues are detected by comparing the initial helical structure, (upper helix in Fig. 5.8 B), and the helical structure at the end of simulation, (lower helix in Fig. 5.8 B). For TM helix 3, except for residues at both termini, a clear peak for the Cα RMSFs is observed between the 8^{th} and 11^{th} residues, (Fig. 5.8 C). This peak is likely caused by the disruptions of helical patterns by the Pro205 at the 12^{th} position. A kink is observed at this region, by comparing initial helix, (upper structure in Fig. 5.8 C), and final structure at the end of simulation, (lower structure in Fig. 5.8 C). For TM helix 4, less internal motions are observed and only residues at the C-terminus endure fluctuations, (Fig. 5.8 D). For TM helix 5, visual comparisons between two helical structures

reveal residues at the C-terminus undergo greater fluctuations, (structures in Fig. 5.8 E), which is confirmed by the distributions of the Cα RMSFs for every residues, (Fig. 5.8 E). For TM helix 6, a sharp peak is seen near the 16^{th} and 17^{th} residues, (Fig. 5.8 F). Visual inspection of the helical structures at the beginning and the end of simulation show a clear deformation at that region, (structures in Fig. 5.8 F).

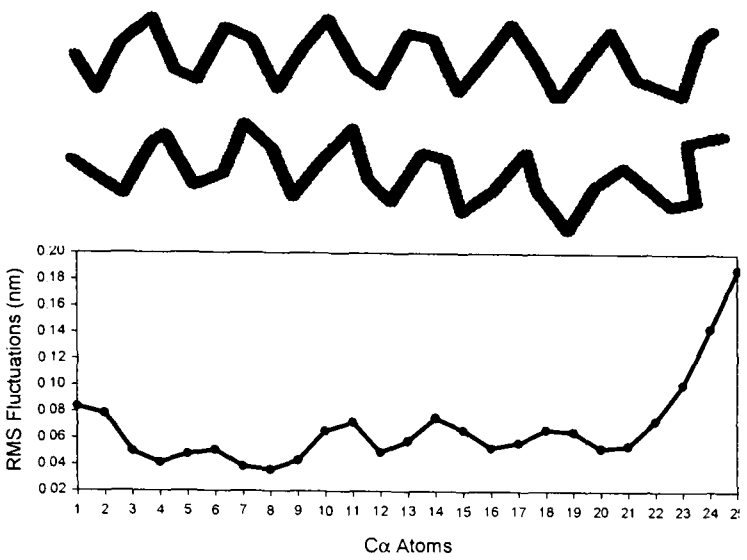
For TM helix 7, a clear peak of the Cα RMSFs is detected from residue 9 to reside 13, (Fig. 5.8 G). The conformation of the TM segment changes from an ideal helical structure, (upper structure in Fig. 5.8 G), to a segment containing two helixes, (lower structure in Fig. 5.8 G). For TM helix 8, residues at both termini have greater fluctuations, (Fig. 5.8 H). In addition, the tilting of the TM helix is observed from comparisons of structures at 0 ps and 1 ns, (structures in Fig. 5.8 H). For TM helix 9, only residues at both termini exhibit fluctuations, (Fig. 5.8 I). For TM helix 10, not only residues at both termini but also residues near the 13^{th} residue, which is Met1028, undergo greater fluctuations, (Fig. 5.8 J). For TM helix 11, residues at both termini endure fluctuations, (Fig. 5.8 K). In addition, the first six residues at the N-terminus show a significant difference between the starting and the ending structures of the simulation, (helices in Fig. 5.8 K). For TM helix 12, only residues at both termini undergo fluctuations, (Fig. 5.8 L).

In general, the RMS fluctuations of the Cα atoms from their time averaged positions, are relatively small, ranging from 0.03 nm to 0.20 nm; however, some variations along the helical segments are detected. These variations can be regarded as characteristic for each helix.

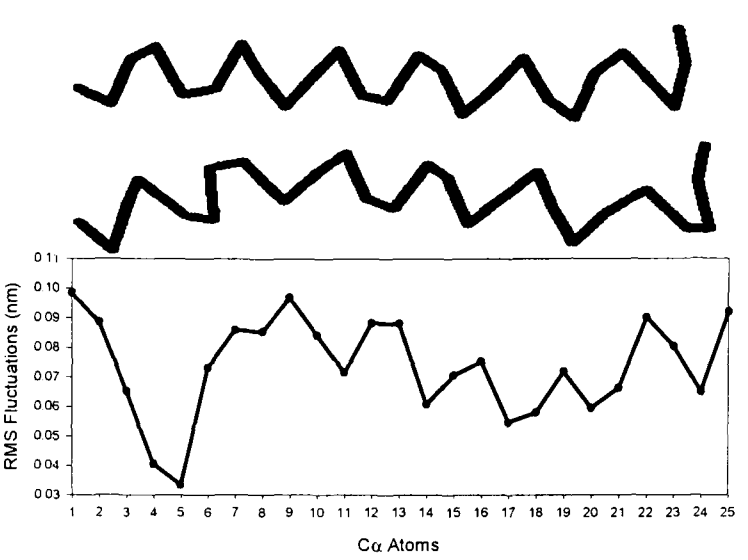
A general trend of greater fluctuations for residues at termini of a helix than residues in the interior of a helix is observed. In addition, the RMS fluctuations of the C-terminal segment are generally larger than the RMS fluctuations of the N-terminal segment. Similar patterns have been observed in other studies as well, (Tieleman et al., 1999a; Tieleman et al., 1999b; Shrivastava et al., 1999). The unexpected mobile residues within the interiors of helices reveal that there might be some usual interactions near the central cores of helices, Section 5.3.9.

For TM helices 1, 5, 6 and 7, the N-terminal residues are comparatively more stable than the internal residues, (Fig. 5.8 A, E, F and G). Some aromatic residues are found at these regions; Phe77 of TM helix 1, Phe305 of TM helix 5 and Ile860/Phe861 of TM helix 7. This reveals that there may be interactions at the interfacial regions between the lipid, water and aromatic residues, which provide an extra force to stabilise the residues or the helices. For TM helix 6, electrostatic interactions provided by the Arg334 at the N-terminus could explain the unusual stability of that region. Similar interactions have been reported from a MD simulation study of a designed amphiphilic helical peptide in a hydrated DMPS system (Belohorcova et al., 1997).

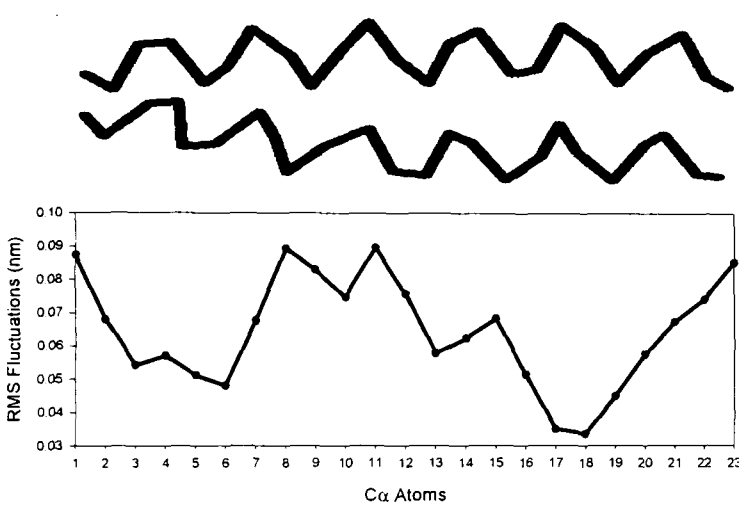
A. TM Helix 1



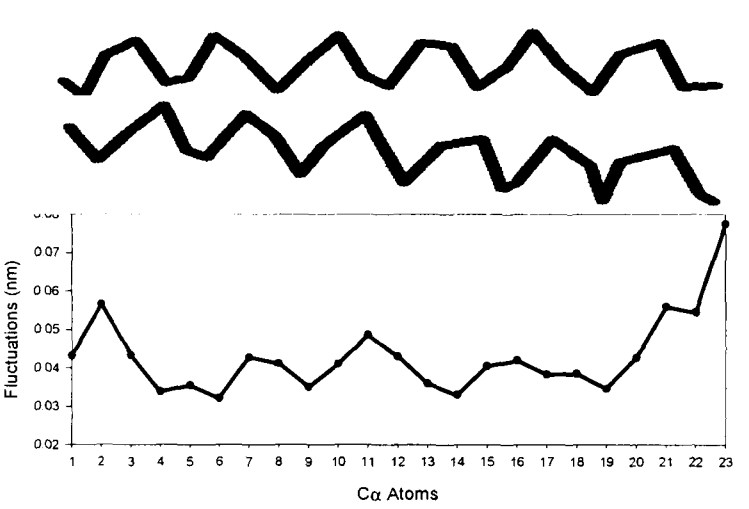
B. TM Helix 2



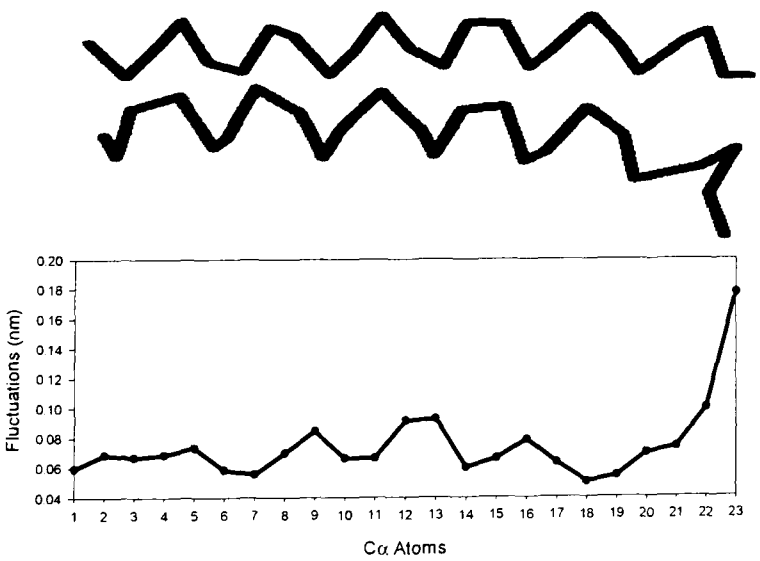
C. TM Helix 3



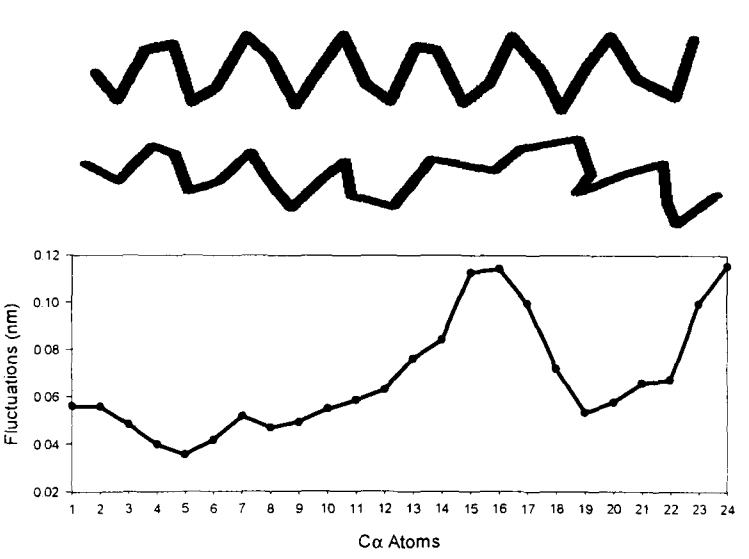
D. TM Helix 4



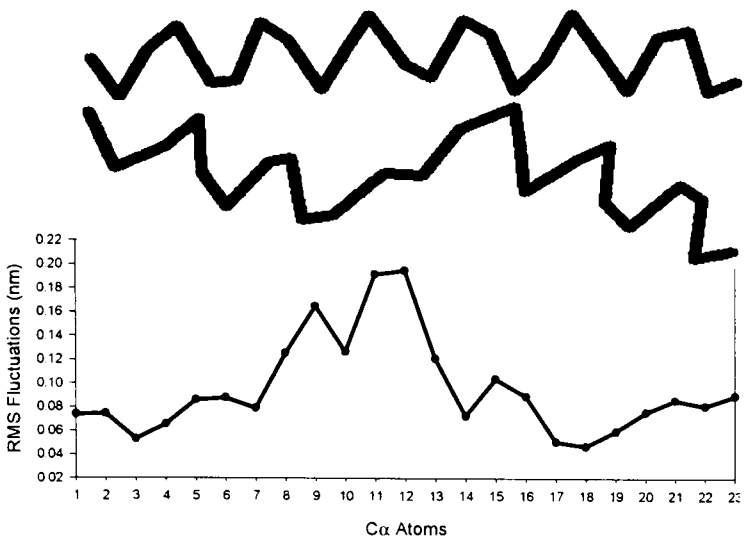
E. TM Helix 5



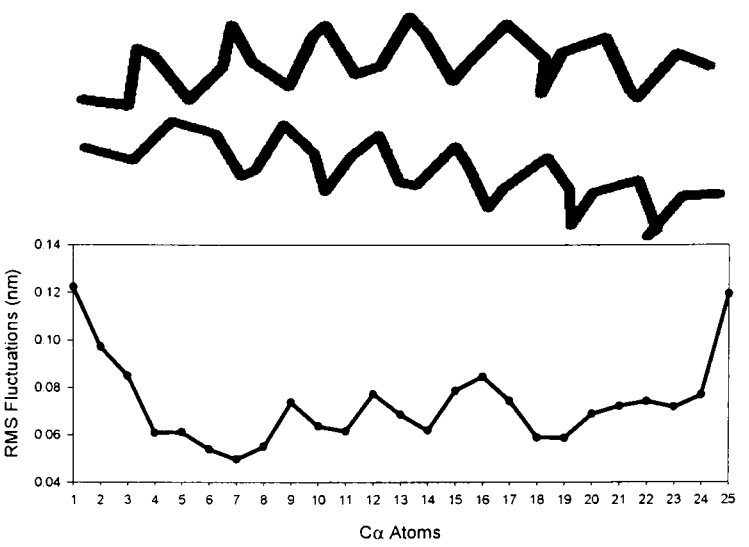
F. TM Helix 6



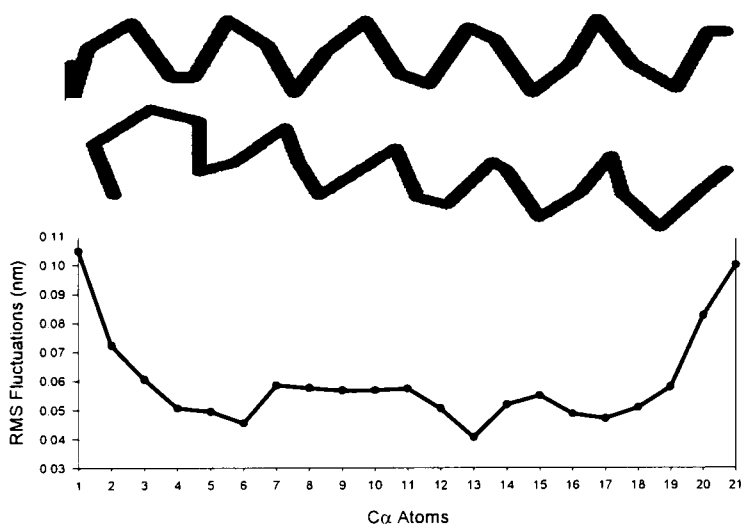
G. TM Helix 7



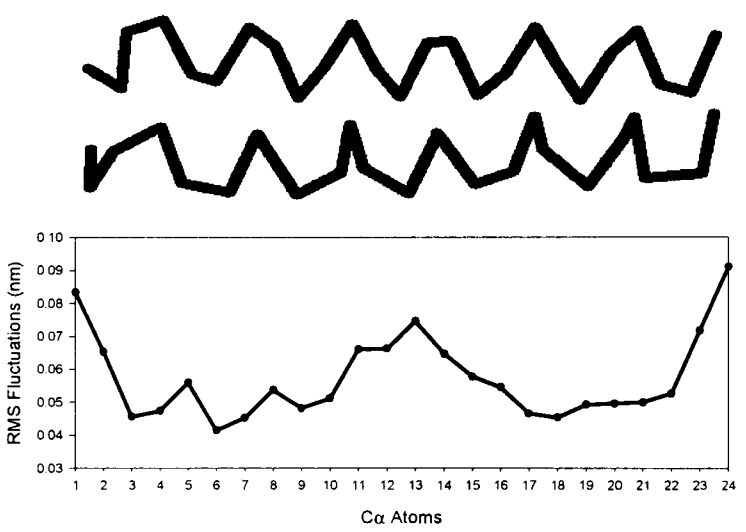
H. TM Helix 8



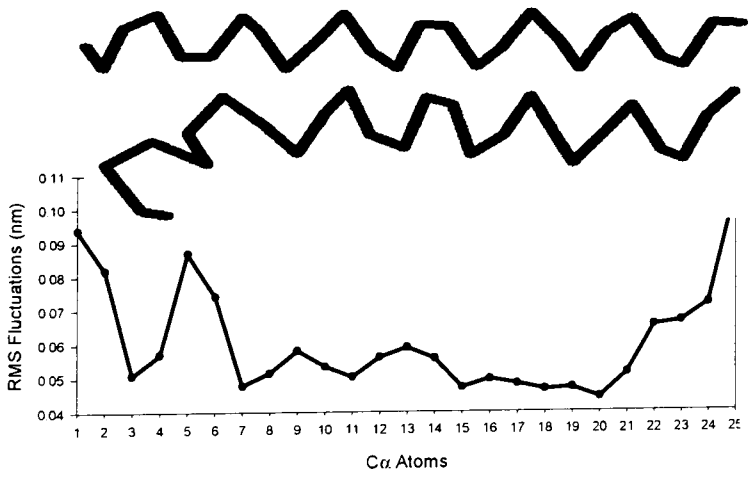
I. TM Helix 9



J. TM Helix 10



K. TM Helix 11



L. TM Helix 12

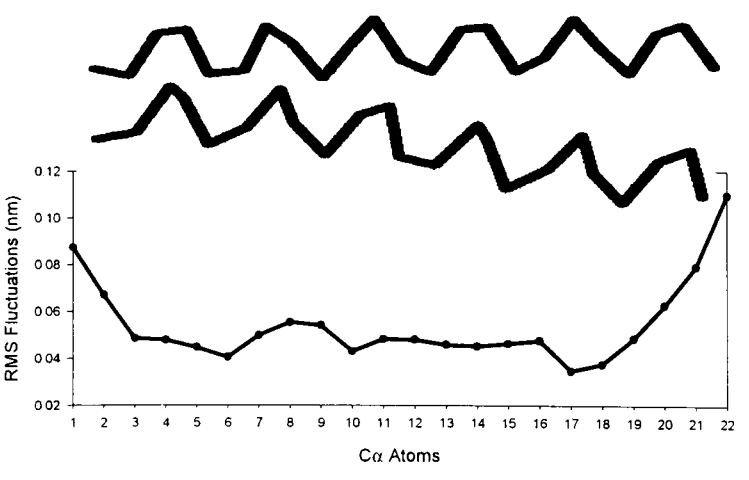


Figure 5.8 Fluctuations of Cα atoms from mean structures of TM helices. The upper and lower helices shown are the starting and final structures of simulations. The position of each Cα atom is roughly aligned with the deviation of each Cα atom.

5.3.4 Dynamics of Secondary Structures.

Consequences of the internal dynamics and fluctuations of a TM segment can be examined in terms of secondary structure as a function of time. DSSP (Kabsch and Sander, 1983) was employed to assign the protein secondary structures, (details see section 2.1.4).

In general, the TM helices of CFTR retain a largely α-helical conformation. The terminal residues are more susceptible to loss of α-helicity than those residues embedded in the hydrophobic core of the bilayer throughout the 1 ns simulations, (Fig. 5.9). However, there are exceptions, TM helices 6 and 7, (Fig. 5.9 F and G), which form two distinct helices within a TM segment.

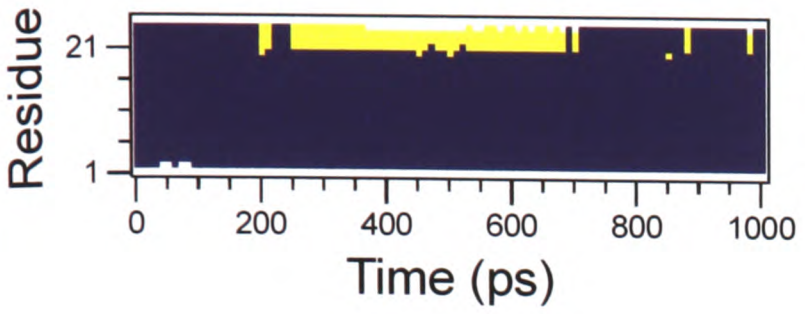
For any TM helix, residues at both N-terminus and C-terminus are inherently more likely to lose their α-helical conformation. One of the reasons is rearrangements of the hydrogen bond patterns at both ends of a TM helix due to dislocations of terminal residues in order to span the lipid bilayer for a short peptide. This mechanism has been suggested in other simulations of TM helices for which α-helical models of different lengths were explored (Fischer et al., 1999). Similar effects are observed in TM helices 9 and 12 in this study. The TM helices 9 and 12 are only 21 residues and 22 residues long respectively. The elongations at both ends in order to span the lipid bilayer are observed for these two helices. This elongation causes the loss of helical conformation at the N- or C- terminal region of each helix, (Fig. 5.9 I and L).

Another frequently seen mechanism for transforming α-helical conformations into other secondary structures for residues at terminal regions is interactions between

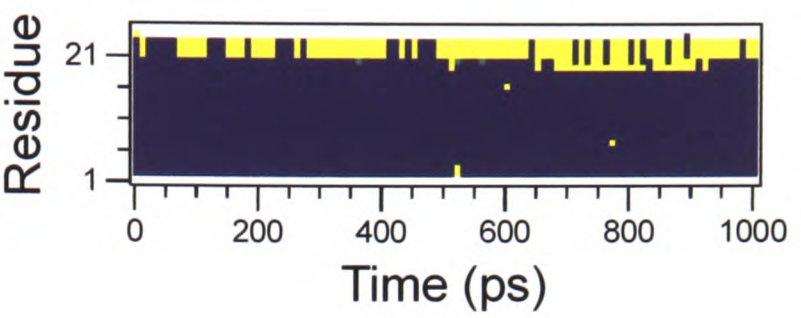
amino acid residues and other interfacial molecules, such as the lipid bilayer or water molecules. For TM helix 1, amino acid residues at the C-terminal region are in a turn conformation during the period of ca. 200 ps to 640 ps, (Fig. 5.9 A). This is due to hydrogen bond formation between the side chain of Trp79 and water molecules penetrating into the hole, which accommodates the TM helix. For TM helix 2, residues at the C-terminal region are largely in the turn conformation during the 1 ns simulation, (Fig. 5.9 B). This is caused by Pro140, a helical destabiliser, at position 22, and a hydrogen bond formed between the carbonyl oxygen of Phe143 at position 25 and a water molecule. For TM helix 8, the residues at the N-terminal region are mostly not in the α-helical conformation, (Fig. 5.9 H). This is caused by three hydrogen bonds formed between Tyr919 at the first position and a water molecule, and one hydrogen bond formed between the amide group of Val920 at the second position and a water molecule. For TM helix 10, residues at the N-terminal region are mainly in either turn or coil conformations during the 1 ns simulation, (Fig. 5.9 J). This is caused by a hydrogen bond formed between the main chain amide group of Phe1016 at the first position and one of the oxygen atoms of a POPC lipid molecule, and a hydrogen bond formed between the main chain amide group of Ala1018 at the third position and a water molecule. For TM helix 11, residues at the N-terminal region are largely not in the helical conformation, which is also caused the extra hydrogen bonds formed between Trp1098, Phe1099, Gln1100 and water or POPC lipid molecules, (Fig. 5.9 K).

The DSSP analysis results are more complex For TM helix 6 and helix 7. For TM helix 6, there is transient loss of α-helicity from residues Val346 to Val350, (Fig. 5.9 F). This leads to a conformation of two helical segments linked by a region of dynamic and flexible turn. This switch from an α-helical to a turn conformation occurs at the very early stage of simulation, less than 50 ps, reveals that TM helix 6 behaves as two α-helices linked by a non-helical region. The DSSP results for TM helix 7 reveals a similar pattern, (Fig. 5.9 G). Two α-helical segments are linked by a turn conformation from Val868 to Ala872 in the TM helix 7. Detailed analysis of these two TM segments are presented in section 5.3.6.

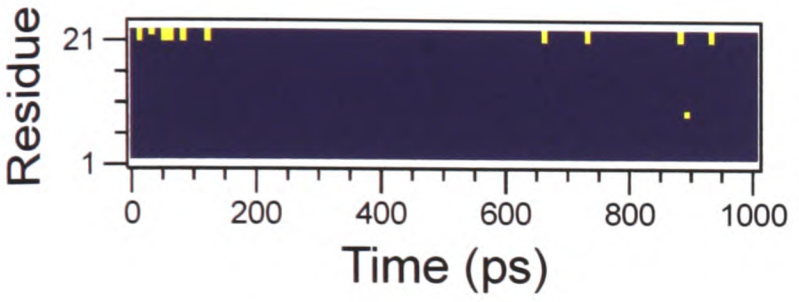
A. TM Helix 1



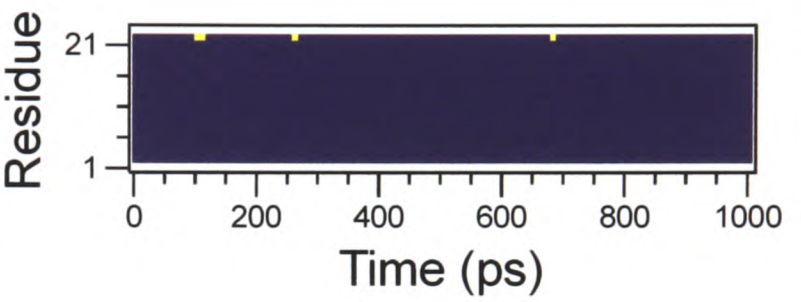
B. TM Helix 2



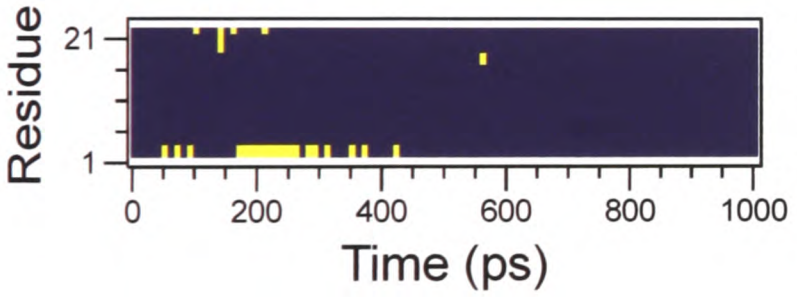
C. TM Helix 3



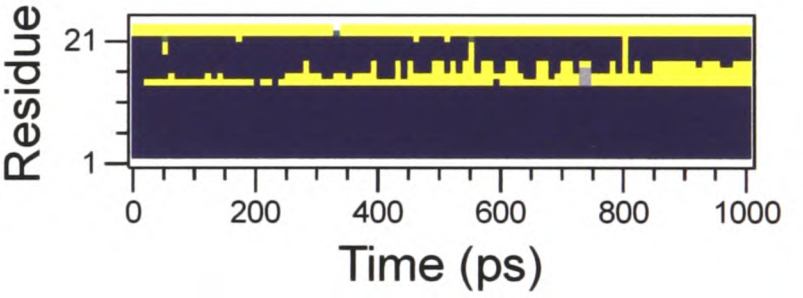
D. TM Helix 4



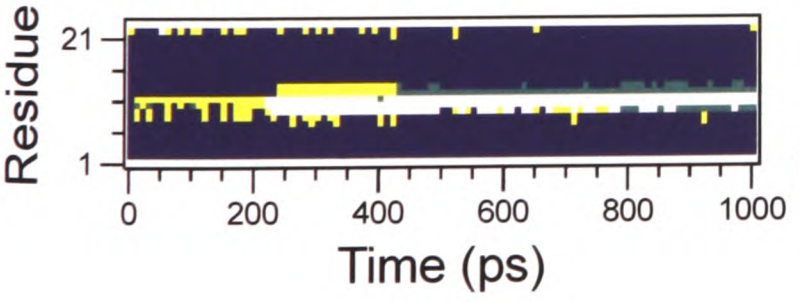
E. TM Helix 5



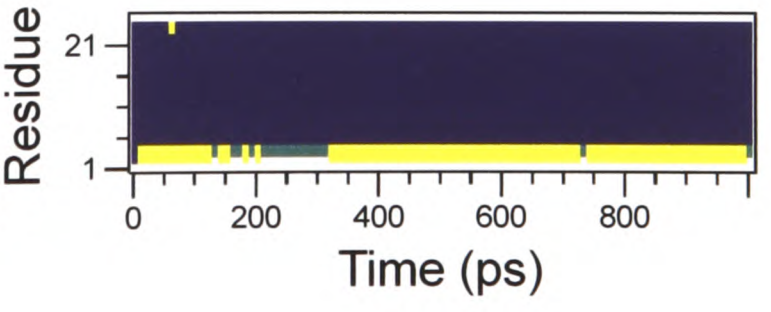
F. TM Helix 6



G. TM Helix 7



H. TM Helix 8



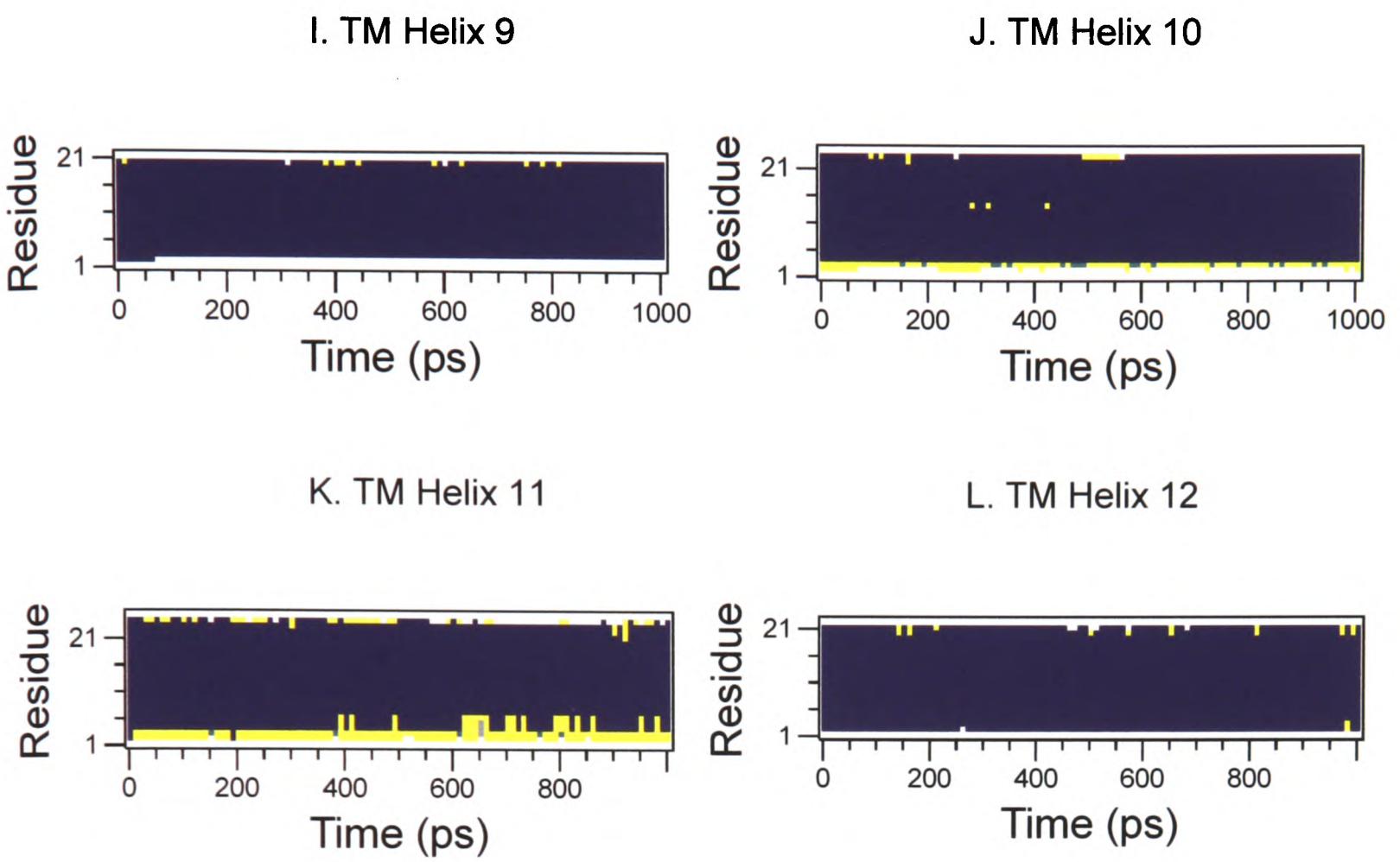


Figure 5.9 Secondary structure assignments by DSSP at different time steps. Regions denoted with different colours represent different secondary structures, see below. Residues numbers denoted at the Y axis represent the position of each residue along the TM helix. Their relative positions on the CFTR primary sequence are shown in Tab. 5.2.

5.3.5 Dynamics of Helical Structures.

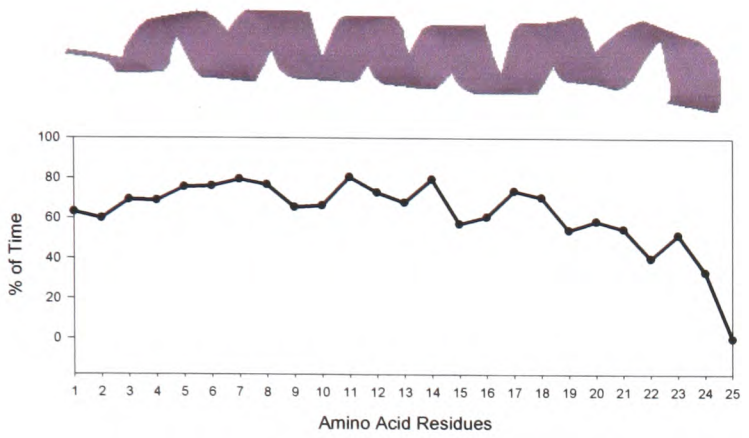
While the DSSP analysis reveals the overall secondary structure arrangements of a TM helix on the basis of the hydrogen bond patterns, it is also of interest to investigate the helical conformation on a residue-by-residue basis in order to describe the helicity of each individual residue during the simulation.

As expected, the helicity profiles of TM helices 6 and 7 indicates low helicity in their centres, (Fig. 5.10 F and G), due to their composition of two separated α-helices connected by a turn. The helicity profiles for other CFTR TM helices show a relatively flat pattern, (Fig. 5.10).

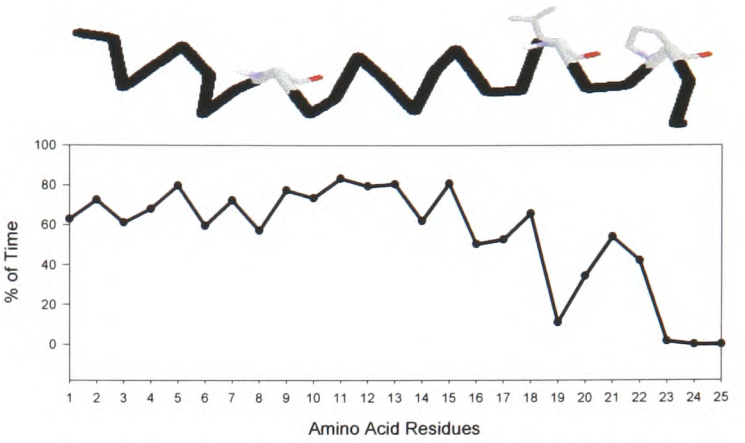
However, there are still dips of helicity observed in TM helices 2 and 3, (Fig. 5.10 B and C). The low helicity for Leu137 in TM helix 2 is explained by the proline residue at position 140, which is just a helical turn behind, (Fig 5.10 B). The same situation happens in the TM helix 3, within which the Val200 has a low helicity because of the Pro205 at the same position in the following helical turn, (Fig. 5.10 C).

The Gly934 was firstly suspected to cause the sharp cliff shown in the helicity profile of the TM helix 8, (Fig. 5.10 H). However, after carefully inspections of simulation systems at different time points, it was found that loses of helicity at Phe932 and Arg933 is a result of a hydrogen bond formed between a water molecule and the side chain of Arg933.

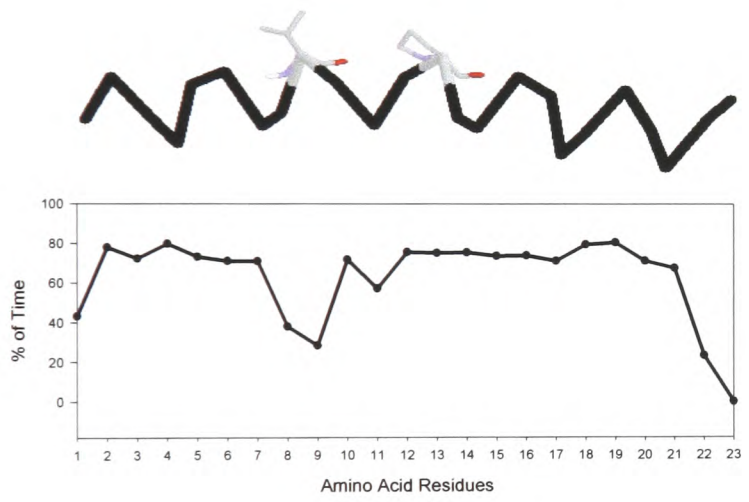
A. TMH 1



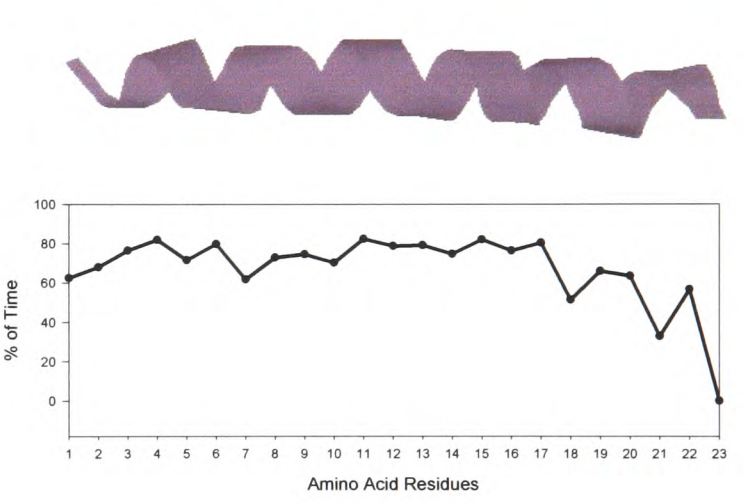
B. TMH 2



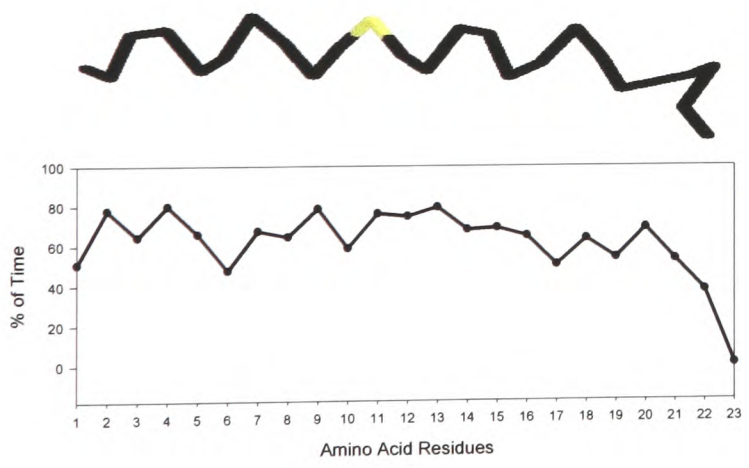
C. TMH 3



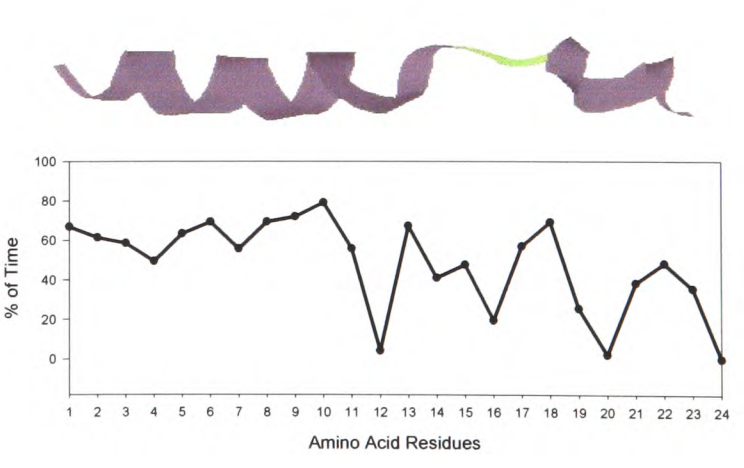
D. TMH 4



E. TMH 5



F. TMH 6



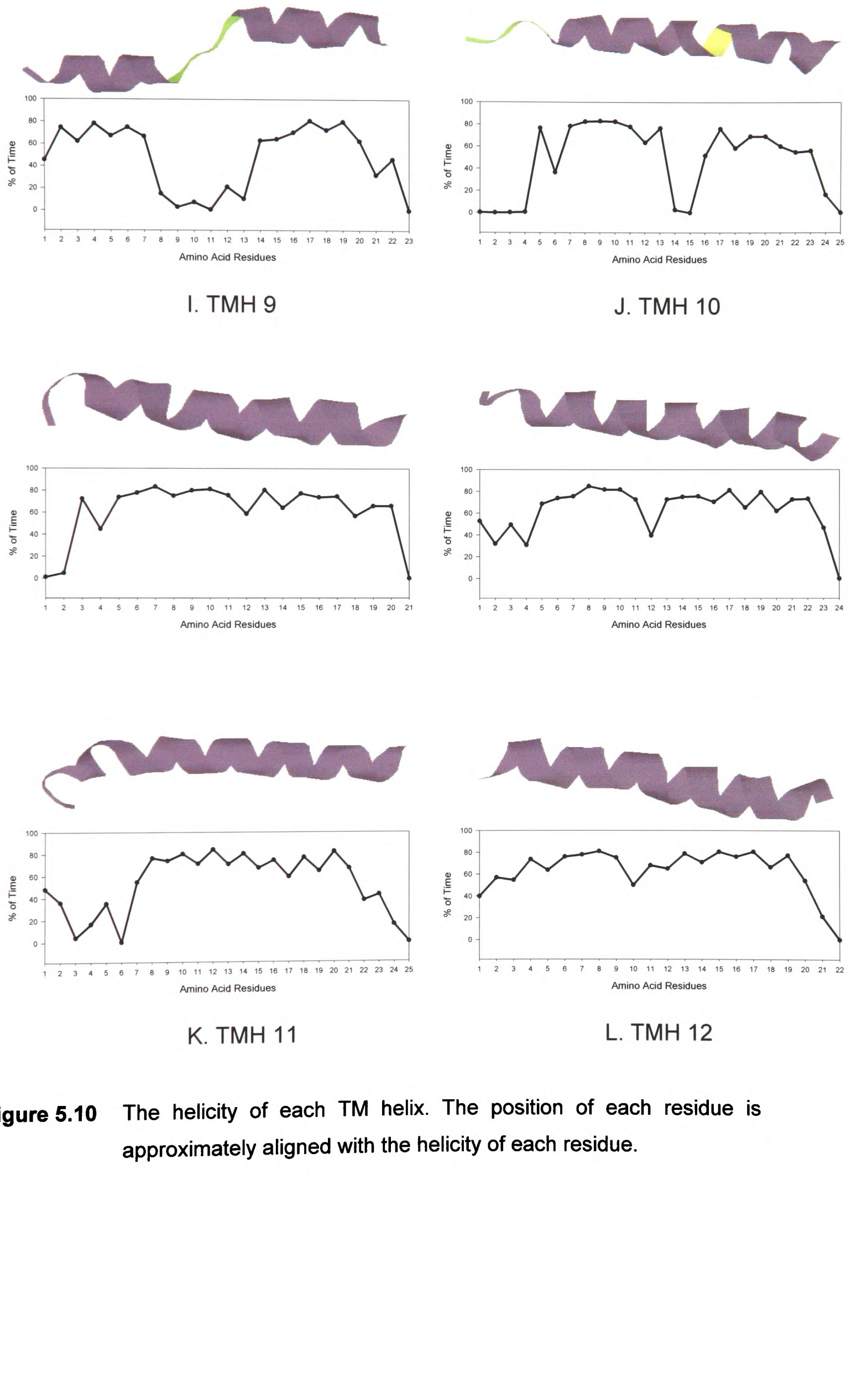


Figure 5.10 The helicity of each TM helix. The position of each residue is approximately aligned with the helicity of each residue.

5.3.6 Dynamics of Hydrogen Bonds.

The unexpected lose of helicity in TM helices 6 and 7 suggest that more detailed analysis for these two TM helices are needed.

From the DSSP analysis, it is observed that the TM helix 6 breaks into two helical segments at around 30 ps. The initial structure of TM helix shows a perfect helical conformation, even though there is an unexpected hydrogen bond between Arg352 and a water molecule, (Fig. 5.11 A). By visualising the structure at 1 ns of this helix, a surprising picture emerges. First, the largely non α-helical conformations appearing on the C-terminal residues of TM helix 6, (Fig. 5.11 F), are caused by the unexpected hydrogen bond between a water molecule and the side chain of Arg352. In addition, an unexpected hydrogen bond is formed between the side chain of Arg347 and the carbonyl group of Cys342, (Fig. 5.11 B). This intra-molecular hydrogen bond interrupts the hydrogen bond patterns to form a helical conformation from Leu346 to Ala349, and thus result in a turn conformation for these 4 residues. However, these two separated helical segments are well aligned with the help of this unexpected hydrogen bond.

For TM helix 7, the initial structure at 0 ps shows a perfect helical conformation, and no unusual interactions between water molecules and protein are detected, (Fig. 5.12 A). However, at the end of simulation, two unusual hydrogen bonds emerge; the two oxygen atoms on the side chain of Glu873 between the amide group of Leu871 main chain and amide group of Ala872 main chain. These two hydrogen bonds not only constrict three involving residues, Leu871, Ala 872 and Glu873, together, but also make a flexible turn conformation for residues from

Val868 to Leu871, (Fig. 5.12 B). The TM helix 7 is then torn apart by the two intra-molecular hydrogen bonds into two helical segments.

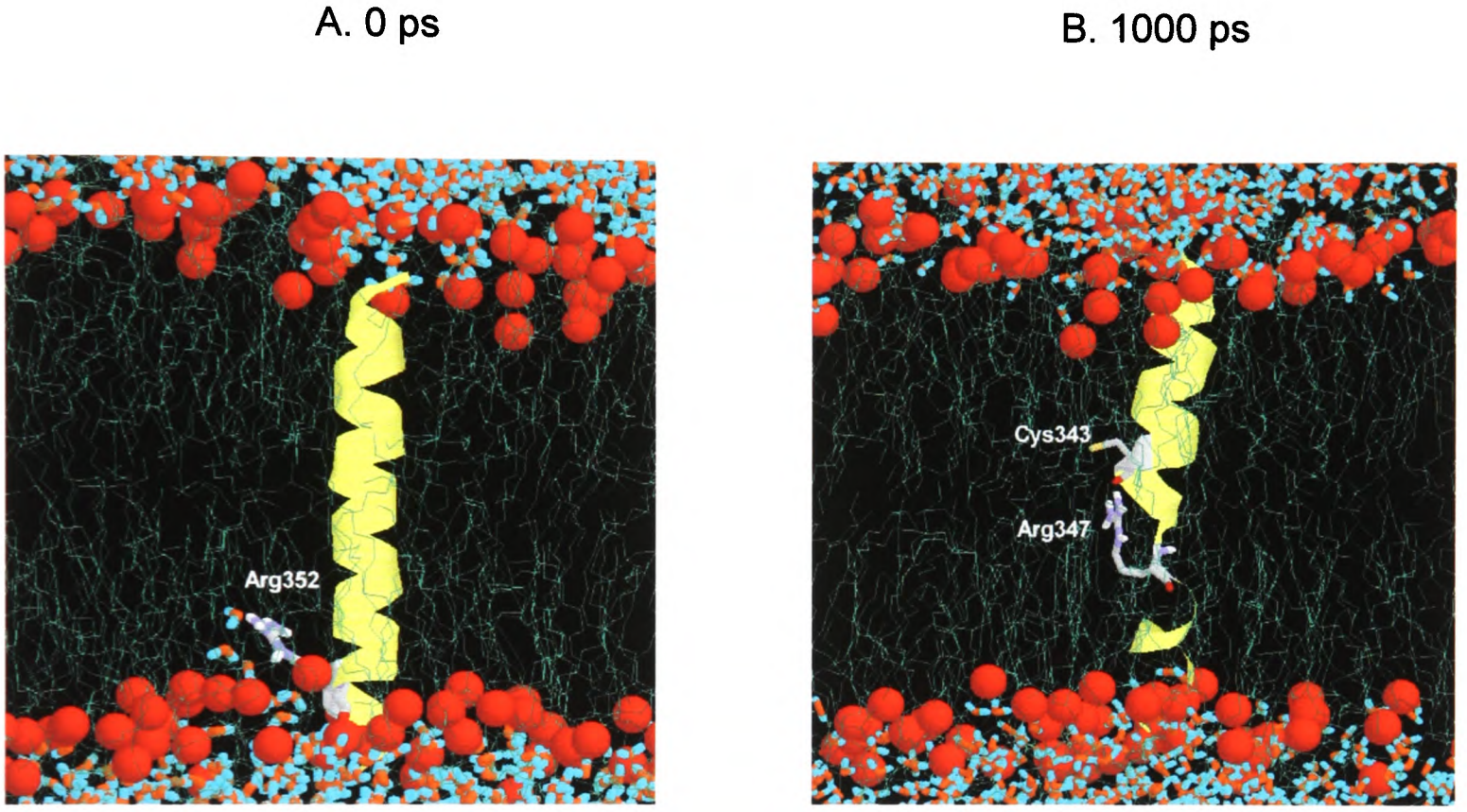


Figure 5.11 Snapshots of structures for TM helix 6 at 0 and 1000 ps. Dynamics of the side chain hydrogen bond are detected between water, Cys342, Arg347 and Arg352. The Arg352 is found to form a hydrogen bond with a water molecule and the TM segment remains an α-helical conformation at the starting point of simulation. However, an unexpected hydrogen bond is formed between the Arg347 and Cys342 at the end of simulation and thus results in a break down of the TM helix into two separated helices.

A. 0 ps

B. 1000 ps

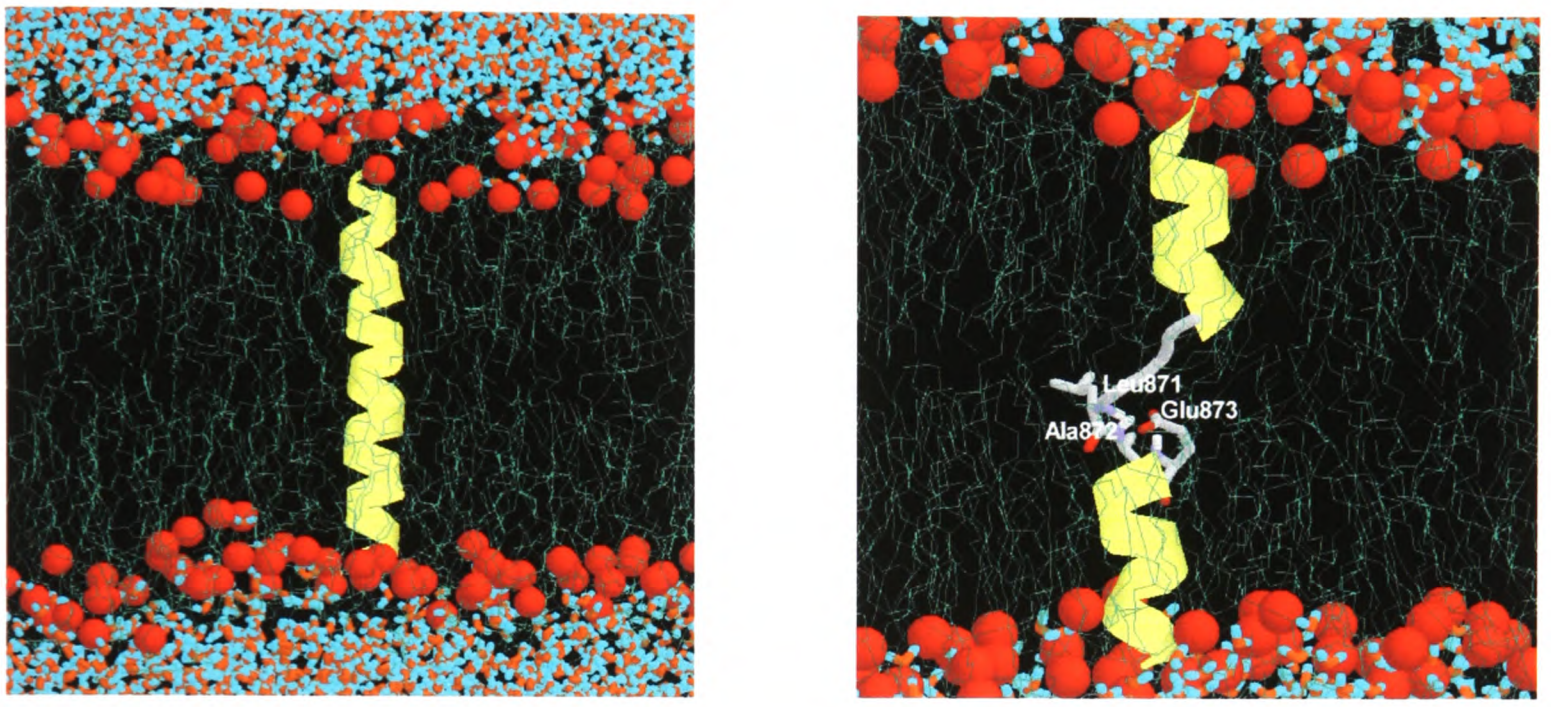


Figure 5.12 Snapshots of structures for TM helix 7 at 0 and 1000 ps. Dynamics of the side chain hydrogen bond are detected between Leu871, Ala872 and Glu873. No unusual hydrogen bond is detected at the beginning of simulation. However, two irregular hydrogen bonds are seen between the Leu871, Glu873 and Ala872 and Glu873. These two unexpected hydrogen bonds change the whole structure of TM helix 7.

5.3.7 Dynamics of Helical Conformations I: Tilt Angles.

It is generally agreed that the helices rarely span the membrane exactly parallel to lipid bilayer normal, which results in a tilt between a TM helix with respect to the lipid bilayer (Bowie, 1997; Walther et al., 1998). The tilt angles of some structure-determined membrane proteins vary from 5° to 39°, depending on the lengths of helices (Bowie, 1997). It is thus of interest to analyse the tilt angle of each TM helix of CFTR with respect to the lipid bilayer.

For TM helix 1, the helical structures are well maintained through the 1 ns simulation, (Fig. 5.13 A), it is also detected that the tilt angle of TM helix 1 steadily increased during the time course of the simulations, (Fig. 5.13 A). The largest tilt angle occurs at around 900 ps with ca. 27°, which is the largest tilt angle observed throughout the simulations of all the 12 helices. The pattern of increasing tilt angles within the 1 ns simulation is also detected in TM helices 5, 8, 10 and 11, (Fig. 5.13 E, H, J and K). These are all relatively long TM helices and their tilt angles at the ends of simulations are relatively large as well; 15° for TM helix 5 (23 residues), 14° for TM helix 8 (25 residues), 17° for TM helix 10 (24 residues) and 16° for TM helix 11 (25 residues).

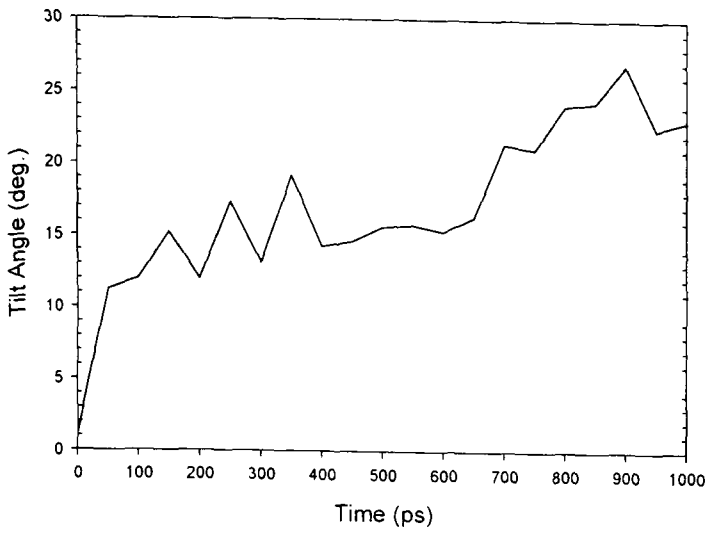
The TM helix 2 has the smallest tilt angle, ca 1°, after 1ns simulation, (Fig. 5.13 B), even though it is 25 residues long. The distortions of the helical conformations at its C-terminal region, (Fig. 5.10 B), could provide extra space to adapt residues within the bilayer. Both TM helices 3 and 4 are 23 residues long and maintain helical conformations well during the simulations, (Fig. 5.10 C and D). Their tilt angles at the end of simulations are estimated at ca. 7° and 9° respectively, (Fig. 5.13 C and D).

The TM helix 9 is only 21 residues long and maintains its helical conformation well during the simulations, (Fig. 5.10 I). However, its tilt angle is still in the range of 10°, (Fig. 5.13 I). One interesting change comes from TM helix 12, which is 22 residues long. This TM helix does not tilt too much at the early stages of the simulation; however, a sharp increase of tilt angle is observed at the time point at ca. 815 ps, (Fig. 5.13 L), and then has a tilt angle of 12° at the end of simulation.

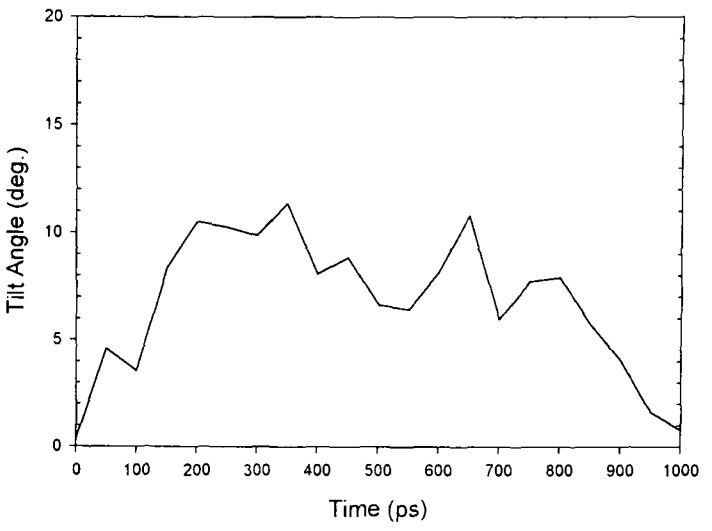
Both TM helices 6 and 7 are divided into two helical segments. However, they behave different within the lipid bilayer. For TM helix 6, the two separated helical elements are still aligned well on the same axis, (Fig. 5.11), this results in a less tilted membrane-spanning segment during the 1 ps simulation, (Fig. 5.13 F), and a small tilt angle of ca. 5° at the end of simulation. For TM helix 7, the two helical segments are pulling away and located in different axes, (Fig. 5.12). This results in an increase of the tilt angle, (Fig. 5.13 G), and the tilt angle at 1 ns is calculated as ca. 13°. However, the tilt angles calculated for TM helices 6 and 7 are heavily biased due to the separations of one helix into two different segments, which result in a poor reflection of how these two helices pack in the lipid bilayer.

With the lipid bilayer of fixed thickness, TM segments with longer lengths and better helicity are expected to have larger tilt angles, and TM segments with shorter lengths or imperfect helical conformations are expected to have smaller tilt angles. This is to some extent supported by observations of the tilt angles of these 12 TM helices over the last 200 ps of simulations. The correlation between longer TM helices and higher tilt angles is generally detected from the tilt angle distributions of these helices, which have different lengths, at different time points, (Fig. 5.14).

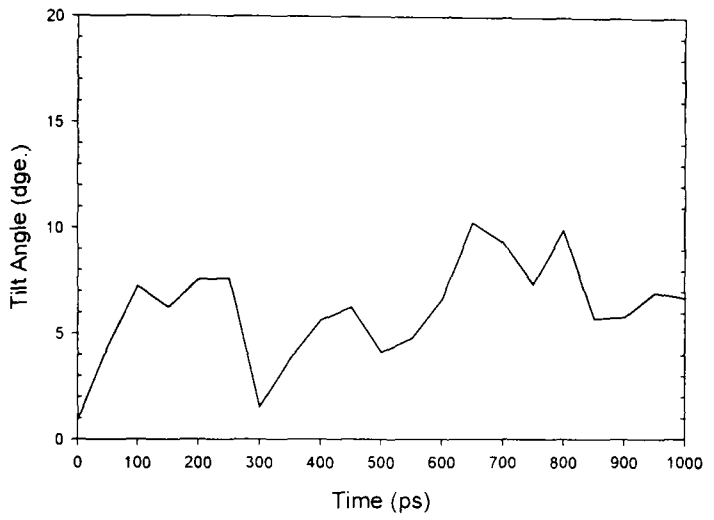
A. TM Helix 1



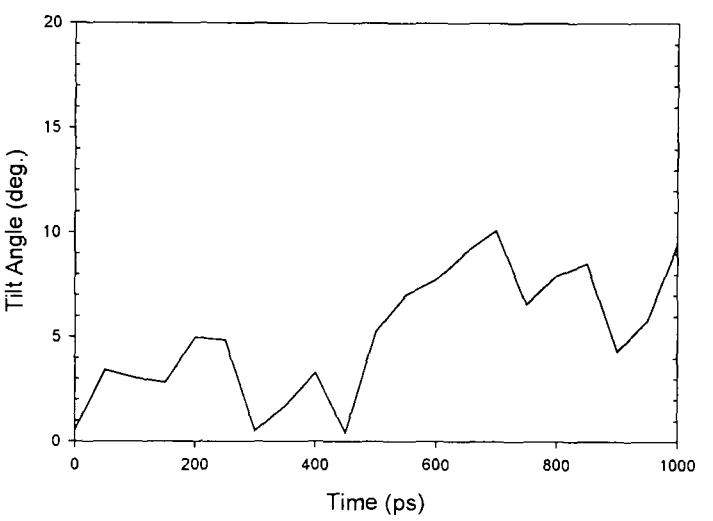
B. TM Helix 2



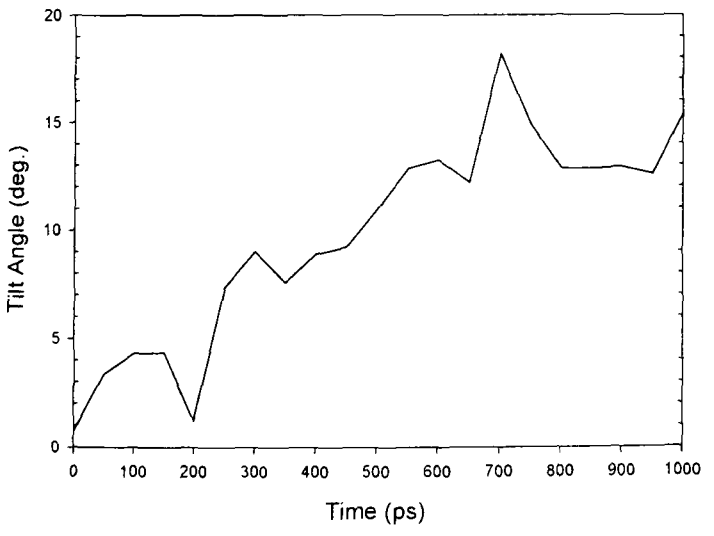
C. TM Helix 3



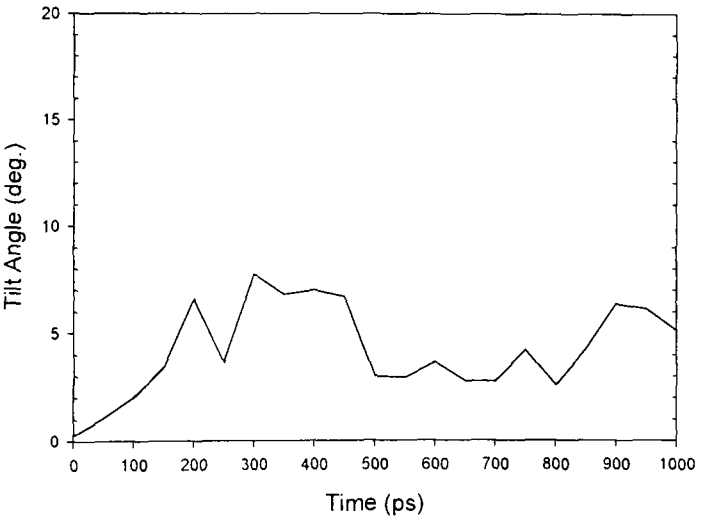
D. TM Helix 4



E. TM Helix 5



F. TM Helix 6



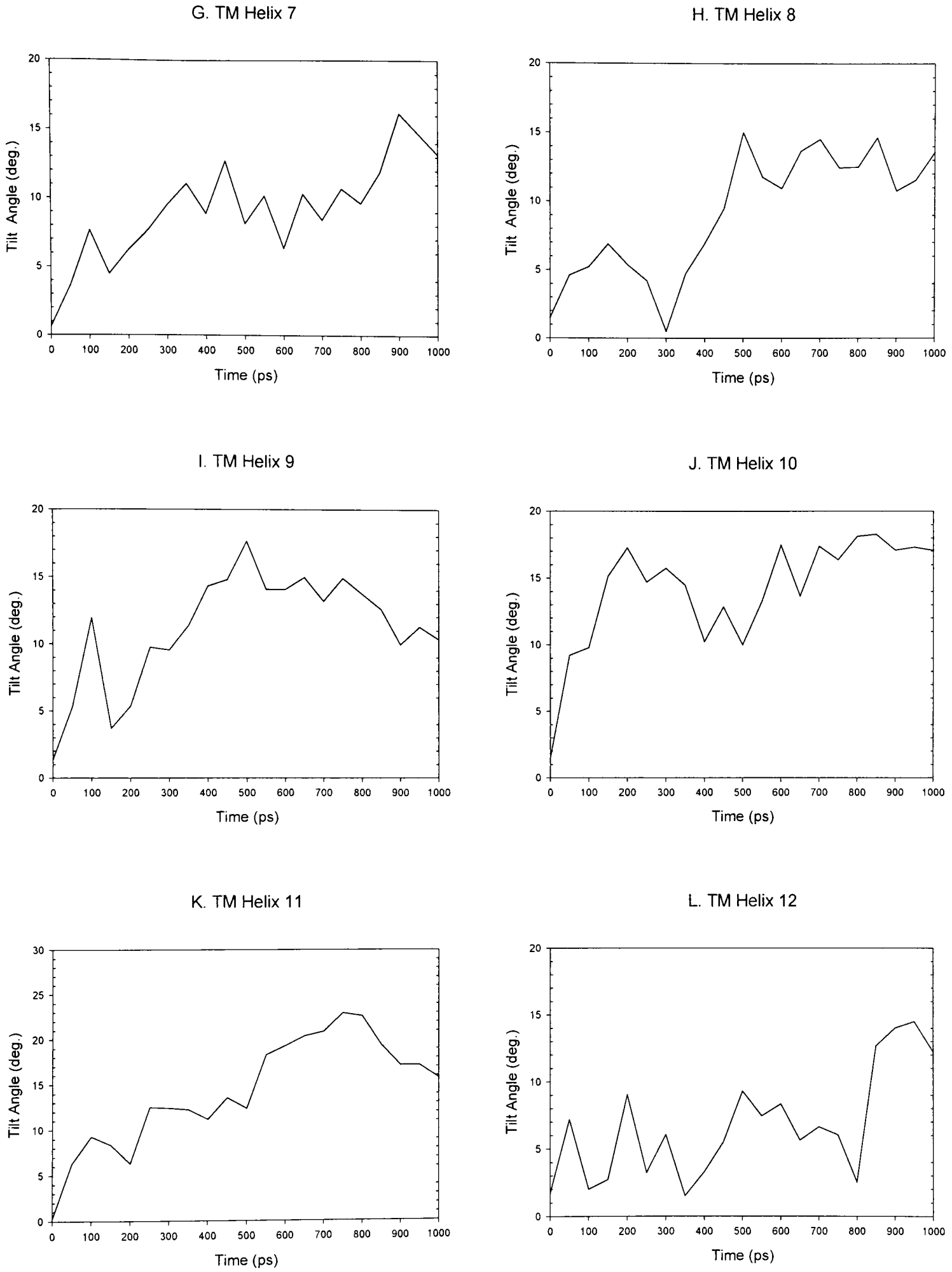


Figure 5.13 Helix tilt relative to the bilayer as a function of time. The helix axis is defined by a vector derived from the centre of the Cα atoms of residues 3 to 9, and the centre of the Cα atoms of residues *n* - 8 to *n* - 2, where *n* represents the length of each helix.

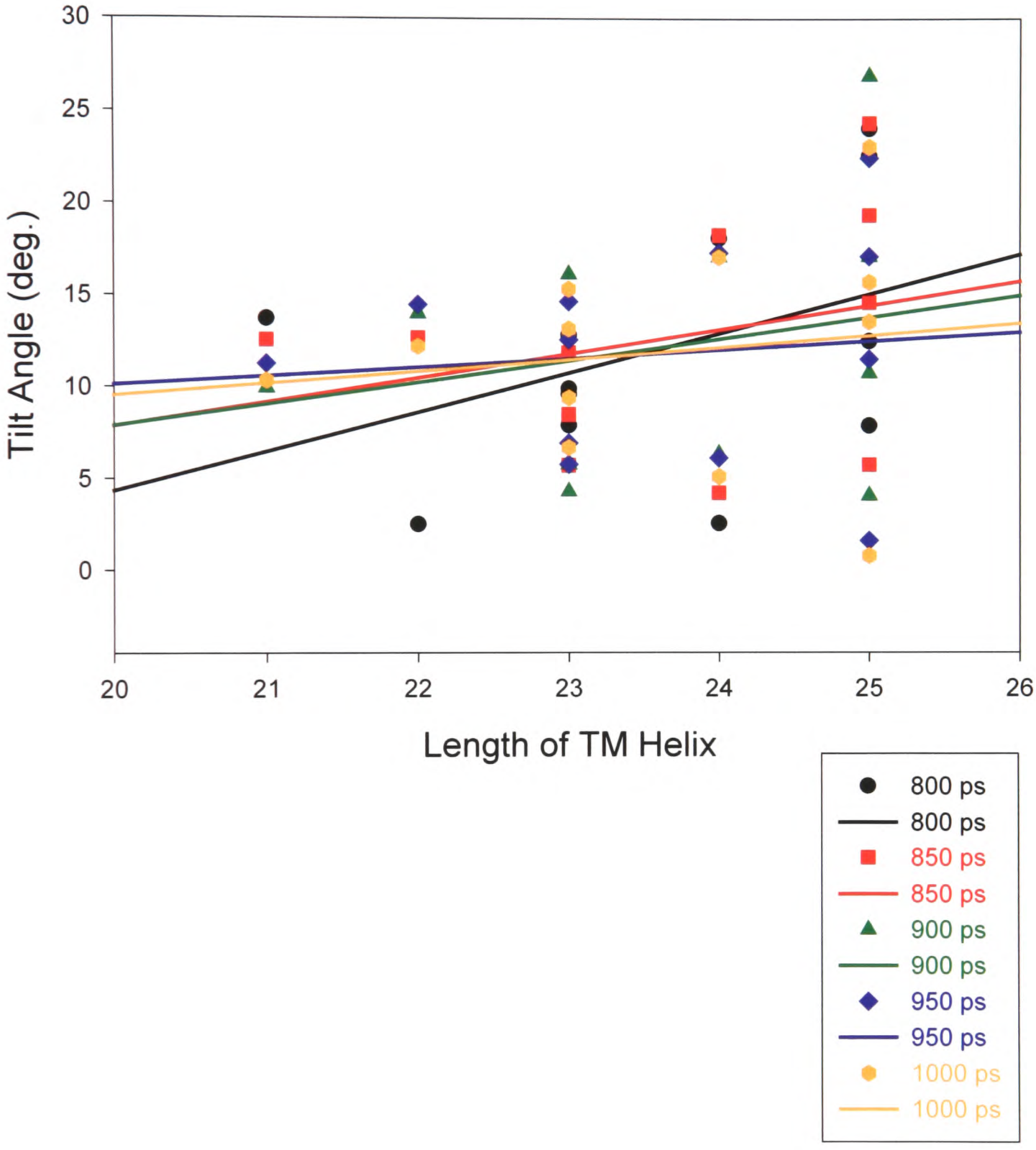


Figure 5.14 Regressions between distributions of tilt angles at last 200 ps for the 12 TM helices of CFTR and their helical lengths. The X axis represents the residue numbers of different TM helices, and the Y axis represents the tilt angles.

5.3.8 Dynamics of Helical Conformations II: Kink Angles.

Proline and glycine residues are known for destabilising the α-helical structure due to the cyclic conformation for proline and high degrees of freedom of peptide bond rotation for glycine. The helical structures are likely to bend due to these two residues, and it has been estimated that prolines bend helices by an angle of ca. 10^{o} to 20^{o} (Richardson and Richardson, 1989). For globular proteins, it has been suggested that proline-induced kinks facilitate the packing of long helical structures (Woolfson and Williams, 1990). For membrane proteins, the proline residues is with surprisingly high recurrence in the putative TM helices (Brandl and Deber, 1986). A variety of structural and dynamical roles has been suggested for the membrane-buried proline residues (Woolfson and Williams, 1990; von Heijne, 1991).

For the CFTR, several proline and glycine residues were identified along its putative TM helices. It is thus of interest to investigate kinks induced by these two amino acid residues. This is achieved by inspecting the snapshots of helical structures taken at different time steps in order to detect possible kinks. The kink angle induced by a proline or glycine residue is defined by the included angle between two separated segments caused by a kink, (Fig. 5.15). With this definition, a kink angle of ca. 180^{o} represents a nearly ideal helix, and a lower kink angle, *i.e.* less than 180^{o}, implies a bent conformation. Helices with kinks were then superimposed at either the N- or the C-terminal half of each helix in order to monitor the consequences of a kink occurring within a helix.

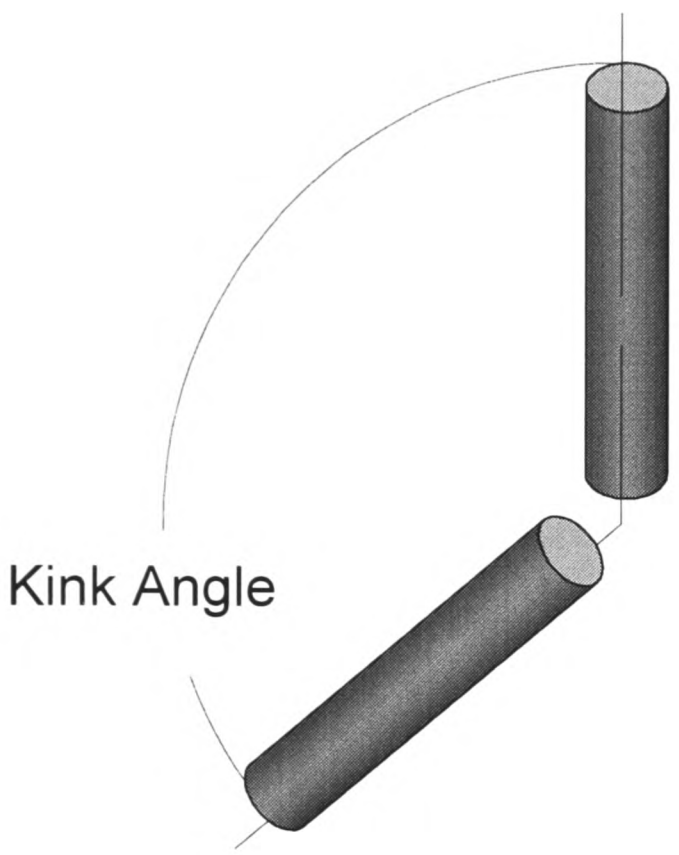


Figure 5.15 The definition of a kink angle used in this study.

For TM helix 2, a sharp kink is induced by the Gly126 at the 8th position within this helix, (Fig. 5.16 A). The least kinked helical conformation is detected at 650 ps with a kink angle of ca. 154°, and the most arched helical conformation is detected at 950 ps with a kink angle of ca. 131°. It is observed that TM helix 2 remains relatively less bent for the first 2/3 time duration of the 1 ns simulation, and adapts more bent conformations for the last 1/3 time of simulation, (Fig. 5.16 A). A contrast was seen for its tilt angle distributions; the helix is more inclined at the first 500 ps of simulation, and then turns less slanting after 650 ps, (Fig. 5.13 B). This sharp kink explains the lowest tilt angle of TM helix 2 within lipid bilayers. Superimposition of helical structures with different bent conformations at their N-termini reveals that Cα backbones moving from the starting positions at 0 ps to the most relaxed conformation at 650 ps, then to the most bend conformation at 950 ps and finally to the ending structure at 1 ns, (structures in Fig. 5.16 A).

For TM helix 3, a kink is observed at Pro205 at the 12th position of this helix, (Fig. 5.16 B). The kink angles at the beginning and the end of the simulation are estimated at 168° and 162° respectively, which is relatively less significant. However, the most relaxed kink angle of 175° is observed at 550 ps, and the most arched kink angle of 158° is observed at 650 ps. A sharp kink angle transition of 17° occurring within a time frame of 100 ps is observed, (Fig. 5.16 B).

For TM helix 5, a kink is observed at Gly314 at the 10th position within this helix, (Fig. 5.16 C). This helix shows an increasing pattern of bending at the first 250 ps, and has the most arched conformation at 250 ps for 140°. Then the kink at Gly314 becomes less bent, and reaches the most flat conformation at 600 ps for 165°. After 600 ps, the kink conformation is getting bent again to the end of simulation.

The difference of kink angles between structure at 0 ps and 1000 ps is less than 2°, and very insignificant. However, the difference of kink angles between the most arched structure and the most flat structure is ca. 25°, (structures in Fig. 5.16 C).

For TM helix 8, a kink is detected at Gly934 at the 16th position of this helix. The kink at this region remains stable and flat throughout the 1 ns simulation, (Fig. 5.16 D). The most flat conformation, ca. 175°, occurs at the end of simulation. The difference of kink angles between the starting and final structures is small and less than 6°. However, a bent kink angle is seen at the 500 ps for ca. 153°, which makes the difference between the most arched and the most flat structures for ca. 22°, (Fig. 5.16 D).

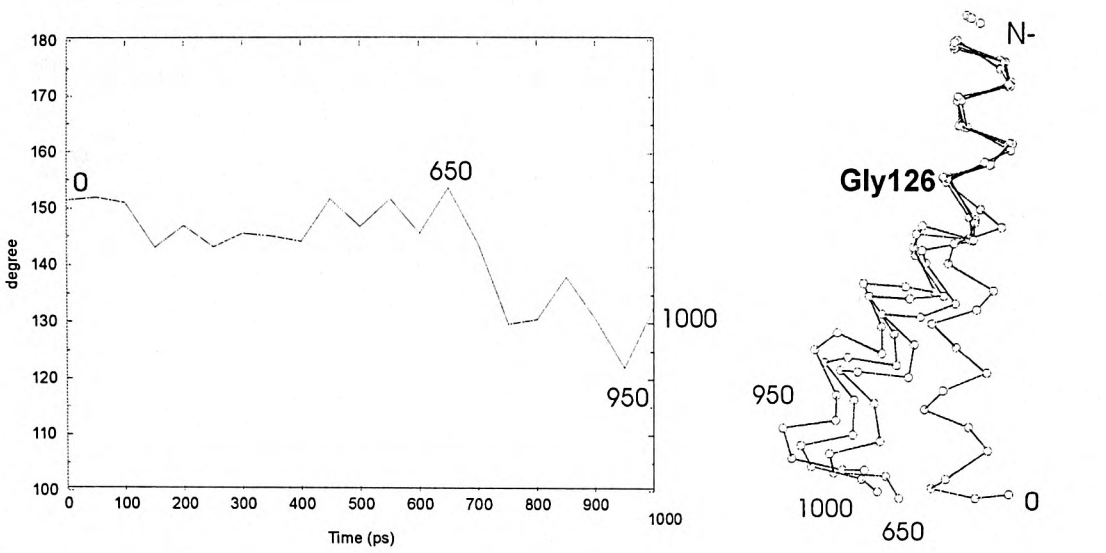
For TM helix 10, a kink is detect at Pro1021 at the 6th position within this helix. In comparison with other kink angles induced by glycine residues, the distribution of kink angles induced by Pro1021 shows a very zigzag pattern, (Fig. 5.14 E), which indicates that this residue is capable of changing its conformation within a short period of time. For example, the most arched conformation is seen at 200 ps for 153°, and the most relaxed kink angle is detected at 300 ps for 170°. A sharp kink angle change of 17° occurs within 100 ps. However, the difference of kink angles at the 0 ps and 1000 ps is still of no significance.

It is found that the differences of kink angle at 0 ps and 1 ns are generally less significant; most of them are less than 6°, except for the kink in TM helix 2 induced by Gly126. There are subtle differences between kink angles induced by proline residue and glycine residue. For kink angles induced by proline resides, Pro205 and Pro1021, a meandering pattern is observed for their kink angle distributions. In addition, both their kink angle transition between the more bent

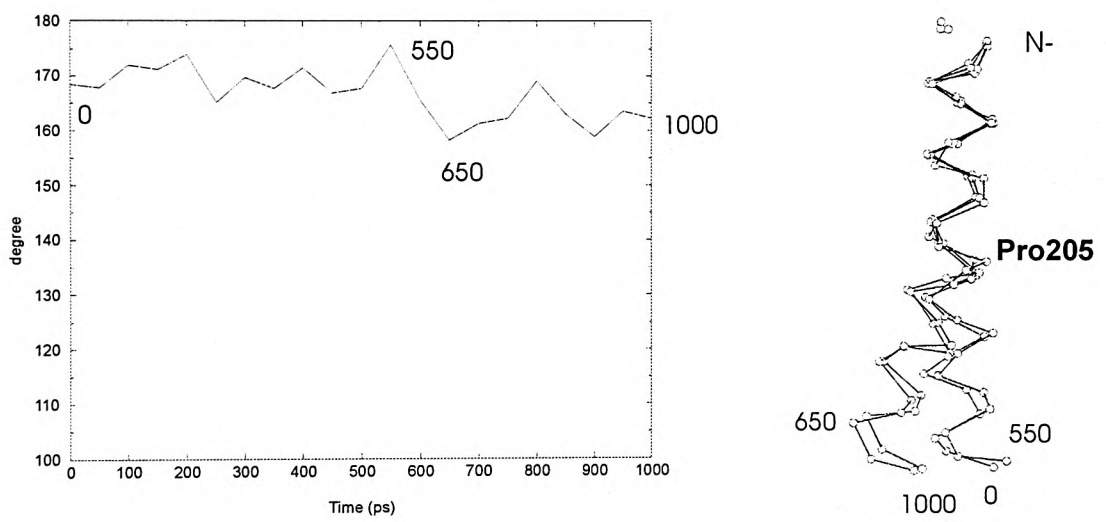
conformation and the most relaxed conformation occur within a period of 100 ps for a change of ca. 17°. This implies that kink angles induced by proline residues are more flexible to change their conformations rapidly from either arched or flat conformation to the opposite conformation.

For kink angles induced by glycine residues, Gly126, Gly314 and Gly934, a relatively smooth pattern is observed for their kink angle distributions. In addition, kink angles induced by glycine residues have larger differences between the most arched and the most relaxed conformations; all kink angle differences are more than 20°. This implies that kinks induced by glycine residues are associated with dynamical flexibility, allowing to adjust their conformation as a function of time.

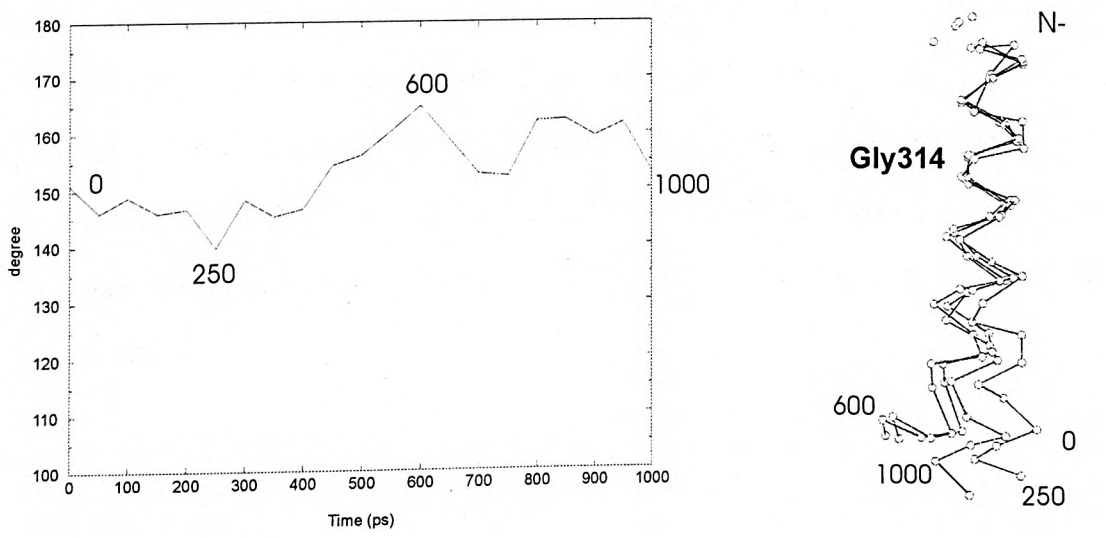
A. Kink angles on Gly126 of TM helix 2 and their Cα traces



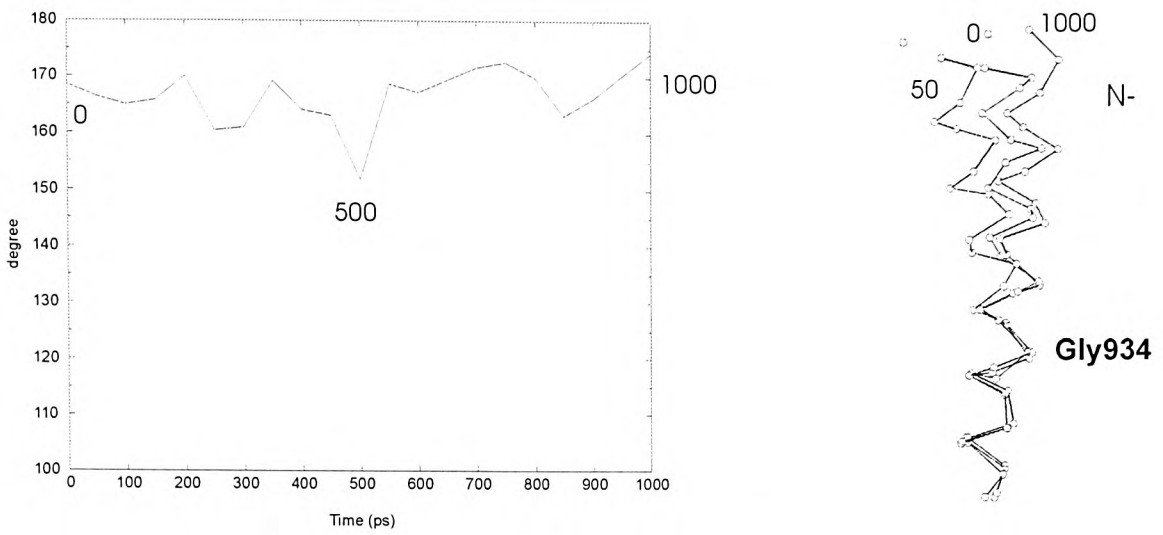
B. Kink angles on Pro205 of TM helix 3 and their Cα traces



C. Kink angles on Gly314 of TM helix 5 and their Cα traces



D. Kink angles on Gly934 of TM helix 8 and their Cα traces



E. Kink angles on Pro1021 of TM helix 10 and their Cα traces

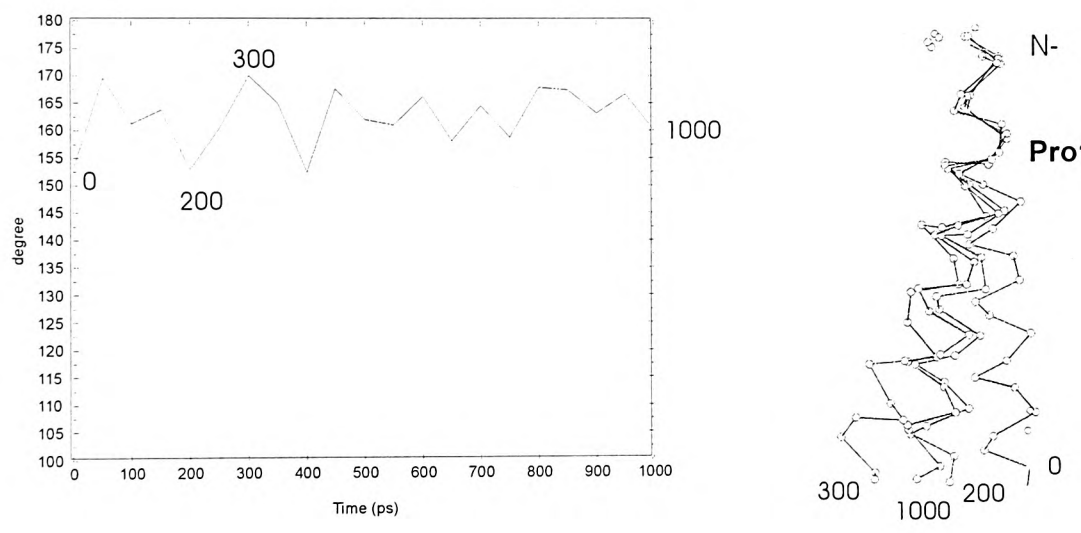


Figure 5.16 Kink angles as a function of time for different helices. The kink angle is derived from the geometric centre of a span of three residues prior to the residue of interest, the geometic centre of the residue of interest, and the geometric centre of a span of three residues following the residue of interest. Helical structures from different time frames are aligned at the N- terminal residues up to the kink regions, except for TM helix 8 which is aligned at the C-termini.

5.3.9 Interactions at the Interface.

Aromatic amino acids residues have been proposed to be favoured amino acids within TM helices at the interfaces of lipid bilayers (Wimley and White, 1996). A statistical analysis also indicates that the amphipathic aromatic amino acids, tryptophan and tyrosine, are found preferentially at the termini of TM helices, and the phenylalanine distributes randomly in membrane proteins (Arkin and Brunger, 1998). The favourable location for tryptophan and tyrosine residues is likely due to the potential hydrogen bonds between amphipathic aromatic side chains with lipid or water molecules, which have been suggested to "lock" a TM helix or protein into the lipid bilayer (Schiffer et al., 1992). A number of computational investigations have already been focussed on these interactions (Woolf and Roux, 1994; Woolf and Roux, 1996; Woolf, 1998).

The hydrogen bonds between the terminal aromatic residues, POPC molecules and water molecules were examined. Furthermore, the orientations of the aromatic ring related to the lipid bilayers were also analysed in this study; two types of the aromatic ring's orientation were defined, floating and anchoring. For the floating conformation, the aromatic ring is "floating" on the surface of the lipid bilayer horizontally, (Fig. 5.17 A); for the anchoring conformation, the aromatic ring is "anchoring" into the lipid bilayer perpendicularly, (Fig. 5.17 B).

There are 9 aromatic residues, out of 24, on either the N-terminal or the C-terminal sides of these 12 CFTR TM helices, (Tab. 5.8). Both anchoring and floating orientations are evenly identified, and no preference for a specific orientation is detected. For the phenylalanine residue, their aromatic rings are predominantly in the floating orientation; 3 out of 4, with the only exception of the Phe77 at the N-

terminus of the TM helix 1, (Tab. 5.8). For the tryptophan and tyrosine residues, their aromatic rings are predominantly in the anchoring orientation; 4 out of 5, with the only exception of the Trp882 at the C-terminus of the TM helix 7, (Tab. 5.8).

The phenylalanine residue is not able to form hydrogen bonds with lipid molecules, due to the nature of its side-chain. However, hydrogen bonds are detected exclusively between the amide nitrogens or carbonyl oxygens and water molecules for the four terminal phenylalanine residues, (Tab. 5.8).

For the amphipathic amino acid residues, most of them form multiple hydrogen bonds with POPC or water molecules. The only exception is Trp1098 at N-terminus of TM helix 11, which forms only one hydrogen bond with the POPC molecule, (Tab. 5.8). Two hydrogen bonds are identified for the Trp882 at the C-terminus of TM helix 7; one with a water molecule and one with a POPC molecule. However, the aromatic ring of Trp882 resides in a floating orientation on the membrane, (Tab. 5.8).

Whether the hydrogen bond plays a role in the orientation of an aromatic ring is not conclusive from these 11 residues. However, it is seen that the phenylalanine tends to have a floating conformation for their aromatic rings. In contrast, the amphipathic aromatic amino acids, which are able to form extra hydrogen bonds with POPC, tend to have an anchoring conformation for their aromatic rings.

The importance of hydrogen bonds between the amphipathic aromatic side chains and lipid molecules have been suggested in a recent MD study of the potassium channel (Shrivastava et al., 1999). However, effects of the hydrogen bond on the orientations of the aromatic ring are not yet understood, since significant changes

of the aromatic ring's orientation were not detected within the time course of 1 ns (Tieleman et al., 1999b; Shrivastava et al., 1999).

◆ **Inter-molecular H-bonds at interfacial aromatic residues:**

TM Helix	Residue	H-bond w/ POPC	H-bond w/ H_2O	Orientation*
1	N-Phe	0	1	A
2	C-Phe	0	1	F
3	C-Trp	2	0	A
5	N-Phe	0	1	F
6	C-Trp	1	1	A
7	C-Trp	1	1	F
8	N-Tyr	0	3	A
10	N-Phe	0	1	F
11	N-Trp	1	0	A

✶ Orientations of aromatic ring toward lipid surface

Table 5.8 Orientations of aromatic side chains.

A. CFTR TM helix 3 (anchoring)

B. CFTR TM helix 10 (floating)

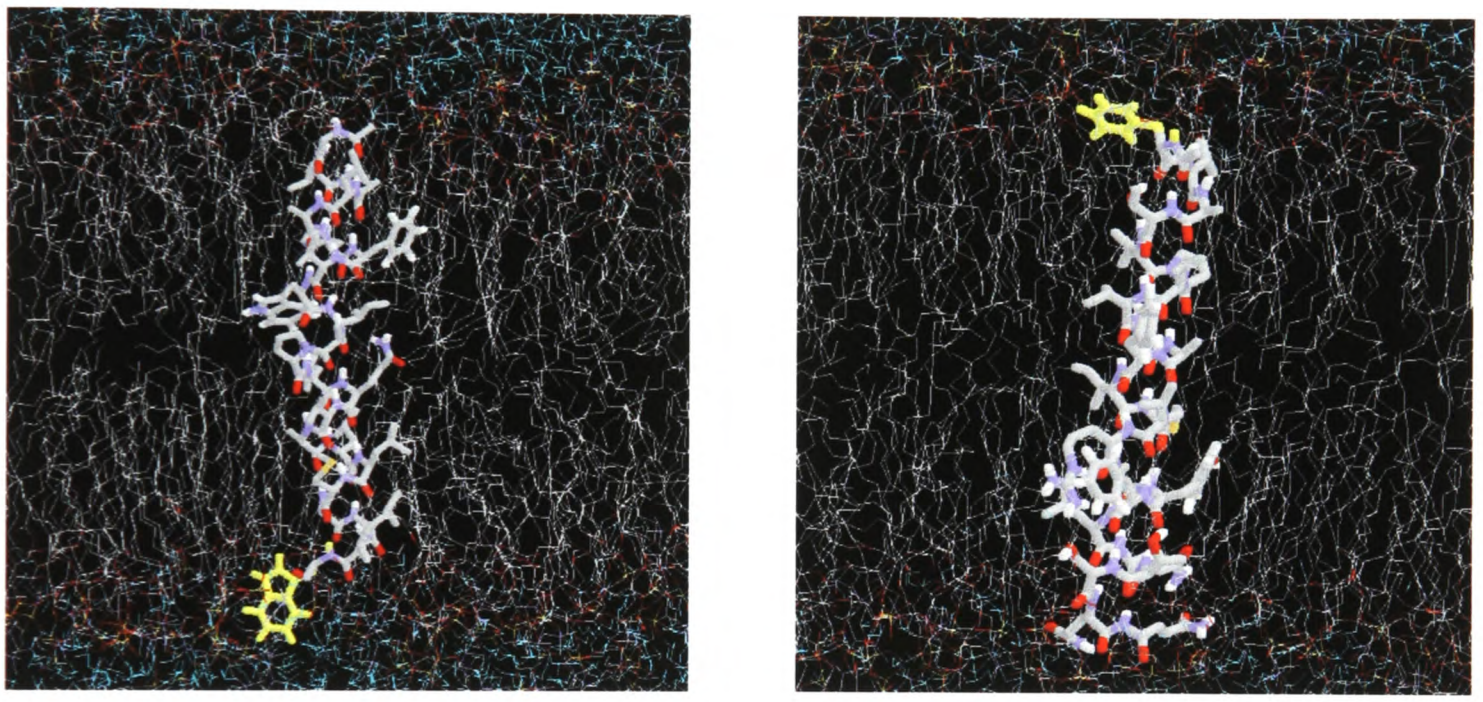


Figure 5.17 Orientations of aromatic side chains.

5.4 Discussion.

5.4.1 Significance of MD Simulations.

In this study, subtle differences in structural dynamics are revealed by the Cα RMSD analysis of each TM helix in CFTR. The overall dynamics of secondary structure are then monitored using DSSP, which reveals some unusual changes of structure. With further analysis of Cα atoms' fluctuations and helicity of each residue, residues with enhanced flexibility are identified.

The structurally unstable residues on a helix may be with functional importance; for example, helical kinks or hinge bending motions. By measuring the tilt and kink angles, behaviours and motions of helices relative to the lipid bilayer are detected. Furthermore, certain residues of interest can be better observed by examining their kink angle fluctuations. Examinations of different interactions provide further simulation data on the nature of residue/helix, helix/phospholipid and helix/water interactions

With applications of MD simulations on isolated helices, the length of a TM helix and the residues within a TM helix can be optimally defined (Forrest et al., 1999). In addition, it may reveal aspects of TM helix dynamics and structural stability which, although present in the corresponding intact proteins, are modulated by the remainder of the intact protein. For example, a pronounced hinge-bending motion of the S6 helix from a voltage-gated shaker (Kv) channel has been demonstrated in MD simulation studies, in both *in vacuo* system (Kerr et al., 1996) and solvated POPC simulations (Shrivastava et al., 1999), which suggest that it is likely that in

an intact membrane protein such conformational transitions may be coupled to changes in the rest of the protein.

5.4.2 Biological Implications of This Study.

In the context of the two-stage folding model for membrane proteins (Popot and Engelman, 1990), it is of interest to examine how interactions between TM helices, lipid bilayer and water affect helical arrangement of the TM domain.

Four cystic fibrosis-associated mutations have been identified at position 347 of TM helix 6: R374C, R347H, R347L and R347P (Dean et al., 1990; Cremonesi et al., 1992; Audrezet et al., 1993). This clearly indicates that Arg347 is important for the structure and function of CFTR. The Arg347 in TM helix 6 has been observed to form an intra-molecular hydrogen bond between its side chain and the backbone carbonyl of Cys342 in this isolated TM helix MD simulations. This unexpected hydrogen bond results in a split down of TM helix 6.

This finding reflects the nature of Arg347 in interacting with a residue or a molecule in order to neutralise its positive side chain. This residue has actually been proposed to be an anion binding site in the CFTR pore (Tabcharani et al., 1993). However, a recent research suggests that the Arg347 plays a structural role in stabilising the conformation of CFTR by forming a salt bridge with Asp924 in TM helix 9 (Cotten and Welsh, 1999).

The helical structure of TM helix 6 at the 0 ps could be regarded as the primary product of the two-stage folding process, (Fig. 5.2). The following process of helical packing and interaction is then facilitated with its positive nature to link with an acidic residue within another TM helix in order to build the tertiary fold in membrane. In the isolated helical system, this residue fails to achieve the inter-

molecular interaction, and then adopts an alternative intra-molecular hydrogen bond, which results in a deformation of the original helical structure.

Similar scenarios were observed in Glu873 of TM helix 7. Whether this implies that the Glu873 also plays a role in helical links in order to stabilise and pack helices within TM domains of CFTR remains to be seen with more functional analysis on this residue. However, the potential importance of Glu873 is hypothesized by this MD simulation study.

The MD studies also provide further information on the nature of interactions between lipid and interfacial aromatic amino acid residues. The side chains of tryptophan are identified to be able to form a hydrogen bond to lipid molecules. This supports the proposed role of amphipathic aromatic sidechains in "locking" integral proteins into a bilayer (Schiffer et al., 1992). These interfacial hydrogen bonds potentially provide a way to stabilise the TM helices with lipid bilayers during the second-stage of the folding process.

In addition, simulations of individual TM helix may also yield insights into the nature of the critical residues for protein structures and activities. For example, the hinge-bending motions, caused by kinks at certain residues along the TM helices, have been proposed to provide possible mechanisms for channel gating, (Kerr et al., 1996). It is therefore of interest to look for possible kinking sites along a TM helix.

Proline residues are more frequently observed in membrane-spanning segments of ion channels and transporters than in the membrane-embedded regions of membrane protein with no transporting function (Brandl and Deber, 1986). Many possible functions of the membrane-buried proline residues have been proposed

which include facilitating the conformational change of a TM helix, and provision of ion liganding (von Heijne, 1991; Sansom, 1992; Jacob et al., 1999; Ri et al., 1999). Several interesting proline residues have been spotted and studied by the MD simulations.

A mutation of proline to serine at position 205 of TM helix 3 is associated with mild cystic fibrosis (Chillon et al., 1993). It has been inferred that a bend or a kink induced by the Pro205 may cause the cytoplasmic end of the TM helix 3 to angle away from the channel wall in toward the central axis of the channel, and this bend provides a possible mechanism to narrow the channel at its cytoplasmic end (Akabas, 1998). This flexibility of Pro205 has been clearly observed in this MD simulations study. It exhibits the potential to adapt a sharp angle change of 17° within 100 ps. The backbone hydrogen patterns disrupted by Pro205 are also detected. The disruptions cause the severe loss of helicity of Val201 and Trp202, which are also observed in this study. The break down of helical hydrogen patterns enable backbone carbonyls and amides to interact with water or ions. Similar effects have been observed from the structure of the potassium channel (Doyle et al., 1998).

The Pro1021 in TM helix 10 is also one of the conserved residues in CFTR across different species (Sheppard et al., 1996). Site-directed mutagenesis data indicate that mutations of Pro1021 change the chloride ion current and the ion selectivity in the HeLa cell (Sheppard et al., 1996). With the flexibility observed for Pro1021, it is reasonable to speculate that this residue may be of importance to maintain the channel architecture in order to regulate the ion transportation of CFTR.

Apart from the proline residue, glycine is another residue with potential to regulate channel activity by altering the helical conformation (Jacob et al., 1999). The glycine residue at position 314 of TM helix 5 has been found very conserved in all CFTR genes cloned in different species (Mansoura et al., 1998). Patients with cystic fibrosis have been reported with mutations at this residue, glycine to aspartic acid (Golla et al., 1994) and glycine to arginine (Nasr et al., 1996). Mutational studies at position 314 propose that TM helix 5 may play a role in conformational changes to control the ion channel gating (Mansoura et al., 1998). It has been demonstrated in this study that Gly314 has the flexibility to adjust the kink angle of TM helix 5 for a range of ca 25°. This provides a structural explaination for possible involvement of Gly314 in the physical process that open and close the channel.

5.4.3 Limitations of the MD Simulations.

It is important to consider, albeit briefly, the limitations of these simulations. One limitation is the use of just a single lipid species, POPC. It will be interesting to study interactions between lipid bilayer and peptides in a range of different lipid molecules. This is becoming foreseeable with the increasing power of computer. In addition, long-range electrostatics interactions have been treated approximately (*i.e.* by use of cutoffs). It will be of interest to see what changes in conclusion, if any, arise from use of more sophisticated approaches to long range electrostatics such as Ewald summation (Tobias et al., 1997; Tieleman et al., 1997). This will be especially useful for investigating residues which are in ionized state, *e.g.* Arg352 and Glu873, to detect the possible stabilising effects for linking helices *via* electrostatics forces or salt bridge provided by their charged nature.

Chapter 6

Concluding Remarks And Future Directions

6.1 Summary.

Computational methods have been applied in this thesis including statistical analysis, algorithms for protein secondary structure assignments, various database mining techniques, algorithms for protein sequence comparison, algorithms for protein secondary structure prediction and molecular dynamics simulations. These computational methods were employed in studying protein structure assignments, sequence phylogenetic analysis, predictions of TM helices for integral membrane proteins and dynamics of isolated TM helices within lipid bilayers.

Chapter 2 described a comparison study for protein secondary structure assignments and descriptors for projecting protein secondary structure elements. Chapter 3 surveyed the peptide bonds in *cis* conformation based on the levels of protein molecules and protein domains, and investigated the possible biological functions for the *cis* peptide bonds. Chapter 4 examined a family of proteins, the ABC transporter, in different genomes, and looked into the possible correlations between their sequences, putative TM helical structures and functions. Chapter 5 explored the structural dynamics of TM helices of the cystic fibrosis transmembrane conductance regulator by MD simulations, and the relationships between biophysical behaviours of each CFTR TM helix and the function of CFTR.

6.2 Future Directions.

6.2.1 Protein Secondary Structure Assignment.

Different programs give results of protein secondary structure definition with subtle differences. Looking for a "best" program to define protein secondary structures seems not appropriate, since this assignment process is very intuitive and subjective. Different programs should be tested for different purposes on the basis of criteria lying behind each method.

Consensus Assignments and Structural Propensity

The idea of consensus assignments by combining assignment results from different methods has been proposed in order to get around the inaccuracies of one secondary structure definition (Colloc'h et al., 1993).

This idea can be extended to the propensity of a residue for a certain secondary structure type. Firstly, a protein structure can be tested by different methods. By evaluating the structural information assigned by different methods, the secondary structure definition of a particular residue can be expressed as a percentage of the different definition given. Employing a weighting scheme, which will be derived from specific requirements relating to each program's unique criteria, will transform the percentage to a propensity of a residue for a specific protein secondary status. With this approach, protein structure definitions can be better applied to describe a residue, and can be fit to different requirements by changing the weighting scheme.

6.2.2 *Cis* Peptide Bonds and Disordered Protein Structures.

Peptide bonds in the *cis* conformation have been observed more abundant than it is expected. However, the biological importance of *cis* peptide bonds and the possible *cis/trans* isomerisation within a protein structure is less clear unless more protein structures are solved the high resolutions.

Applications of Molecular Dynamics Simulations

One different approach is to apply combinations of homology modelling and MD simulations to this problem. Many protein structures were discussed in section 3.3.2 for the potential functional and structural roles of *cis* peptide bonds. On the basis of templates provided by these structure-solved proteins, similar structure models can be built up; however, the *cis* conformation will be assigned to targeted residues in these constructed structures. This will generally achieve a homology modelling with different *cis/trans* conformations at certain microenvironments.

Followed by applications of MD simulations on these artificial structures, detailed dynamics for residues of interest and for the overall protein structures can be detected. This will hopefully provide a way to deduce the structural and functional implications of *cis* peptide bonds in protein structures.

6.2.3 ABC Transporters.

Strategies were shown here to scrutinise ABC transporters from their primary sequences. Correlations have been detected between their sequence relationships, arrangements of putative TM helices and their functions. However, these genome databanks are constantly updated. It will be useful to automate and improve methods employed in this study, and then expand to the following directions.

Functional Studies

It will be of interest to extend this study to understand the correlations between structural similarities and functional similarities in a broad view; *i.e.*, how structures differ from exporting system and importing systems of ABC proteins, the structural similarities between periplasmic binding proteins and fold arrangements of helical structures. This is much dependent on an improved functional characterisation of ABC transporters.

Comparison of Transmembrane Helical Folds

Comparison of putative TM helices either based on the linear arrangements of TM helices or based on their hydrophobic/conservative periodicity were achieved with human inspection in this study. Developing a mechanism to carry out the whole process automatically will be of immense help to speed up and to extent this strategy to the many other emerging sequencing-completed genomes. In addition, with the methods integrated here, a long-term goal of this study is to devise a mechanism for studying membrane proteins on the basis of the predicted TM topologies and structures from their primary sequences.

Variations in Sequence and Their Impact

One useful extension of this study is to construct a database to accommodate a variety of transporters in the context of their primary sequences and their functions in order to understand how variation in sequence correlates to different phenotypes. This will speed up the screening process for drug-development targets and be of great help for identifying potential target for medical treatment as well.

6.2.4 Cystic Fibrosis Transmembrane Conductance Regulator.

From the MD simulation study, some critical residues, which have been proven to be involved in the regulation of channel gating experimentally, have been identified to have unusual dynamics along the TM helix. This also suggests a mechanism for those critical residues that they may carry out their function by changing the helical conformations through hinge-bending motions or kinks. It has also been detected that a positive charged residue within the pore-lining helix forms an unexpected hydrogen bond and then alters the helical structure seriously in an isolated helix environment. This residue has been suggested to provide an anion-binding residue, or form a salt bridge with a negative charged residue on the other helix in order to maintain the tertiary helical arrangement of TM domains.

In silico Site-directed Mutagensis

With the capacity of the MD simulations to investigate dynamics of each residue, it will be of immense interest to expand this study to examine the hot spots of mutation within TM helices for cystic fibrosis. In other words, construct helical models based on mutant protein sequences, and then monitor the dynamical behaviour of the mutant TM helices in comparison with the normal helices.

Detections of Inter-helical Interactions

One alternative approach to analyse the interaction between helices is to construct a larger membrane system containing a wider hole to accommodate more than one helix, which has been reported in previous studies (Tieleman et al., 1999a; Tieleman et al., 1999b).

With this expansion, it will be of great help in identifying residues important for maintaining channel architecture, channel gating and formation of channel pore.

For example, dock both TM helices 6 and 9 into a lipid bilayer system and then examine the possible interactions between Arg347 and Asp924. This will make it clear whether the Arg347 is an anion-binding residue as proposed (Tabcharani et al., 1993), or is responsible for inter-helical interactions between TM helix 6 and 9 (Cotten and Welsh, 1999).

Explore the intact topological model of TM domains of CFTR. This will provide insights on how the 12 TM helices pack and arrange within the lipid bilayer, and enhance our understanding of the mechanism of ion selectivity for CFTR, and ion permeation/transportation of CFTR

Simulations of Intact CFTR Molecule

Simulate the intact CFTR protein. This is the long term goal of this study and in need of more experimental data and structure-determined structures to support it. By examining interactions between membrane-spanning domains, two nucleotide-binding domains, regulatory domain and loop regions, it will provide insights of how CFTR functions as a cAMP-regulated chloride channel. Hopefully, it will lead to mechanisms to cure the genetic disease, cystic fibrosis.

6.3 Final Remarks.

This thesis demonstrates the applications of a variety of computational methods to the study of biological systems, and has illustrated where computational studies can make contribution to our understanding of a biological problem.

Biological studies used to be descriptive. However, the constant accumulation of biological information from experiments and rapid expansions of sequence information from genome projects have changed the way of doing biology dramatically. Furthermore, with help from powerful computers and well-designed algorithms, a lot of biological questions can be answered now without doing a "wet" experiment but just by computational work. Examples of different aspects have been presented in chapter 3, 4, 5 and 6 of this thesis.

How well shall we believe these computational studies and how reliable the results coming out from the "non-wet" experiments are should not be the only questions asked. Emphasis should be placed on the use of computational methods to model biological systems in order to generate data which are complementary to experimental data, and provide insights in designing further experiments. In addition, computational approaches in many cases can yield information, which is still inaccessible by "wet" experiments. It seems likely that, in time, computational approaches will be as an invaluable tool as other experimental methods used daily in laboratories are.

Appendices

A.1 Introduction to Molecular Dynamics.

Many process of functional interest in protein structures involve the motion of atoms at relatively low frequency, which can be well described by classical dynamics [McCammon, 1987 #377]. Hence, by implementing classical mechanics at the atomic scale, the motion of atoms in a protein can be simulated by *Newton's Second Law of Motion*. This technique is referred as "molecular dynamics" (MD).

Molecular dynamics has a wide range of uses in studying proteins, such as protein-protein interactions and interactions between different biological macro-molecules [Karplus, 1990 #73]. With this technique, the motion of atoms in a protein of known structure can be better interpreted. The insights between these interactions and their implications toward structure/function relationships may be acquired.

The method of molecular dynamics simulations are employed extensively in the studies of this thesis. Details of this method are summerised concisely here.

A.1.1 Applications of Classical Mechanics.

In the classical mechanics, the motion of an object is governed by Newton's second law of motion (equation A.1), where F is the force on this object, m its mass, and a its acceleration.

$$F = ma \tag{A.1}$$

Expansion of this equation A.1 into a system containing N objects; for example, a system of N interacting atoms at the atomic scale, where $i = 1 \ldots N$.

$$F_i = m_i a_i \qquad \text{(A.2-1)}$$

$$a_i = \left(\frac{\partial^2 r_i}{\partial t^2}\right) \qquad \text{(A.2-1)}$$

While r_i is the position of atom i, combinations of equations A.2-1 and A.2-2 will give equation A.2.

$$F_i = m_i \left(\frac{\partial^2 r_i}{\partial t^2}\right) \qquad \text{(A.2)}$$

In molecular dynamics simulations, atomic positions are determined as a function of time. To begin a simulation at time $t = 0$, an initial set of atomic coordinates and velocities are required. Velocities are assigned to atoms at random, *e.g.* using a Maxwell distribution. The position of atoms after a time-step Δt, may be determined using equation A.3 derived from equation A.2.

$$x_i(t + \Delta t) = x_i(t) + v_i \Delta t + \frac{1}{2} a_i (\Delta t)^2 \qquad \text{(A.3)}$$

where, x_i is the position of atom i, v_i its velocity and a_i its acceleration. In order to determine the acceleration of an atom it is necessary to know the force acting on it. The force can be derived from the potential energy (U) of the system (equation A.4).

$$F_x = -\frac{\partial U}{\partial x} \qquad \text{(A.4)}$$

A.1.2 Force Field.

The typical potential force field, *i.e.* energy function, commonly used in molecular dynamics can be subdivided into non-bonded interactions, bonded interactions and restraints [Goodfellow, 1992 #74]. Thus, the force field is expressed as

$$E_{TOTAL} = E_{NON-BONDED} + E_{BONDED} + E_{RESTRAINT} \qquad \text{(A.5)}$$

• *Non-Bonded Interactions*:

There are two types of non-bonded interactions between two atoms, the van der Waals potential and the electrostatic interactions. The van der Waals potential can be represented by two different models; the Lennard-Jones interaction and the Buckingham interaction.

In this thesis, the Lennard-Jones interaction is used to address the van der Waals potential. Thus, the non-bonded interactions can be expressed as:

$$E_{NONBONDED} = \sum_{i=1,natoms} \sum_{\substack{j=1,natoms \\ i<j}} \left(\frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}} \right) + \left(\frac{q_i q_j}{4\pi\varepsilon_o \varepsilon_r r_{ij}} \right). \qquad \text{(A.6)}$$

The first term is the Lennard-Jones potential, composed of a repulsive term representing repulsive nuclear and electronic interactions and an attractive term representing favourable dipole-dipole interactions, where r_{ij} is the distance between pairs of atoms. A and B are constants obtained from consideration of number of outer electrons and polarisability. The second term is the Columb

potential, which describes electrostatic interactions, where q is the charge on each atom, ε_o is the permittivity of free space and ε_r is the dielectric constant of the medium between the two atoms. The dielectric constant between a pair of atoms, *i.e.* ε_r, is set to be r_{ij}. With the limitations of computer power, the long-range electrostatic interactions are always ignored.

• *Bonded Interaction*:

Bonded interactions do not only include the bond stretching energies between a pair of atoms, but also the conformational interactions of 3-body bond angle energies and 4-body dihedral angle energies. A special type of dihedral interaction, improper dihedral, is also included in the bonded energies in order to force atoms to remain in a plane or to prevent transition to a configuration of opposite chirality [Brooks, 1983 #75; Brunger, 1992 #76]. The forms of bonded energy are expressed as following:

$$E_{BONDED} = \sum_{bonds} k_B (r - r_x)^2 + \sum_{angles} k_\vartheta (\theta - \theta_x)^2 + \sum_{dhedrals} \sum_{=1,m} k_\phi (1 + \cos(n\phi + \delta)) + \sum_{mpropers} \sum_{=1,m} k_\phi (\phi - \delta)^2 \quad \text{(A.7)}$$

where k specifies force constants.

Each term effectively restrains the particular quantity in question to lie close to an equilibrium value specified in a parameter file. These equilibrium values are obtained from analysis of experimentally derived structures and force constants from spectroscopic properties of small molecules [Brooks, 1983 #75].

• *Restraints*:

Some extra energies are used in order to impose restraints on the motion of the system, either to avoid unreasonable deviations, or to include knowledge from experimental data. Strictly speaking, these restraint energies are not part of the force field. There are two kinds of restraints frequently used. The position restraint is used to restraint a particle to fixed reference position, and the distance restraint is used as an additional penalty to the potential when the distance between specified pairs of atoms exceeds a threshold value [Brooks, 1983 #75].

A.1.3 Building Ideal α-helical Structures by MD Simulations.

Idealised transmembrane α-helical models were generated with a two-stage *in vacuo* SA/MD method [Kerr, 1994 #177; Kerr, 1994 #178]. At the first stage, a Cα helical template is generated with an ideal structure of 3.6 residues/turn followed by the superimposition of side-chain atoms on the corresponding Cα atom. Then, the simulated annealing is applied to construct the bond lengths, bond angles, peptide planarity and peptide chirality gradually. The annealing starts at 1000K and ends at 300K. The *van der Waals* repulsion is slowly introduced in the annealing process. Five α-helical templates are generated from this stage.

At the second stage, each structure from stage 1 is applied into 5 runs of molecular dynamics. Firstly, positional restraints are imposed on Cα atoms and gradually relaxed during the process of temperature reduction dynamics from 500K to 300K. Afterwards, distance restraints are introduced at a 5 *ps* burst of constant

temperature dynamics and the final 1,000 steps of conjugate gradient energy minimization. During the latter burst of dynamics and energy minimization, no positional restraints are imposed on the positions of the Cα atoms; however, distance restraints are applied between the carbonyl O of residue $i + 4$ and the amide H of residue i in order to maintain α-helical geometry.

Positional restraints are imposed using the NOE energy term of the X-PLOR potential function [Brunger, 1992 #76], equation A.8

$$E_{NOE} = \min(E_{MAX}, \frac{Sk_BT}{2\rho^2}(d - d_{OPT}))^2 \qquad \text{(A.8)}$$

where E_{MAX} = 50 kcal mol^{-1}, S = 2.5, d is the actual distance between a pair of atoms, d_{OPT} is the optimum distance and ρ is the allowed deviation from d_{OPT}.

Intra-helical distance restraints for maintaining α-helical structure were derived from α-helical H-bonding geometries observed in crystal structures of globular proteins [Baker, 1984 #180] and are implemented using a harmonic potential energy function, equation A.9

$$E_{HARM} = H\sum_i (r_i - r_i^{REF})^2 \qquad \text{(A.9)}$$

where, r_i is the coordinate of a Cα atom and r_i^{REF} is the coordinate of the atom in the target structure and H is a force constant with value 4.0 kcal mol^{-1} $Å^{-2}$ at the start of stage 2 and 0.5 kcal mol^{-1} $Å^{-2}$ at the end.

Electrostatic interactions are implemented into the potential energy function in order to be applied in the process of temperature reduction dynamics, the 5 *ps* constant temperature dynamics and the final energy minimization. For main-chain atoms, partial charges are assigned with the definitions of CHARMM PARAM19 parameter set [Momany, 1992 #179]. For side-chain atoms of polar residues, partial charges are gradually scaled up at the ranges from 0.05 to 0.4 times of their full charges, while the temperature is reduced from 500 to 300K. The final scale factor of 0.4 for polar side-chain partial charges is also applied during the 5 *ps* constant temperature dynamics at 300K and the final energy minimization.

A flow chart for model building by the SA/MD protocol is shown, (Fig. A.1).

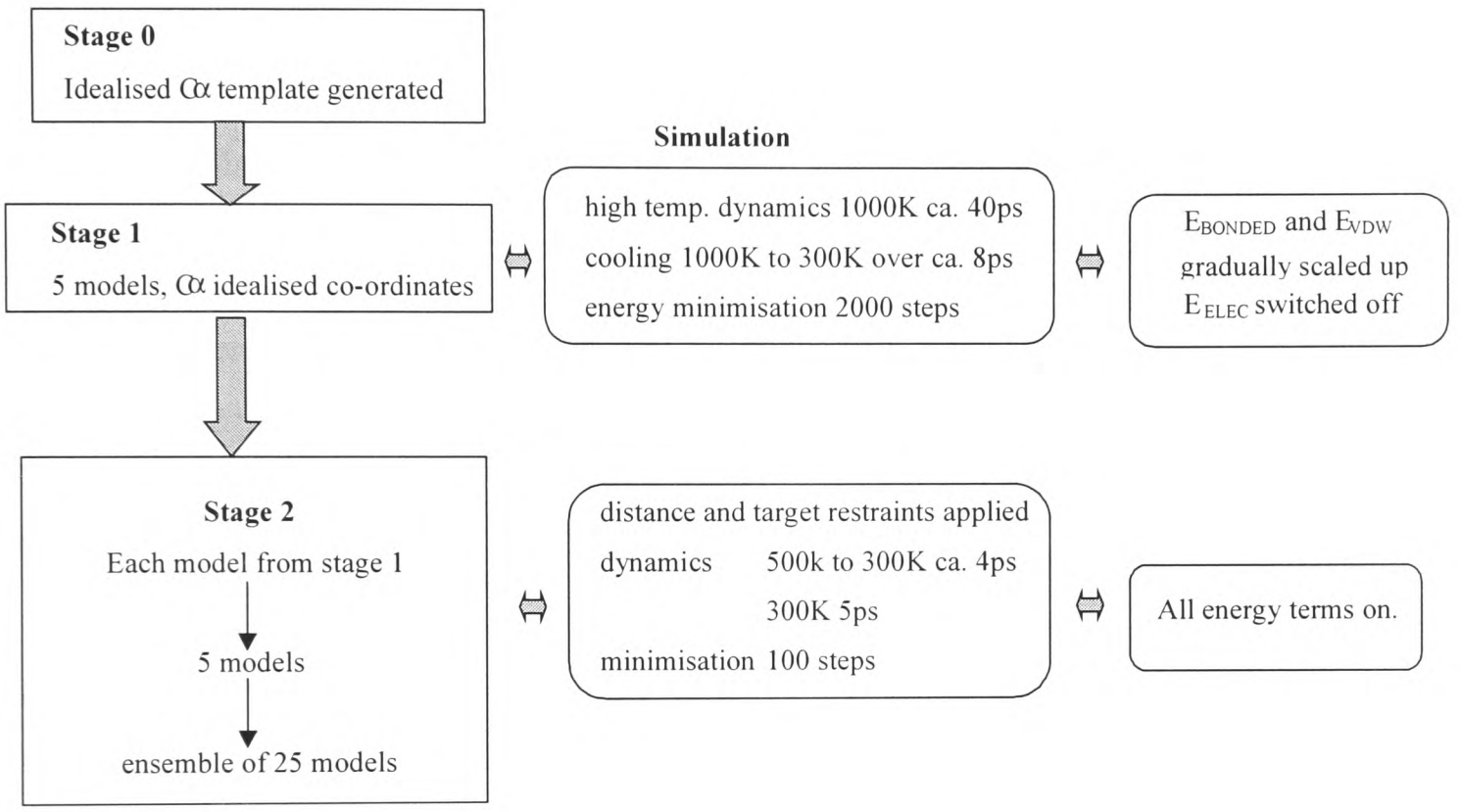


Figure A.1 Flow diagram summarising the restrained SA/MD protocol used for model building.

A.2 Publications Arising from Work in this Thesis.

A.2.1 Conference Abstracts.

Hong, R.-I, and Sansom, M. S. P. (1999) Sequence and Structure Studies of ABC Transporter, 2nd FEBS Advanced Lecture Course, Gosau, Austria.

Hong, R.-I, M. J. Adams. and Johnson, L. N. (1999) A Possible Role of Peptidyl-prolyl *cis/trans* Isomerisation in Protein Structures, Pacific Symposium on Biocomputing 99, Hawaii, U.S.A.

Hong, R.-I, Kerr, I. D. and Sansom, M. S. P. (1998) ABC Transporter: From Sequences to Structures, BBSRC Functional Genomics Workshop, Warwick, U.K.

A.2.2 Research Papers.

Hong, R.-I and Sansom, M. S. P. (1999) Sequence Analysis of ABC Transporter in Different Genomes. (In Preparation).

Hong, R.-I and Sansom, M. S. P. (1999) Molecular Dynamics Simulations of CFTR Transmembrane Helices in a POPC bilayer System. (In Preparation).

Hong, R.-I, Kerr, I. D. and Sansom, M. S. P. (In Press) Computational approaches to transporter Structure (invited chapter for Membrane Transporter - A Practical Approach ed S. A. Baldwin), IRL Press.(In Press)

References

Adams, P. D., I. T. Arkin, D. M. Engleman, and A. T. Brunger. 1995. Computational searching and mutagenesis suggest a structure for the pentameric transmembrane domain of phospholamban. *Nature Struct. Biol.* 2:154-162.

Akabas, M. H. 1998. Channel-lining residues in the M3 membrane-spanning segment of the cystic fibrosis transmembrane conductance regulator. *Biochemistry.* 37:12233-12240.

Aleksandrov, A. A., and J. R. Riordan. 1998. Regulation of CFTR ion channel gating by MgATP. *FEBS Lett.* 431:97-101.

Allen, J. P., G. Feher, T. O. Yeates, H. Komiya, and D. C. Rees. 1987. Structure of the reaction center from *Rhodobacter sphaeroides* R-26: the protein subunits. *Proc. Natl. Acad. Sci.* 84:6162-6166.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* 25:3389-3402.

Anderson, M. P., H. A. Berger, D. P. Rich, R. J. Gregory, A. E. Smith, and M. J. Welsh. 1991a. Nucleoside triphosphates are required to open the CFTR chloride channel. *Cell.* 67:775-784.

Anderson, M. P., R. J. Gregory, S. Thompson, D. W. Souza, S. Paul, R. C. Mulligan, A. E. Smith, and M. J. Welsh. 1991b. Demonstration that CFTR is a chloride channel by alteration of its anion selectivity. *Science.* 253:202-205.

Anfinsen, C. B. 1973. Principles that govern the folding of protein chains. *Science.* 181:223-230.

Arkin, I., and A. Brunger. 1998. Statistical analysis of predicted transmembrane a-helices. *Biochem. Biophys. Acta.* 1429:113-128.

Arkin, I. T., A. T. Brunger, and D. M. Engelman. 1997. Are there dominant membrane protein families with a given number of helices. *Prot: Struct. Func. Gen.* 28:465-466.

Audrezet, M. P., B. Mercier, H. Guillermit, I. Quere, C. Verlingue, G. Rault, and C. Ferec. 1993. Identification of 12 novel mutations in the CFTR gene. *Hum. Mol. Genet.* 2:51-54.

Bairoch, A., and R. Apweiler. 1998. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998. *Nucl. Acids Res.* 26:38-42.

Baker, E. N., and R. E. Hubbard. 1984. Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* 44:97-179.

Bakos, E., T. Hegedus, Z. Hollo, E. Welker, G. E. Tusnady, G. J. R. Zaman, M. J. Flens, A. Varadi, and B. Sarkadi. 1996. Membrane topology and glycoslation o fthe human multidrug resistance-associated protein. *J. Biol. Chem.* 271:12322-12326.

Barker, W. C., J. S. Garavelli, D. H. Haft, L. T. Hunt, C. R. Marzec, B. C. Orcutt, G. Y. Srinivasarao, L.-S. L. Yeh, R. S. Ledley, H.-W. Mewes, F. Pfeiffer, and A. Tsugita. 1998. The PIR-International Protein Sequence Database. *Nucl. Acids Res.* 26:27-32.

Barton, G. J. 1995. Protein secondary structure prediction. *Curr. Opin. Struct. Biol.* 5:372-376.

Barton, G. J., and M. J. E. Sternberg. 1987. A strategy for the rapid multiple alignment of protein sequences: confidence levels from tertiary structure comparisons. *J. Mol. Biol.* 198:327-337.

Bash, P. A., M. J. Field, R. C. Davenport, G. A. Petsko, D. Ringe, and M. Karplus. 1991. Computer simulation and analysis of the reaction pathwa of triosephosphate isomerase. *Biochemistry.* 30:5826-5832.

Baum, B. R. 1998. PHYLIP: phylogeny inference package. *Quarterly Rev. of Biol.* 64:539-541.

Bax, A., and L. Lerner. 1986. Two-dimensional nuclear magnetic resonance spectroscopy. *Science.* 232:960-967.

Beamer, L. J., D. Fischer, and D. Eisenberg. 1998. Detecting distant relatives of mammalian LPS-binding and lipid transporter proteins. *Prot. Sci.* 7:1643-1646.

Bear, C. E., C. Li, N. Kartner, R. J. Bridges, T. J. Jensen, M. Ramjeesingh, and J. R. Riordan. 1992. Purification and functional reconstitution of the cystic fibrosis transmembrane conductance regulator (CFTR). *Cell.* 68:809-818.

Belohorcova, K., J. H. Davis, T. B. Woolf, and B. Roux. 1997. Structure and dynamics of an amphilic peptide in a lipid bilayer: a molecular dynamics study. *Biophys. J.* 73:3039-3055.

Belrhali, H., P. Nollert, A. Royant, C. Menzel, J. P. Rosenbusch, E. M. Landau and E. Pebay-Peyroula. 1999. Protein, lipid and water organization in bacteriorhodopsin crystals: a molecular view of the purple membrane at 1.9 Å resolution. *Structure* 7:909-917.

Benson, D. A., M. S. Boguski, D. J. Lipman, J. Ostell, and B. F. F. Ouellette. 1998. GenBank. *Nucl. Acids Res.* 26:1-7.

Berendsen, H. J. C. 1998. A glimpse of the Holy Grail. *Science.* 282:642-634.

Berendsen, H. J. C., J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684-3690.

Berendsen, H. J. C., D. van der Spoel, and R. van Drunen. 1995. GROMACS:A message-passing parallel molecular dynamics implementation. *Comp. Phys. Comm.* 95:43-56.

Berger, O., O. Edholm, and F. Jahnig. 1997. Molecular dynamics simulations of a fluid bilayer of dipalmitoylphosphatidycholine at full hydration, constant pressure and constant temperature. *Biophys. J.* 72:2002-2013.

Bernstein, F. C., T. F. Williams, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. 1977. Protein data bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* 112:535-542.

Biggin, P. C., and M. S. P. Sansom. 1999. Interaction of a-helices with lipid bilayers: a review of simulation studies. *Biophys. Chem.* 76:161-183.

Blake, C. C. F., D. F. Koenig, G. A. Mair, A. C. T. North, D. C. Phillips, and V. R. Sarma. 1965. Structure of hen egg-white lysozyme. *Nature.* 206:757-761.

Blattner, F. R., et. al. 1997. The complete genome sequence of *Escherichai coli* K-12. *Science.* 277:1453-1462.

Bleasby, A. J., D. Akrigg, and T. K. Attwood. 1994. OWL - A non-redundant, composite protein sequence database. *Nucl. Acids Res.* 22:3574-3577.

Bode, W., P. Schwager, and R. Huber. 1978. The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. The refined crystal structures of the bovine typsinogen-pancreatic trypsin inhibitor complex and of its ternary complex with Ile-Val at 1.9 Å resolution. *J. Mol. Biol.* 118:99-112.

Boobbyer, D. N. A., P. J. Goodford, P. M. McWhinnie, and R. C. Wade. 1989. New hydrogen-bond potentials for use in determining energetically favorable binding sites in molecules of known structure. *J. Med. Chem.* 32:1083-1094.

Bordo, D., and P. Argos. 1994. The role of side-chain hydrogen bonds in the formation and stabilization of secondary structure in soluble proteins. *J. Mol. Biol.* 243:504-519.

Bork, P., A. K. Downing, B. Keiffer, and I. D. Campbell. 1996. Structure and distribution of modules in extracellular proteins. *Quart. Rev. Biophys.* 29:119-167.

Bowie, J. U. 1997. Helix packing in membrane proteins. *J. Mol. Biol.* 272:780-789.

Boyd, D., C. Schierle, and J. Beckwith. 1998. How many membrane proteins are there? *Prot. Sci.* 7:201-205.

Brandl, C. J., and C. M. Deber. 1986. Hypothesis about the function of membrane-buried proline residues in transporter proteins. *Proc. Natl. Acad. Sci. USA.* 83:917-921.

Brandts, J. F., H. R. Halvorson, and M. Brennan. 1975. Consideration of the Possibility that the slow step in protein denaturation reactions is due to *cis-trans* isomerism of proline residues. *Biochemistry.* 14:4953-4963.

Brandts, U., and B. Trumpower. 1994. The protonmotive Q-cycle in mitochondria and bacteria. *Crit. Rev. Biochem. Mol. Biol.* 29:165-197.

Brent, R., and R. L. Finley Jr. 1997. Understanding gene and allele function with two-hybrid methods. *Annu. Rev. Genet.* 31:663-704.

Brillantes, A.-M. B., K. Ondrias, A. Scott, E. Kobrinsky, E. Ondriasova, M. C. Moschella, T. Jayaraman, M. Landers, B. E. Ehrlich, and A. R. Marks. 1994. Stabilization of calcium release channel (ryanodine receptor) function by FK506-binding protein. *Cell.* 77:513-523.

Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. 1983. CHARMM: A program for macromolecular energy, minimisation, and dynamics calculations. *J. Comp. Chem.* 4:187-217.

Brunger, A. T. 1992. X-PLOR Version 3.1. a system for X-ray crystallography and NMR. Yale University Press, New Haven, Ct.

Brunger, A. T., and M. Karplus. 1991. Molecular dynamics simulations with experimental restraints. *Acc. Chem. Res.* 24:54-61.

Brünger, A. T., J. Kuriyan, and M. Karplus. 1987. Crystallographic R factor refinement by molecular dynamics. *Science.* 235:458-460.

Bult, C. J., et. al. 1996. Complete genome sequence of the Methanogenic archaeon, *Methanococcus jannaschii. Science.* 273:1058-1073.

Campbell, I. D. 1996. Structural biology, yesterday, today and tomorrow. *In* Foundations of Modern Biochemistry. JAI Press, Inc. 133-151.

Cantiello, H. F., G. R. Jackson Jr., C. F. Grosman, A. G. Prat, S. C. Borkan, Y. Wang, I. L. Reisin, C. R. O'Riordan, and D. A. Ausiello. 1998. Electrodiffusional ATP movement through the cystic fibrosis transmembrane conductance regulator. *Am. J. Physiol.* 274:C799-C809.

Chance, B., et. al. 1994. Synchrotron radiation in the biosciences. Oxford University Press, New York, USA.

Chang, X.-B., Y.-X. Hou, T. J. Jensen, and J. R. Riordan. 1994. Mapping of cystic fibrosis transmembrane conductance regulator membrane topology by glycosylation site insertion. *J. Biol. Chem.* 269:18572-18575.

Cheng, S. H., D. P. Rich, J. Marshall, R. J. Gregory, M. J. Welsh, and A. E. Smith. 1991. Phosphorylation of the R domain by cAMP-dependent protein kinase regulates the CFTR chloride channel. *Cell.* 66:1027-1036.

Chillon, M., T. Casals, V. Nunes, J. Gimenez, E. Perez Ruiz, and X. Estivill. 1993. Identification of a new missense mutation (P205S) in the first transmembrane domain of the CFTR gene associated with a mild cystic fibrosis phenotype. *Hum. Mol. Genet.* 2:1741-1742.

Chou, P. Y., and G. D. Fasman. 1974. Prediction of protein conformation. *Biochemistry.* 13:222-245.

Cole, S. T., et. al. 1998. Deciphering the biology of the *Mycobacterium tuberculosis* from the complete genome sequence. *Nature.* 393:537-544.

Collins, F. S. 1992. Cystic fibrosis: molecular biology and therapeutic implications. *Science.* 256:774-782.

Colloc'h, N., C. Etchebest, E. Thoreau, B. Henrissat, and J.-P. Mornon. 1993. Comparision of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignmnet. *Prot. Eng.* 6:377-382.

The *C. elegans* Consortium. 1998. Genome sequence of the Nematode *C. elegans*: a platform for investigating biology. *Science.* 282:2012-2018.

Cornette, J. L., K. B. Cease, H. Margalit, J. L. Spouge, J. A. Berzovsky, and C. DeLisi. 1987. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* 195:659-685.

Cotten, J. F., and M. J. Welsh. 1999. Cystic fibrosis-assoiated mutation at argine 347 alter the pore architecture of CFTR. *J. Biol. Chem.* 274:5429-5435.

Cowan, S. W., T. Schirmer, G. Rummel, M. Steiert, R. Ghosh, R. A. Apuptit, J. N. Jansonius, and J. P. Rosenbusch. 1992. Crystal structures explain functional properties of two *E. coli* porins. *Nature.* 358:727-733.

Cremonesi, L., M. Ferrari, E. Belloni, C. Magnanai, M. Seia, P. Ronchetto, M. Rady, M. P. Russo, G. Romeo, and M. Devoto. 1992. *Hum. Mutat.* 1:314-319.

Cserzo, M., E. Wallin, I. Simon, G. von Heijne, and A. Elofsson. 1997. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface methods. *Prot. Engng.* 10:673-676.

Daune, M. 1999. Molecular biophysics: structures in motion. Oxford University Press, Oxford.

Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. 1978. A modle of evolutionary change in proteins. *In* Atlas of protein sequence and structure. National Biomedical Research Foundation, Washington, DC. 345-352.

Dean, M., M. B. White, J. Amos, B. Gerrard, C. Stewart, K. T. Khaw, and M. Leppert. 1990. Multiple mutations in highly conserved residues are found in mildly affected cystic fibrosis patients. *Cell.* 61:863-870.

Deckert, G., et. al. 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature.* 392:353-358.

Decottignies, A., and A. Goffeau. 1997. Complete inventory of the yeast ABC proteins. *Nature Gent.* 15:137-145.

Deisenhofer, J., O. Epp, I. Sinning, and H. Michel. 1995. Crystallographic refinement at 2.3 Å resolution and refined model of the photosynthetic reaction centre from *Rhodopseudomonas viridis*. *J. Mol. Biol.* 246:429-457.

Derreumaux, P., and T. Schlick. 1998. The loop opening/closing motion of the enzyme triosephosphate isomerase. *Biophys. J.* 74:72-81.

Devidas, S., H. Yue, and W. B. Guggino. 1998. The second half of the cystic fibrosis transmembrane conductance regulator forms a functional chloride channel. *J. Biol. Chem.* 273:29373-29380.

Doniach, C. W. a. S. 1989. A computer model to dynamically simulate protein folding: Studies with crambin. *Proteins: Struct., Funct., Genet.* 6:193-209.

Donnelly, D., M. S. Johnson, T. L. Blundell, and J. Saunders. 1989. An analysis of the periodicity of conserved residues in sequence alignments of G-protein coupled receptors. *FEBS Lett.* 25:109-116.

Donnelly, D., J. P. Overington, S. V. Ruffle, J. H. A. Nugent, and T. L. Blundell. 1993. Modelling a-helical transmembrane domains: the calculation and use of substitution tables for lipid-facing residues. *Protein Science.* 2:55-70.

Doyle, D. A., J. M. Cabral, R. A. Pfuetzner, A. Kuo, J. M. Gulbis, S. L. Cohen, B. T. Chait, and R. Mackinnon. 1998. The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science.* 280:69-77.

Drenth, J. 1994. Principles of protein X-ray crystallography. Springer-Verlag, New York.

Dreusicke, D., P. A. Karplus, and G. E. Schulz. 1988. Refined structure if porcine cytosolic adenylate kinase at 2.1 Å resolution. *J. Mol. Biol.* 199:359-371.

Duan, Y., and P. A. Kollman. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science.* 282:740-744.

Dunbrack Jr, R. L., and M. Karplus. 1993. Backbone-dependent rotamer library for proteins: application to side-chain prediction. *J. Mol. Biol.* 230:543-574.

Dunker, A. K. 1982. A proton motive force transducer and its role in proton pump, proton engines, tobacco mosaic virus assembly and hemoglobin allosterism. *J. Theor. Biol.* 97:95-127.

Eddy, S. R. 1995. Multiple alignment using Hidden Markov Models. *In* Proc. Third Int. Conf. Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park. 114-120.

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

Field, M., and C. E. Semard. 1993. Toxigenic diarrheas, congenital diarrheas, and cystic fibrosis: disorders of intestinal ion transport. *Annu. Rev. Physiol.* 55:631-655.

Fischer, D., and D. Eisenberg. 1997. Assigning folds to the proteins encoded by the genome of *Mycoplasma genitalium*. *Proc. Natl. Acad. Sci. USA.* 94:11929-11934.

Fischer, G., and F. X. Schmid. 1990. The mechanism of protein folding. Implication of *in vitro* refolding model for *de novo* protein folding and translocation in the cell. *Biochemistry.* 29:2205-2213.

Fischer, G., B. Wittmann-Liebold, K. Lang, T. Ketfhaber, and F. X. Schmid. 1989. Cyclophilin and peptidyl-prolyl *cis-trans* isomerase are probably identical proteins. *Nature.* 337:476-478.

Fischer, W., L. R. Forrest, G. R. Smith, and M. S. P. Sansom. 1999. Single transmembrane helices of viral ion channels in a lipid/water environment - a molecular dynamic simulation study. *Proc. Natl. Acad. Sci. USA.* submitted.

Fleischmann, R. D., et. al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science.* 269:496-511.

Forrest, L. R., D. P. Tieleman, and M. S. P. Sansom. 1999. Defining the transmembrane helix of M2 protein from influenza A by molecular dynamics simulations in a lipid bilayer. *Biophys. J.* 76:1886-1896.

Forstner, M., M. Kriechbaum, P. Laggner, and T. Wallimann. 1998a. Structure changes of creatine kinase upon substrate binding. *Biophys. J.* 75:1016-1023.

Forstner, M., A. Muller, D. Rognan, M. Kriechbaum, and T. Wallimann. 1998b. Mutation of *cis*-proline 207 in mitochondrial creatine kinase to alanine leads to increased acid stability. *Prot. Eng.* 11:563-568.

Fraser, C. M., et. al. 1995. The minimal gene complement of *Mycoplasma genitalium*. *Science.* 270:397-403.

Fraser, C. M., et. al. 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature.* 390:580-584.

Fraser, C. M., et. al. 1998. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science.* 281:375-388.

Fraternali, F., and W. F. van Gunsteren. 1996. An efficient mean solvation force model for use in molecular dynamics simulations of proteins in aqueous solution. *J. Mol. Biol.* 256:939-948.

Freedman, N. J. 1996. Desensitization of G protein-coupled receptors. *Recent progr. in Hormone Res.* 51:319-353.

Freiberg, C., et. al. 1997. Molecular basis of symbiosis between *Rhizobium* and legmes. *Nature.* 387:394-401.

Freskgard, P.-O., N. Bergenhem, B.-H. Jonsson, M. Svensson, and U. Carlsson. 1992. Isomerase and chaperone activity of prolyl isomerase in the refolding of carbonic anhydrase. *Science.* 258:466-468.

Frishman, D., and P. Argos. 1995. Knowledge-based protein secondary structure assignmnet. *Proteins: Struct., Funct., Genet.* 23:566-579.

Frishman, D., and P. Argos. 1997a. 75% accuracy in protein secondary structure prediction. *Prot: Struct. Func. Gen.* 27:329-335.

Frishman, D., and P. Argos. 1997b. The future of the secondary structure prediction accuracy. *Folding & Design.* 2:159-162.

Fritz-Wolf, K., T. Schnyder, T. Wallimann, and W. Kabsch. 1996. Structure of mitochondrial creatine kinase. *Nature.* 381:341-345.

Gadsby, D. C., G. Nagel, and T.-C. Hwang. 1995. The CFTR chloride channel of mammalian heart. *Annu. Rev. Physiol.* 57:387-416.

Garnier, J., D. J. Osguthorpe, and B. Robson. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* 120:97-120.

Gerstein, M. 1997. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.* 274:562-576.

Gerstein, M. 1998. Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins: Struc., Funct., Genet.* 33:518-534.

Gerstein, M., and W. Krebs. 1998. A database of molecular motions. *Nucl. Acid. Res.* 26:4280-4290.

Gerstein, M., and M. Levitt. 1997. A structural census of the current population of protein sequences. *Proc. Natl. Acad. Sci. USA.* 94:11911-11916.

Gerstein, M., G. Schulz, and C. Chothia. 1993. Domain closure in adenylate kinase - joints of either side of 2 helices close like neighboring fingers. *J. Mol. Biol.* 229:494-501.

Golla, A., A. Deufel, C. Aulehla-Scholz, I. Bohm, B. Hilz, T. Meitinger, and T. Deufel. 1994. Identification of a novel missense mutation (G341E) in exon 7 of the cystic fibrosis transmembrane conductance regulator gene identified in a CF patient with pancreatic sufficiency. *Hum. Mutat.* 3:67-68.

Goodfellow, J. M., and D. S. Moss. 1992. Computer modelling of biomolecular processes. Ellis Horwood Limited.

Grandori, R. 1998. Systematic fold recognition analysis the sequence encoded by the genome of *Mycoplasma pneumoniae. Prot. Eng.* 11:1129-1135.

Grathwohl, C., and K. Wuthrich. 1981. NMR studies of the rate if proline *cis/trans* isomerisation in oligopeptides. *Biopolymers.* 20:2623-2633.

Greer, J. 1976. Automated interpretation of electron density maps of proteins: derivation of atomic co-ordinates for the main chain. *J. Mol. Biol.* 100:427-458.

Hediger, M. A. 1994. Structure, function and evolution of solute transporters in prokaryotes and eukaryotes. *J. Exp. Biol.* 196:15-49.

Hegyi, H., and M. Gerstein. 1999. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* 228:147-164.

Helekar, S. A., D. Char, S. Neff, and J. Patrick. 1994. Prolyl isomerase requirement for the expression of functional homo-oligomeric ligand-gated ion channels. *Neuron.* 12:179-189.

Henikoff, S., and J. G. Henikoff. 1992. Amino acid subsititution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA.* 89:10915-10919.

Hermans, J., H. J. C. Berendsen, W. F. van Gunsteren, and J. P. M. Postma. 1984. A consistent empirical potential for water-protein interactions. *Bioploymers.* 23:1513-1518.

Hess, B., H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. 1997. LINCS: A linear constraint solver for molecular simulations. *J. Comp. Chem.* 18:1463-1472.

Heymann, J. B., D. J. Muler, K. Mitsuaoka, and A. Engel. 1997. Electron and atomic force microscopy of membrane proteins. *Curr. Opin. Struct. Biol.* 7:543-549.

Higgins, C. F. 1995. The ABC of channel regulation. *Cell.* 82:693-696.

Higgins, C. F., I. D. Hiles, G. P. C. Salmomd, D. R. Gill, J. A. Downie, I. J. Evans, I. B. Holland, L. Gray, S. O. Buckel, A. W. Bell, and M. A. Hermodson. 1986. A family of related ATP-

binding subunits coupled to many distinct biological processes in bacteria. *Nature.* 51:448-450.

Hilbert, H., et. al. 1996. Sequence analysis of 56 kb from the genome of the bacterium Mycoplasma pneumoniae comprising the dnaA region, the atp operon and a cluster of ribosomal protein genes. *Nucl. Acids. Res.* 24:628-639.

Hille, B. 1992. Ionic channels of excitable membranes. Sunderland.

Hirokawa, T., S. Boon-Chieng, and S. Mitaku. 1998. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics.* 14:378-379.

Ho, K. 1998. The ROMK-cystic fibrosis transmembrane conductance regulator connection: new insights into the relationship between ROMK and cystic fibrosis transmembrane conductance regulator channels. *Curr. Opin. Nephrol. Hypertens.* 7:49-58.

Hodges, P. E., A. H. Z. McKee, B. P. Davis, W. E. Payne, and J. I. Garrels. 1999. Yeast Protein Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucl. Acids Res.* 27:69-73.

Holm, L., and C. Sander. 1993. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 233:123-138.

Holm, L., and C. Sander. 1996. Mapping the protein universe. *Science.* 273:595-602.

Hong, R.-I., and M. S. P. Sansom. 1999. Computer prediction of transporter topology and structure. *In* Membrane transporter. IRL Press, Oxford, England.

Huber, R., and W. S. Bennett Jr. 1983. Functional significance of flexibility in proteins. *Bioploymers.* 22:261-279.

Humphery-Smith, I., S. J. Cordwell, and W. P. Blackstock. 1997. Proteome research: complementarity and limitations with respect to the RNA and DNA worlds. *Electophoresis.* 18:1217-1242.

Hung, L.-W., I. X. Wang, K. Kikaido, P.-Q. Liu, G. F.-L. Ames, and S.-H. Kim. 1998. Crystal structure of the ATP-binding subunit of an ABC transporter. *Nature.* 396:703-707.

Huynen, M., T. Doerks, F. Eisenhaber, C. Orengo, S. Sunyaev, Y. Yuan, and P. Bork. 1998. Homology-based fold predictions for *Mycoplasma genitalium* proteins. *J. Mol. Biol.* 280:323-326.

Ikura, M., G. M. Clore, A. M. Gronenborn, G. Zhu, C. B. Klee, and A. Bax. 1992. Solution structure of a calmodulin-target peptide complex by multidimensional NMR. *Science.* 256:632-638.

Iwata, S., J. W. Lee, K. Okada, J. K. Lee, M. Iwata, B. Rasmussen, T. A. Link, S. Ramaswamy, and B. K. Jap. 1998. Complete structure of the 11-subunit bovine mitochondrial cytochrome bc1 complex. *Science.* 281:64-71.

Iwata, S., C. Ostermeier, B. Ludwig, and H. Michel. 1995. Structure at 2.8 Å resolution of cytochrome c oxidase from *Paracoccus denitrificans*. *Nature.* 376:660-669.

Jacob, J., H. Duclohier, and D. S. Cafiso. 1999. The role of proline and glycine in determining the backbone flexibility of a channel-forming peptide. *J. Biophys.* 76:1367-1376.

Jayaraman, T., A.-M. B. Brillantes, A. P. Timerman, H. Erdjument-Bromage, S. Fleischer, P. Tempst, and A. R. Marks. 1992. FK506-binding protein associated with the calcium release channel (ryanodine receptor). *J. Biol. Chem.* 267:9474-9477.

Jiang, Q., D. Mak, S. Devidas, E. M. Schwiebert, A. Bragin, Y. Zhang, W. R. Skach, W. B. Guggino, J. K. Foskett, and J. F. Engelhardt. 1998. Cystic fibrosis transmembrane conductance regulator-associated ATP release is controlled by a chloride sensor. .

Jones, D. T. 1998. Do transmembrane protein superfolds exist? *FEBS Letters.* 423:281-285.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochem.* 33:3038-304.

Joseph-McCarthy, D., D. Lolis, E. A. Komives, and G. A. Petsko. 1994. Crystal structure of the K12M/G15A triosephosphate isomerase double mutant and electrostatic analysis of the active site. *Biochemistry.* 33:2815-2823.

Kabsch, W., and C. Sander. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features *Biopolymers.* 22:2577-2637.

Kaneko, T., et. al. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis sp.* strain PCC6803. II. sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* 3:109-136.

Karpen, M. E., and C. L. Brooks III. 1996. Modelling protein conformation by molecular mechanics and dynamics. *In* Protein structure prediction. M. J. E. Sternberg, editor. IRL press, Oxford. 1-30.

Karplus, M. 1997. The Levinthal paradox: yesterday and today. *Folding and Design.* 2:569-575.

Karplus, M., and G. A. Petsko. 1990. Molecular dynamics simulations in biology. *Nature.* 347:631-639.

Kawarabayasi, Y., et. al. 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Res.* 5:55-76.

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

Kendrew, J. C., R. E. Dickerson, B. E. Strandberg, R. G. Hart, D. R. Davies, D. C. Phillips, and V. C. Shore. 1960. Structure of myoglobin, a three-dimensional Fourier synthesis at 2 Å resolution. *Nature.* 185:422-427.

Kerr, I. D., R. Sankararamakrishnan, and M. S. P. Sansom. 1994a. Simplified models of the pore domain of the nicotinic acetylcholine receptor. *Biochem. Soc. Trans.* 22:158S.

Kerr, I. D., R. Sankararamakrishnan, O. S. Smart, and M. S. P. Sansom. 1994b. Parallel helix bundles and ion channels: molecular modelling *via* simulated annealing and restrained molecular dynamics. *Biophys. J.* 67:1501-1515.

Kerr, I. D., H. S. Son, R. Sankararamakrishnan, and M. S. P. Sansom. 1996. Molecular dynamics simulations of isolated transmembrane helices of potassium channels. *Biopolymers.* 39:503-515.

King, R. D., and M. J. E. Sternberg. 1996. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Prot. Sci.* 5:2298-2310.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science.* 220:671-680.

Klenk, H.-P., et. al. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature.* 390:364-370.

Knighton, D. R., J. Zheng, L. F. Teneyck, V. A. Ashford, N. Xuong, S. S. Taylor, and J. M. Sowadski. 1991. Structure of a peptide inhibitor bound to the catalytic subunit of cyclic adenosine-monophosphate dependent protein-kinase. *Science.* 253:414-420.

Komiya, H., T. O. Yeates, D. C. Rees, J. P. Allen, and G. Feher. 1987. Structure of the reaction centre from *Rhodobacter sphaeroides* R- 26 and 2.4.1: Symmetry relations and sequence comparisons. *Proc. Natl. Acad. Sci. USA.* 85:9012-9016.

Krauss, N., W. Hinrichs, I. Witt, P. Fromme, W. Pritzkow, Z. Dauter, C. Betzel, K. S. Wilson, H. T. Witt, and W. Saenger. 1993. Three dimensional structure of system 1 of photosynthesis at 6 Å resolution. *Nature.* 361:326-331.

Kreusch, A., A. Neubueser, W. Schiltz, J. Wekesser, and G. E. Schulz. 1994. The structure of the membrane channel porin from Rhodopseudomonas blastica at 2.0 Å resolution. *Prot. Sci.* 3:58-63.

Kuhlbrandt, W., D. N. Wang, and Y. Fujiyoshi. 1994. Atomic model of plant light harvesting complex by electron crystallography. *Nature.* 367:614-621.

Kukol, A., P. D. Adams, L. M. Rice, A. T. Brunger, and I. T. Arkin. 1999. Experimentally based orientational refinement of membrane protein models: a structure for the Influenza A M2 H+ channel. *J. Mol. Biol.* 286:951-962.

Kunst, F., et. al. 1997. The complete genome sequence of the Gram-positive bacterium *Aquifex aeolicus*. *Nature.* 390:249-256.

Kunzelmann, K., G. L. Kiser, R. Schreiber, and J. R. Riordan. 1997. Inhibition of epithelial Na+ currents by intracellular domains of the cystic fibrosis transmembrane conductance regulator. *FEBS Lett.* 400:341-344.

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

Labesse, G., N. Colloc'h, J. Pothier, and J.-P. Mornon. 1997. P-SEA: a new efficient assignment of secondary structure from Ca trace of proteins. *Comput. Appl. Biosci.* 13:291-295.

Laiter, S., D. L. Hoffman, R. K. Singh, I. I. Vaisman, and A. Tropsha. 1995. Pseudotorsional OCCO backbone angle as a single descriptor of protein secondary structure. *Prot. Sci.* 6:1633-1643.

Landolt-Marticorena, C., K. A. Williams, C. M. Deber, and A. F. Reithmeir. 1993. Non-random distribution of amino acids in the transmembrane segments of human type I single span membrane proteins. *J. Mol. Biol.* 229:602-608.

Lang, K., F. X. Schmid, and G. Fischer. 1987. Catalysis of protein folding by prolyl isomerase. *Nature.* 329:268-270.

Levitt, M., and J. Greer. 1977. Automatic identification of secondary structure in globular proteins. *J. Mol. Biol.* 114:181-293.

Li, C., M. Ramjeesingh, and C. E. Bear. 1996. Purified cystic fibrosis transmembrane conductance regulator (CFTR) does not function as an ATP channel. *J. Biol. Chem.* 271:11623-11626.

Lin, L. N., H. Hasumi, and J. F. Brandts. 1988. Catalysis of proline isomerization during protein-folding reactions. *Biochim. Biophys. Acta.* 956:256-266.

Linton, K. J., and C. F. Higgins. 1998. The *Escherichia coli* ATP-binding cassette (ABC) proteins. *Mol. Microbiol.* 28:5-13.

Lipman, D. J., S. F. Altschul, and J. D. Kececioglu. 1989. A tool for multiple sequence alignment. *Proc. Natl. Acad. Sci. USA.* 86:4412-4415.

Loo, M. A., T. J. Jensen, L. Cui, Y.-X. hou, X.-B. Chang, and J. R. Riordan. 1998. Perturbation of Hsp90 interaction with nascent CFTR prevent maturation and accelerates its degradation by the proteasome. *EMBO J.* 17:6879-6887.

Mansoura, M. K., S. S. Smith, A. D. Choi, N. W. Richards, T. V. Strong, M. L. Drumm, F. S. Collins, and D. C. Dawson. 1998. Cystic fibrosis transmembrane conductance regulator (CFTR) anion binding as a probe of the pore. *Biophys. J.* 74:1320-1332.

Marrink, S., D. P. Tieleman, A. R. van Buuren, and H. J. C. Berendsen. 1996. Membranes and water: an interesting relationship. *Faraday Discuss.* 103:191-201.

Marrink, S. J., O. Berger, D. P. Tieleman, and F. Jahnig. 1998. Adhesion forces of lipids in a phospholipid membrane studied by molecular dynamics simulations. *Biophys. J.* 74:931-943.

Martonosi, A. N. 1998. No crystals, no grant, revisited. *Nature Struct. Biol.* 5:11-12.

McCammon, J. A., and S. C. Harvey. 1987. Dynamics of proteins and nucleic acids. Cambridge University Press, Cambridge.

McDermott, G., S. M. Prince, A. A. Freer, A. M. Hawthornthwaite-Lawless, M. Z. Papiz, R. J. Codgell, and N. W. Isaacs. 1995. Crystal structure of an integral membrane light-harvesting complex from photosynthetic bacteria. *Nature.* 374:517-521.

McNicholas, C. M., W. B. Guggino, E. M. Schwiebert, S. C. Hebert, G. Giebisch, and M. E. Egan. 1996. Sensitivity of a renal K+ channel (ROMK2) to the inhibitory sulfonylurea compound glibenclamide is enhanced by coexpression with the ATP-binding cassette

transporter cystic fibrosis transmembrane regulator. *Proc. Natl. Acad. Sci. USA.* 23:8083-8088.

Meacham, G. C., Z. Lu, S. King, E. Sorscher, A. Tousson, and D. M.Cyr. 1999. The Hdj-2/Hsc70 chaperone pair facilitates early steps in CFTR biogenesis. *EMBO J.* 18:1492-1505.

Mewes, H. W., K. Albermann, M. Bahr, D. Frishman, A. Gleissner, J. Hani, K. Heumann, K. Kleine, A. Maierl, S. G. Oliver, F. Pfeiffer, and A. Zollner. 1997. Overview of the yeast genome. *Nature.* 387:7-8.

Mewes, H. W., J. Hani, F. Pfeiffer, and D. Frishman. 1998. MIPS: a database for protein sequences and complete genomes. *Nucl. Acid Res.* 26:33-37.

Mingarro, I., G. von Heijne, and P. Whitley. 1997. Membrane-protein engineering. *Trends in Biotech.* 15:432-437.

Mitsuoka, K., T. Hirai, K. Murata, A. Miyazawa, A. Kidera, Y. Kimura, and Y. Fujiyoshi. 1999. The structure of bacteriorhodopsin at 3.0 Å resolution base on electron crystallography: implication of charge distribution. *J. Mol. Biol.* 286:861-882.

Mizuguchi, K., and N. Go. 1995. Comparision of spatial arrangements of secondary structural elements in proteins. *Prot. Eng.* 8:353-362.

Mogi, T., L. J. Stern, B. H. Chao, and H. G. Khorana. 1989. Structure-function studies on bacteriorhodopsin VIII. substitutions of the membrane-embeded prolines 50, 91 and 186: the effects are determined by the substituting amino acids. *Proc. Natl. Acad. Sci. USA.* 264.

Momany, F. A., and R. Rone. 1992. Validation of the general purpose QUANTA 3.2/CHARMm Forcefield. *J. Comp. Chem.* 13:888-900.

Montelione, G. T., and S. Anderson. 1999. Structural genomics: keynote for a Human Proteome Project. *Nature Struct. Biol.* 6:11-12.

Morris, A. L., M. W. MacArthur, E. G. Huntchinson, and J. M. Thornton. 1992. Stereochemical quality of protein-structure coordinates. *Proteins: Struct., Funct., Genet.* 12:345-364.

Muggleton, S., R. D. King, and M. J. E. Sternberg. 1992. Protein secondary structure prediction using logic-based machine learning. *Prot. Eng.* 5:647-657.

Muller, C. W., and G. E. Schulz. 1992. Structure of the complex between adenylate kinase from *Escherichia coli* and the inhibitor ap5a refined at 1.9 Å resolution - a model for a catalytic transition-state. *J. Mol. Biol.* 224:159-177.

Murzin, A. G. 1996. Structural classification of proteins: new superfamilies. *Curr. Opin. Strct. Biol.* 6:386-394.

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

Nasr, S. Z., T. V. Strong, M. K. Mansoura, D. C. Dawson, and F. S. Collins. 1996. Novel missense mutation (G314R) in a cystic fibrosis patient with hepatic failure. *Hum. Muata.* 7:151-154.

Ng, K. K. S., K. Drickamer, and W. T. Weis. 1996. Structural analysis of monosaccharide recognition by rat liver mannose-binding protein. *J. Biol. Chem.* 271:663-674.

Nikaido, H., and J. A. Hall. 1998. Overview of bacterial ABC transporters. *Methods in Enzymology.* 292:3-20.

Noble, M. E. M., R. K. Wierenga, A.-M. Lambeir, R. R. Opperdoes, A.-M. W. H. Thunnissen, K. H. Kalk, H. Groendijk, and W. G. J. Hol. 1991. The adaptability of the active site of trypanosomal triosephosphate isomerase as observed in the crystal structures of three different complexes. *Proteins: Struct., Funct., Genet.* 10:50-69.

Notredame, C., and D. G. Higgins. 1996. SAGA: sequence alignment by genetic algorithm. *Nucl. Acids Res.* 24:1515-1524.

Oesterhelt, D., and W. Stoeckenius. 1973. Functions of a new photoreceptor membrane. *Proc. Natl. Acad. Sci. USA.* 70:2853-2857.

Opella, S. J. 1997. NMR and membrane proteins. *Nature Struc. Biol.* 4:845-848.

Orengo, C. A., N. P. Brown, and W. R. Taylor. 1992. Fast structure alignment for protein data bank search. *Prot: Strut. Func. Gent.* 14:139-177.

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

Pao, S. S., I. T. Paulsen, and M. H. Saier Jr. 1998. Major facilitator superfamily. *Microbiol. and Mol. Biol. Rev.* 62:1-34.

Pauling, L., and R. B. Corey. 1951. The pleated sheet, a new layer configuration of polypeptide chains. *Proc. Natl. Acad. Sci. USA.* 37:251-256.

Pauling, L., R. B. Corey, and H. R. Branson. 1951. The structure of proteins:two hydrogen-bonded helical configuration of polypeptide chian. *Proc. Natl. Acad. Sci. USA.* 37:205-210.

Paulsen, I. T., M. K. Sliwinski, and M. H. Saier Jr. 1998. Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. *J. Mol. Biol.* 277:573-592.

Pautsch, A., and G. E. Schulz. 1998. Structure of the outer membrane protein A transmembrane domain. *Nature Struct. Biol.* 5:1013-1017.

PDBmanual. 1996. Atomic coordinates entry format description version 2.0. *In* Protein Data Bank Contents Guide. Brookhaven National Laboratory, Protein Data Bank.

Pearson, W. R. 1996. Effective protein sequence comparison. *Meth. Enzymol.* 266:227-258.

Pearson, W. R. 1998. Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* 276:71-84.

Pearson, W. R., and D. J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA.* 276:71-84.

Pebay-Peyroula, E., G. Rummel, J. P. Rosenbusch, and E. M. Landau. 1997. X-ray structure of bacteriorhodopsin at 2.5 Å from microcrystals grown in lipidic cubic phases. *Science.* 277:1676-1681.

Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA.* 96:4285-4288.

Persson, B., and P. Argos. 1994. Prediction of transmembrane segments in proteins utilising multiple sequence alignments. *J. Mol. Biol.* 237:182-192.

Pertuz, M. 1992. Protein Structure. W. H. Freeman and Company, New York.

Phillips, D. C. 1967. The hen egg-white lysozyme molecule. *Proc. Natl. Acad. Sci. USA.* 57:484-495.

Phillips, D. C. 1970. British Biochemistry, Past and Present. Academic Press, London, UK.

Ponder, J. W., and F. M. Richards. 1987. Tertiary templates for proteins: using of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* 193:775-791.

Popot, J.-L. 1993. Integral membrane protein structure: transmembrane a helices as autonomous folding domains. *Curr. Opin. Strc. Biol.* 3:532-540.

Popot, J.-L., and D. M. Engelman. 1990. Membrane protein folding and oligomerization: the two-state model. *Biochem.* 29:4031-4037.

Preusch, P. C., J. C. Norvell, J. C. Cassatt, and M. Cassman. 1998. Progress away from 'no crystals, no grant'. *Nature Struct. Biol.* 5:12-14.

Quentin, Y., G. Fichant, and F. Denizot. 1999. Inventory, assembly and analysis of *Bacillus subtilis* ABC transport systems. *J. Mol. Biol.* 287:467-484.

Raghava, G. P. S., and G. J. Barton. 1998. Comparison and limitations of techniques for multiple alignment of protein sequences. *(submitted)*.

Ramachandran, G. N., and A. K. Mitra. 1976. An explanation for the rare occurence of *cis* peptide units in proteins and polypeptides. *J. Mol. Biol.* 107:85-92.

Ramachandran, G. N., G. Ramakrishnan, and V. Sasisekharan. 1963. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* 7:95-99.

Ramachandran, G. N., and V. Sasisekharan. 1968. Conformation of polypeptides and proteins. *Adv. Prot. Chem.* 28:283-437.

Reddy, M. M., P. M. Quinton, C. Haws, J. J. Wine, R. Grygorczyk, J. A. Tabcharani, J. W. Hanrahan, K. L. Gunderson, and R. R. Kopito. 1996. Failure of the cystic fibrosis transmembrane conductance regulator to conduct ATP. *Science.* 271:1876-1879.

Rees, D. C., L. DeAntonio, and D. Eisenberg. 1989. Hydrophobic organization of membrane proteins. *Science.* 245:510-513.

Reimer, U., G. Scherer, M. Drewello, S. Kruber, M. Schutkowski, and G. Fischer. 1998. Side-chain effects on peptidyl-prolyl *cis/trans* isomerisation. *J. Mol. Biol.* 279:449-460.

Ri, Y., J. A. Ballesteros, C. K. Abrams, S. Oh, V. K. Verselis, H. Weinstein, and T. A. Bargiello. 1999. The role of a conserved proline residue in mediating conformational changes associated with voltage gatting of Cx32 gap junctions. *J. Biophys.* 76:2887-2898.

Rich, D. P., H. A. Berger, S. H. Cheng, S. M. Travis, M. Saxena, A. E. Smith, and M. J. Welsh. 1993. Regulation of the cystic fibrosis transmembrane conductance regulator Cl- channel by negative charge in the R domain. *J. Biol. Chem.*,. 268:20259-20267.

Richards, F. M., and C. E. Kundrot. 1988. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins: Struct., Funct., Genet.* 3:71-84.

Richardson, J. S., and D. C. Richardson. 1989. Principles and patterns of protein conformation. *In* Prediction of protein structure and the principle of protein conformation. Plenum Press, New York, U.S.A.

Riordan, J. R. 1993. The cystic fibrosis transmembrane conductance regulator. *Annu. Rev. Physiol.* 55:609-630.

Riordan, J. R., J. M. Rommens, B. S. Kerem, N. Alon, R. Rozmahel, Z. Grzelczak, J. Zielenski, S. Lok, N. Plavsic, J.-L. Chou, M. T. Drumm, M. C. Iannuzzi, F. S. Collins, and L.-C. Tsui. 1989. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science.* 254:1066-1073.

Rosenberg, M. F., R. Callagham, R. C. Ford, and C. F. Higgins. 1997. Structure of the multidrug resistance P-glycoprotein to 2.5 nm resolution determined by electron microscopy and image analysis. *J. Biol. Chem.* 272:10685-10694.

Rost, B. 1998. Marrying structure and genomics. *Structure.* 6:259-263.

Rost, B., R. Casadio, P. Fariselli, and C. Sander. 1995. Prediction of helical transmembrane segments at 95% accuracy. *Prot. Sci.* 4:521-533.

Rost, B., P. Fariselli, and R. Casadio. 1996. Topology prediction for helical transmembrane proteins at 86% accuracy. *Prot. Sci.* 5:1704-1718.

Rost, B., and C. Sander. 1993. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl. Sci. Acad. USA.* 90:7558-7562.

Rowland, P., A. K. Basak, S. Gover, H. R. Levy, and M. J. Adams. 1994. The three-dimensional structure of glucose 6-phosphate dehydrogenase from *Leuconostoc mesenteroides* refined at 2.0 Å resolution. *Structure.* 2:1073-1087.

Russell, R. B., and G. J. Barton. 1992. Multiple sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins: Struc., Funct., Genet.* 14:309-323.

Russell, R. B., and G. J. Barton. 1993. The limits of protein secondary structure prediction accuracy from multiple sequence alignment. *J. Mol. Biol.* 234:951-957.

Saier Jr, M. H. 1994. Computer-aided analyses of transport protein sequences: gleaning evidence concerning function, structure, biogenesis and evolution. *Microbiol. Rev.* 58:71-93.

Saier Jr, M. H. 1998. Molecular phylogeny as a basis for the classfication of transporter proteins from bacteria, archaea and eukarya. *Adv. in Microbial Physiol.* 40:81-136.

Sali, A. 1998. 100,000 protein structures for the biologist. *Nature Struct. Biol.* 5:1029-1032.

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

Samatey, F. A., G. Zaccai, D. M. Engelman, C. Etchebest, and J. L. Popot. 1994. Rotational orientation of transmembrane α-helices in bacteriorhodopsin. A neutron diffraction study. *J. Mol. Biol.* 236:1093-1104.

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

Sanchez, R., and A. Sali. 1998. Large-scale protein structure modeling of *Saccharomyces cerevisiae* genome. *Prot. Natl. Acad. Sci. USA.* 95:13597-13602.

Sander, C., and R. Schneider. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Func. Genet.* 9:56-68.

Sansom, M. S. P. 1992. Proline residues in transmembrane helices of channel and transporter proteins: a molecular modelling study. *Prot. Eng.* 5:53-60.

Sansom, M. S. P. 1998. Peptides and lipid bilayers: dynamic interactions. *Curr. Opin. Coll. Intf. Sci.* 3:518-524.

Sansom, M. S. P., R. Sankararamakrishnan, and I. D. Kerr. 1995. Modelling membrane proteins using structural restraints: the closed conformation of the nicotinic acetylcholine receptor. *Nature Struct. Biol.* 2:624-631.

Sarkadi, G. E. T. a. E. B. a. A. V. a. B. 1997. Membrane topology distinguishes a subfamily of the ATP-binding cassette ABC transporters. *FEBS Letters.* 402:1-3.

Saurin, W., and E. Dassa. 1994. Sequence relationships between integral inner membrane-proteins of binding protein-dependent transport-systems - evolution by recurrent gene duplications. *Prot. Sci.* 3:325-344.

Saurin, W., M. Hofnung, and E. Dassa. 1999. Getting in or out: early segregation between importers and exporters in the evolution of ATP-binding cassette (ABC) transporters. *J. Mol. Evol.* 48:22-41.

Schertler, G. F. X. 1992. Overproduction of membrane proteins. *Curr. Opin. Struct. Biol.* 2:534-544.

Schiffer, M., C. H. Chang, and F. J. Stevens. 1992. The functions of tryptophan residues in membrane proteins. *Prot. Eng.* 5:213-214.

Schmid, F. X. 1993. Peptidyl prolyl isomerase: enzymatic catalysis of slow protein-folding reactions. *Annu. Rev. Biophys. Biomol. Struct.* 22:123-143.

Schmid, F. X., and R. L. Baldwin. 1978. Acid catalysis of the formation of the slow-folding species of RNase A: evidence that the reaction is proline isomerization. *Proc. Natl. Acad. Sci. USA.* 75:4764-4768.

Schultz, B. D., A. Takahashi, C. G. Liu, R. A. Frizzell, and M. Howard. 1997. FLAG epitope positioned in an external loop preserves normal biophysical properties of CFTR. *Am. J. Physiol.* 273:C2080-C2089.

Schultz, G. E., and R. H. Schirmer. 1979. Springer Advanced Texts in Chemistry. Springer-Verlag, New York, Heidelberg, Berlin.

Schwiebert, E. M., M. E. Egan, T. H. Hwang, S. B. Fulmer, S. A. Allen, G. R. Cutting, and W. B. Guggino. 1995. CFTR regulates outwardly rectifying chloride channels through an autocrine mechanism involving ATP. *Cell.* 81:1063-1073.

Schwiebert, E. M., M. M. Morales, S. Devidas, M. E. Egan, and W. B. Guggino. 1998. Chloride channel and chloride conductance regulator domains of CFTR, the cystic fibrosis transmembrane conductance regulator. *Proc. Natl. Acad. Sci. USA.* 95:2674-2679.

Shaiu, W.-L., T. Hu, and T.-S. Hsieh. 1999. The hydrophilic, protease-sensitive terminal domains of eucaryotic DNA topoisomerases have essential intracellular functions. *Proc. Pacific Symp. on Biocomputing*:578-589.

Sheng, X. R., H. J. Zhang, X. M. Pan, X. F. Li, and J. M. Zhou. 1997. Domain movement in rabbit muscle adenylate kinase might involve proline isomerization. *FEBS Letters.* 413:429-432.

Sheppard, D. N., S. M. Travis, H. Ishihara, and M. J. Welsh. 1996. Contribution of proline residues in the membrane-spanning domains of cystic fibrosis transmembrane conductance regulator to chloride channel function. *J. Biol. Chem.* 271:14995-15001.

Shrivastava, I., C. Capener, L. R. Forrest, and M. S. P. Sansom. 1999. Structure and dynamics of K^+ channel pore-forming helices:- simulation studies. *Biophys J (Submitted).*

Shrivastava, I., and M. S. P. Sansom. 1999. Simulating ion permeation through a K^+ channel. *(Submitted).*

Siddiqui, A. 1997. Computer analysis and classification of protein structural domains. University of Oxford, Oxford.

Siekierka, J. J., S. H. Y. Hung, M. Poe, C. S. Lin, and N. H. Sigal. 1989. A cytosolic binding protein for the immunosuppressant FK506 has peptidyl-prolyl isomerase activity. *Nature.* 341:755-757.

Sklenar, H., C. Etchebest, and R. Lavery. 1989. Describing protein-structure - a general algorithm yielding complete helicoidal parameters and s unique overall axis. *Proteins: Struct., Funct., Genet.* 6:46-60.

Skolnick, J., and A. Kolinski. 1990. Simulation of the folding of globular protein. *Science.* 250:1121-1125.

Smith, D. R. 1997. Complete genome sequence of Methanobacterium thermoautotrophicum: functional analysis and comparative genomics. *J. of Bact.* 179:7135-7155.

Smith, P. E., H. D. Blatt , and Pettitt, B. M. 1997. A simple two-dimensional representation for the common secondary structural elements of polypeptides and proteins. *Proteins: Struct., Funct., Genet.* 27:227-234.

Smith, S. O., and O. B. Peersen. 1992. Solid state NMR approaches to study membrane protein structures. *Ann. Rev. Biophys. Biomol. Struc.* 21:25-47.

Smith, T. 1999. A banner year for membranes. *Nature Struct. Biol.* 6:1-2.

Sonnhammer, E. L. L., S. R. Eddy, and R. Durbin. 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins: Struct. Func. Genet.* 28:405-420.

Sonnhammer, E. L. L., G. von Heijne, and A. Krogh. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. *In* Proc. Sixth Int. Conf. Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park. 175-182.

Srinivasan, B. N., and T. L. Blundell. 1993. An evaluation of the performance of an automated procedure for comparative modellingof protein tertiary structure. *Prot. Eng.* 6:501-512.

Srinivasan, N., K. Guruprasad, and T. L. Blundell. 1996. Comparative modelling of proteins. *In* Protein structure prediction. M. J. E. Sternberg, editor. IRL press, Oxford. 111-140.

Srinivasan, R. 1976. Helical length distribution from protein crystallographic data. *Ind. J. Biochem. Biophys.* 13:192-193.

Srinivasan, R., R. Balasubramanian, and S. S. Rajan. 1975. Some new methods and general results of analysis of protein crystallographic structural data. *J. Mol. Biol.* 98:739-747.

Stein, R. L. 1993. Mechanism of enzymatic and nonenzymatic prolyl *cis-trans* isomerization. *Adv. in Prot. Chem.* 44:1-24.

Sternberg, M. J. E. 1996. Protein structure prediction-principles and approaches. *In* Protein structure prediction. M. J. E. Sternberg, editor. IRL press, Oxford. 1-30.

Sternberg, M. J. E., and M. J. Thornton. 1977. On the conformation of proteins: an analysis of β-pleated sheets. *J. Mol. Biol.* 110:285-296.

Stewart, D. E., A. Sarkar, and J. E. Wampler. 1990. Occurence and role of *cis* peptide bonds in protein structures. *J. Mol. Biol.* 214:253-260.

Stoeckenius, W. 1999. Bacterial rhdopsins: evolution of a mechanistic model for the ion pumps. *Prot. Sci.* 8:447-459.

Stoesser, G., M. A. Moseley, J. Sleep, M. McGowran, M. Garcia-Pastor, and P. Sterk. 1998. The EMBL nucleotide sequence database. *Nucl. Acids Res.* 26:16-20.

Stoesser, G., P. Sterk, M. A. Tuli, P. J. Stoehr, and G. N. Cameron. 1997. The EMBL nucleotide sequence database. *Nucl. Acids Res.* 25:7-14.

Stutts, M. J., C. M. Canessa, J. C. Olsen, M. Hamrick, J. A. Cohn, B. C. Rossier, and R. C. Boucher. 1995. CFTR as a cAMP-dependent regulator of sodium channels. *Science.* 269:847-851.

Stutts, M. J., B. C. Rossier, and R. C. Boucher. 1997. Cystic fibrosis transmembrane conductance regulator inverts protein kinase A-mediated regulation of epithelial sodium channel single channel kinetics. *J. Biol. Chem.* 272:14037-14040.

Sugita, M., Y. Yue, and J. K. Foskett. 1998. CFTR Cl- channel and CFTR-associated ATP channel: distinct pores regulated by common gates. *EMBO J.* 17:898-908.

Sutcliffe, M. J., I. Haneef, D. Carney, and T. L. Blundell. 1987. Knowledge based modelling of homologous protein, part I: three-dimensional frameworks derived from the simultaneous superpostion of multiple structures. *Prot. Eng.* 1:377-384.

Tabcharani, J. A., J. M. Rommens, Y.-X. Hou, X.-B. Chang, L.-C. Tsui, J. R. Riordan, and J. W. Hanrahan. 1993. Multiion pore behavior in the CFTR chloride channel. *Nature.* 366:79-82.

Tateno, Y., K. Fukami-Kobayashi, S. Miyazaki, H. Sugawara, and T. Gojobori. 1998. DNA data bank of Japan at work on genome sequence data. *Nucl. Acids Res.* 26:8-15.

Taylor, W. R., D. T. Jones, and N. M. Green. 1994. A method for α-helical integral membrane protein fold prediction. *Proteins: Struct. Func. Genet.* 18:281-294.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W:improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* 22:4673-4680.

Tieleman, D. P., and H. J. C. Berendsen. 1996. Molecular dynamics simulations of a fully hydrated dipalmitoylphosphatidycholine bilayer with different macroscopic boundary conditions and parameters. *J. Chem. Phys.* 105:4871-4880.

Tieleman, D. P., and H. J. C. Berendsen. 1998. A molecular dynamics study of the pores formed by *E.coli* OmpF porin in a fully hydrated POPE bilayer. *Biophys. J.* 74:2786-2801.

Tieleman, D. P., J. Breed, H. J. C. Berendsen, and M. S. P. Sansom. 1999a. Alamethicin channels in a membrane: molecular dynamics simulations. *Faraday Disc.* 111:209-223.

Tieleman, D. P., S. J. Marrink, and H. J. C. Berendsen. 1997. A computer perspective of membranes: molecular dynamics studies of lipid bilayer systems. *Biochim. Biophys. Acta.* 1331:235-270.

Tieleman, D. P., M. S. P. Sansom, and H. J. C. Berendsen. 1999b. Alamethicin helices in a bilayer and in solution:- molecular dynamics simulations. *Biophys J.* 76:40-49.

Tobias, D. J., K. C. Tu, and M. L. Klein. 1997. Atomic-scale molecular dynamics simulations of lipid membranes. *Curr. Opin. Coll. Interface Sci.* 2:15-26.

Tomb, J.-F., et. al. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature.* 388:539-547.

Tsukihara, T., H. Aoyama, E. Yamashita, T. Tomizaki, S. Yamaguchi, K. Shinzawa-Itoh, R. Nakashima, R. Yaona, and S. Yoshikawa. 1996. The whole structure of the 13-subunit oxidised cytochrome c oxidase at 2.8 A. *Science.* 272:1136-1144.

Vogt, G., T. Etzold, and P. Argos. 1995. An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *J. Mol. Biol.* 249:816-831.

von Heijne, G. 1991. Proline kinks in transmembrane α-helices. *J. Mol. Biol.* 218:499-503.

von Heijne, G. 1992. Membrane Protein Structure Prediction, Hydrophobicity Analysis and the Positive-inside Rule. *J. Mol. Biol.* 225:487-494.

von Heijne, G. 1994. Membrane protein: from sequence to structure. *Annu. Rev. Biophys. Biomol. Struct.* 23:167-192.

von Heijne, G. 1996. Principles of membrane protein assembly and structure. *Prog. Biophys. Mol. Biol.* 7:113-139.

Vriend, G. 1990. What If: A molecular modeling and drug design program. *J. Mol. Graph.* 8:52-56.

Wade, R. C., K. J. Clark, and P. J. Goodford. 1993. Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. *J. Med. Chem.* 36:140-156.

Wako, H., and T. L. Blundell. 1994. Use of amino-acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. 2. secondary structures. *J. Mol. Biol.* 238:693-708.

Wallimann, T., M. Wyss, D. Brdiczka, and H. M. Eppenberger. 1992. Intracellular compartmentation, structure and function of creatine-kinase isoenzymes in tissues with high and fluctuating energy demands - the phosphocreatine circuit for cellular-energy homeostasis. *Biochem. J.* 281:21-40.

Wallin, E., and G. von Heijne. 1998. Genome-wide analysis of integral membrane proteins from eubacterial, archean, and eukaryotic organisms. *Prot. Sci.* 7:1029-1038.

Walther, D., C. Springer, and F. E. Cohen. 1998. Helix-helix packing angle preferences for finite helix axes. *Prot: Struct. Funct. and Gent.* 33:457-459.

Wasinger, V. C., S. J. Cordwell, A. Cerpa-Poljak, J. X. Yan, A. A. Gooley, M. R. Wilkins, M. W. Duncan, R. Harris, K. L. Williams, and I. Humphery-Smith. 1995. Progress with gene product mapping of the Mollicutes - *Mycoplasama Genitalium*. *Electrophoresis.* 16:1090-1094.

Watts, A., A. S. Ulrich, and D. A. Middleton. 1995. Membrane protein structure: the contribution and potential of novel solid-state NMR approaches. *Mol. Memb. Biol.* 12:233-246.

Weiss, M. S., A. Jabs, and R. Hilgenfeld. 1998. Peptide bonds revisited. *Nature Struct. Biol.* 5:676-676.

Weiss, M. S., A. Kreusch, E. Schiltz, U. Nestel, W. Welte, J. Weckesser, and G. E. Schulz. 1991. The structure of porin from *Rhodobacter capsulatus* at 1.8 Å resolution. *FEBS Lett.* 280:379-382.

Welte, W., U. Nestel, T. Wacker, and K. Diederichs. 1995. Structure and function of the porin channel. *Kidney International.* 48:930-940.

Wetlaufer, D. B. 1973. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. USA.* 70:697-701.

Wilkins, M. R., K. L. Williams, R. D. Appel, and D. F. Hochstrasser. 1997. Proteome research: new frontiers in functional genomics. Springer-Verlag, Berlin and Heidelberg.

Wimley, W. C., and S. H. White. 1996. Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nature Struct. Biol.* 3:842-848.

Winter, M. C., and M. J. Welsh. 1997. Stimulation of CFTR activity by its phosphorylated R domain. *Nature.* 389,:294-296.

Woolf, T. B. 1997. Molecular dynamics of individual α-helices of bacteriorhodopsin in dimyristoyl phosphatidylcholine. I. structure and dynamics. *Biophys. J.* 73:2376-2392.

Woolf, T. B. 1998. Molecular dynamics of individual α-helices of bacteriorhodopsin in dimyristoyl phosphatidylcholine. II. interaction energy analysis. *Biophys. J.* 74:115-131.

Woolf, T. B., and B. Roux. 1994. Molecular dynamics simulation of the gramicidin channel in a phospholipid bilayer. *Proc.Natl.Acad.Sci. USA.* 91:11631-11635.

Woolf, T. B., and B. Roux. 1996. Structure, energetics, and dynamics of lipid-protein interactions - a molecular-dynamics study of the gramicidin-A channel in a DMPC bilayer. *Proteins: Struct. Func. Genet.* 24:92-114.

Woolfson, D. N., and D. H. Williams. 1990. The influence of proline residues on α-helical structure. *FEBS Lett.* 277:185-188.

Yeates, T. O., H. Komiya, D. C. Rees, J. P. Allen, and G. Feher. 1987. Structure of the reaction centre from *Rhodobacter sphaeroides* R- 26: Membrane-protein interctions. *Proc. Natl. Acad. Sci. USA.* 84:6438-6442.

Yuan, X., A. K. Downing, V. Knott, and P. A. Handford. 1997. Solution structure of the transforming growth factor b-binding protein-like module, a domain associated with matrix fibrils. *EMBO J.* 16:6659-6666.

Yuan, X., J. M. Werner, V. Knott, P. A. Handford, I. D. Campbell, and A. K. Downing. 1998. Effect of proline *cis-trans* isomerization on TB domain secondary structure. *Prot. Sci.* 7:2127-2135.