Title Development of Genetic Algorithm for Optimisation of Predicted Membrane Protein Structures

Name Noushin Minaji-Moghaddam

DEVELOPMENT OF GENETIC ALGORITHM FOR OPTIMISATION OF PREDICTED MEMBRANE PROTEIN STRUCTURES

by

Noushin Minaji-Moghaddam

A thesis submitted for the degree of Doctor of Philosophy
Of the University of Bedfordshire

March 2007

DECLARATION

I declare that this thesis is my own unaided work. It is being submitted for the degree of (*name award*) at the University of Bedfordshire.

It has not been submitted before for any degree or examination in any other University.

Name of candidate: Noushin Minaji-Moghaddam Signature: N. Minaji

Date: 14/09/2007

DEVELOPMENT OF GENETIC ALGORITHM FOR OPTIMISATION OF PREDICTED MEMBRANE PROTEIN STRUCTURES

Noushin Minaji-Moghaddam

ABSTRACT

Due to the inherent problems with their structural elucidation in the laboratory, the computational prediction of membrane protein structure is an essential step toward understanding the function of these leading targets for drug discovery.

In this work, the development of a genetic algorithm technique is described that is able to generate predictive 3D structures of membrane proteins in an *ab initio* fashion that possess high stability and similarity to the native structure. This is accomplished through optimisation of the distances between TM regions and the end-on rotation of each TM helix. The starting point for the genetic algorithm is from the model of general TM region arrangement predicted using the TMRelate program. From these approximate starting coordinates, the TMBuilder program is used to generate the helical backbone 3D coordinates. The amino acid side chains are constructed using the MaxSprout algorithm. The genetic algorithm is designed to represent a TM protein structure by encoding each alpha carbon atom starting position, the starting atom of the initial residue of each helix, and operates by manipulating these starting positions. To evaluate each predicted structure, the SwissPDBViewer software (incorporating the GROMOS force field software) is employed to calculate the free potential energy.

For the first time, a GA has been successfully applied to the problem of predicting membrane protein structure. Comparison between newly predicted structures (tests) and the native structure (control) indicate that the developed GA approach represents an efficient and fast method for refinement of predicted TM protein structures. Further enhancement of the performance of the GA allows the TMGA system to generate predictive structures with comparable energetic stability and reasonable structural similarity to the native structure.

ACKNOWLEDGEMENTS

The research is dedicated to my parents who have given me their understanding support and the deepest motivation for the work.

The writing of a thesis is a long and arduous task. Only with the support of others can anyone survive the ordeal. Fortunately, I have been blessed with an abundance of support from kindness advisors.

I wish to express my deepest gratitude to Dr Jonathan Mullins for his most invaluable advice and fatherly personal encouragement. This thesis would not be possible were it not for support of Dr Jonathan Mullins. I must thank, a thousand times over, to him. His support has covered an overwhelming expertise in the methodological issues as well as in the philosophical underpinnings of the field without it, this work would not have been achieved.

I would not be where I am today without the sage advice of Professor Angus Duncan, I owe many thanks to Professor Angus Duncan who have supported, inspired, motivated, me throughout my graduate studies. I would like to thank him for his invaluable and constructive comments on the work and suggestions for improvements.

I would like to thank to Roberto Togawa for his support and cooperation with my work.

I would like to thank to my brother Mehdi for all his support these many years and a thousand times thanks for his assistance in order to finish my thesis.

Last but not least I would like to thank my husband Mojtaba for all his support and everything he has done for me.

Noushin Minaji Moghaddam

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of Equations

GLOSSARY OF TERMS AND SYMBOLS

Symbol	Meaning	Page No
GA	Genetic Algorithm	page 2
TM	TransMembrane protein	page 2
TMHMM	TransMembrane protein topology with Hidden Markov Model	page 7
HMMTOP	Hidden Markov Model for Topology Prediction	page 7
TMRelate	The name of a software for predicting of general topology of TM regions representation	page 7
EC	Evolutionary Computation	page 12
SA	Simulated Annealing	page 13
TS	Tabu Search	page 13
ES	Evolutionary Strategy	page 15
EP	Evolutionary Programming	page 16
Cα	Central Carbon atom in amino acid	page 35
DNA	Deoxyribonucleic acid	page 35
ϕ	Phi angle	page 44
ψ	Psi angle	page 44
ω	Omega angle	Page44
NMR	Nuclear Magnetic Resonance	Page 51
PDB	Protein Data Bank	Page 59
GROMOS	GROningen Molecular Simulation Software	Page 60
BR	Bacteriorhodopsin	Page 63
TMGA	prediction of TransMembrane protein with Genetic Algorithm	Page 69
TMBuilder	TransMembrane region Builder predict the arrangement of helices in membrane	Page 95
MaxSprout	The name of a software for building on the side chains to the alpha carbon	Page 98
SwissPDBViewer	The name of a software for calculating the free energy	Page 98
R	Rotation	Page 103
Z	The value for providing information about similarity	Page 112
RMSD	Root Mean Square Deviation	Page 112
HP	Hydrophobic / Polar	Page 134

CHAPTER 1

1 Introduction-Genetic Algorithms and Protein Structure

A protein is a chain of amino acid residues that folds into a specific tertiary structure under certain physiological conditions. One of the most significant problems in biology is the protein folding problem and our underlying understanding of the biological importance of the information contained in genes. The functional properties of a protein depend on the spatial arrangement (tertiary structure) of its primary sequence.

In most genomes, 20-30% of protein-coding genes code for membrane proteins and they represent less than 0.3% of currently available structures and comprise a majority of the targets of currently marketed drugs (Zoonens et al., 2005).

These proteins at least partly reside in the membranes of cells and organelles, possessing transmembrane regions (TM).

There are appreciable difficulties in obtaining information relating to the 3D structure of membrane proteins from X-ray crystallography and NMR, (Rost, 1997) as both of these techniques require isolation, purification and / or crystallisation which may be difficult or impossible for membrane proteins due to the hydrophobic nature of the membrane lipid bilayer.

There are several different methods currently being used for protein structure prediction, these include homology modelling based on the existence of similar, known structure and *ab initio* modelling used to predict the structure of individual folds and their subsequent folding in tertiary conformation.

There are a variety of optimisation techniques that may be used to address the problem of protein structure prediction.

Genetic algorithms (GA) represent a highly efficient optimisation and search technique, and have proved to be a robust method for solving problems with more than one feasible solution (Holland, 1975, Goldberg, 1989).

In this study, the GA is utilised in a number of ways, in the manipulation of a number of parameters. Firstly, the distance matrix that specifies the amino acid proximities between TM regions in membrane proteins and associations inferred from inter-amino acid distance and secondly, the rotation of each helix and their subsequent association through residue side chain pairs.

This ability to manipulate and test the orientation of such structures in determined membrane proteins is a significant proofing step along the pathway to tertiary structure prediction for membrane proteins of hitherto unknown structure.

There are several approaches currently being used for soluble protein structure prediction by GA (Patton et al. 1995), but none for membrane proteins.

In this study, a GA is presented as an optimisation and search technique for investigation and prediction of the optimum spatial arrangement and orientation of TM regions in membrane proteins.

In the GA, a population of current solutions is maintained. The solutions evolve by mutations and crossovers. Technically, the operation consists of exchanging parts of strings between pairs of solutions (here, the spatial positions and angular rotation of the TM regions), so as to produce new solutions.

Through such iterations, good features from one solution can be transferred to the others and further explored. The solutions are evaluated by calculating the free energy (Gunsteren *et al.*, 1996) of the resultant structures, a measure of their feasibility in terms of energetic stability. The results are closely scrutinised by the GA and a population is gradually improved by selection.

1.1 Applications of genetic algorithm

The literature to date portrays the development and application of many types of genetic algorithm for a wide range of applications. Examples of such GA approaches that have been applied to such diverse fields such as Timetable scheduling systems, Medical and industry applications, are presented in the following sections.

1.1.1 Timetable scheduling

Timetable scheduling is a common problem for all institutions of higher education and also for all factories. Genetic algorithms have been used effectively to solve the problem of timetable and scheduling. Vorac et al., (2002) used a GA for solving a timetable problem. They used special mutation in order to move the GA to more promising areas in the search space and generate efficient timetables for big schools with a complicated teaching plan.

Boyd and Savory (2001) applied a GA to scheduling laboratory personnel. The GA appears to be useful for scheduling in highly technical work environments that employ multiskilled workers.

1.1.2 Medical Applications

Medical applications have used GA to address complex problems. Durant (2002) utilised a GA for the fitting of hearing aids. The author configures the GA to vary six parameters that control a frequency-selective dynamic range expansion system. The purpose of this system was to provide less gain to unwanted background sounds such as motor hum and breathing, while amplifying speech and other desired sounds sufficiently.

Bevilacqua *et al.* (2004) applied GA, combined with a maximum likelihood method to assist a physician in performing a diagnosis of some retina pathologies. They implemented an automatic process of matching different images of the same retina to study the development of some pathologies.

Rajapakse et al. (2005) used GA for seeking a motif with higher generalization ability that can elucidate rules that govern peptide binding to medically important receptors is important for screening targets for drugs and vaccines.

1.1.3 Industry applications

Genetic algorithms have been employed in many industrially based applications. Engelhardt *et al.* (2000) addressed the problem of the water industry in England and Wales that has changed dramatically over recent decades, where private companies now have responsibility for setting the prices that their customers are charged for their water supply, under the auspices of a regulatory body.

There were concerns that regulatory-based performance measures are backward looking and unable to be translated directly into a pro-active planning strategy. The GA was used to address the suitability of these measures in a decision - making framework

Patel *et al.* (2005) utilised a GA for the problem of scheduling oil production by cyclic steaming at an oil field in the San Joaquin Valley. The objective was to maximize cumulative production.

1.2 Motivation of present work

The main objective of the work of this thesis is to develop and apply a genetic algorithm to the prediction of the 3D structure of the transmembrane portions of membrane proteins. The work was primarily motivated by the difficulties encountered in attempting to reliably predict membrane protein structure from sequence.

The development of reliable prediction programs for membrane protein topology, such as TMHMM (Krogh et al., 2001) and HMMTOP (Tusnadt and Simon, 1998), that provide information as to the probable amino acid positions of the beginning and end of a given TM region and which side of the membrane the N-terminus of the protein is located, has allowed the development of reliable approaches for the prediction of TM region arrangement and adjacency.

The foremost of these is the approach encapsulated in a program developed by other members of the group, called TMRelate (Roberto Togawa, John Antoniw and Jonathan Mullins,2003). The TMRelate program is described in the draft manuscript that may be found in Appendix C. The output of TMRelate provides a prediction of the overall arrangement and relative positioning of the individual TM regions for any multi-spanning membrane protein, and allows the generation of approximate 3D models of transmembrane domains of membrane proteins.

However, the energetic stability of the resulting models is not considered in the TMRelate algorithm and the resulting structures are therefore not feasible in terms of energetic stability.

The main point of the work of this thesis was to provide an efficient and powerful algorithmic framework for the refinement of general model structures of TM regions of membrane proteins to energetically viable and chemically stable structural models.

In so doing, the work contributes to the closure of the final predictive gap between primary sequence and viable 3D structures for membrane proteins.

1.3 Overview of the thesis

The structure of this thesis is as follows: Chapter 2 introduces optimisation techniques for solving non linear problems. Genetic algorithms (GAs) as a stochastic search and heuristic and optimisation technique are introduced in the context of the intended application, namely the prediction of the 3D structure of the transmembrane regions of membrane proteins.

Chapter 3 reviews the rationale for the development of membrane protein structure prediction. The specific issues pertaining to membrane protein structure and the crucial nature of the prediction of membrane protein structure in biology are explained.

Chapter 4 the design of a GA technique for the prediction of membrane protein structure is discussed. The rationale and theories that are used to design candidate solutions for the problem will be discussed. The GA operators used for effective exploration of the conformation search space and selection policies applied in each generation are described.

Chapter 5 presents details of the development and program implementation of the TMGA system in order to address this specific problem, one of the many challenges that lie at the interface between computer science and biological systems, in the field known as bioinformatics or computational biology. The application of biologically oriented concepts such as genetic algorithms into computer science and the reciprocal application of adaptive computer algorithms to biological problems are described.

Chapter 6 discusses series of experiments conducted to validate the proposed TMGA system, and the resulting generation of energetically stable 3D structures. The TMGA results for the processed structure derived from the sequence for bacteriorhodopsin and comparison of these predicted structures with the experimentally determined bacteriorhodopsin structure are presented.

Chapter 7 is a discussion of the work as a whole, its merits and weaknesses, its context in relation to the work of others, and a synopsis of the contribution of this work to where we are now in terms of the ultimate goal of being able to derive reliable molecular structures for membrane proteins from primary sequence. Future improvements to the GA technique for predicting membrane protein structure are also discussed.

Chapter 8 outlines the main conclusions of the study.

CHAPTER 2

2 Genetic Algorithm as a search technique

This chapter introduces optimisation techniques for solving non-linear problems. Genetic algorithms (GAs) as a stochastic search and heuristic and optimisation technique are introduced in the context of the intended application, namely the prediction of the 3D structure of the transmembrane regions of membrane proteins. In this chapter, the following questions are considered:

- What is a GA?
- What are the advantages of using GA as a search technique?
- What are the components of a GA?
- How does a GA work?
- How have GAs been used to address the prediction of protein folding in other work?

The following sections of this chapter introduce genetic algorithms. Section 2.2 introduces other optimisation techniques.

In section 2.3, the background of the GA is reviewed as part of Evolutionary Computation (EC) and other areas of EC are discussed. This is followed by a definition of GA and a description of the components of GA structure in section 2.4. Section 2.5 discusses how a GA works. In section 2.6, other GA applications for prediction of protein structure are reviewed and in section 2.7 the rational for using the GA technique for this purpose is explained. In section 2.8, the advantages of using GA as a search technique are summarised.

2.1 Optimisation techniques

In order to build an algorithmic system to solve a particular problem, the problem first needs to be examined and then the optimum search technique selected to find the best solution for that particular problem. This kind of problem can be regarded as a global optimisation problem that contains many local optima in the region of interest. Such problems are defined as non-linear problems.

There are two categories of search methods, namely stochastic and deterministic methods. Stochastic methods attempt to efficiently cover the search space and one of the local optima will be selected as the global optimum.

Deterministic methods attempt to create the points, which belong to a close neighbourhood of global optimum. Modern heuristic techniques have been developed in order to escape from local optima and they include:

- Simulated annealing (SA) sometimes called stochastic searching, which was proposed by Kirkpatrick et al. (1983). This is a variation on "hill-climbing" where random guesses are introduced.

- Tabu search (TS), which was developed by Glover (1994), to solve combinatorial optimisation problems. The basic idea is to impose restrictions on the search process to guide it to investigate difficult regions. This technique uses a deterministic rather than stochastic search.

- Genetic algorithms (GAs), which were developed by Holland (1975) use past information to direct their search. A GA uses a pool of solutions and neighbourhood function is extended to act on pairs of solutions using the so called crossover operator (Theodore et al., 1998).

These search methods are generic techniques for resolving search and optimisation problems with large space searching. They are specific techniques that are very efficient when the solution or even the problem is unknown. They use a random approach to continue processing where optimal solutions contain the highest probability of being found in the search space. They are therefore able to provide good working solutions in a reasonable amount of time (Reeves, 1995).

2.2 The background of genetic algorithms

The so-called Genetic Algorithm (GA) (Holland,1975) is a flexible optimisation technique that operates efficiently on pieces of information. It is a robust method that borrows biological concepts such as natural selection of genes in the course of evolution. GAs were made famous by Goldberg (1989) who developed the first working GA.

GAs form a part of the field of evolutionary computation. They are inspired by natural selection and evolution in nature by mimicking the processes of Darwinian evolution. Evolutionary computation or evolutionary algorithms apply our knowledge about evolution in the form of algorithms and computer programs that attempt to solve complex problems.

As Smith (1989) elaborated, upon Darwin's theory and natural evolution, improved individuals will gradually develop in successive generations, and Darwin believed that the main force driving these evolutionary changes was natural selection. Those organisms with characteristics most favourable for survival and reproduction will not only have more offspring, but will pass their characteristics on to those offspring.

The theory of natural selection not only predicts evolutionary change, it also predicts the emergence of organisms, which are able to survive and reproduce in the environment better than those before. In natural evolution, parents who are able to survive best in their environment will be selected for producing children.

The children inherit genetic features from their parents. The unit of inheritance is the gene and the genes are contained within chromosomes. Genes in new generations of individuals, which adapt well, are also selected more often for reproducing.

Therefore, a new population will adapt better chromosomes containing better genes, in order to provide better features to survive in their environment.

In biological terms, fitness is usually measured in terms of breeding success. Genes indirectly contribute to an organism's fitness by providing better features that confer some advantage in breeding.

Using similar principles in computer science evolutionary computation was introduced as a part of the field of artificial intelligence, which includes neural networks, fuzzy systems and machine learning.

Evolutionary computation comprises the four main areas as follows:

1. Evolution strategy (ES), developed in Germany by Rechenberg (1973) and further developed by Schwefel (1981). The method was designed for parameter optimisation problems in order to generate a number of strategy parameters. The strategy parameters are used to control the behaviour of the mutation operators. ES is a deterministic search method.

2. Evolutionary Programming (EP) this is originally developed by Fogel et al. (1966). The difference between this method and the others is that there are no recombination operators and it depends on mutation operators. Among the evolutionary computation methods EP is better in obtaining global optimum which relies on mutation rather than crossover (Gnanadass et al., 2004).

3. Genetic algorithms (GA) were invented to mechanistically mimic principles of natural evolution. The GA is used as a search technique to efficiently optimise a set of candidate solutions to a problem in a short time, accurately and with good reliability. The GA is particularly well suited to optimising many problems which contains many local optima solutions within the search space and where more traditional methods fail (Haupt et al.,2003). The GA are able to find optimal or near optimal solution by using natural mechanism such as selection, crossover and mutation (Tian-li et al., 2005).

However, predicting the final stable three dimensional structure of the protein is a very complex and non-linear problem (Calabretta et al.,1995). The GA is attractive for addressing such problems. A genetic algorithm performs a stochastic search by randomly choosing solutions and randomly selecting the search direction about a solution or between solutions. The GA is able to work with many solutions at once which are represented randomly in an initial population.

2.3 Definition of GA terms

The following terms associated with GAs may be defined as follows:

- Offspring – a candidate solution to a problem is represented as a chromosome and it is usually a bit string or some other encoding
- Encode – to convert a phenotype to the corresponding genotype
- Phenotype – solution parameters corresponding to a particular offspring
- Genotype – the representation of an offspring, such as a bit string or list of values
- Population – a collection of a fixed number of offspring
- Parent – an offspring that is then input to a further GA further operation
- Generation – replacing the entire population to allow new offsprings to reproduce
- Fitness - a number representing the quality of a particular offspring
- Fitness function – a method of determining the fitness of a given offspring
- Selection – a method to produce an intermediate generation
- Crossover – a type of reproduction operator that exchanges information between two offspring to produce two new offspring
- Mutation – a type of reproduction operator that modifies a single offspring to produce a new offspring

2.3.1 Encoding

In building a GA to address a particular problem, the first task is to decide how to encode the possible solutions. The technique for encoding solutions may vary from problem to problem and depends on the nature of the problem variables

One of the important design variables for GA is the encoding which determines the possible mutation and crossover operators (Butter et al., 2006).

In many GA applications, a fixed length bit string is been used to encode the offspring. The binary encoding is often unable to represent frequent integer numbers as the neighbouring integer numbers differ in several bits values.

There are other applications that used integer string or other representation to encoding the candidate solutions (Davis, 1991). The integer encoding is able to change within a small range especially for predicted structures. The predicted structure is effectively altered by changing a single variable.

2.3.2 Evaluation

Each solution needs to be evaluated by measurement of a fitness function to reflect how good it is as a solution to the problem (Mardle and Pascoe, 1999),. The idea is that fitter candidates are in some way more likely to be selected and the GA allows the solutions to be sorted from best to worst.

The GA boosts the overall fitness of the population by keeping the best representation at each generation and produce new candidates using the selected population (Segonne et al., 2005)

2.3.3 Selection

There are several ways of picking which parents will be used to generate offspring in the next generation:

1. "Roulette Wheel" selection or fitness-based selection which is used in this work. This is the original and perhaps standard method. In this method, fitness values are normalised so that each individual is responsible for a certain proportion of the total fitness of the population. These values are converted to percentages and used to probabilistically pick parents. The probability that a bit string is chosen as a parent is equal to the percentage of the total fitness for which that the particular bit string accounts for. This method is refereed to as "Roulette Wheel" selection because each bit string can be thought of as a wedge on a roulette wheel (whitely, 1993).

2. The second method is called steady state selection. This way, a large percentage of the population is carried over from generation to generation. All individuals have equal probability of being selected as parents. This selection helps maintain variety in the population unlike in proportional selection, one single individual with a much higher fitness than the rest of the population will be chosen much more frequently as a parent and its characteristics will soon become dominant in the population (Vail, 2001).

Natural selection models nature's survival of the fittest mechanism. Fitter solutions survive while weaker ones are eliminated. Generally, in the selection step, the search focuses on the promising area of the search space (Butter, 2006).

2.3.4 Genetic operators

To produce new individuals (offspring), genetic operators are applied to on the individual chosen in the selection step. There are two main kinds of operators: Crossover and Mutation.

2.3.4.1 Crossover

Crossover enables the algorithm to extract the best genes from different individuals and recombine them into potentially superior children.

The idea of using crossover is to recombine useful components of the members of a breeding pairs and produce two individuals that inherit traits of both parents. Two new offspring will be created that will replace their parents.

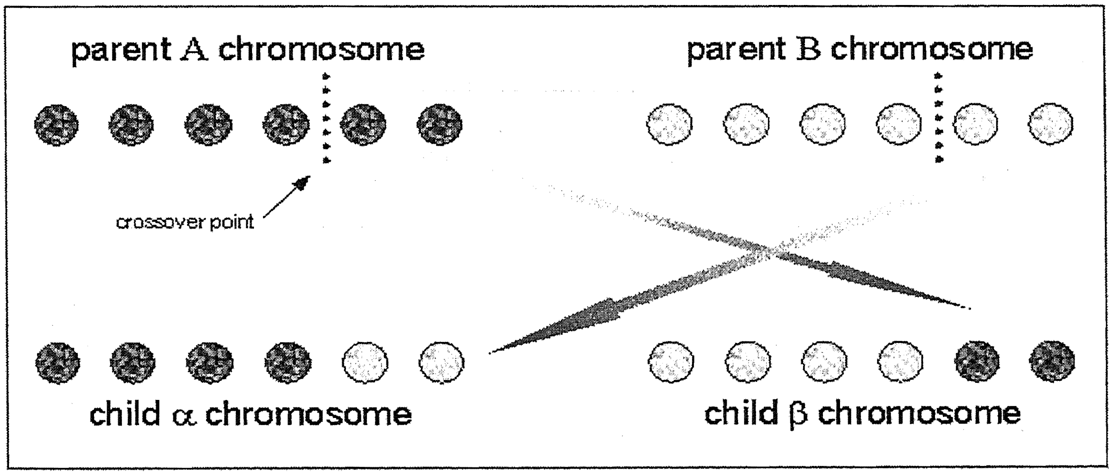


Figure 2.1 Complementary Crossover

Single cross-point, as shown there are two parents (A= dark chromosome and B= light chromosome), with a cross point selected randomly between genes 4 and 5. After crossover new children carry the parents genes but child A has 4 dark and 2 light and child B has 4 light and 2 dark

There are several different ways of combining genes via crossover.

- In single point crossover, a point is picked in the chain is randomly picked, and, all bits or values before that point are taken from one parent and exchanged with those of other parent.
- Two point crossover extends this idea and two points are randomly chosen. All of the values up to the crossover point are taken from the first parent then the subsection of the second parent that begins at the crossover point and continues for the random length is copied into the child, and then the remaining values are taken from the first parent.
- In uniform crossover, more than two points are chosen randomly. Each value from the first parent has a 0.5 probability of swapping with the corresponding gene of the second parent.

Each type of crossover is controlled by probability of crossover (Pc). This probability controls the rate at which solutions are subjected to crossover (Srinivas, 1994). When the solutions are not subjected to crossover, they remain unmodified.

2.3.4.2 Mutation

Mutation adds to the diversity of a population and thereby increases the likelihood that the algorithm will generate individuals with better fitness values. Without mutation, the algorithm can only produce individuals whose genes are a subset of the combined genes of the initial population. This genetic operator alters one or more components of the solution i.e. the gene in a chromosome. Mutation has the effect of ensuring that all possible chromosomes are reachable. This process is carried out randomly according to the probability of mutation (Pm) that defines the expected rate of mutation.

This is useful since crossover may not be able to produce new alleles or feature if they do not appear in the initial generation. The mutation operator is able to randomly select any bit position in a string and change it. For instance, the bit will unconditionally change from 0 to 1 or vice versa (Goldberg, 1989).

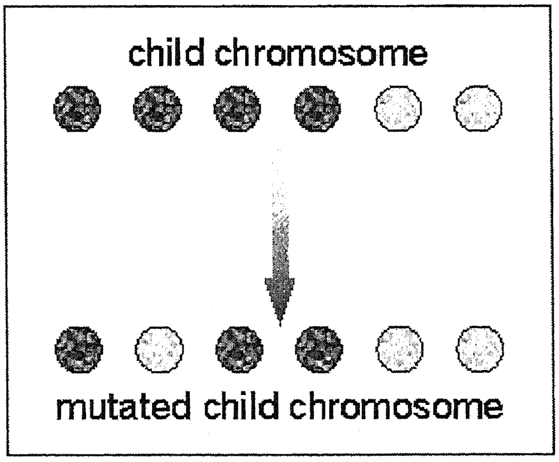


Figure 2.2 Mutation-one gene has been changed as shown when an offspring is created, one or more genes can be randomly changed for instance here, dark gene (No 2) changed to light gene

The purpose of crossover is to search globally (between solutions). On the other hand, the purpose of crossover is to combine features of different genes. For this reason the same offspring are not allowed to be chosen to be both members of the pair. By contrast, the purpose of mutation is to search locally (around a solution).

2.4 Genetic algorithm theory

In order to implement a directed search, it is helpful to understand the theory behind genetic algorithms. In analysing genetic algorithms each individual is cut and manipulated by crossover. Sub-strings define regions of the search space and are called schemas. A schema is a template that identifies a subset of strings with similarities at certain string positions [Holland, 1975]. A schema matches a particular string, for instance consider binary strings of length 6. The schema 1**0*1 describes the set of all strings of the length 6 with 1s at positions 1 and 6 and a 0 at position 4.

The " * " denotes a " do not care " or wild card, which means that positions 2,3 and 5 can be either a 1 or a 0. The order of a schema is defined as the number of fixed positions in the template, while the defining length is the distance between the first and last fixed positions. For example, the order of 1**0*1 is 3 and its defining length is 5. The fitness of a schema is the average fitness of all strings matching the schema.

The number of unique schemata in a particular population depends on the number of schemata contained in an individual string. The schemata with high fitness values and small defining lengths are appropriately called building blocks.

This is the essence of the schema theory, which was proposed by Holland (1975) as the "fundamental theorem of genetic algorithms". The notion that strings with high fitness values can be located by sampling building blocks with high fitness values and combining the building blocks effectively is called the building block hypothesis.

The capacity of GAs to simultaneously process a large number of schemata is called implicit parallelism. A search algorithm balances exploration of the search space with exploitation of areas of that space. Exploration points out new areas to search in, while exploitation concentrates the search in a particular area.

Genetic algorithms dynamically balance between exploration and exploitation through the recombination and selection of operators respectively. With the GA operators, the schema theorem proves that relatively short, low order, above average schema are expected to yield an exponentially increasing number of trials or copies in subsequent generations [Goldberg, 1989]. Expressed mathematically

$$\boxed{m(h,t+1) \geq \frac{m(h,t)\, f(h)}{\overline{f}_t}\left[1 - p_c \frac{\delta(h)}{l-1} - o(h)\, p_m\right]}$$

Equation 1. Representation of Schema theorem mathematically.

Where m(h,t) is the expected number of schema h at generations t, f(h) is the fitness of schema h and f(t) is the average fitness at generation t. The genotype length is 1 , δ (h) is the defining length and o (h) the order of schema h. P_c and p_m are the probabilities of crossover and mutation respectively.

2.5 Applications of GA in protein structure prediction

Genetic algorithms have been employed as an optimisation technique for many applications. The recent developments in their use in predicting protein structure are discussed in this section.

An early application of genetic algorithms to protein structure prediction by Unger and Moult (1993) is widely based on a 2D lattice. The GA was implemented by generating population of conformation themselves, which were not encoded as binary. This conformation is a 2D model of linear sequence of amino acids and evaluated by an energy function in order to find lower an energy conformations.

A standard GA approach to protein structure prediction by Patton et al. (1995) was based on the work of Unger and Moult in the area of energy function optimisation but instead they used a 3D lattice. Each peptide is represented as a single point in the lattice model.

In this application, each individual is represented by a single relative movement for each peptide which contains five possible values (up, down, right, left, forward) and is encoded by 7 bits. The objective function is based on adjacent hydrophobes in the protein's primary sequence. This evaluation function is different that of Unger and Moult (1993).

Schulze-Kremer (1996) investigated the prediction of main chain folding patterns of small proteins by using a GA. The author configured the GA to operate on numbers, not bit strings and represented each individual by torsion angles. A feature of the torsion angle representation is the fact that even small changes in the angles can result in large changes in the overall conformation.

The GA in this application used a simple potential energy function to evaluate each individual. The total energy is the sum of the expressions for bond length potential, bond, torsion and improper torsion angle potential, van der Waals pair interactions, electrostatic potential and hydrogen bonds.

Cui et al. (1998) used a GA for predicting protein structure The initial population, size of 500 was generated by randomly selecting the backbone and side chain torsion angles (φ and ψ) in constrained regions. Each individual was evaluated by calculating a fitness scale as specified in a formula. The probability of crossover was estimated by dividing the fitness of an individual by total fitness.

Mutation was operated by two kinds of mutation, which were used to change the conformation dramatically and to make more local searches. The GA process was stopped only if the decrease of the lowest energy in the population is less than 1 unit during the last 20 generations.

Krasnogor et al. (1999) employed GAs to address the protein structure prediction problem in the “HP model”,which is an acronym for hydrophobic-hydrophilic model. HP models abstract the hydrophobic interaction process in protein folding by reducing a protein to a heteropolymer that represents a predetermined pattern of hydrophobicity in the protein.

The authors used a simple lattice model in order to capture global aspects of protein structures. In this work, the embedding of a sequence in a lattice is represented by internal coordinates by means of relative moves and also a modified energy potential assist the GA search while preserving the ranking of the standard HP model.

Their results supported the use of the relative encoding over the absolute one, although the modified energy potential was unable to improve optimisation performance.

Szustakowski and Weng (2000) developed a method for aligning the three dimensional structures of two proteins. The basic approach is to determine the protein secondary structure element alignment by used of a GA then extend the alignment to include any equivalent residues found in loops or turns. Alignments are evaluated using intermolecular distance matrices. The structure alignment problem becomes a search for regions of similarity shared by the two distance matrices. This search could be used for efficient database searching.

The authors found the GA to be fast and efficient, capable of generating correct alignments from a small number of randomly generated secondary structure elements alignments.

Braden (2002) used a GA to address the prediction of short peptid structure prediction. In this application, each offspring is represented as a group of five bits, arranged in a string, where each group codes for one residue and describes the characteristics of hydrophobicity, charge and side chain size.

There are L genes for a protein consisting of L residues. The fitness function is based upon non-connected residues that are adjacent in cardinal directions. Characteristics of the protein are stored in static arrays and used to evaluate fitness. This method improves upon that presented by Unger and Moult(1993) but is unable to work with a more complex protein structure.

Unger (2004) used a GA for protein structure prediction. In his work, each solution was represented as a set of pairs of values for the two dihedral angles (φ, ψ) along the main chain. The free energy for each conformation was evaluated in order to select those structures with lower energy values.

This approach was unable to ensure that the encoded structure is free of collisions and crossover operation is a very risky in the sense that it is likely to lead to conformations with internal collisions.

Bui and Sundarraj (2005) developed a genetic algorithm for the protein-folding problem following the HP model in a 2D square lattice. In the HP model each amino acid is classified as an H for hydrophobic and P for hydrophilic. The algorithm combines the concept of secondary structure prediction with a genetic algorithm. Their results showed that it outperforms existing evolutionary and Monte Carlo algorithms. They intend to advance their algorithm to a 3D HP model.

2.6 Advantages of using a Genetic algorithm

The GA technique has advantages over traditional non-linear solution techniques that cannot always achieve an optimal solution. The GA works with its own rules and manipulates the population of candidate solutions that ultimately provide a solution to the problem. GAs use a selective process to encourage over achieving and discourage under achieving solutions in the population.

The GA procedures require no knowledge of the nature of the problem. This is very useful for complex or loosely defined problems or problems, with many local optima. This method produces robust searches and performs efficient searches on poorly defined spaces (Goldberg, 1989).

The method is able to find new solutions through a process of natural evolution, although this evolution can be inductive in some conditions. The GA is a cooperative computational method which has been previously successful in many different computational tasks, including protein folding (Unger,2004).

2.7 Summary

This chapter opened by reviewing optimisation techniques for solving non-linear problems and then concentrated on introducing the genetic algorithm (GA) technique, which forms part of the field of evolutionary computation. The basic structure of a GA is defined as being composed of the following processes:

1. Encoding the candidate solutions
2. Evaluation
3. Selection
4. Operation of GA operators

To describe how GAs work, the theory behind GA was elaborated. This theory introduces schema, which are templates and define regions of search space. These templates contain similarities at certain positions and are able to make building blocks. This is referred to as the fundamental theorem of genetic algorithms.

Other applications of GAs are reviewed to illustrate how the GA is highly suited to addressing the protein structure prediction problem and to assess the effectiveness and suitability of GAs for the specific application of this study.

CHAPTER 3

3 Prediction of membrane Protein Structure

In this chapter, initially the summary of this study is illustrated in figure 3.1 and a number of issues will be discussed in order to explain the rationale for the work presented in this thesis. These issues could be addressed by the following questions:

- What is special about membrane proteins?
- Why will the prediction of membrane protein structure be useful?
- What will it allow to be done?
- What are the different approaches to predicting protein structure?

In the following sections of this chapter, the general principles of membrane protein structure are discussed, leading to a discussion of the prediction of membrane protein structure. Section 3.2 contains a general introduction to protein structure. In section 3.3, aspects specific to membrane protein structure are discussed, and this is followed by an explanation of membrane protein folding in section 3.4. Section 3.5 discusses approaches to the prediction of the 3D structure of proteins. Section 3.6 reviews optimisation methods, which are used in this study in order to find the best predictive models of protein structure and section 3.7 introduces the electronic structural database that is used in this work, from which atomic coordinates of determined structures are taken for comparison in this study. In section 3.8, the energy force field function is discussed along with the interaction energies between molecules and within conformations.

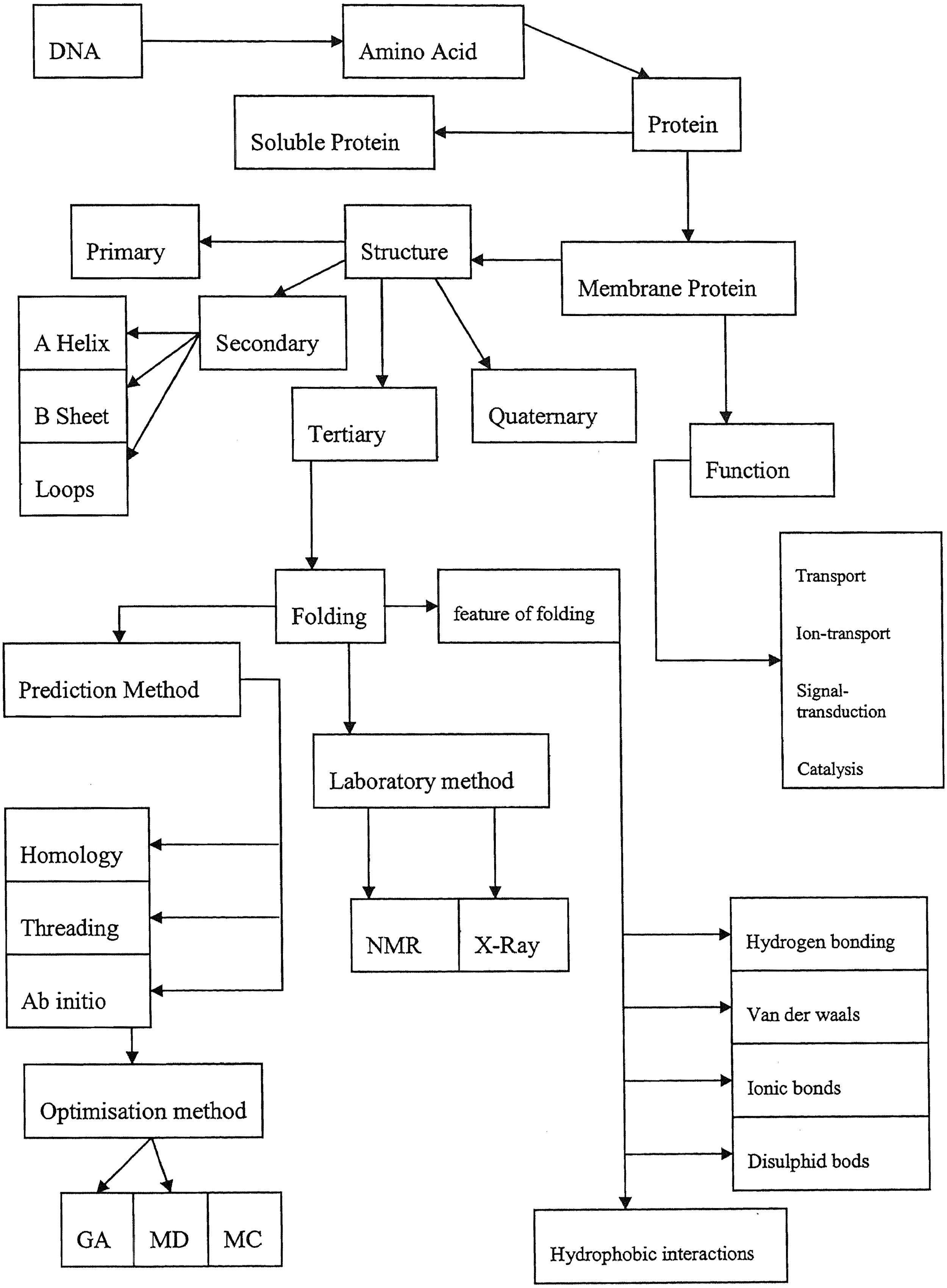


Figure 3.1 Logic follow diagram, describing the consideration of this study

3.1 Introduction of protein structure

Genes contain DNA (Deoxyribonucleic acid), the primary genetic molecule of life, which carries information for all biological organisms. DNA contains information in its sequence called the genetic code which is translated to a sequence of amino acids, the building blocks of proteins (Kendrew,1994).

There are twenty standard types of amino acid that may be found in a protein. Each amino acid contains a central carbon atom (Cα) to which is attached a hydrogen atom, an amino group (NH_2), a carboxyl group (COOH) and side chains as shown in figure 3.1A. The differences between the twenty amino acids lie in the side chains, which are small chemical groups that give each amino acid its unique characteristics and properties. Together, a group of side chains may confer critical functional properties on a protein such as the ability to bind ligands and catalyse biochemical reactions.

They direct the folding of the developing polypeptide and stabilise its final conformation. Amino acids are linked together end to end via peptide bonds. Condensing the carboxyl group of one amino acid with the amino group of the next to eliminate water generates the peptide bond (figure 3.1 B).

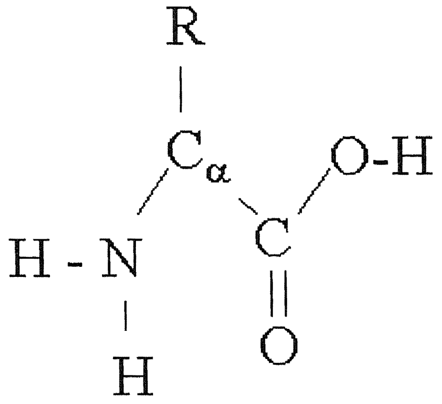


Figure 3.2 (A) Showing the components of an amino acide where R is the side chain.

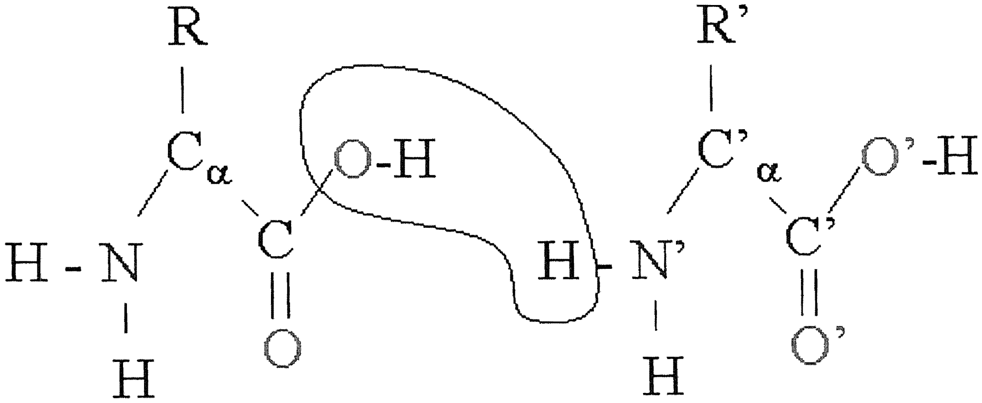


Figure 3.2 (B) Joining two amino acids by generating a peptide bond between them.

Amino acids can be divided into three classes as defined by the chemical nature of their side chain. The first class comprises those with hydrophobic side chains, that amino acids are not able to mix well with water (By contrast, polar amino acids that are able to mix with water are called hydrophilic). Hydrophobic amino acids include Valine, Alanine Leucine, Isoleucine, Phenylalanine, Proline and Methioninė.

The second class are the charged amino acid residues such as, Aspartate, Glutamate, Lysine, Arginine, and the third class comprises those amino acids with polar side chains:.Serine, Threonine, Tyrosine, Histidine, Cysteine, Asparagine, Glutamine and Tryptophan. The amino acid glycine, which has only hydrogen atom as a side chain does not fit into the above classification (Branden and Tooze ,1999).

Proteins are constructed from one or more chains of amino acids. There are two types of protein as follows:

- Globular or soluble proteins which contain polar residues on the surface and hydrophobic residues on the interior of the protein.

- Membrane proteins, which contain an external portion that is water soluble at each end and a hydrophobic section in the middle.

3.2 Aspect of membrane protein structure

Cells and the organelles within them are enveloped by membranes, which consist largely of lipids and protein molecules. The lipids form a bilayered a sheet structure that is hydrophilic on its two outer surfaces and hydrophobic in between two surfaces (Lodish et al., 2000) as shown in figure 3.2.

Hydrophilic head groups have maximum contact with water and the hydrophobic tails are forced into minimum contact with water within the membrane core. The contact between the molecules of the lipid bilayer serves to minimise the free energy of the structure and thereby maximising stability (Elliott et al., 2001).

In the lipid bilayer, the hydrophobic region layer is almost 30 Å thick and provides the distinctive environment occupied by the transmembrane regions of membrane proteins (Popot and Engelman, 2000).

3.2.1 Definition of membrane proteins

The proteins that are embedded in this layer are called membrane proteins. Membrane proteins can be associated stably with a lipid bilayer membrane in two general ways as shown in figure 3.2:

- Proteins that bind to the surface of the membrane, and often to the extracellular loops of a transmembrane protein, and are called peripheral membrane proteins
- Proteins that are covalently bonded to a lipid prosthetic group , are called integral membrane proteins.

Integral membrane proteins can be divided into two groups: those are embedded partially in the lipid bilayer, without crossing the membrane and proteins that contain regions that traverse the membrane are called transmembrane proteins (Matthews *et al.*, 1997). TM proteins aggregate and precipitate in water. They require detergents or nonpolar solvents for extraction (De Brevern et al., 2005).

Bacteriorhodopsin is an integral membrane protein usually found in 2D crystalline patches known ***as*** "purple membrane" which can occupy up to nearly 50% of the surface area of the archael cell. The repeating element of the hexagonal lattice is composed of three identical protein chains each rotated by 20 degrees relative to the other. Each chain has seven TM alpha helices and contains one molecule of retinal buried deep within (Grisshammer et al., 2006).

Bacteriorhodopsin (BR) is particularly abundant in the membrane of the purple bacterium Halobacterium halobiom (see appendix C) and belongs to the seven TM receptor family of proteins. The polypeptide crosses the membrane seven times, forming a cluster of seven α-helices spanning the membrane, connected by hydrophilic loops.

The cluster has a light absorbing pigment at its centre to capture light energy and converts it to a proton gradient, which in turn is used by a second membrane protein called ATP synthase to generate chemical energy in the form of ATP. The cell uses ATP to drive a multitude of vital processes (Elliott et al., 2001).

They can be grouped into one of four basic categories depending on whether they function in order to carry out:

- Transport of substrate in and out of the cell
- Ion transport and nerve impulse conductance
- Signal transduction (transmission of signals across a membrane from outside the cell to the inside)
- Catalysts (enzymes) for metabolic reaction (Kleinsmith and Kish, 1995)

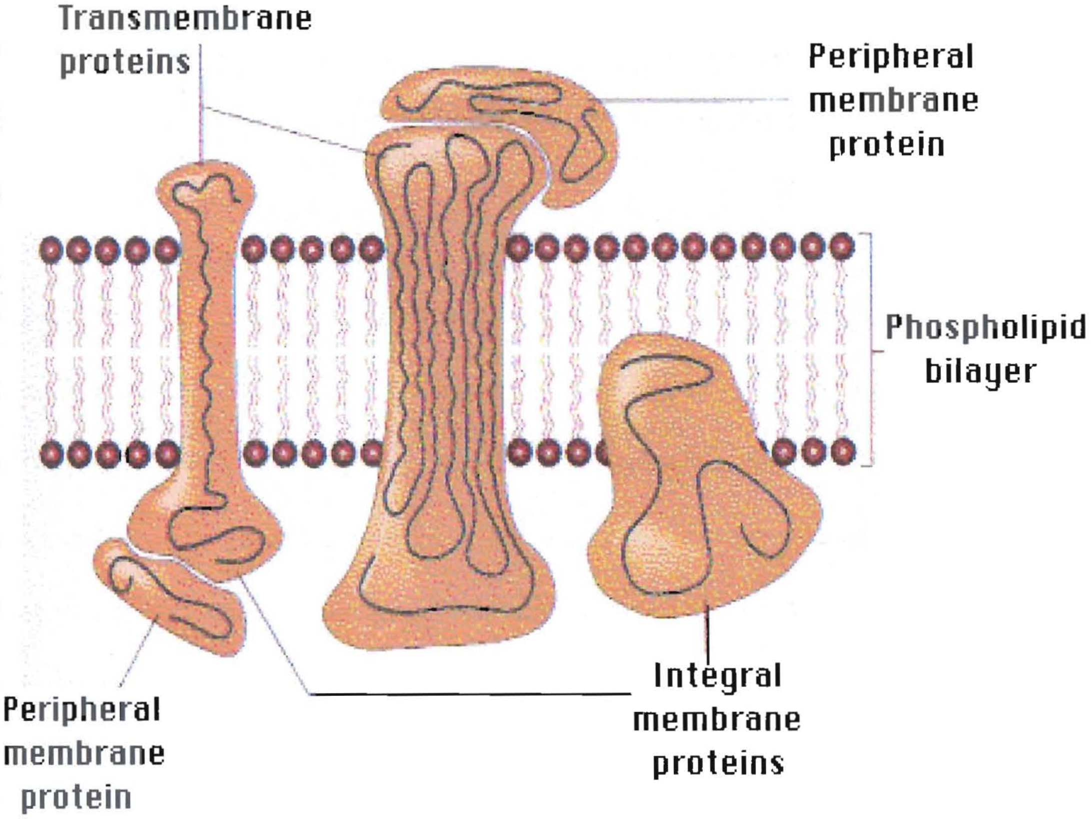

Transmembrane proteins
Peripheral membrane protein
Phospholipid bilayer
Peripheral membrane protein
Integral membrane proteins

Figure 3.3 Model of the lipid bilayer (cell membrane) with membrane proteins (Kimball, 1994). Taken from (http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/C/CellMembranes.html)

3.2.2 Primary structure

The series of amino acids, which are linked together into a polypeptide chain, or sequence, are regarded as the primary structure of a protein. The amino acid sequence of a protein ultimately determines the higher levels of structure of the molecule. A short section of amino acid sequence with recognised function is referred to as a motif (Harry *et al.*,1997).

The primary structure of a protein contains all the necessary information required for determining three dimensional structure. These types of information relating to proteins are stored in databases. As the techniques for sequencing proteins have improved and the genetic codes of whole genomes become fully known, and from them, the amino acid sequence, and whole proteomes, the amount of information has inexorably increased.

A number of databases have been established (Elliott *et al.*, 2001). The various databases contain specific kinds of information such as the Swiss-Prot database that is used to obtain information relating to the amino acid sequences and functions of proteins from all sequenced organisms. The last version of the Swiss-Prot database (version 48, May 2005) contained over 217,000 protein sequences.

3.2.2.1 Secondary structure

Secondary structure refers to the twisting and turning of a segment of polypeptide main chain without regard to the conformation of its side chains. There are three classes of secondary structure:

- The α-helix

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

In the membrane protein structure, the α-helix structure buries the polar amide groups inside the helix, surrounding them with hydrophobic side chains, giving a cylinder coated with hydrophobic groups. This hydrophobic α helix can immerse itself in the hydrophobic lipid bilayer and is then known as a transmembrane helix (Harry et al., 1997).

Most transmembrane helices are characterised by length of 15 -20 hydrophobic amino acids (Engelman and Popot, 2000). The α-helix structure consists of 3.6 amino acid units per turn, and the rise along the helix for each amino acid is 1.5Å.

Helices in the TM region of membrane proteins and helices in soluble proteins have differencing amino acid components abundances and distributions, such as Proline which has a more diverse pattern of interactions in membrane than in soluble proteins (Adamian et al., 2003).

The regularity of α-helix structure depends on the specific bond lengths and bond angles between amino acid units. The bond lengths considered are those for the peptide bond and hydrogen bond. The bond angles relate to the angle of rotation around the N- C_α bond which is termed Phi(Φ) and is -60°, and the angle of rotation around the C_α-C' bond which is termed psi (ψ) and is characteristically -50°(Branden and Tooze, 1999). The final angle to be considered is the omega angle (ω) between N- C', which is very close to 180°.

- The β-sheet

In 1951, the same year Pauling proposed the α-helix, Pauling and Corey postulated the existence of a different polypeptide secondary structure, the β-sheet. In the secondary structure extended polypeptide backbones are side by side, and use the full hydrogen bonding between neighbouring polypeptide chains rather than within a chain as in an α-helix. Sheets come in two varieties: The anti parallel β-sheet and the parallel β-sheet.

- Connecting loops α-helices and sections of β-sheet are connected together by unstructured polypeptides called connecting loops (Elliott. et al., 2001).

3.2.2.2 Tertiary structure

The arrangement of the various secondary structures into the compact structure of a globular or membrane protein is referred to as the 3D or tertiary polypeptide structure.

The polypeptide chain becomes folded into its proper tertiary structure by a number of bonds and interaction between amino acids side chains. These bonds and interactions include:

- Hydrogen Bonds are non-covalent bond and form between an amino group of one residue (n) and a carboxyl group three residues away (n+3). The α-helix is characterised by hydrogen bonds forming along the chain. The hydrogen bonds make a positive contribution to protein stabilisation only when there is an absence of accessible competing water. In the unfolded state, all potential hydrogen bonding partners in the extended polypeptide chain are satisfied by hydrogen bonds to water although in the protein fold these protein-to-water hydrogen bonds are broken and only some are replaced by intra-protein hydrogen bonds. (Lesk, 2004).

- Van der Waals Forces occur when atoms are very close together. These forces involve both attraction and repulsion. The nuclei of atoms or molecules attract the electrons of other atoms or molecules. Van der Waals forces are much weaker than hydrogen bond and also have a repulsive component when two nuclei are squeezed together, the electrons in their orbital repel each other. This repulsion increases as the closer the atoms are together. Van der Waals forces become accumulatively significant when many atoms within a biological membrane permits the formation of a large number of these weak interactions (Harton et al., 2002).

- Ionic bonds are most likely to be formed by atoms that tend to lose their outermost electron, and by atoms that tend to acquire electrons to complete their outmost shells. These atoms can generally attain a completely filled outer electron shell most easily by giving electrons to or accepting electron from another atom rather than by sharing them. When an electron jumps from an atom to another, both atoms become electrically charged ions because of their opposite charges, are attracted to each other and are thereby held together by ionic bonding (Alberts et al., 2004). This is a non-covalent bond and occur between basic amino acids with positive charge and acidic amino acids with negative charge. This interaction is potentially the strongest non-covalent forces and play a role in the recognition of one molecule by another (Harton et al., 2002)

- Disulphide bonds are covalent bonds and are found in secreted proteins only. The reducing environment inside cells readily disrupts these bonds. The disulphide bonds form between the side chains of two cysteine residues. Two SH groups from cystein residues, which may be in different parts of the amino acid sequence but adjacent in the three dimensional structure, are oxidized to form one S-S (disulphide) group according to the following reaction scheme:

$$\boxed{\text{2-CH}_2\text{SH} + 1/2\ \text{O}_2 \quad \Leftrightarrow \quad \text{-CH}_2\text{-S-S-CH}_2\ \text{+H}_2\text{O}}$$

Equation 2 Chemical structure of disulphide bond

(Branden and Tooze, 1999).

Hydrophobic interaction is the association of nonpolar molecules with other nonpolar molecules rather than with water. This interaction can have a significant cumulative effect on the stability of a macromolecule and in determining the three dimensional structure of most membrane proteins (Harton et al., 2002).

- The helix-helix, lipid-lipid and lipid-helix interactions have specific energy terms associated with each of them, with particular consequences for the overall stability of the membrane proteins.

3.2.2.3 Quaternary structure

Some proteins have a quaternary structure; that is, they are comprised of two or more polypeptide chains. Each polypeptide chain in such a protein is called a subunit.

The forces that hold the subunits together are of the same general types, which are encountered, in tertiary structure. These interactions occur between amino acid side chains located in different polypeptide subunits (Kleinsmith and Kish, 1995)

The quaternary structure of bacteriorhodopsin consists of three bacteriorhodopsin molecules that associate to a trimmer configuration, which is organised into a two-dimensional crystalline array, named purple membrane (Muller et al., 1999).

Cytochrom oxidase is an integral membrane protein and is a magnificent enzyme that is composed of three core subunits (Hofacker and Schulten, 1998).

3.3 Membrane protein folding problem

Folding of membrane proteins is thought to be similar to soluble proteins in that the secondary structural elements first form and then these elements come together to form the final tertiary structure. The functional property of a membrane protein depends on the tertiary structure of its primary sequence.

However, since the membrane environment imposes constraints on the peptide chain, secondary and tertiary structural features are quite different from those imposed by an aqueous environment and the folding and features of amino acids in membrane proteins differs much from that of soluble proteins (Edman, 2001).

Several theoretical studies (Orlandini et al., 2000) have suggested that folding of membrane proteins occurs in four steps, partitioning, folding, insertion and association. They are based on structural and thermodynamic measurements of the partitioning of small hydrophobic peptides and proteins between aqueous and membrane phases.

The bilayer interface provides a free energy for initial binding and folding of hydrophobic peptides. The formed helices are inserted across the membrane and then associated with other transmembrane helices.

The four-step model is a process along an interfacial path, a water path, or a combination of the two. Determination of the free energies for each of these steps along a path allows thermodynamic stabilities to be computed. Since most membrane proteins are helical, this may be a general pathway (Wimley and White,1999).

Solving the protein folding problem has been accepted as a significant step in understanding the importance of the information contained in genes and is a difficult problem due to the size and complexity of the search space.

There are two main laboratory methods for determining the three-dimensional structure of proteins:

3.3.1 X-ray

X-ray diffraction of protein crystals that has determined most of the known protein structures. In this method, it is essential to purify a protein and this often proves to be an extremely difficult process, particularly for membrane proteins.

3.3.2 Nuclear magnetic resonance (NMR)

Nuclear magnetic resonance (NMR) provides structural information on proteins in concentrated solutions and therefore investigates the protein in an environment more closely resembling that of the cell, permitting certain conformational changes to be observed (Elliott et al., 2001).

NMR spectroscopy is now a major tool for solving protein structure. The introduction of NMR spectroscopy as an alternative method for protein structure determination at atomic resolution has led to a significant increase in the number of known protein structures. Although it has limitations in terms of the size of proteins whose structures can be analysed it has the great advantage of permitting structural determination in the lipid environment. It is therefore better suited to studying membrane proteins than X-ray crystallography (Opella, 1997). However, to date it has only been possible to subject a few integral membrane proteins to NMR analysis. Very few membrane proteins can be studied by high-resolution NMR, partly because of their size and complexity (Schwaiger et al., 1998).

Recently, Gao et al. (2006) applied solution NMR spectroscopy to the structural determination of small and medium sized α helical membrane proteins.

In most genomes, 20-30% of protein-coding genes code are available for membrane proteins (Arkin et al., 1997). These proteins reside at least partly in the membranes of cells and organelles, possessing transmembrane regions (TM).

Despite their significance in the genomes of all organisms and their functionality, progress in determining their 3D structures has been much slower compared to the soluble proteins. There are appreciable difficulties in obtaining information relating to the 3D structure of membrane proteins from X-ray crystallography and NMR, (Rost, 1997) as both of these techniques require isolation, purification and crystallisation which may be difficult or impossible for membrane proteins because they dissolve only in fat not water.

However, recent years have been particularly fruitful for the structural biology of membrane proteins. Tamm et al. developed a technique in 2003 to determine structures of small membrane proteins, which are all β-barrels in the molecular mass range of about 20 kDa, by solution nuclear magnetic resonance (NMR).

3.4 Tertiary structure prediction methods

The biological role of a protein is determined by its chemical function, which depends on its structure. Although more structures are determined experimentally, it is impossible to determine all the protein structures from experiment. There are three major theoretical methods for predicting the structure of proteins as follows:

3.4.1 Homology Modelling

The most reliable method of determining the shape of a protein is to search for a close homologue in the database of solved protein structures. This method has been highly successful with soluble proteins.

In order to identify structures similar to the target protein, homology modelling involves a variety of sequence comparison techniques and requires the existence of homologous proteins, for which a structure has been solved. For this reason, homology modelling has been most useful for the few TM protein families, for which at least one member has been crystallized (Fleishman and Ben-Tal, 2006).

However, because at present only few representative atomic-resolution structures of TM protein families are available, homology modelling cannot serve as a general purpose approach for structural modelling of membrane proteins.

3.4.2 Threading or fold recognition

When homology modelling is unable to recognise the correct fold for the target sequence, the next method that is frequently attempted is fold recognition or threading. In fold recognition, a target sequence is aligned with all structures in a fold library.

Modelling by threading is independent of sequence comparison. Therefore, it can be used to identify the relationships among proteins even if sequence similarity is extremely low or non-existent (Domingues et al., 2000).

Threading methods are limited by the high computational cost, since each entry in the whole library of thousands of possible folds needs to be aligned in all possible ways on each occasion to select the folds (Zhang, 2003) and this will be even more problematic for membrane proteins because these proteins are large in comparison with soluble proteins.

3.4.3 *Ab Initio* prediction

Most *ab initio* prediction methods use reduced representations of the protein to limit the conformational space and use empirical energy functions that capture the most important interactions that drive the folding of the protein sequence toward its native structure (Hardin et al., 2002).

It could therefore be expected that *ab initio* structure prediction, whereby the protein structure is predicted without resorting to homology with other proteins or to experimental data, should be a more feasible goal for membrane proteins than for soluble proteins (Fleishman and Ben-Tal, 2006).

Ab initio prediction is often divided into two components: devising a score function that can distinguish between native structures and non-native ones and the subsequent method of searching the conformational space. There are numerous approaches for evaluating *ab initio* predictions, including Simulated Annealing (SA), Molecular Dynamics (MD), Monte Carlo Simulations (MC) and Genetic Algorithms (GA).

Molecular Dynamics (MD) simulations of proteins and protein-substrate complexes provide a detailed and dynamic picture of the nature of inter-atomic interactions with regards to protein structure and function. The MD approach is computationally expensive and needs improvement in β strand and loop matching (Crivellie et al., 2002).

Monte Carlo (MC) simulations do not use forces but rather compare energies, via the use of Boltzmann probabilities and based on random walks, ignoring the information obtained in previous steps (Brunette and Brock, 2005).

Simulated Annealing (SA) usually considers solutions, their cost, neighbours, and moves, and can only have one population size one and make just one mutation per cycle, and so the key difference between SA and GA is that the SA creates a new solution by modifying only one solution at a time with a local move.

Genetic Algorithms attempt to improve on the sampling and the convergence of MC approaches and possess the following advantages:

- The GA considers individuals, their fitness, selection, crossover and mutation.
- The GA creates solutions by combining two different solutions and undertaking crossover, which is the key element that distinguishes the GA from hill climbing and simulated annealing.
- Easy implementation, requiring minimum mathematical effort.
- Ability to manipulate many parameters simultaneously.
- They can search a solution space for which the fitness landscape is complex or has many local optima.

- The GA is parallel and explores the solution to a problem in many directions at a time.

- This parallelism allows them to implicitly evaluate many schema at once and so they are suited to where the space of all potential solutions is truly extensive.

- If a GA is unable to deliver a perfect solution, it can at least deliver a good working solution.

The disadvantages of genetic algorithms include the following:

- The GA lacks a sound mathematical description that allows designers to calculate (and subsequently apply) the best optimisation parameters.

- The GA has the limitation of defining the representation of a problem which must be robust, and the appropriate fitness function, size of population, the rate of crossover or mutation must all be chosen with care.

- Premature convergence - if an individual that is more fit than most of its competitors emerges early on in the course of the run it may reproduce so abundantly that it drives down the population diversity too soon.

- The fitness function may prove to be deceptive, where the locations of improved points give misleading information about where the global optimum is likely to be found (Marczyk, 2004).

3.5 Electronic Database

There are two protein structure databases that are used extensively in this study in order to obtain information about protein sequence and structure.

3.5.1 Swiss-Prot

There are huge numbers of protein sequences, which need to be stored in an electronic database in order to extract different information for different users. SWISS-PROT is a manually curated protein sequence database, which is distributed with a large number of documentation files and contains over 200,000 entries from all organisms containing information such as sequence information, references, the biological source of the protein, function of the protein, secondary structure of the protein, transmembrane regions and similarity to other proteins.

It also provides for ready incorporation and integration with other databases (Bairoch et al., 2000). SWISS-PROT is free for academic users and available from http://www.expasy.ch/, and is becoming the standard protein sequence database of the bioinformatics field.

3.5.2 The Protein Data Bank

The Protein Data Bank (PDB) is a repository for the processing and distribution of 3D macromolecular structure data primarily determined experimentally by X-ray crystallography and NMR, and contains over 20,000 protein structures. A PDB file contains an organism name (header) with a brief description (title), equivalence between the PDB file and protein sequence databases and the conversion between these databases, stored in the DBREF line. It also includes details of regions of α-helix and β turns and sheets defined experimentally by X-ray crystallography and NMR (Helix and Sheet tags) information about each residue in the structure in three-letter code, the residue number, and the atom number and element, and xyz co-ordinates (ATOM) (Berman et al., 2000).

3.6 Force field

A force field is a set of mathematical functions designed to reproduce intermolecular interaction energies and intermolecular conformational energies, as accurately as possible, in order to discriminate between plausible and implausible conformations in a search space. In chemistry, a force field is defined as a potential function and in the context of molecular mechanics refers to the functional form and sets of parameters used to describe the potential energy of a system of particles.

Different force fields are designed for different purposes, for molecular dynamics of macromolecules and also for energy minimisation, such as in the case of GROMOS (GROningen Molecular Simulation Software).

GROMOS is a computer program (Gunsteren van et al., 1996), which has been developed for the dynamic modelling of molecules in the study of bimolecular systems. It has the following basic capabilities: the simulation of proteins or arbitrary molecules using the molecular dynamic or stochastic dynamic, energy minimisation of these molecules and analysis of molecular conformations obtained by experiment (X-ray, NMR), by model building or by other methods of computer simulation (Gunsteren van et al., 1996).

3.7 Summary

This chapter reviewed the three levels of protein structure. The α-helix as a secondary protein structure was described as the main unit of transmembrane protein structure. In order to find out how transmembrane protein structures are folded in the membrane, a number of interactions and forces, which exist between amino acids molecules in tertiary structure are defined.

The bacteriorhodopsin protein is introduced as a known transmembrane protein of well-determined structure and function. The importance of the biological functionality of membrane proteins and the imperative for the prediction of membrane protein structure for science and medicine is discussed.

Approaches used to predict 3D structure of protein from sequence are reviewed in order to compare and contrast the different methods. The evaluation of predicted conformations by calculation of free energy using force fields is described in its role as the foremost form of evaluation applied to molecular structures minimising the force field. Finally the different optimisation techniques that have been used for protein prediction are outlined.

CHAPTER 4

4 Design of a GA for predicting membrane protein structure

In this chapter, the design of a GA technique for the prediction of membrane protein structure is discussed. The rationale and theories that are used to design the candidate solutions in this will be discussed. The GA operators used for the effective exploration of the conformation search space and the selection policies applied in each generation will be described.

4.1 The design of candidate solutions

This system was developed to automatically predict the spatial arrangement of TM regions in membrane protein structures by utilising genetic algorithms (GA) as an optimisation technique.

The candidate solution is designed as a template of conformation for a given membrane protein, the focus of this study being the well characterised bacteriorhodopsin protein. The GA has been designed to work with membrane protein structures that contain alpha helical TM regions but not proteins that have any significant degree of beta-stranded structure such as the porin family. A given structure is specified by the position {x, y, z} of the atoms in the protein, defined by their Protein Data Bank (PDB) file.

In this work, Bacteriorhodopsin (BR) is selected as the native known structure to be tested, because high resolution electron crystallography has been extensively used to determine the 3D structure of the protein. Bacteriorhodopsin structures are imported as PDB files such as 1AP9, 1AT9, 1BRR or 1F88, etc.

The PDB file 1AT9 was selected in this study as it that contains Cartesian coordinates for a 7 helix bundle without any breakage in the helices. The interconnecting loops between the TM regions from the Bacteriorhodopsin PDB file are ignored (as demonstrated in figure 4.1) and removed for the purpose of free energy comparisons.

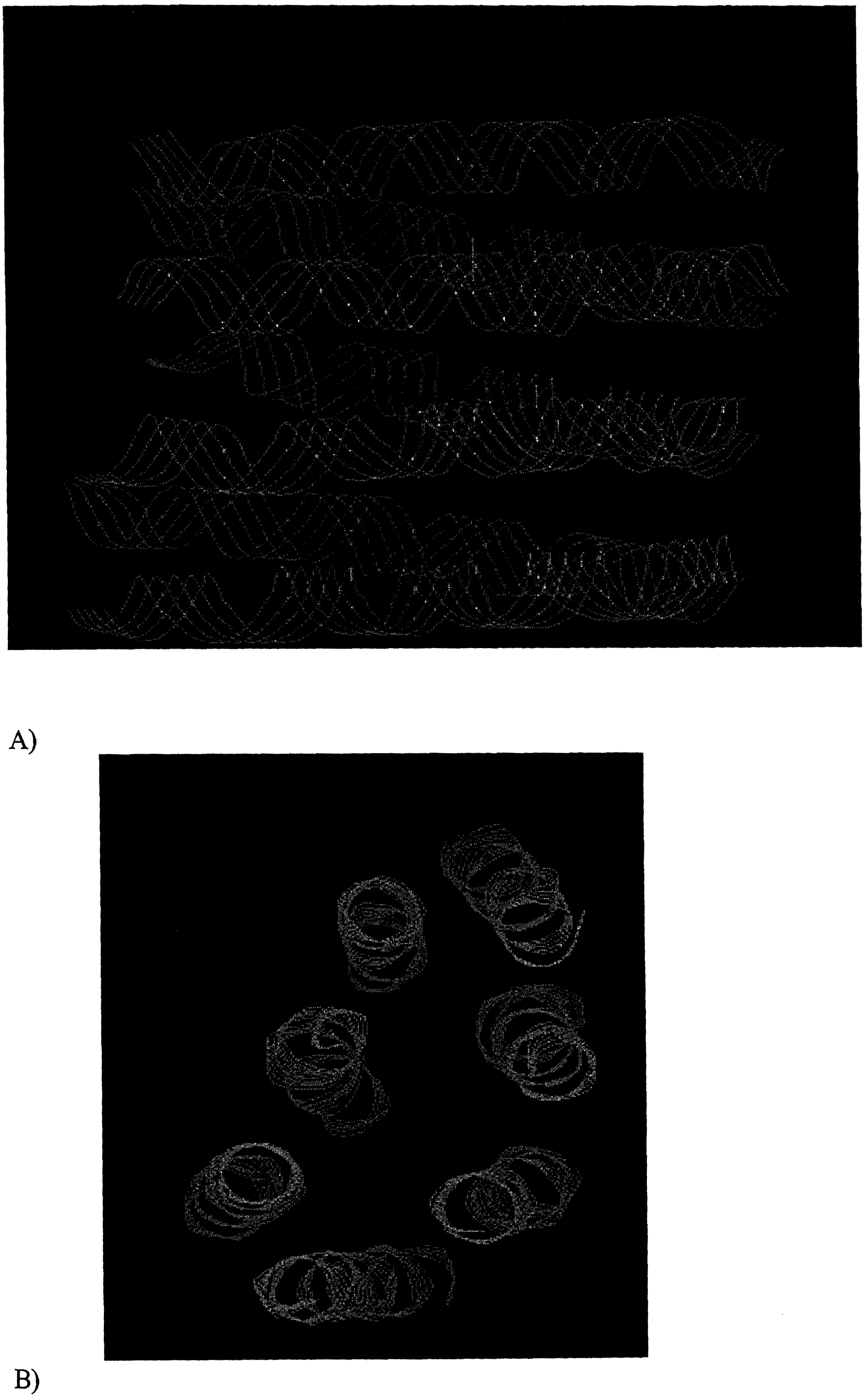


Figure 4.1 The bacteriorhodopsin structure without interconnecting loops, A) side-on B) end-on

In order to create a template, which is an approximation of the native structure, the average distance between pairs of TM regions was estimated and also their rotational orientation in relation to one another was approximated, along with an approximation of their general arrangement with respect to one another obtained by using the TMRelate program. In building this 3D template, the Psi (ψ) angle and Phi(φ) angle for the alpha helix are applied as the general angles. Each angle by rotation about two single bonds can adjust the structure of each amino acid in a polypeptide (figure 4.2)

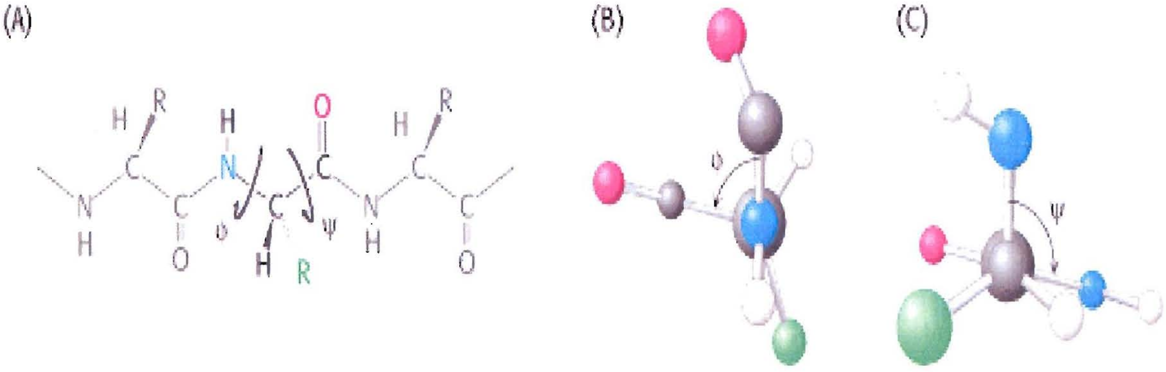


Figure 4.2 Molecular and Atom structure of Amino Acid (A) Phi (ϕ) is the angle of rotation about the bond between the nitrogen and the α-carbon atoms, whereas psi (ψ) is the angle of rotation about the bond between the α-carbon and the carbonyl carbon atoms. (B) A view down the bond between the nitrogen and the α-carbon atoms, showing how ϕ is measured. (C) A view down the bond between the α-carbon and the carbonyl carbon atoms, showing how ψ is measured (Berg et al., 2002).

The template consists of an arrangement of 7 alpha carbon atom starting positions, the starting atom of the initial residue of each helix (figure 4.3). Each Cα atom is represented by three axis coordinates (x, y and z) that the value of y is based on the membrane depth and the rise of approximately 1.5 Å per amino acid along the vertical axis of helix in order to generate each helix running straight through the membrane from one face to the other. This y value is calculated in the TM-Builder program and inherited by the TMGA program, which is involved in manipulating the other two axis coordinates (x,z) and the rotation angle (r) as input data for TM-Builder.

The starting location of Cα atoms in the template is based on the average distance between helices in the bacteriorhodopsin structure. The position of each Cα atom can be varied within a certain range in different directions, but in a way that maintains the distance between TM regions within the defined range thereby avoiding unravelled structures, or the problem of side chain conflicts, where side chains on adjacent transmembrane regions are predicted to occupy the same space.

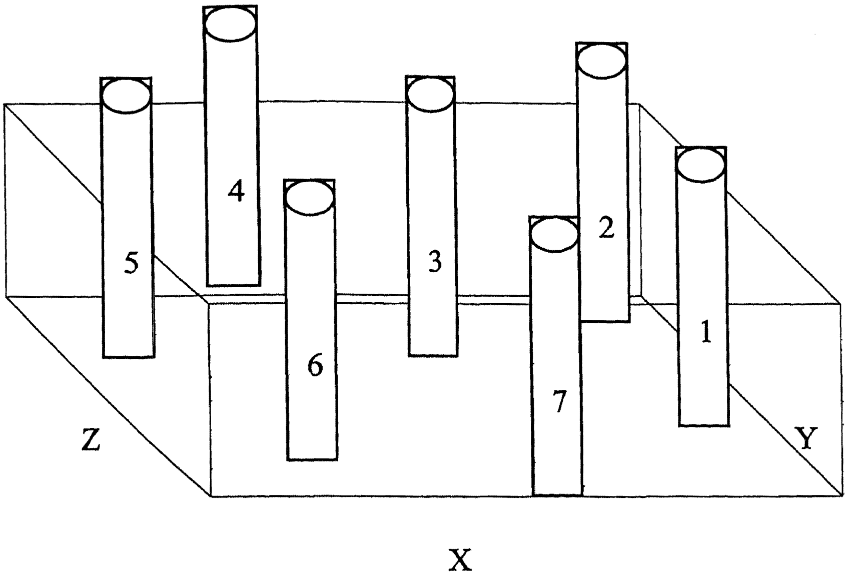


Figure 4.3 A created template for TMGA system that contains a segment of cell membrane with an approximate arrangement of 7 TM helices (the external loops between TM region are not considered).

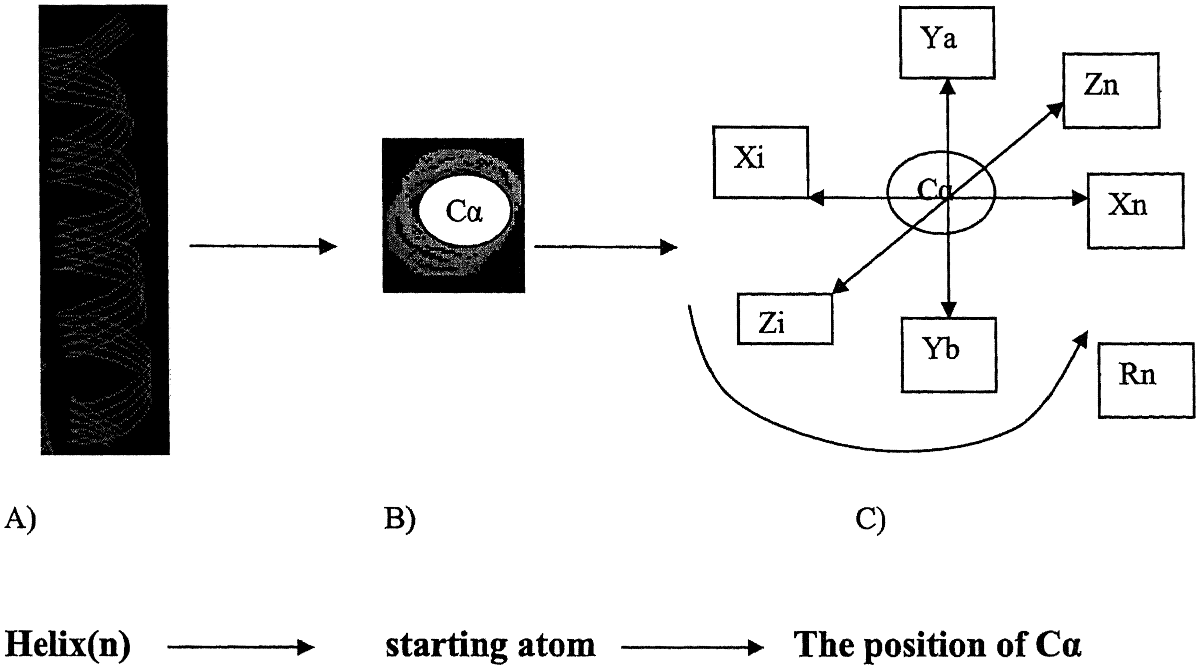


Figure 4.4 TM region structure A) Representation of the structure of the helix B) End-on considering the starting atom of the initial residue C) Representation of the end on position of a starting Cα atom within a certain range along two axes (x,z) as y value is constant with a rotation angle value(r).

As figure 4.4 shows, the position of each Cα atom is based on three axis coordinates. The y value will be constant for all starting Cα atoms and the subsequent y values are estimated in the TM-Builder program. The GA method is based on manipulating the distances between TM regions in the membrane by manipulating the x and z value and also performing a rotation of each helix by choosing the r value for each Cα atom in order to bring about different spatial associations and stabilising interactions between helices.

4.2 Choice of random method

The TMGA system is based on the GA technique. One of the qualities of GAs is that they can be applied to unknown problems and the generation of solutions do not rely on prior knowledge. In this work, the GA randomly generated two axis coordinates (x,z) and rotation value (r) for each alpha carbon atom and created a string of integer numbers for each candidate solution. The x and z axis coordinates provide the location of each TM region in a template and they need to be specified in a certain range in order to maintain the average distance between helices.

The range of each variable (x and z) is based on the average distance between helices in the bacteriorhodopsin protein. The range of the r coordinate is based on the rotation of each amino acid in a helix and varied from 0 to 359°.

Therefore a function was designed in the TMGA system that randomly generates an integer number within a specified range. All the integer numbers in the specified range have equal probability of being generated.

The GA is able to manipulate the position of each helix as well as the rotation of the helices by making random changes to the variables and by using the fitness function to determine if there is an improvement in the new conformation.

4.3 Encoding method

The approach of encoding the GA was based on using a bit string to encode each particular solution of the problem domain. This is inefficient in domains described by several parameters that might each require a range of values. In this study, predicting the 3D structure of membrane proteins is described by calculating coordinate values for each residue. These values have to be set in a specific range for each residue in order to generate realistic distances between adjacent helices. In this application, the genetic algorithm is configured to operate on numbers, not bit strings as in the original genetic algorithm, and the so-called hybrid approach is taken.

A hybrid representation is usually easier to implement and also facilitates the use of domain specific operators. However, there are some disadvantages, such as:

- The mathematical foundation of GA holds only for binary representations
- Binary representations also run faster in many applications
- An additional encoding / decoding process may be required to map numbers on to the bit strings.

The program used the "black box" technique (Goldberg, 1989) and set a bank of 21 input switches, three switches to be considered for each helix, as in this work, a membrane protein structure with seven TM regions was selected.

For every setting of 21 switches, there is an output of predicted transmembrane region structure. In theoretical terms, the spatial arrangement of seven TM regions as helix wheels is considered as a TM protein structure template. Each template defines an individual in the population.

The objective of the problem is to set switches to obtain the r value and the x and z coordinates for each helix wheel in order to predict an energetically feasible membrane protein structure, which is approximate to the native structure. The black box in the TMGA system is designed with 21 switches, which will be altered within a certain range for each TM region (figure 4.5).

A simple code is generated by considering a string of 21 integer values, where each of the 21 switches is represented by one integer value. This method uses blind selection in order to generate possible solutions for the search space. As the problem is specified in the black box, this technique avoids generating unwanted solutions and which can be time consuming.

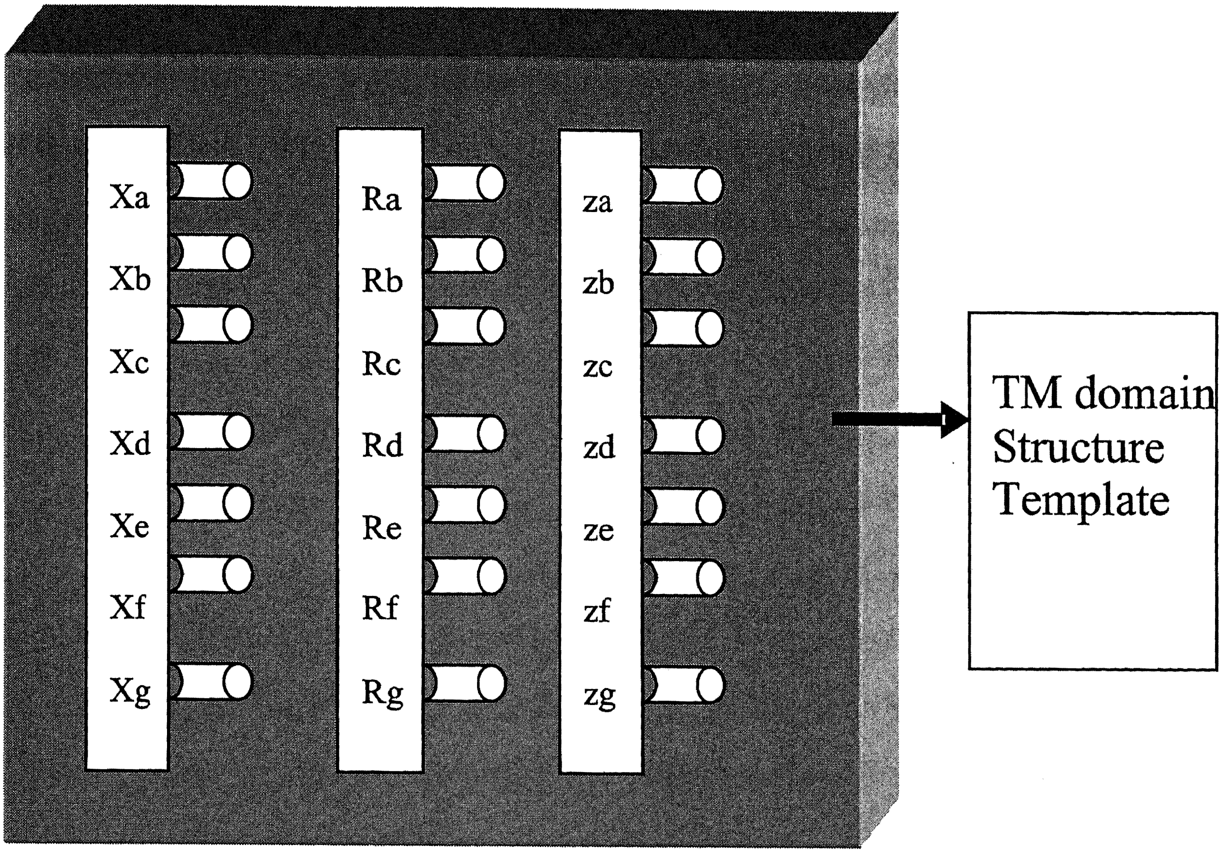


Figure 4.5 A black box optimisation problem with 21 switches illustrates the idea of generating a structure template for the transmembrane region of a 7 helix membrane protein.

4.4 GA data structure

The data structure of the TMGA system consists of a population of individuals or solutions, which are represented as a prediction of the 3D structure of a membrane protein. Each individual has a set of characteristics. The device that encodes the characteristics of that individual is represented as a genotype.

Each chromosome consists of a number of genes and the value of the gene is known as an allele that represents an integer value. All the genotypes within the population have the same number of chromosomes, the same number of genes in each chromosome and the same number of alleles in each gene (figure 4.6).

The functions performed namely selection, evaluation, crossover and mutation operators are part of the genetic data structure. The parent selection is performed on the population of individuals, and crossover and mutation operators are performed at the chromosome level.

As shown in figure 4.6, the TMGA system generates a population of offspring. Each offspring represents a 3D structure of a membrane protein. The genetic algorithm operates on the two dimensional coordinates of the starting atom of each TM region (as the y axis value will be constant) and also rotation angle value. There are several advantages to working in a space defined by Cartesian coordinates.

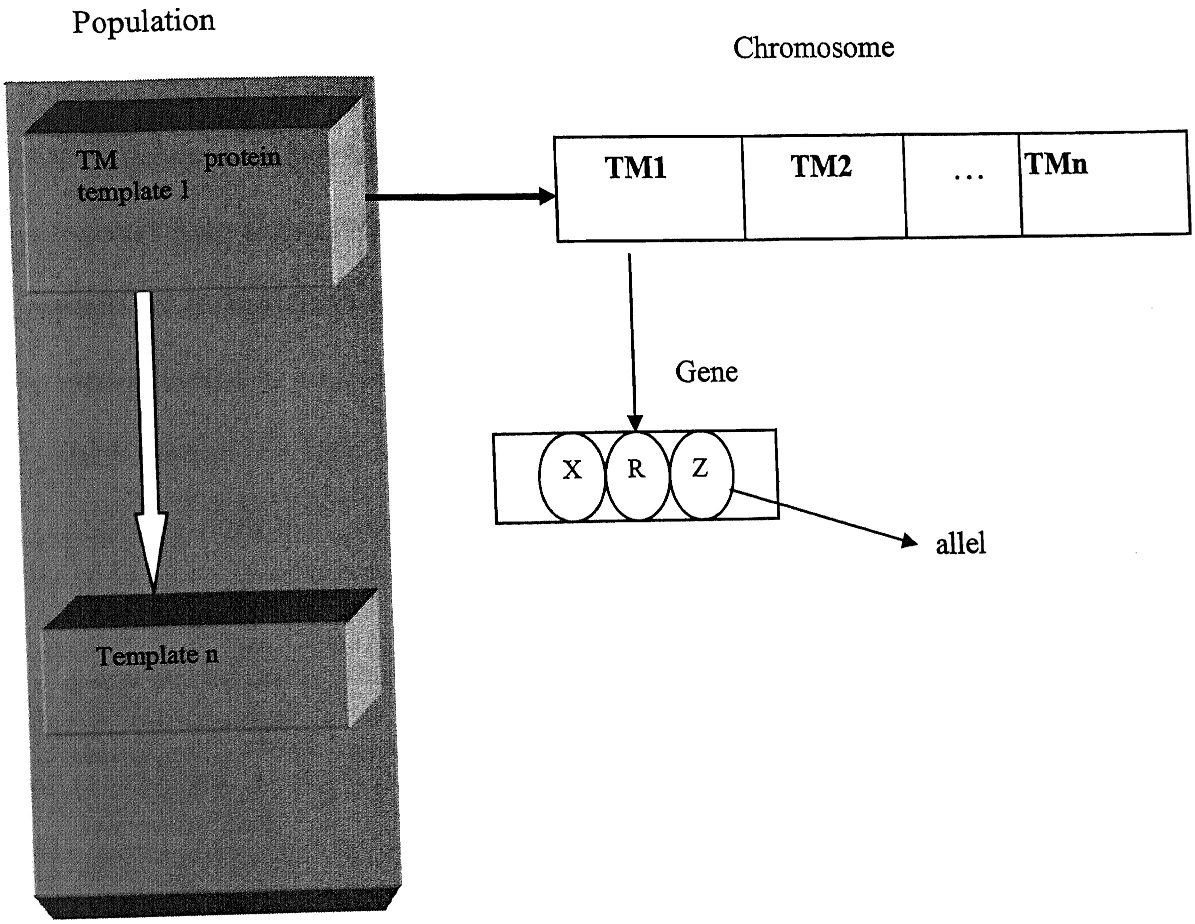


Figure 4.6 Representation of the population of solutions, considering an individual as a chromosome that consists of number of genes and each gene consisting of three alleles.

The first is that the initial distance between two TM regions can be set up in terms of considering the average inter-helical distances from determined structures, and then applied to the TM protein structure template in the system. The second is that the GA is able to optimise the rotation of each helix as well as the distance between adjacent helices. The third advantage lies within the simplicity of constructing the TM protein itself.

4.5 Fitness function

In order to test and visualise the predicted transmembrane protein structure, the SwissPDBViewer software was used. This software calculates the free energy, and performs energy minimisation between the atoms of the amino acid side chains of the protein loaded as a PDB file, through using an external program, the GROMOS force field (Gunsteren and Berendsen, 1987). These force fields or potential energy functions allow the energy of a structure to be evaluated as well as repairing distorted geometries. The energy functions contain the bond, angle, torsion and non-bonded pairs or external terms that include the electrostatic and van der Waals terms (MacKerell, 1998). There are many parameters considered in the energy functions that even individually can greatly affect the calculated energy.

The aim of this evaluation is to manipulate the conformation of a protein in order to find the minimum free energy function for the predicted 3D structure of a membrane protein, which is a measure of the stability of the protein structure.

In this work, the calculated energy of the native bacteriorhodopsin structure (1AT9) with the extramembrane loops removed is estimated and is compared with the energy of the predicted membrane protein structure, in order to eventually find the 3D structure of the transmembrane portion that is the nearest to the native structure.

4.6 Choice of reproduction method

The genetic algorithm focuses on the most promising parts of a solution space in order to direct the combination of strings containing good partial solutions. This process is introduced as a reproductive method. The reproduction contains three processes: selection, crossover and mutation.

4.6.1 Selection

The selection operator is based on the Roulette Wheel technique. The roulette wheel selection technique involves a random selection, as of a number slot from a roulette wheel. Fitter individuals of a population are represented by a higher number of slots compared to other individuals with lower fitness. Hence the process of selection is random, but there is greater probability that fitter individuals will be selected more often. After the first parent is selected using the roulette wheel selection, a second one is selected, but making sure that it is a different individual to the first one so that processing time is not wasted on reproducing from identical parents.

In our system, the process for selecting a different second parent makes sure that parents with the same energy value are not selected together. The selection operator is designed to discard those solutions with higher energy value from the population so those solutions with lower energy values will have more chance to be selected for the next generation.

4.6.2 Crossover

The crossover operator used in this work is the basic crossover. Past work by Konig and Dandekar suggests that use of two point crossover leads to lower energy structures (Konig & Dandekar, 1999). The crossover in this system operates at two points along the chain of two axis coordinates, and the rotation value starting points for each TM region. As shown in figure 4.7 the arrangement of TM regions is varied by manipulating the three values. The crossover operator is employed to randomly place each cross point along the chain where a TM region is ended and the next TM region is started.

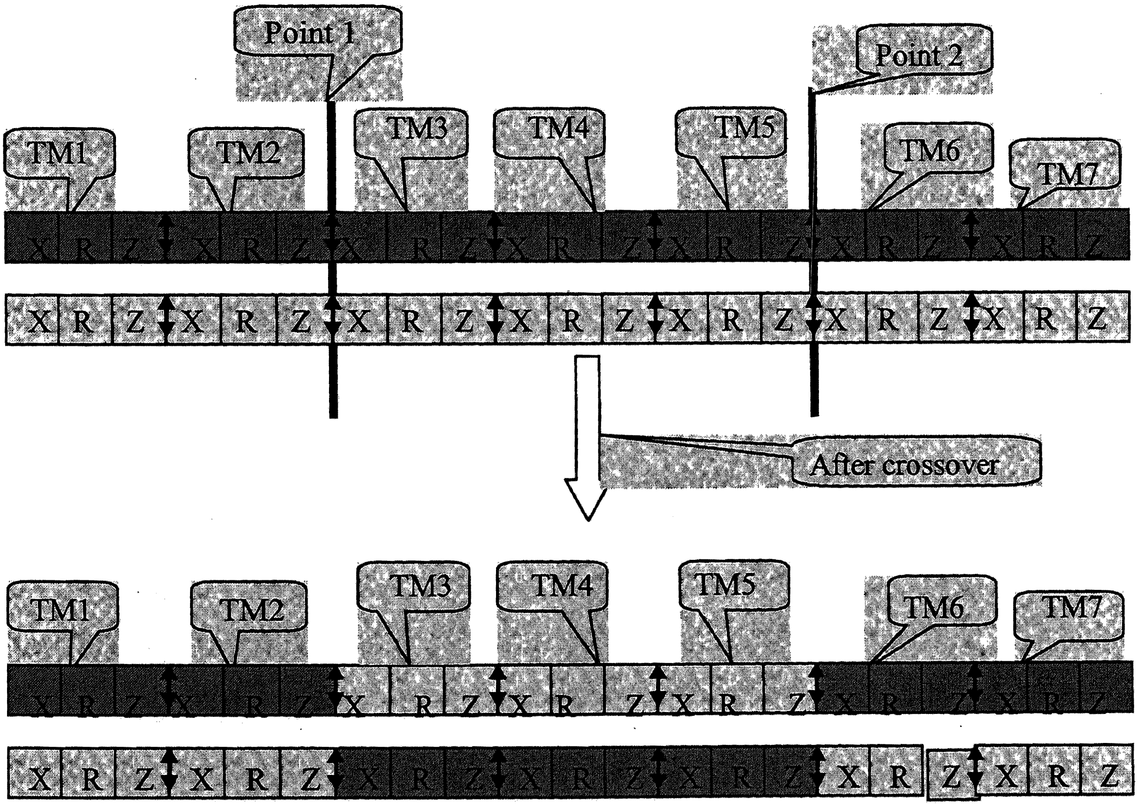

Point 1
Point 2
TM1
TM2
TM3
TM4
TM5
TM6
TM7
X R Z X R Z X R Z X R Z X R Z X R Z X R Z
X R Z X R Z X R Z X R Z X R Z X R Z X R Z
After crossover
TM1
TM2
TM3
TM4
TM5
TM6
TM7
X R Z X R Z X R Z X R Z X R Z X R Z X R Z
X R Z X R Z X R Z X R Z X R Z X R Z X R Z

Figure 4.7 Crossover with two cross points.

4.6.3 Mutation

The mutation operator used is the basic mutation but there are differences in the way the mutation operator is applied in different experiments. In this application, three point mutation is used. As shown in figure 4.8, the arrangement of the two axis coordinates and the rotation value for each offspring can be modified by the mutation operator. The system is able to change the arrangement of the three values for each offspring selected for mutation. This allows the distance between TM regions to be randomly altered by mutating x and z and by also changing the rotation for different helices.

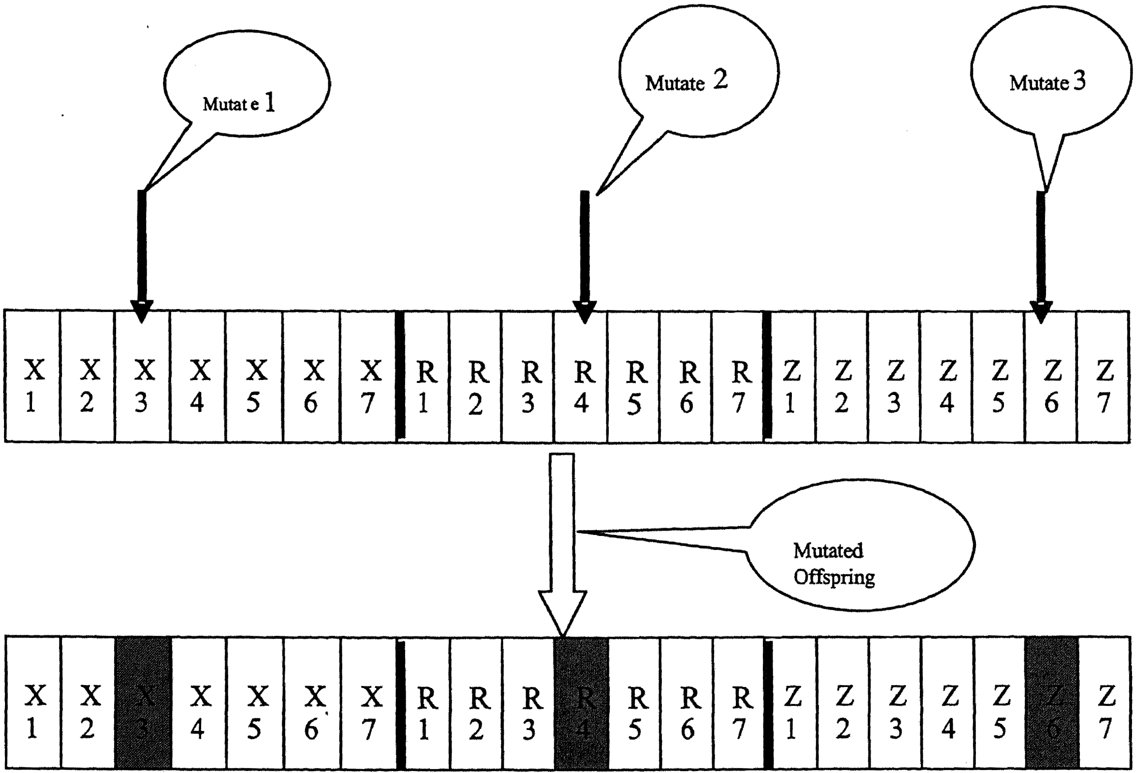

Mutat e 1
Mutate 2
Mutate 3
X 1 | X 2 | X 3 | X 4 | X 5 | X 6 | X 7 | R 1 | R 2 | R 3 | R 4 | R 5 | R 6 | R 7 | Z 1 | Z 2 | Z 3 | Z 4 | Z 5 | Z 6 | Z 7
Mutated Offspring
X 1 | X 2 | X 3 | X 4 | X 5 | X 6 | X 7 | R 1 | R 2 | R 3 | R 4 | R 5 | R 6 | R 7 | Z 1 | Z 2 | Z 3 | Z 4 | Z 5 | Z 6 | Z 7

Figure 4.8 Mutation with three points.

4.7 Stopping criteria

There are a variety of criteria that may be used to decide when a GA is terminated that are situational and set according to individual performance. The criteria are as follows:

1. The aim of a given genetic algorithm is to find a solution for a given problem. So when such a solution is found, the genetic algorithm stops.

2. When the best genotype (solution) has been reached or the target solution fitness has been surpassed (smaller than that fitness for minimising fitness function and larger for maximising fitness function), the GA terminates.

3. The GA is stopped when the difference between the best and worst genotypes in the population becomes smaller than a specified percentage of that same difference in the initial population.

4. The termination may occur after the maximum number of generations or the maximum time, which are specified by the algorithm.

5. The GA is stopped if there is no improvement in the best fitness function value

The difference between the average energy in each population becomes smaller than between the early generations and gradually diminishes until there is no further change.

In the algorithm used in this study, the process is stopped when there is no significant improvement in the calculated energy force fields over a given number of generations. In this experiment, the given number of generations was five and the system been determined after that.

Summary

The design of the proposed TMGA system has been presented as a series of functions and operators. The rational concepts that are used to apply the GA technique in this system have also been discussed. Enhancements were made to the basic GA method in terms of the design of candidate solutions to the problem of the prediction of membrane protein structures.

The candidate solution is designed to represent the arrangement of TM regions in membrane. The loops between helices are ignored in order to manipulate the position of each TM region in the membrane and concentrate upon improving the computational efficiency of the system.

The energy force field is utilised for the evaluation of each candidate solution. The GROMOS force field is applied to calculate the free energy. The application of GA operators for effective exploration of the conformation search space is described that includes the selection of solutions for the next generation. The selection method is based on the Roulette Wheel technique. The process of selection is designed in a way as to provide a greater chance for fitter individuals to be further subjected to the GA operators. The GA operators used to generate new solutions include crossover and mutation.

CHAPTER 5

5 Implementation of TMGA system

The TMGA system was developed in order to address a specific problem that it is at the interface between computer science and biological systems, known as bioinformatics or computational biology. This area projects biologically oriented concepts such as genetic algorithms into computer science and adapts computer science algorithms like pattern matching and computational geometry to addressing real biological problems.

The TMGA system is intended to predict the 3D structure of membrane proteins and also provide a framework for the improved application of the genetic algorithm technique to structural biology problems. This is accomplished through consideration and manipulation of the distances between TM regions in membrane proteins, and their orientations in relation to one another, which were varied and constrained within the conformation search process. This ability to manipulate and test the resultant orientation of such structures in predicted membrane protein structures is a significant step along the pathway to tertiary structure prediction for membrane proteins of unknown structure.

The following sections of this chapter introduce the software development in the TMGA system. Section 5.3 describes the acquisition of data and section 5.4 will present an overview of the TMGA system representative units and subunits and specific PDB software, which work with the TMGA system. In section 5.5, the output is explained and a summary is given in section 5.6.

5.1 Software development

Outlined below are the steps taken in this research in developing the software to its current level. The experimental genetic algorithm is implemented using a PC, which is easy to use, and a common computer system. The TMGA (Trans Membrane Protein Genetic Algorithm) system is implemented using Visual Basic (as part of the Microsoft visual studio development environment).

This enables the system to utilise the benefits of the Windows 95/98 operating system with features such as 32 bit variable addressing, continuous memory model architecture, colour depth desktop management, event driven processing, industrial standard user interface and the facilities for multi-threading and multi-tasking. It is a complete development environment for building Internet applications.

Visual Basic 6.0 is a powerful tool for developing applications in the Basic programming language with the ability to control the most powerful databases such as Microsoft Access 97, part of the MS office suite, which is a Windows, based database system (Microsoft Access is used to provide a knowledge based software for the TMGA system).

Object oriented design was chosen for development of the software tools as opposed to a functional approach, because it is considered to produce more maintainable and more easily understood system architecture and code.

5.2 Overview of the TMGA system

The major goal was to predict 3D structure of TM proteins by using a genetic algorithm. The TMGA system is based on the GA technique that is a good method for solving problems where the range of combinations of parameters is so large that it is impossible to search comprehensively.

One of the features of a GA is in the way data are manipulated. The structure of data takes the form of a population of membrane protein structures. The amount of data that needs to be processed by a GA should be kept to a minimum, so that the behavioural complexity of the system can be managed effectively.

The TMGA system needed specific items of data in order to implement the GA. The following inputs are selected for each experiment:

- Offspring; the representation of TM protein structure
- Length of offspring; representing the number of TM regions
- Population size; generating the number of offspring
- Phenotype; defining a solution for the specified problem in each experiment
- Fitness Function; providing a method to evaluate the quality of each offspring
- Fitness; representing a value for the quality of an offspring
- Selection; estimating the quality of each offspring
- GA operators; indicating the type of crossover and mutation and specifying the probability of each operator in different experiments.

The data selected are varied in order to provide different results (possible solution) for each experiment. Also, TMGA is able to change the behaviour of the GA by altering the parameters listed above for optimisation of results.

5.2.1 Acquisition of data

An important aspect in describing this study is to specify the methods used to acquire data. In chapter 4, the predicted template was described. Each helix is assumed to run straight through the membrane from one face to the other and the location of TM regions in a template can be changed within the specified range in order to keep the distance between TM regions at a realistic value and avoid unfeasible models of 3D structure of TM proteins being generated during the of running the GA program.

This value has been calculated by considering the average distance between TM regions in a dataset of 25 TM proteins of determined structures, using the TMDistance program as described below. In the TMGA system, the distance between helices is set by using the atomic coordinates for the alpha carbon of the starting residue in each helix.

The TMDistance program (Togawa et al., 2006) reads the spatial co-ordinates from PDB file for the atom in each TM region and was used to calculate the distance between residues located on adjacent TM regions in 25 different membrane proteins. With each residue pair in different TM regions, if the distances between the side chains of two residues is less than a user-selected distance (3.0Å, 3.5Å 4.0Å, 4.5Å or 5.0Å), the relevant residue-pair is added to the internal bi-dimensional array (matrix counter).

After all the PDB file(s) are read, *TMDistance* creates the 20x20 association matrix output with the average distances represented in the internal bi-dimensional array. The distances between the alpha carbon backbone atoms of the proximal residue pairs were then calculated and recorded, resulting in the working average of 8 Angstroms.

The allowable synthesised 3D coordinate values are then stored in a database and allow the GA program to select values from particular ranges which are illustrated in table 5.1. This approach ensures that the entire structural template can be predicted by the GA program. These 3D coordinate values are used to create a population of "chromosomes" upon which the GA technique is run.

Axis	X	R	Z
helix			
TMa	-3<x<3	0<r<359	-3<z<3
TMb	-14<x<-8	0<r<359	-3<z<3
TMc	-20<x<-14	0<r<359	4<z<10
TMd	-27<x<-21	0<r<359	12<z<18
TMe	-24<x<-18	0<y<359	25<z<31
TMf	-9<x<-3	0<r<359	20<z<26
TMg	-4<x<2	0<r<359	11<z<17

Table 5.1 TM region parameters employed in TMBuilder, demonstrates the range of X and Z coordinates values (R is fixed) starting for each TM region for helical transmembrane protein such as bacteriorhodopsin. The R value relates to the range of starting rotation angles available for each helix.

5.3 Design of the TMGA system

The TMGA system is comprised of two units that fulfil specific tasks. Figure 5.1, shows the flow of information between the different units. The main unit is represented as the TMGA engine that contains three subunits, these are GA data structure, data storage and GA tools, whereas the second unit is the PDB software. The complete TMGA system enables the TMGA engine and PDB software to cooperate with each other.

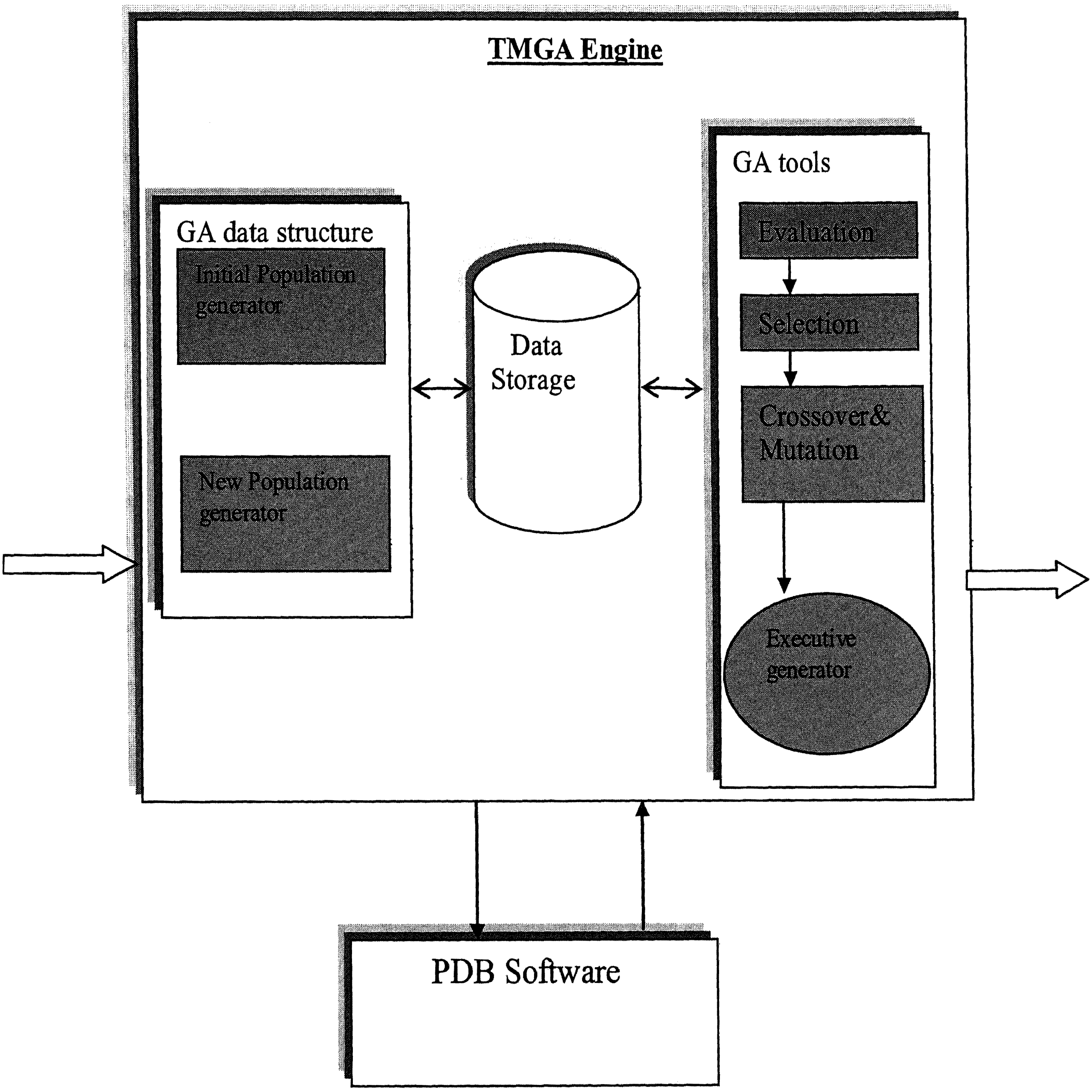


Figure 5.1 Design of the TMGA system.

5.3.1 TMGA engine

The TMGA engine is the main component of the TMGA system and is based on a standard GA method and varies GA parameters in order to achieve the best result. It is composed of three elements: data storage, GA data structure and GA tools. Before the TMGA is started all its parameters need to be set and validated so that the genetic algorithm executes correctly without causing any errors.

The data storage unit is responsible for storing the data acquired from the user and the data generated by the TMGA system. In this system the data storage unit utilises Microsoft Access to manipulate the data. The GA data structure subunit contains two processes. The first process carried out by the initial population generator which is responsible for the generation of an initial population.

In this process, a random function is used in order to select x, z and r values coordinate from the data storage for each helix and then these values are stored. The second process is carried out by the new population generator and creates a new population for the next generation. The output of these two processes is a text file that represents the population of predicted templates. The GA data structure subunit is responsible for making the connection between the data storage subunit and PDB software unit. The GA tools unit utilises GA operators in order to evaluate and select individuals with the best fitness and to perform crossover and mutation and also is able to execute other software to generate PDB file which is carried out by the next unit. The PDB software unit contains three software elements, which cooperate with the TMGA engine unit in order to calculate the potential energy and generate a PDB file for the predicted membrane protein structure.

5.3.2 The processing algorithms

The TMGA system utilises a modified version of a standard GA as a method of predicting new TM protein structures applying the following stages of operations:

- Initialise the random population
- Evaluate the potential energy of each new TM protein structure
- Convert the energy to an evaluation of fitness
- Select the individual with energy less than 10000. This value was selected as cut-off because The calculated free energy cut-off of 10,000 kJ/mol was applied in Order to impose a degree of selective pressure by eliminating outlying individuals that possess free energies that are more than two orders of magnitude (2 x log10) away from being energetically feasible in life. Elimination of these individuals with extremely high energy values resulted in earlier successful termination of the GA and improved free energy values at termination.
- Perform reproduction/ crossover/ mutation
- Generate a new population
- Repeat from second step until a new structure is no longer generated

The initialising operation in the system that benefited from a pre-processing stage included the generation of a range of x and z axis coordinates and rotation values for fixing the starting atomic coordinates of each helix. A degree of pre-processing needs to be performed in order to reduce the CPU time required to process the data and hence reduce the overall computational cost. The x and z coordinates values were pre-processed and the range of 3D coordinate values for each starting atom of the helix were stored in a database. The database was connected to the program for using these data and storing each population of predicted membrane protein structures. A given predicted membrane protein structure is represented in the GA program as a string of 3D coordinate values for each TM region. Finally, the MaxSprout program (Holm and Sander, 1991) was applied to the carbon backbone structures. MaxSprout constructs residue side chains in the appropriate positions according to a database of structures. The MaxSprout algorithm is described more fully on page 97.The integration of these data in the database is shown in appendix A. The following is a summary of the algorithm illustrating the processes carried out by the TMGA engine unit in being used to apply a GA technique in the prediction of membrane protein structure.

```
Step 1  Initialise the population of predicted membrane protein structures

Step 2  end of loop

Step 3  Generate PDB text file

Step 4  Execute TM-Builder software

Step 5  Execute MaxSprout software
```

The results of this stage are transferred to the SwissPDBViewer software in order to calculate the potential energy of each predicted membrane protein structure.

The next operation provided the fitness evaluation that benefited from pre-processing and included a method of inverting the calculated energy and storing the result for each individual in a database. The following approach illustrates the method used to invert the energy value for each individual;

1) Load force field value for all individuals in database

2) Organise by ascending force field values and select highest value in order to invert the highest number to become the lowest and vice versa, by subtraction of a given value from the highest energy value attained.

3) Calculate the final energy value for each individual

4) End of process

This inverted energy value is used to evaluate the fitness of an individual, and guides selection for the next generation in the GA program.

Two further processes, namely crossover and mutation are performed in the operation of the GA. The following algorithm summary illustrates the GA tools unit developed in the software:

```
Step1   randomly Select two individuals from the solutions (read from data storage)

Step2   Operate two point crossover

Step3   Select an individual of solutions (read from data storage)

Step4   Operate three points mutation

Step5   End of loop

Step6   Generate PDB text file

Step7   Execute TM-Builder software

Step8   Execute MaxSprout software
```

The experimental parameters are summarised in table 5.2. The parameters described were adapted especially for addressing the problem of membrane protein structure.

Table 5.2 GA parameters employed in the experiment runs.

Representation	Integer number string of Cartesian coordinates
Population size	It is set at either 50 or 100
Fitness	Evaluation based upon the potential energy of each TM protein structure
Selection	Roulette wheel technique And choose all individuals with energy values<10000 and others will be discarded
Crossover	Two points cross over and crossover rate = 0.8
Mutation	It is varied from two points to three points mutation with mutation rate = 0.02
Termination	When the lowest energy is successively observed without change over several generations

A valid range for energy is set in order to select the predicted membrane protein structures that are nearest to native structure.

In this study, the system is instructed to check the energy of all individuals in each population. If the minimum energy observed does not improve after three generations, the GA will be terminated.

5.3.3 PDB software

To develop an automated approach suitable for the prediction of membrane proteins, the GA approach is finally executed through 3 component pieces of software that together convert the starting coordinates and orientations generated by the GA into 3D TM helical structures. These are the as TMBuilder program (Togawa, et al., 2002), MaxSprout (Holm and Sander, 1991) and the energy calculation protocol of the Swiss-PDB Viewer (Guex and Peitsch, 1997). Here, each package is described briefly as follows:

TM-Builder

The TM-Builder program is used to generate a PDB file for each predicted TM protein structure, by generating 3D structures for each individual helix in turn, when taken together provide the structure of the whole TM domain of the protein. The atomic coordinates of the alpha carbon for the starting residue of each helix are used in the placing of each helix. The output of the TM-Building software is a PDB format data file that contains Cartesian coordinates for the alpha carbons of all the helical residues. At this stage, this PDB file will represent a new predicted TM protein structure but without considering the side chains of residues. The approach assumes complete helical structure for all TM regions, which was a necessary simplification at this stage. Using this information an atomic co-ordinate for the alpha carbon of each residue is calculated, according to the mathematic representation of an alpha helix, described below.

The program builds each alpha carbon in the corresponding angular position as it would be illustrated in a helix wheel representation.

For the calculation of α-helix co-ordinates, the following parameters are considered:

Value of Y: The y axis is taken to run essentially straight through the membrane from one face to the other. The value of y is based on the membrane depth, and the rise of approximately 1.5 Å per amino acid along the vertical axis of the helix. This value has been refined following sampling of a series of unbroken TM α-helices. Based on the average distance covered by 5 full turns of the helix (3.6 amino acids per turn), it has been calculated at 27.14 / 18 = 1.5. An average of a sample of alpha helices from 20 membrane proteins in the PDB database where the distance was measured over 18 amino acids in continuous alpha helices.

The value of X and Z: The X and Z coordinates for each helix describe the rotation around the central axis (or spindle) of the helix, but of course, are 90° out of phase with each other. The alpha carbon (CA) of each residue is assumed to be placed in a circle around the central spindle (y axis), with a 100° increment in the rotation angle (RA) for each residue, corresponding to the known periodicity of 3.6 residues per helix

For the x co-ordinate:

$$x = r*SIN(RA*\pi/180))$$

Equation 3. Formula for X coordinate

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

$$r = \text{radius of the helix} = (3.817719/3.6478)*(((3.817719*3.6))/\pi)/2$$

Equation 4. Formula for Radius of the helix

again obtained from sampling helical structures. In the event that the helix has a starting end on rotation (R) which is manipulated by the GA, then the equation is:

$$x = r*SIN((R+RA)*\pi/180)$$

Equation 5. Formula for X coordinate with end on rotation

Similarly, for z

$$z = r*SIN((R+RA-90)*\pi/180)$$

Equation 6. Formula for Z coordinate

The only difference between the x and z equations being the prior subtraction of 90° from the rotation angle. The starting coordinates for each helix are generated by the GA. The program generates pdb files as an output.

The pdb file consists of a list of CA atom. The atom records present the atomic coordinates for standard residues. They also present the occupancy and temperature factor for each atom.

MaxSprout

The second package involves building on the side chains to the alpha carbon backbone for each predicted TM region. The MaxSprout software is available through the Internet (http://www.ebi.ac.uk/maxsprout/) and uses the PDB coordinates of the alpha carbon backbone as input data and generates a PDB file for the predicted membrane protein helices with side chains added (see Appendix A).

MaxSprout is a fast database algorithm for generating side chain coordinates from a C (alpha) trace where there are gaps in the backbone (though this is not applicable here). When gaps are present, the backbone is assembled from fragments taken from known structures. Side chain conformations are optimised using a rough potential energy function to avoid clashes (http://www.ebi.ac.uk/maxsprout). The input must be in PDB format and the program generates a PDB file with side chain coordinates.

SwissPDBViewer

The third package is used to calculate the free energy for each predicted structure. The Swiss-PDB viewer program takes the PDB file with the side chain coordinates as input (see Appendix A). The software automatically implements a script, which is written to calculate the potential energy of each predicted membrane protein structure.

Deep View (formerly called Swiss-PdbViewer) is a user friendly but powerful molecular graphics program. It is designed for use with computing tools available from the expert protein analysis system, or ExPASy (http://www.expasy.org/spdbv/).

DeepView is simple to use for viewing structures and it can detect hydrogen bonds within proteins and between proteins and ligands and also examine electron-density maps from crystallographic structure determination, to judge the quality of maps and models, and to identify many common types of problems in protein models. It allows the user to view several proteins simultaneously and superimpose them to compare their structures and sequences. It computes electrostatic potentials and molecular surfaces, and carries out energy calculations and energy minimisation

DeepView or Swiss-PdbViewer is not in itself a force field calculating program but a tool that is most commonly used to apply a protein primary sequence onto a 3D template and obtain immediate feedback of how well the threaded protein will be accepted by the reference structure before submitting a request to build missing loops and refine side chain packing.

The program was written and tested by the Swiss group of Nicolas Guex, Alexandre Diemand, Torsten Schwede and Manuel C. Peitsch (Guex and Peitsch, 1997).

Swiss-PdbViewer includes a version of the GROMOS 43B1 force field calculation (Gunsteren et al., 1996). This force field allows the evaluation of the energy of a structure as well as repairing distorted geometries through energy minimization. In this implementation, all computations are done *in vacuo*, without reaction field.

5.4 Output

The output of the TMGA system is divided into three parts: starting coordinates file, PDB file without side chain and PDB file with side chain.

All the files are stored in a directory to allow the user to compare the results of different generation of the predicted membrane protein structures.

The starting coordinates file contains a list of x, y and z coordinates for each starting alpha carbon atom for each TM region in the predicted membrane protein structures. The text file is generated with a specified format, to be read by the TM-Builder software.

The output of the TM-Builder software represents a new predicted membrane protein structure in PDB file format. This output consists of predicted coordinates for the alpha carbon backbone of the TM helices.

The PDB file with side chains is an output of the MaxSprout software, and consists of the same structure as generated by TM-Builder but with the side chains constructed on the backbone, and is utilised by SwissPDBViewer software in order to calculate the energy of each predicted membrane protein structure.

5.5 Summary

This chapter has addressed many of the technical issues associated with the implementation of the TMGA system.

Firstly, the methods used to acquire data in order to generate the entire template for the GA program were discussed. The components of the TMGA system were defined. This provided a description of all the processes that are used to implement the GA technique in order to predict 3D structures of TM proteins.

To improve the genetic algorithm technique for the specific problem of prediction of membrane protein structure, a number of enhancements were developed. These included the design of the selection method, calculating the potential energy and the performance of modified GA operators for generating populations of predicted structures. The incorporation within the TMGA system of other programs enables the system to generate PDB files for each predicted TM protein structure and calculate the free energy in order to evaluate the new predicted structure, in comparison with the native structure. The output of the TMGA system allows the user to compare the results in various experiments in order to make appropriate modifications to improve the solutions to the specified problem.

CHAPTER 6

6 Results of experiments

In this chapter, the TMGA results and newly generated structures (in effect, predicted models) of bacteriorhodopsin are presented. Bacteriorhodopsin (see Appendix B) is a transmembrane protein that contains 7 TM helices and its 3D structure is perhaps the best characterised structure of any membrane protein. In these experiments, bacteriorhodopsin is used as an appropriate test case in order to compute the difference between the predicted structure obtained by utilising a GA as an optimisation technique, and the known structure. This approach, however, could be applied to any membrane protein, as the starting helix arrangements and end-on orientations are inherited from the TMRelate program (Roberto Togawa, John Antoniw and Jonathan Mullins, 2002) predicted from amino acid sequence.

6.1 GA parameters

In the first set of experiments, the construction of TM proteins is based on the 2 positional variables (x and z) associated with each helix. The value of R is based on the rotation of each helix so in these experiments is assumed to be constant.

The distance between helices is manipulated and in this experiment the extent of interaction between helices is based on their distance apart. In the first instance, the GA was designed to randomly generate 50 TM protein structures per generation, selected as the population size for a small range of input. Each helix is assumed to run straight through the membrane from one face to the other, which is a necessary approximation for the methodical application of the GA.

Figure 6.1 shows the average energy for each generation. The first 10 generations of the GA are associated with random testing and in this run, after 18 generations the GA was terminated. These results indicate that the GA is unable to generate structures with sufficiently low (i.e. negative) energy as the average energy for each generation remained high. This also indicates that manipulating the distances between helices is unable on its own to generate feasible interactions between helices in order to generate structures with high stability that is near to native structure.

Using two input variables generally requires a larger population size as large populations in the GA allow the search space to be sampled more (Jones et al., 1998) and generate more possible solutions for the specific problem.

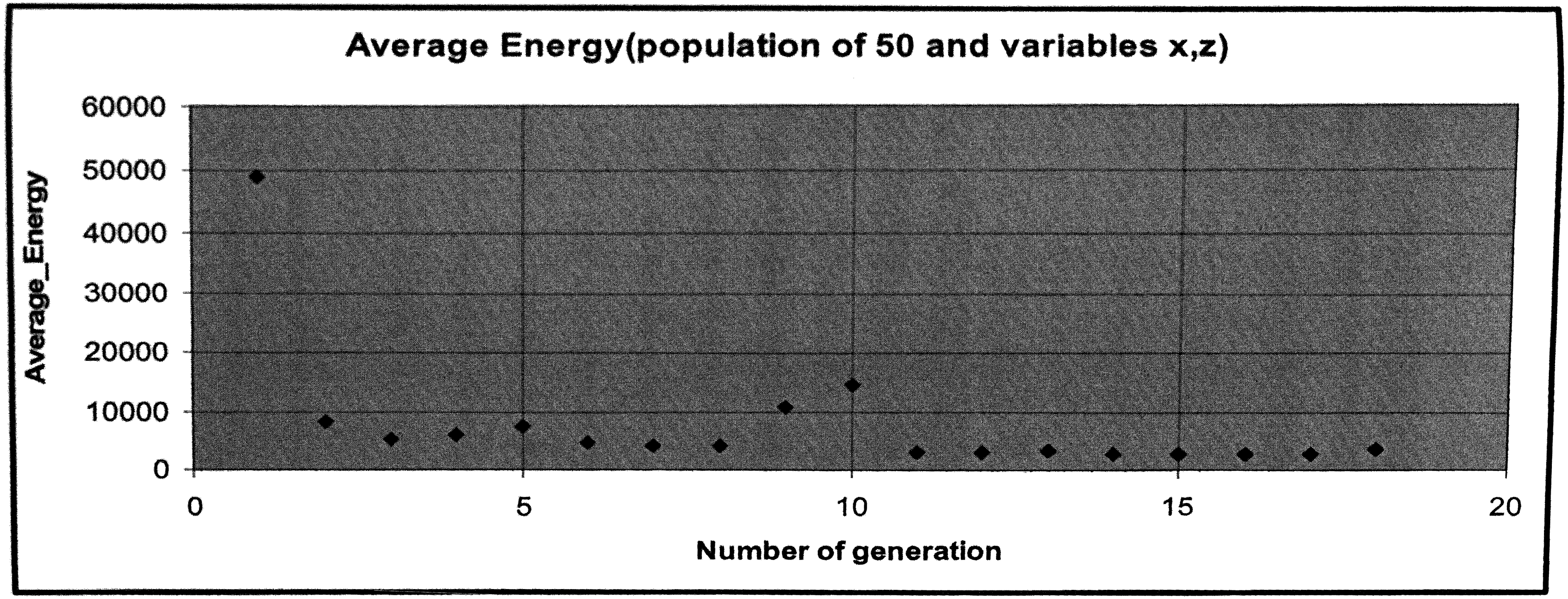


Figure 6.1 Average energy for each generation of 50 structures, varying only x and z

In the next set of experiments, the GA was designed to randomly generate 50 TM protein structures per generation and increase the data input to three variables, as the construction of TM helices was based on the 3 positional variables (x, z and R) associated with each helix.

As figure 6.2 shows, the average energy values were reduced after 10 generations and the GA was terminated after 22 generations. This experiment indicates that increasing the number of input variables that are manipulated generates a higher proportion of new structures with low energy.

The average energy is decreased compared with the first set of experiments but the GA is unable to significantly improve the quality of the better structures as using the same population size kept the search space small. However, the effectiveness of a GA can be improved by utilising a larger search space.

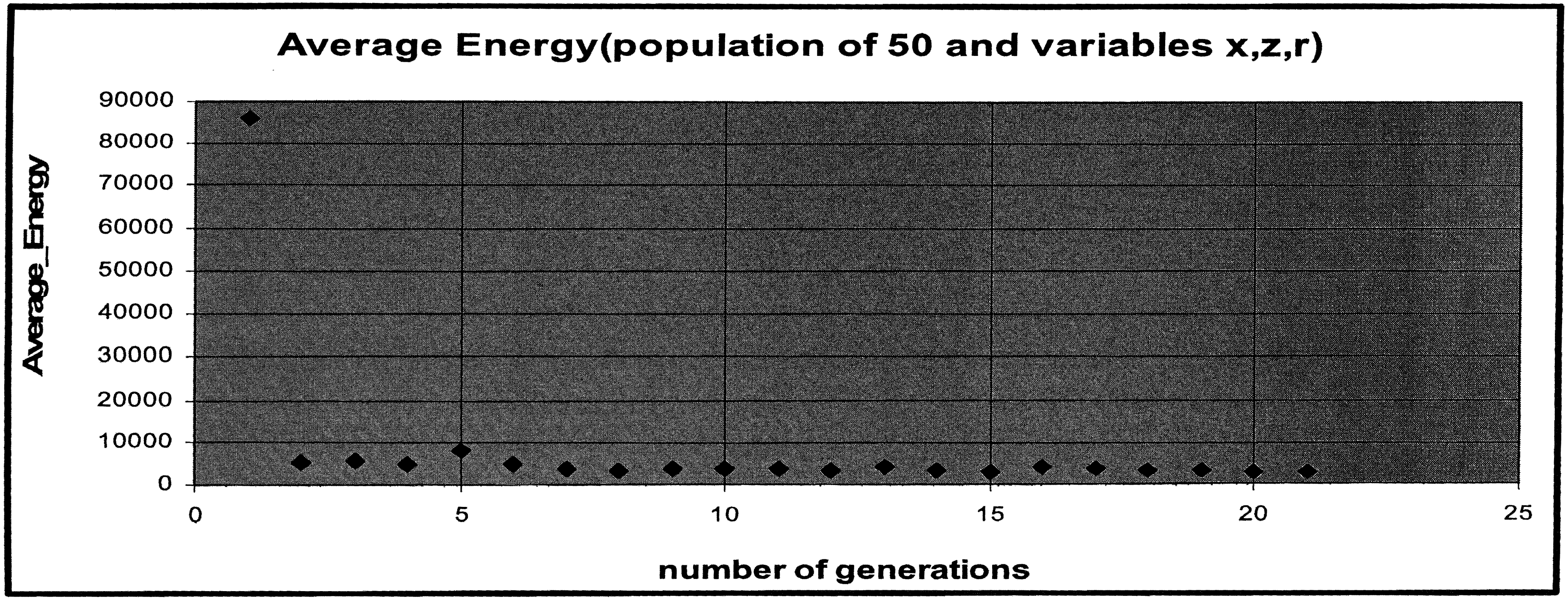


Figure 6.2 Average energy for each generation of 50 individuals, with GA manipulation of three variables

In order to compare the effectiveness of the GA for predicting 3D structures of membrane proteins, with different population sizes, a run with a population of 100 was made with the three input variables (x,z and r).

The results indicate that introducing a larger population of 100 increases the search space and the possible combinations in the generation of new TM protein structures.

As figure 6.3 shows, the GA was terminated after 56 generations and the average energy was markedly reduced after 11 generations. This experiment generated structures with high stability and low free energy.

This low (negative) energy indicates that manipulating the rotation of each helix as well as the distances between helices results in the generation of new structures that are more similar to the native structure.

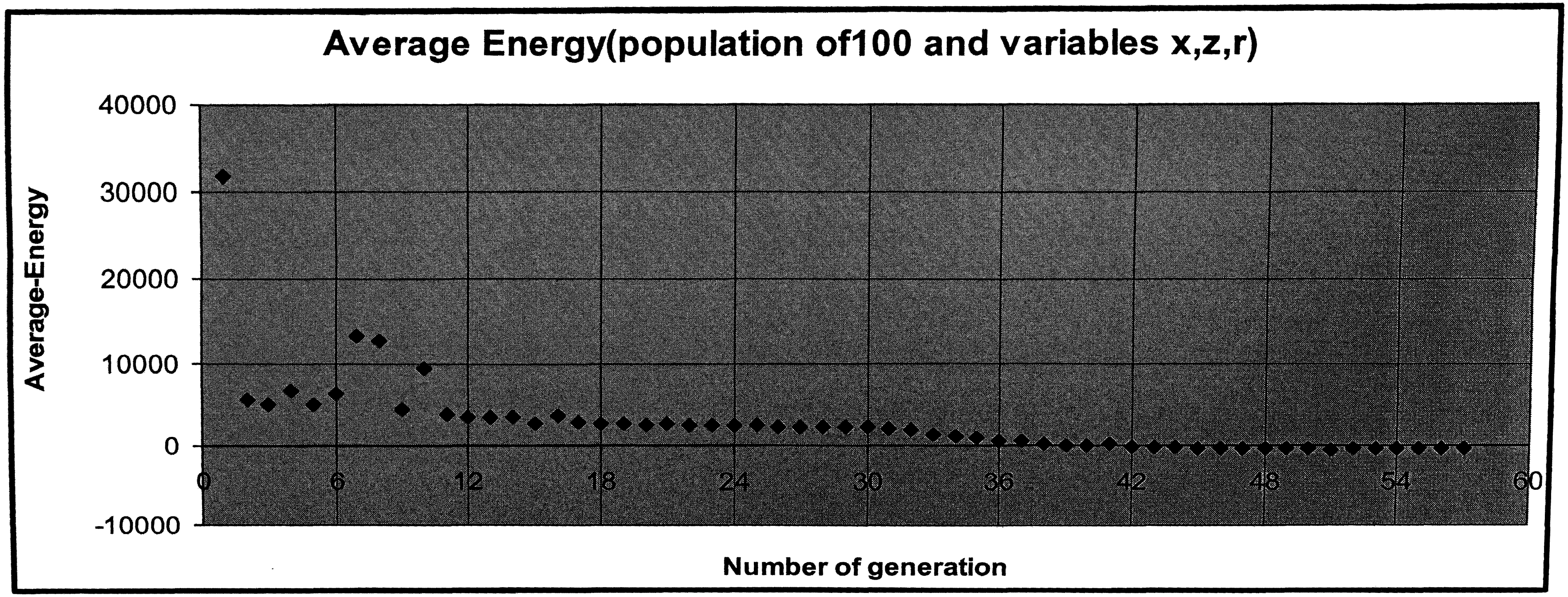


Figure 6.3 Average energy for each generation of 100 individuals with manipulation of three variables

6.2 Comparison of individual structures obtaining the lowest energies in the different experiments

Figure 6.4 shows the lowest energy obtained in each generation in the three main sets of experiments. As individuals in the first 10 generations in each run are based on random selection, the energy values are high.

In the first set of experiments with a population size of 50 and exploiting two variables (x and z), the energy values are elevated and the similarity between the energies of the predicted and native structure is low.

In the second set of experiments with a population size of 50 and three variables (x, z and r) the calculated energy is decreased more quickly, but the GA is unable to predict structures any closer to that of the native structure in terms of greater structural similarity and higher stability as the lowest energy obtained is still relatively high. The program was terminated after 20 generations.

In the final set of experiments with a population size of 100 and manipulating three variables (x, z and r), by increasing the population size, the GA was able to introduce more new solution structures into the search space in order to predict a conformation which is near to the native structure.

As the results show, though the stability of the new structures is lower than the native structures (as their energies are still higher than the native structure), this experiment does show that the TMGA system is able to generate 3D structures of membrane proteins with feasible molecular stability in that they possess negative values of free energy.

The comparison between the lowest energy in each generation with the average energy of each generation indicates that the average energy illustrates the general evolutionary improvement in the TMGA system as the selection technique in each generation is not based on the individual with the lowest energy on its own.

The results of the lowest energy of each generation represent the ultimate efficiency of the GA in the TMGA system.

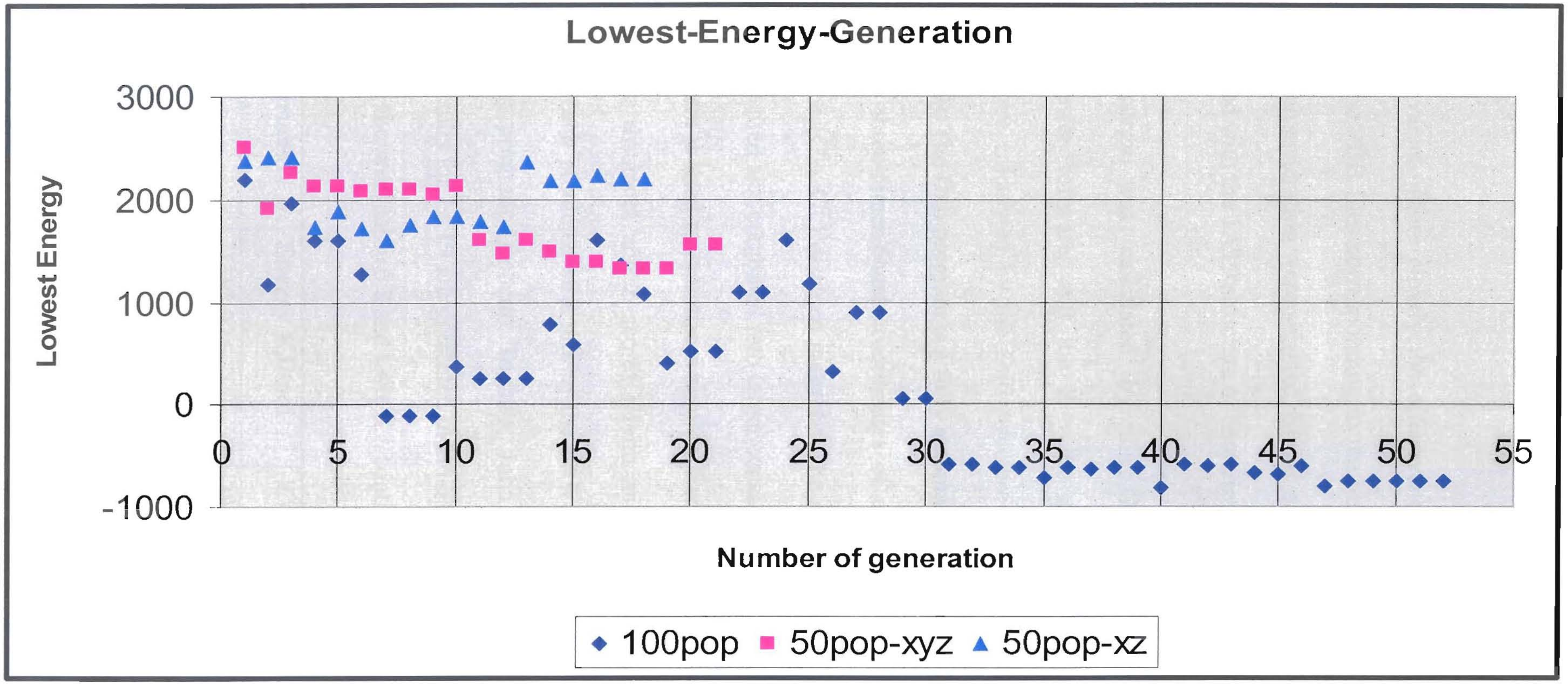


Figure 6.4 The lowest energy attained in each generation

6.3 Evaluation of generated structures

In order to evaluate more closely the effectiveness of genetic algorithms for this application, the PDB files corresponding to the best result (that with the lowest free energy) of each generation were selected. In this analysis, all the predicted structures were compared with the 1AT9 structure of bacteriorhodopsin (Kimura et al., 1997).

The free energy was calculated by the GROMOS algorithm within the SwissPDBViewer software, and the Z value and RMSD (Root Mean Square Deviation) were calculated by the Dali structure comparison program (Wood and Pearson, 1999). The Dali program is a fully automated structure comparison algorithm that uses the three-dimensional co-ordinates of each protein in order to calculate residue-residue (C^{α}—C^{α};) distance matrices. The similar contact patterns in the two matrices are paired and combined into larger consistent sets of pairs. In Dali, a Monte Carlo procedure is used to optimize a similarity score defined in terms of equivalent intramolecular distances and Z scores (Holm and Sander, 1993). In addition, the similarity of each predicted structure to the native structure was assessed using the TMEvaluation program (Roberto Togawa, John Antoniw and Jonathan Mullins, 2003). The results for 1AT9 are presented in table 6.1.and the results of the generated PDB files are shown in table 6.2 and table 6.2(continued 1, 2).

Native structure	**Energy KJ/mol**	**Z**	**RMSD**
1AT9	-1583.924	29.7	0

Table 6.1 Free energy, Z and RMSD values for the “standard” native structure

Generations	Lowest Energy	Z	RMSD	Percentage correctly Predicted at-5Å	Percentage correctly Predicted at-8Å	Percentage correctly Predicted at 10Å
0	2190	3.2	3.2	12.68	23.49	33.65
1	1169	2	12.2	6.34	18.06	22.99
2	1961	1.8	1.9	2.11	8.95	17.37
3	1594	5.8	4.5	0.7	18.83	32
4	1594	5.8	4.5	0.7	18.83	32
5	1273	3.9	3.6	3.52	13.43	25.55
6	-110	6.2	3.6	3.52	17.28	28.69
7	-110	6.2	3.6	3.52	17.28	28.69
8	-110	6.2	3.6	3.52	17.28	28.69
9	359	6.1	3	4.23	16.67	29.09
10	255	5.3	4.2	4.93	17.75	31.45
11	255	5.3	4.2	4.93	17.75	31.45
12	255	5.3	4.2	4.93	17.75	31.45
13	785	4.5	4.4	1.41	12.96	23.51
14	578	3	3.2	0	10.83	21.62
15	1605	5.3	4.3	5.63	15.12	26.89
16	1353	4.8	11.1	6.34	16.36	26.42
17	1074	4.5	8.6	1.41	14.35	23.58
18	406	5.5	10	0	16.05	30.27
19	510	4.5	4.4	1.41	12.96	22.48
20	521	4.1	9.4	1.41	14.66	25.47
21	1085	4.2	3.7	2.82	13.43	23.98
22	1085	4.2	3.7	2.82	13.43	23.98
23	1605	5.3	4.3	5.63	15.12	26.89
24	1165	6.2	4	4.23	18.21	29.48
25	315	3.3	10.7	0	10.65	21.31

Table 6.2 Analysis of the structures with the lowest free energy in each generation, RMSD and Z value and also percentage similarity by residue pairs of each predicted structure to the native structure.

Generations	Lowest Energy	Z	RMSD	Percentage correctly Predicted at-5Å	Percentage correctly Predicted at-8Å	Percentage correctly Predicted at 10Å
26	890	5.1	2.5	4.23	20.22	33.96
27	890	5.1	2.5	4.23	20.22	33.96
28	51	5	2.7	2.11	16.36	27.75
29	51	5	2.7	2.11	16.36	27.75
30	-582	5.4	4	3.52	16.2	28.62
31	-582	5.4	4	3.52	16.2	28.62
32	-621	5.4	3.8	3.52	16.05	28.93
33	-621	5.4	3.8	3.52	16.05	28.93
34	-717	5.8	5.4	4.93	17.44	23.17
35	-621	5.4	3.8	3.52	16.05	28.93
36	-634	5.1	3.7	4.23	19.14	32.39
37	-621	5.4	3.8	3.52	16.05	28.93
38	-621	5.4	3.8	3.52	16.05	28.93
39	-817	5.8	4.6	5.63	20.68	33.33
40	-591	5.1	2.7	3.52	15.74	29.87
41	-607	5.4	4.5	2.11	16.82	30.97
42	-591	5.1	2.7	3.52	15.74	29.87
43	-665	5	5.4	4.23	17.9	30.27
44	-684	5.4	4	6.34	19.75	32.55
45	-607	5.4	4.5	2.11	16.82	30.97

Table 6.2 (continued_1) Analysis of the structures with the lowest free energy in each generation, RMSD and Z value and also percentage similarity by residue pairs of each predicted structure to the native structure.

Generations	Lowest Energy	Z	RMSD	Percentage correctly Predicted at-5Å	Percentage correctly Predicted at-8Å	Percentage correctly Predicted at 10Å
46	-802	5.1	3.6	1.41	16.98	31.6
47	-752	5.2	2.6	3.52	16.82	29.72
48	-752	5.2	2.6	3.52	16.82	29.72
49	-752	5.2	2.6	3.52	16.82	29.72
50	-752	5.2	2.6	3.52	16.82	29.72
51	-752	5.2	2.6	3.52	16.82	29.72
52	-752	5.2	2.6	3.52	16.82	29.72
53	-655	5	4.4	4	16.54	29.32
54	-655	5	4.4	4	16.54	29.32
55	-752	5.2	2.6	3.52	16.82	29.72
56	-655	5	4.4	4	16.54	29.32

Table 6.2(continued_2) Analysis of the structures with the lowest free energy in each generation, RMSD and Z value and also percentage similarity by residue pairs of each predicted structure to the native structure.

The results of the analysis of the structural similarity of the transmembrane regions of bacteriorhodopsin as determined using the Dali structure comparison program are displayed in figures 6.5 and 6.6.

When the predicted protein structures share significant similarity, the Z values are increased and vice versa. Z values can provide reliable information about similarity in the same way as RMSD (the root-mean-square deviation)(Zhang and Skolnick, 2005), but with the difference that Z values are based on a Monte Carlo simulation technique and the overall Z score is the comparison of actual alignment score with the scores obtained on a set of random sequences (Aude and Comet, 1996)

Figure 6.5 shows that the Z values are higher in the new structures with low energy and as the energy increases the Z value decreases.

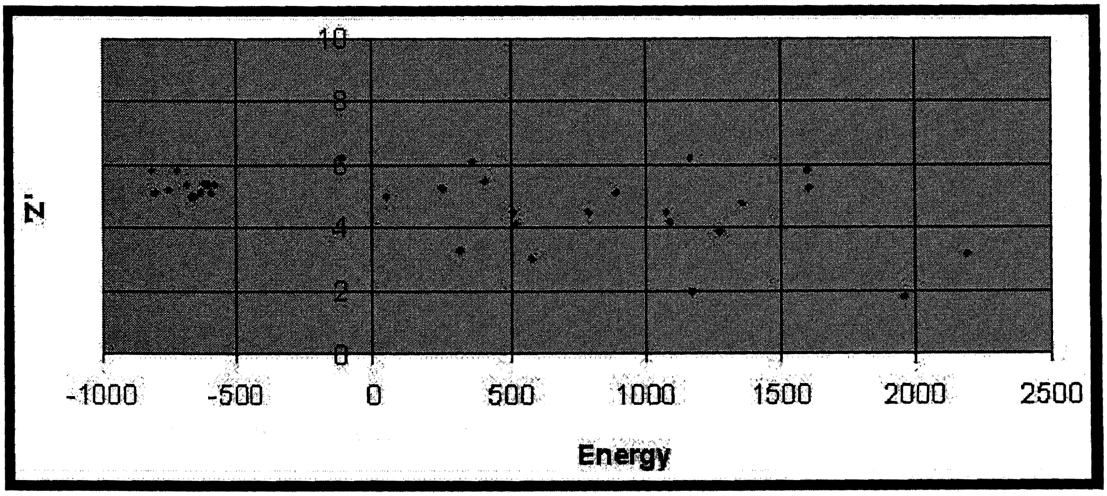


Figure 6.5 Z scores reported by the Dali program for the most stable structure of each generation

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.445925
R Square	0.198849
Adjusted R Square	0.184283
Standard Error	0.815268
Observations	57

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	9.07345	9.07345	13.65124	0.000508
Residual	55	36.55637	0.664661		
Total	56	45.62982			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	5.052207	0.108845	46.41661	8.45E-46	4.834077	5.270337	4.834077	5.270337
X Variable 1	-0.00045	0.000123	-3.69476	0.000508	-0.0007	-0.00021	-0.0007	-0.00021

Table 6.3 Summary of regression analysis, On regression analysis, the relationship between Z scores and free energy gave an 'r 'value of 0.45 and a significance (F) of 0.0005, showing that there is a statistically significant relationship between higher Z scores and lower values of free energy for the most stable structures of each generation. This relationship supports the use of free energy as an evaluation function in the GA.

Table 6.3 shows that there is a relation between Z scores and free energy and the GA is able to predict new structures with high similarity (higher Z scores) and high stability (lower energy).

The most frequently used measure of structural similarity is the RMSD, and lower values of RMSD indicate higher similarity (Carugo, 2003). The RMSD calculated between equivalent atoms in two structures *is* defined as:

$$rmsd = \sqrt{\frac{\sum_i d_i^2}{n}}$$

Equation 7. Formula for RMSD

Where d is the distance between each of the n pairs of equivalent atoms in two optimally superimposed structures (Carugo and Pongor, 2001). The RMSD is based on the total number of atoms included in the structural alignment.

As figure 6.6 shows, the GA is able to generate structures that are similar to the native structure with reasonably low RMSD values and low energy and also generate some new structures with rather higher energies but with low RMSD.

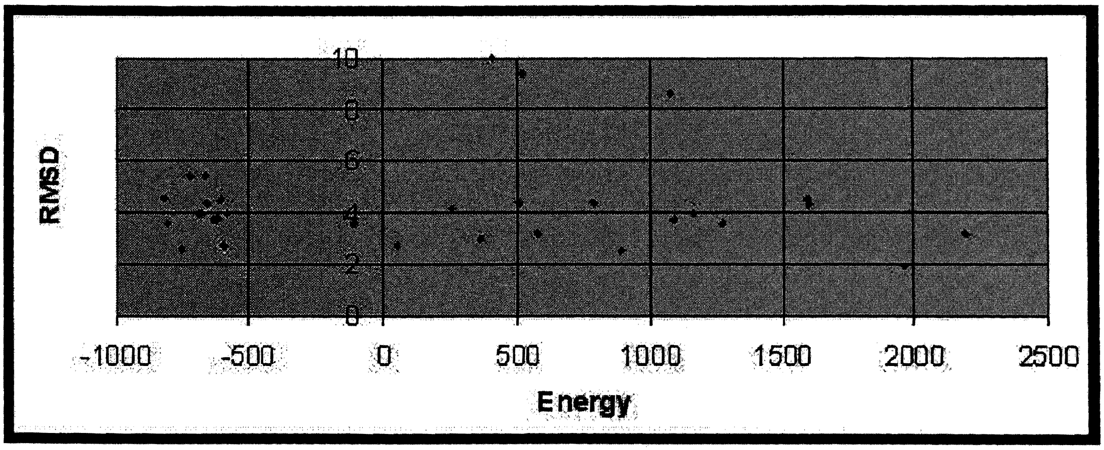


Figure 6.6 RMSD values reported by the Dali program for the most stable structure of each generation with the lowest free energy.

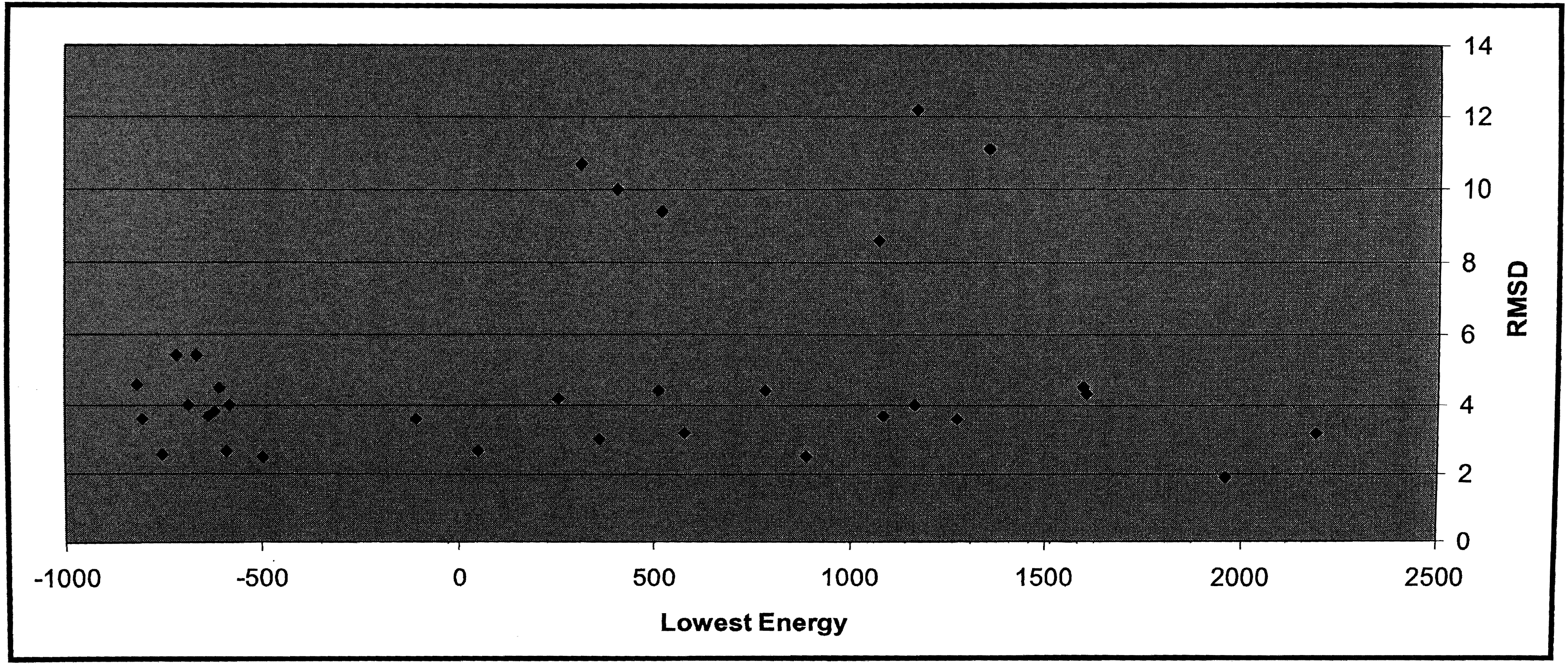


Figure 6.7 The similarity to native structure (RMSD values) compared with the lowest energies of generated structures.

The figure 6.7 shows the most stable predicted structures (with the lowest energy) in each generation against the RMSD value. These results indicate that although the most stable predicted structure in each generation is not correlated with lower RMSD values (high similarity with the native structure), the most consistently low RMSD values are observed with the series of lowest energies.

The graph demonstrates that for the GA to generate a predicted structure with low energy and high similarity requires running a higher number of generations. As the number of generations is increased the number of predicted structures with low free energy and also low RMSD values will be increased.

This indicates that random selection initially drives the efficiency of the GA, then after 30 generations the GA effectively selects those structures that are nearest to native structure for the next generation. This will drive the system to predict structures with low energy and high similarity.

Figure 6.8 shows the comparison between the two measures of similarity for the most stable newly generated TM protein structures.

As the TMGA system is based on evaluation by the free energy of generated structures and with the observation that the GA successfully generates predicted 3D structures with stability near to that of the native structure, this analysis addresses the effectiveness of the evolutionary process in the TMGA system in terms of selection of structures of viable structural similarity to the native structure. This consideration is of great importance with respect to the capacity of GA approaches to generate model structures that are of genuine relevance to the study of the structure / function relationships of a given protein.

As the results show, the system generated several structures with low stability and high similarity to the native structure. This indicates that despite being formulated primarily to select by free energy, the GA operators are able to drive the system significantly well to generate a range of new TM protein structures that attain varying levels of similarity or stability (some with comparable stability but low similarity to the native structure, others with inferior stability but reasonable similarity to the native structure, and a notable proportion with comparable stability and reasonable similarity to native structure).

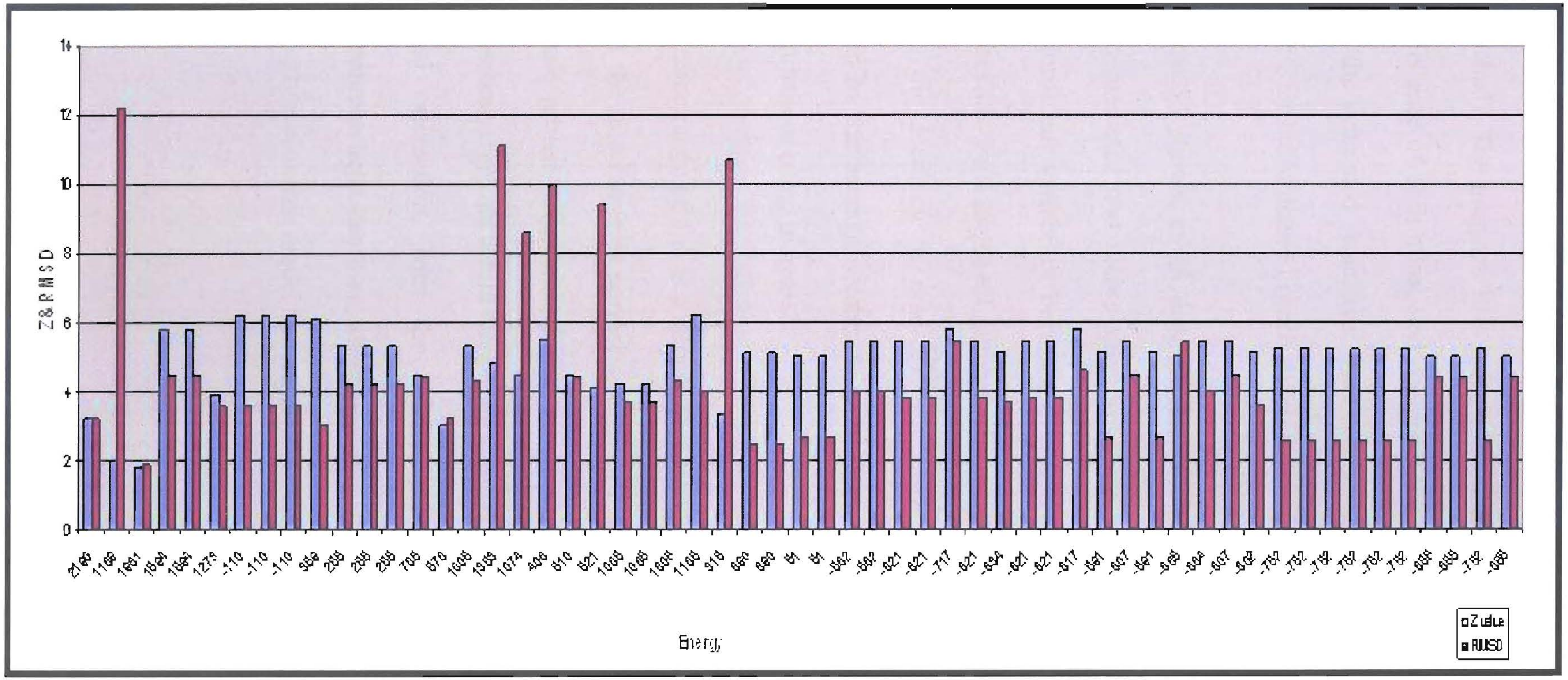


Figure 6.8 The distribution of the Z value and RMSD for the predicted structure compared with the lowest energy in each generation.

The accuracy of the TMGA system was also tested by use of a program called TMEvaluation that was developed by other members of the group (Togawa, unpublished). The program lists pairs of residues that are calculated to be within a given distance of each other, and compares the list obtained for the native structure against the generated structures. This comparison was performed at three different distances, for the structure in each generation with the best energy calculation.

As figure 6.9 shows, at 5Å distance, the percentage of coincident pairs of residues between predicted and native structures remains at a similar level throughout the generations, but variability between consecutive generations decreases as the run progresses i.e. with increased number of generations, the results become tighter. In general, the accuracy of predicted structures after 30 generations is increased, indicated by the average number of coincident residue pairs at close range between different helices being increased.

The distance range of 5 Å is a rather unforgiving criterion, and was selected as this is close to the mid range of normal side chain / side chain proximities in determined membrane protein structures. For a residue pair to be listed at this range, the transmembrane regions would have to be arranged with highly accurate adjacency and rotational orientation, with the residue pairs also appearing at the correct transmembrane helix depth on adjacent helices. In this regard, scores of 5% represent surprisingly reasonable model structures.

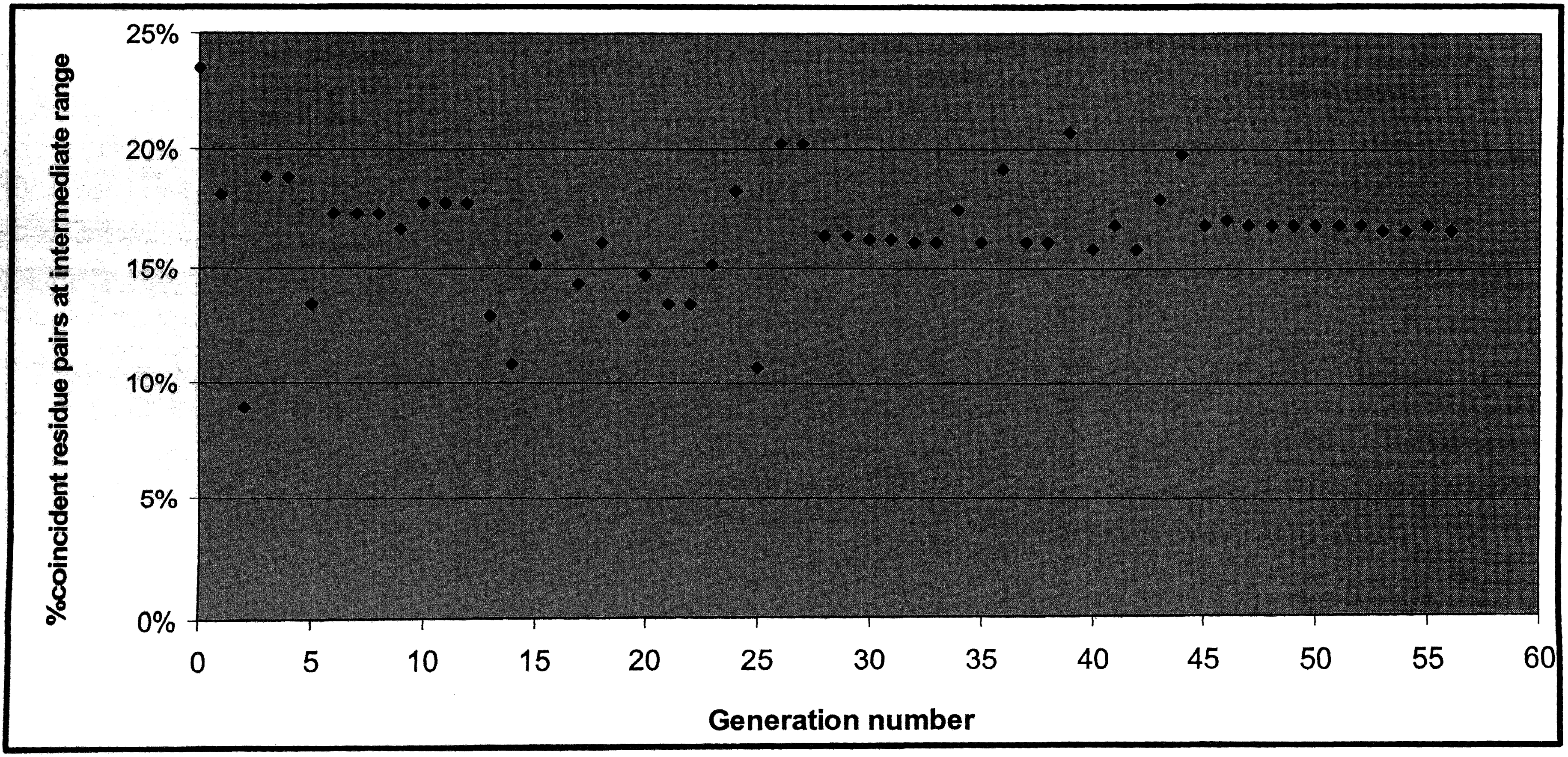


Figure 6.9 Percentage of coincident residue pairs within 5Å in the test and native structures.

In figure 6.9, the percentage of coincident residue pairs is presented for the structures with the lowest energy score in the 56 generations, but this time within a greater distance range of 8 Å. As the graph shows, by increasing the allowable distance limit the percentage of coincident residue pairs is generally increased and the variability between consecutive generations is less than 5 Å. Again, by the later generations, there is very little variability.

The distance range of 8 Å was selected as this is close to the average distance between the alpha carbon backbones of tightly packed transmembrane regions in determined membrane protein structures.

For a residue pair to be listed at this range, the transmembrane regions would have to be placed at comparable distance apart to those of the native structure and with the correct end-on rotation and relative helix depth. Scores in the region of 20% are very promising.

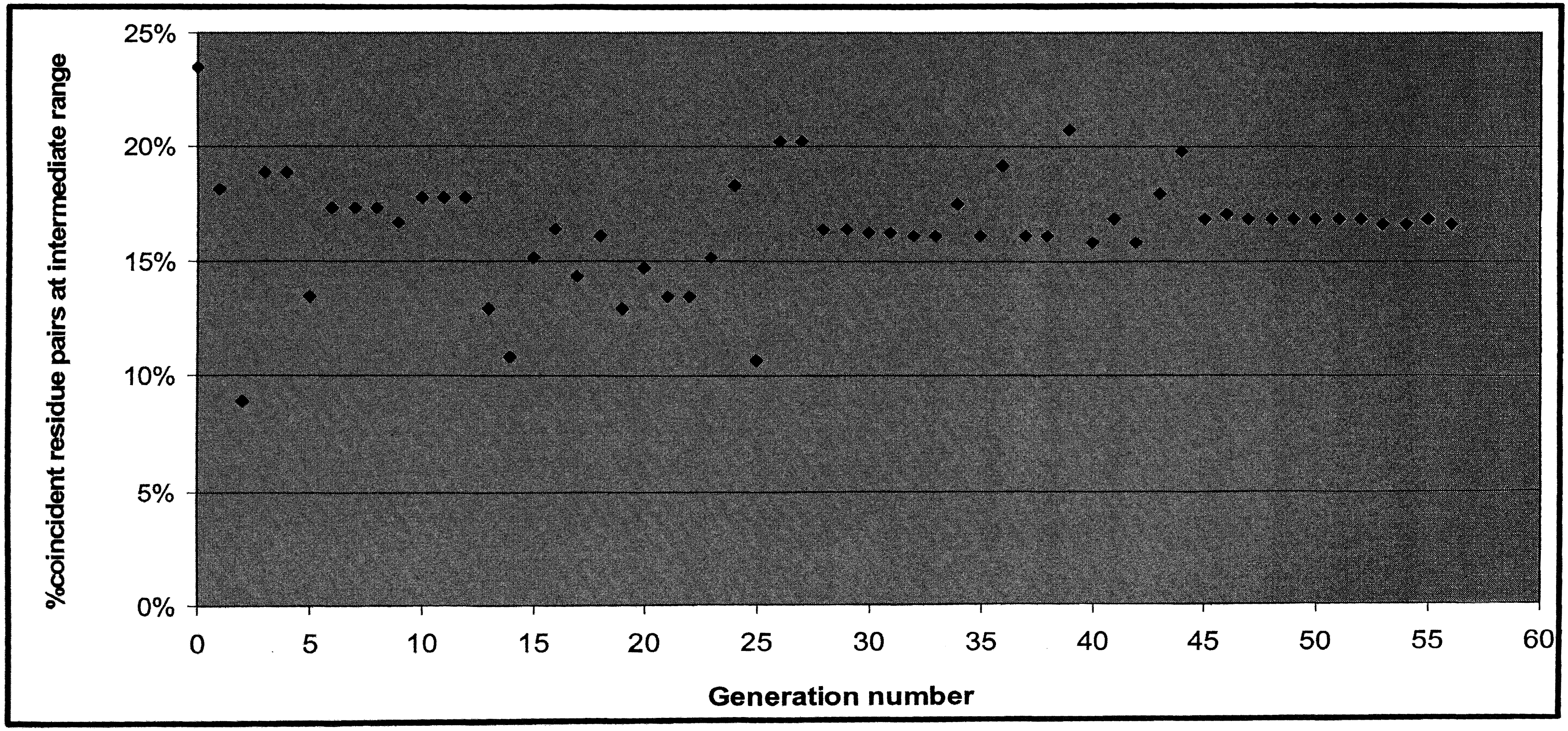


Figure 6.10 Percentage of coincident residue pairs within 8Å in the test and native structures.

The percentage of coincident residues within a 10 Å distance in each new structure and the native structure is shown in figure 6.11. The 10 Å range is a somewhat more generous criterion, based on the upper limit of inter-helical distances between alpha carbon backbones found in determined membrane protein structures, and allows for all residue pairs that are found in both structures with comparable TM region adjacency, comparable end-on rotation and at comparable membrane depth.

These results, showing that approximately one third of the amino acid side chains involved in side chain / side chain residue pairs possess comparable positioning in three dimensions indicate that the GA generated plausible 3D structures for membrane proteins of energetically stable structure and that were sufficiently similar to the native structure to allow comparable helix packing.

However, it is clear that the GA could be further improved, perhaps by the incorporation of other evaluation methods such as consideration of weighted scores for specific residue-residue associations and the formation of recognised 3D structural motifs and packing patterns.

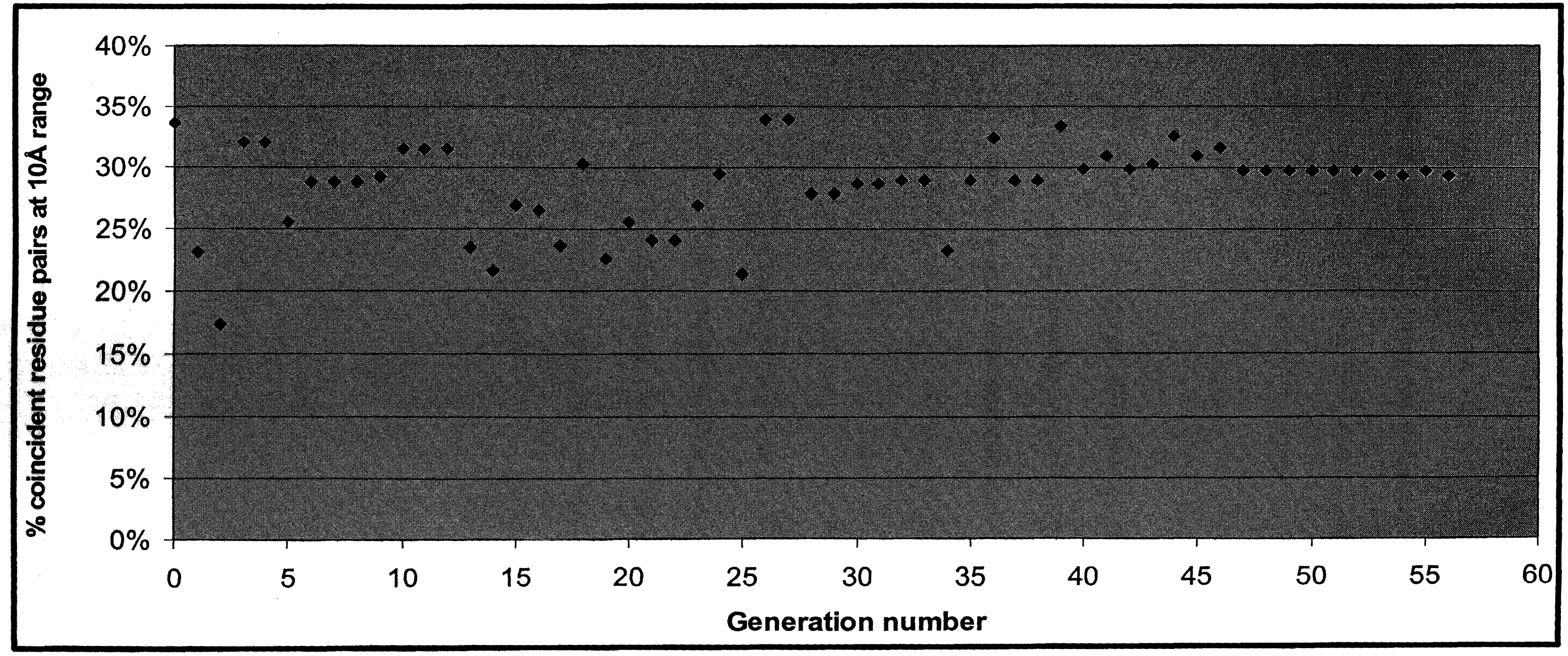


Figure 6.11 Percentage of coincident residue pairs within 10Å in the test and native structures.

In theory, as the free energy decreases the accuracy of predicted structures by all measures should increase. However, as shown in figure 6.12, the results achieved show that high percentages of coincident residue pairs can be attributed to structures with both high and low energy.

The GA is able to generate plausible 3D structures for the transmembrane regions of membrane proteins with high speed and efficiency on the basis of selecting those with the lowest energies for the next generation. In general, the TMGA system proved successful in using an evolutionary approach to generate candidate protein structures, which bear good resemblance to the native structure in terms of stability and packing of TM helical regions.

This work confirms that the consideration of individual model solutions by their free energy for selection in the next generation is an appropriate strategy, suggesting that there is a broad relationship between theoretical energetic stability and actual 3D structure, but that further improvements can be made to the application of the genetic algorithm, involving greater consideration of the intricate structural and functional characteristics of membrane proteins.

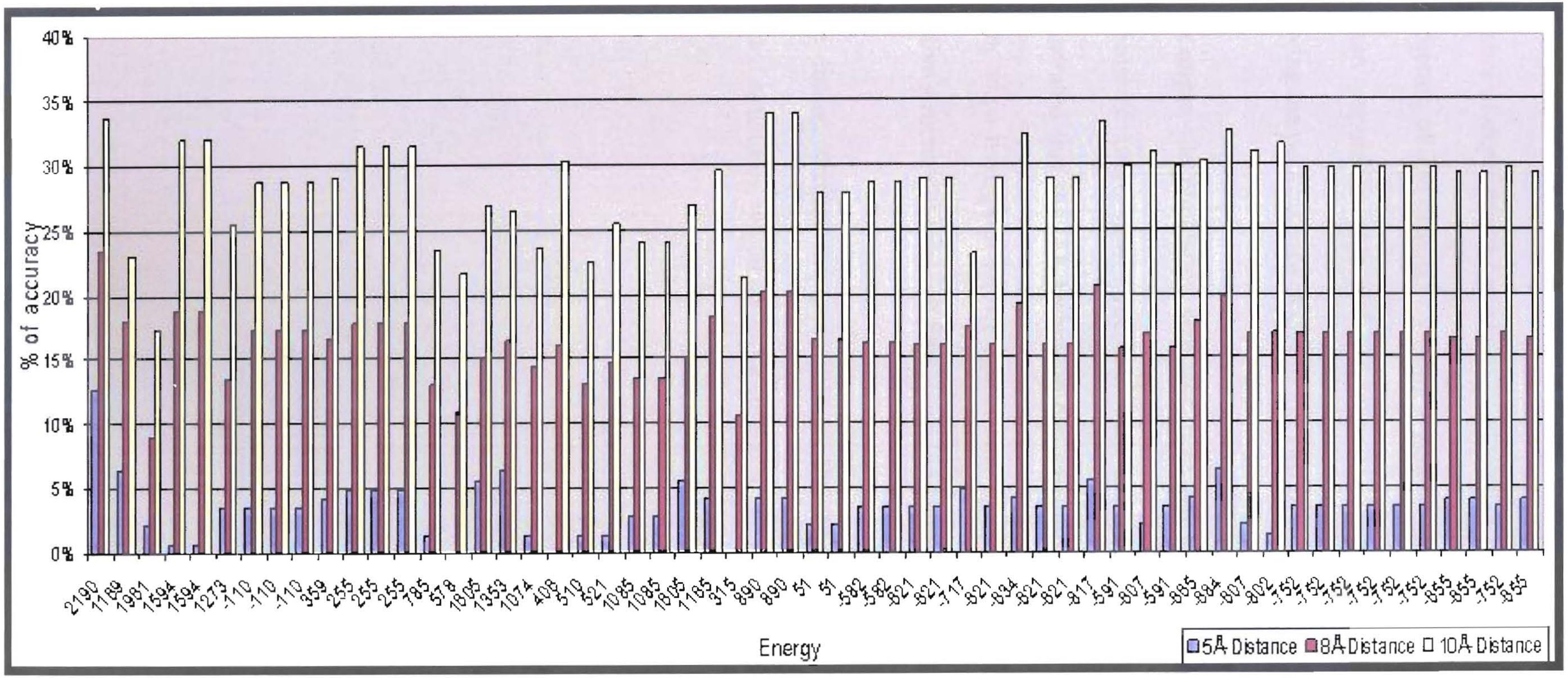


Figure 6.12 The percentage of correctly predicted side chain pairs for the structure attaining the lowest free energy in each generation.

6.4 Summary

A number of experiments have been conducted in the work of this thesis to predict the 3D structure of TM proteins by using a genetic algorithm. Experiments to derive the optimum population size and GA operators were conducted for the purpose of improving the performance of GA in the TMGA system.

For example, the population size was increased to improve the predicted structures and generate more new structures with new features. These experiments have demonstrated that the GA technique can capably predict the 3D structure of a TM protein, taken from prior knowledge of only the sequence, to a level that approaches the native structure in terms of free energy and structural similarity.

In the future, further improvements to the TMGA system will generate better predicted structures, which are more similar to the native structure.

CHAPTER 7

7 Discussion

“Artificial life” is an attempt at understanding all biological phenomena through their reproduction in artificial systems, such as in computer simulations. Computational approaches to the protein folding problem are often addressed by alternative techniques for predicting the tertiary structure of proteins given their amino acid sequence. Predicting the three dimensional structure of a protein from its amino acids is one of the most important problems of modern biology.

Integral membrane proteins play important roles in living cells. Among the 30,000+ solved protein structures in the Protein Data Bank (PDB), less than 2% are membrane proteins while on average, 20-30% of the genes in a genome encode membrane proteins. Thus, computational approaches for the prediction of membrane protein structures have become an attractive alternative (Chen and Xu, 2006).

GAs are efficient general search algorithms and as such are appropriate for any optimization problem, including problems related to protein folding. Many studies show that GAs are superior to Monte Carlo and other search methods for protein structure prediction (Unger, 2004).

7.1 Comparison of TMGA system with others

The GA is appropriate for any Optimization problem, including problems related to protein folding (Unger, 2004).

As the conformation space for possible membrane protein structure is challengingly large, this is a suitable problem for the application of a GA. In this work, for the first time, a GA has been applied to predicting membrane protein structure and it has been shown that it is possible to optimise the prediction of TM protein structure.

Faulon et al (2003) present a computational technique for the assembly of helix bundle of membrane protein matching a predefined set of distance constraints. They take a set of helices in pdb format and a set of distances between pairs of atoms on those helices. The output of this method is all the possible helix arrangements that match the provided distances. They did not, however, undertake energetic calculations on each conformation and compare each solution with the others by a predefined RMSD. They used the structure of Rhodopsin as model.

In this study, the arrangement of TM regions is optimised in order to arrive at a reliable distance between helices that generate a new structure with high stability which possess low free energy and more similarity to the native structure (1AT9).

There are different approaches available for predicting soluble protein structures by using GA. Braden (2002) used a GA to develop a protein structure prediction technique by modelling proteins in three dimensional integer space and residues with characteristics of hydrophobicity, charge and side chain size. Each residue consists of a group of five bits and each group is decoded to an integer number.

The group represents a potential location at which a residue can be found. In the TMGA system, the TM protein is initially modelled from an arrangement of 7 alpha carbon atom starting positions, the starting atom of the initial residue of each helix in three dimensional spaces. Each initial residue of each helix consists of a group of three integers that represent the location of a TM region in space with respect to the other helices.

Bui et al. (2005) investigated the protein folding problem in the HP model in which each amino acid is classified, based on its hydrophobicity, as H (Hydrophobic or non-polar) or a P (Hydrophilic or polar) in a 2D square lattice by finding a lowest energy conformation. Their GA method used secondary structures as individuals of the population and one point crossover and uniform mutation. The secondary structure is mutated by replacing it with another structure selected at random.

In the TMGA system, the conformation of the TM protein structure is based on the arrangement of 7 TM regions placed in a limiting 2D grid and the helix breakage and kinking in TM regions are ignored in each template.

The GA operator in this system was modified by using two points crossover and three points mutation. The structure is mutated by selecting values (x or y or z) randomly from the initial residue of helices.

The candidate solution in this GA method is designed as a template of the conformation for a given membrane protein that contains alpha helical TM regions.

Each helix is specified by the position {x, y, and z} of the atoms in protein that is identified by their Protein Data Bank (PDB) file. The template consists of an arrangement of 7 alpha carbon atom starting positions, the starting atom of the initial residue of each helix. The coordinate of the starting atom of the initial residue of each helix defines the distance between two helices and this distance can be changed by varying the coordinate.

In this work, the GA is able to optimise the distance between helices as well as the rotation of each helix in order to arrive at the most reliable distance that results in the highest stability. This method assumes the optimal value for all bond angles for residues within the carbon backbone, and the side chains are constructed using the MaxSprout algorithm based on a database of structures. The method considers all helices as passing straight from one side of the membrane to the other without breakage or kinking. The focus of this work is therefore to investigate the optimal arrangement and fine positioning of transmembrane regions in membrane protein structures

The significance of this work is that the GA predicted new structures near to the native structure (1AT9) in terms of low free energy and high stability as well as high similarity. New structures could be generated in the same way for proteins of undetermined structure, to serve as approximate molecular models.

As shown in the results presented in figure 6.7, the lowest energy obtained (-817 KJ/mol) belongs to a new predicted structure that is less similar in structural terms, as its RMSD value is higher than certain other predicted structures with energy -752 KJ/mol and RMSD 2.6Å. This indicates that although the structure with the lowest energy is more stable than the other structures, it does not resemble the native structure as closely as some other less stable structures. This is likely to be due to "over-fitting" of the optimised structures to the evaluation function of free energy, at the expense of accuracy of structural prediction compared to native structures, which must allow for conformational stresses and change.

There is another generated structure with a substantially higher free energy (890 KJ/mol) but more similar (RMSD 2.5 Å) to the native structure (1AT9). The comparison between TM regions in these structures indicated that subtle differences in the distance, between helices containing the same amino acids, significantly affected the free energy or stability of each structure. The comparison between the positioning and end-on orientation of TM regions in the generated structures with those of the native structure (1AT9) is shown in the following figures and also are provided as PDB files with Rasmol software on CD (see appendix D).

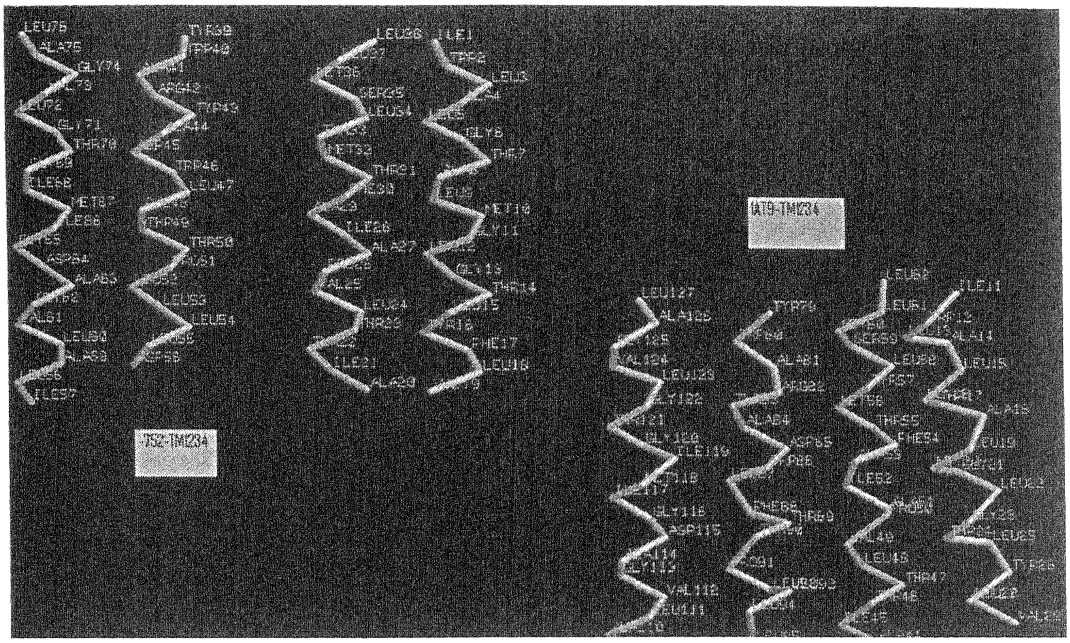


Figure 7.1 The comparison between TM1,TM2,TM3 and TM4 (-752KJ/mol) with the same TM regions in 1AT9

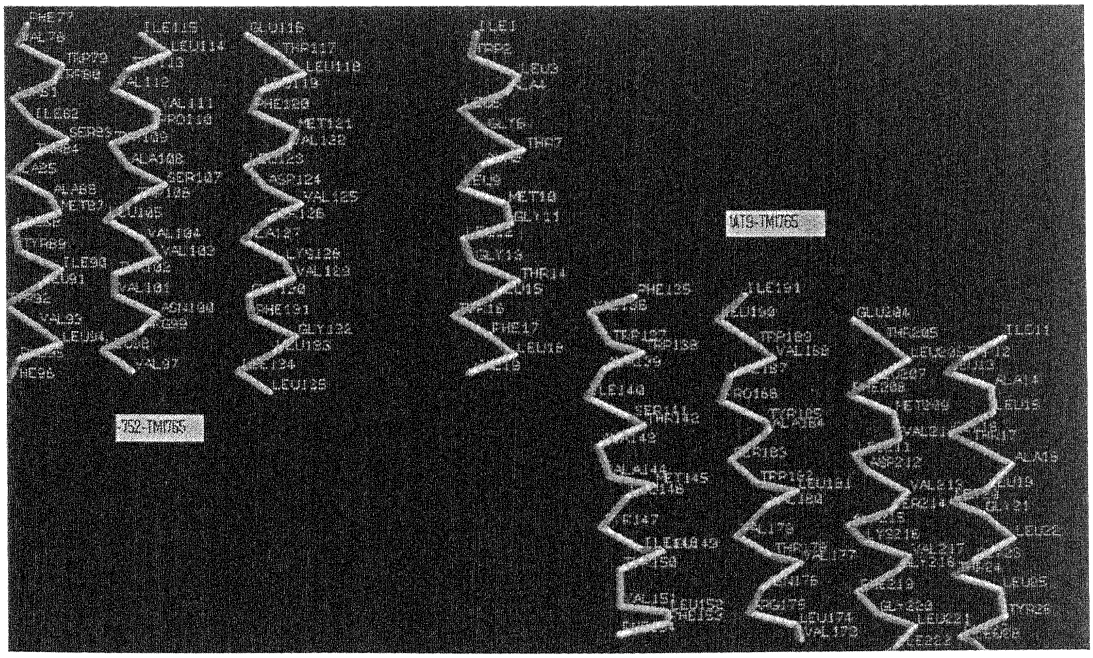


Figure 7.2 The comparison between TM1,TM7,TM6 and TM5 (-752KJ/mol) with the same TM regions in 1AT9

Figure 7.1 and 7.2 shows TM regions of a generated structure (-752 KJ/mol). TM5 and TM7 compare very well in terms of end on rotation but TM1 and TM6 do not compare so well. It appears that the residues of TM1 do not map on to similar helical structure in the predicted helix compared to the native TM1.

TM1 and TM 2 would need only minor adjustment to resemble the native structure and TM4 compares extremely well but TM3 is lightly different in predicted structure. This due to helix breaks on native in TM3 of the native structure, as in this application we ignored any breaks in TM regions.

The comparison of TM regions in terms of distance is favourable, as well as the free energy being low, but further refinement would be needed to attain a perfect match with the native structure such as the distance between TM 2 and TM 3 although this is partly due to the angling of the helices affecting the distance.

These results are promising in terms of predicting membrane protein structure and indicate the potential application for this approach in further work.

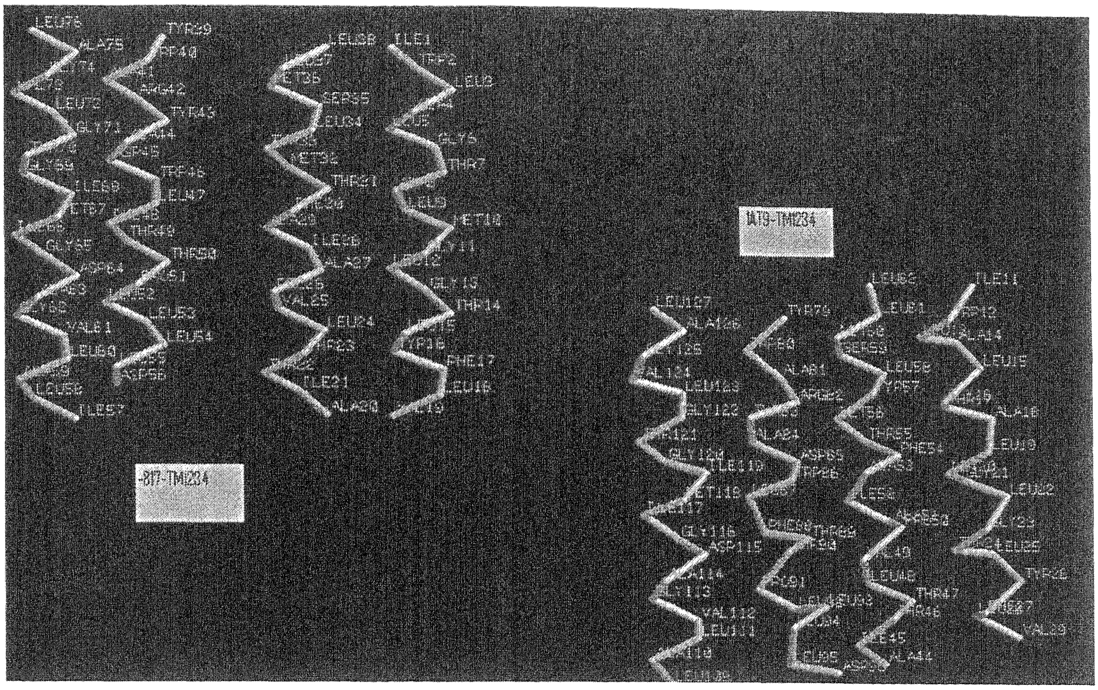


Figure 7.3 The comparison between TM1,TM2,TM3 and TM4 (-817KJ/mol) with the same TM regions in 1AT9

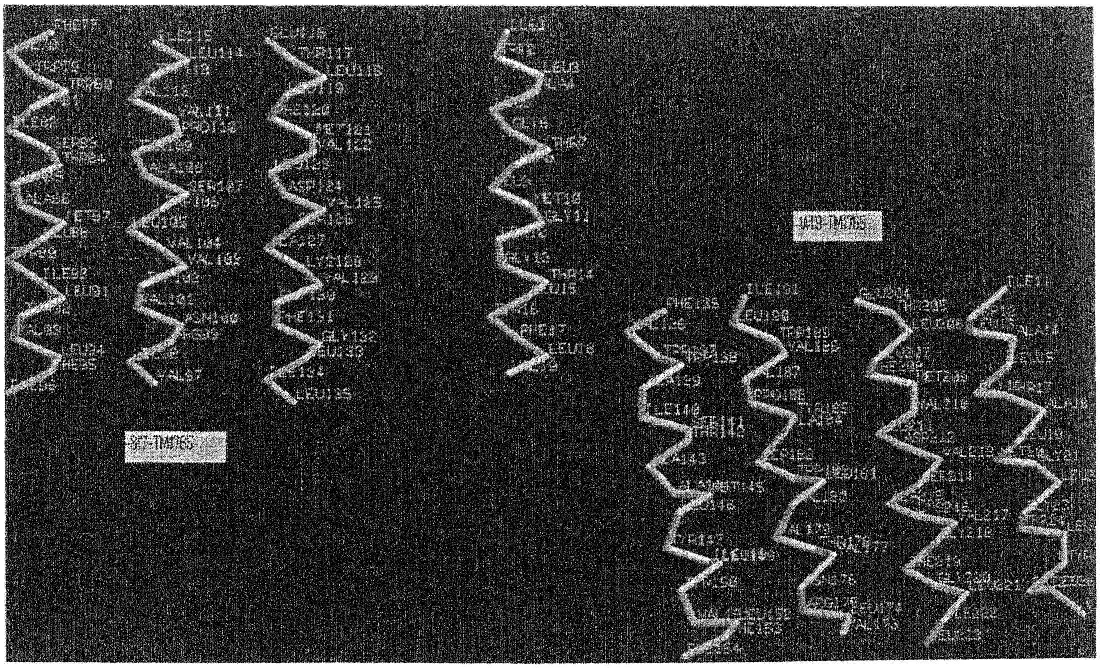
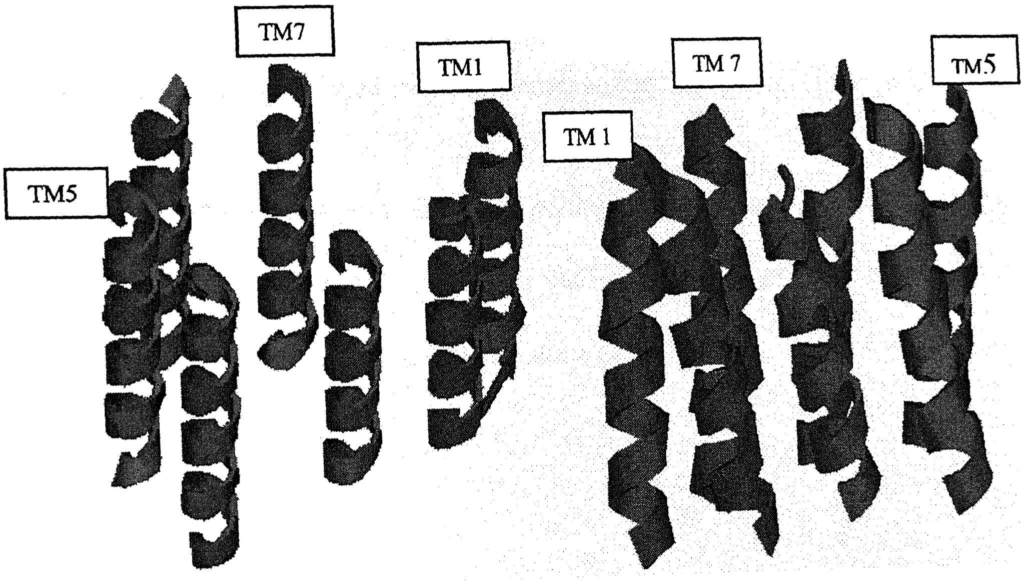


Figure 7.4 The comparison between TM1,TM7,TM6 and TM5 (-817KJ/mol) with the same TM regions in 1AT9

Figures 7.3 and 7.4 show the TM regions of the predicted structure (-817 KJ/mol), although the similarity between the predicted structure (-817 KJ/mol) and the experimentally determined structure is reduced (RMSD increased) but the stability of the new predicted structure is higher than other predicted structure.

This indicated that distances between amino acids on adjacent helices are similar to the native structure, suggesting that in this predicted structure, the amino acids make interactions with each other as in the native structure, but the structure is not sufficiently similar for this to be reflected in a lower RMSD value.

This result indicates that the interactions between helices closely depend on the distances and angles between TM regions, and the stability of TM protein structure might be disproportionately increased by virtue of the arrangement of helices in this template, where they are all assumed to pass straight through the membrane with constant angle.



-817-predictd

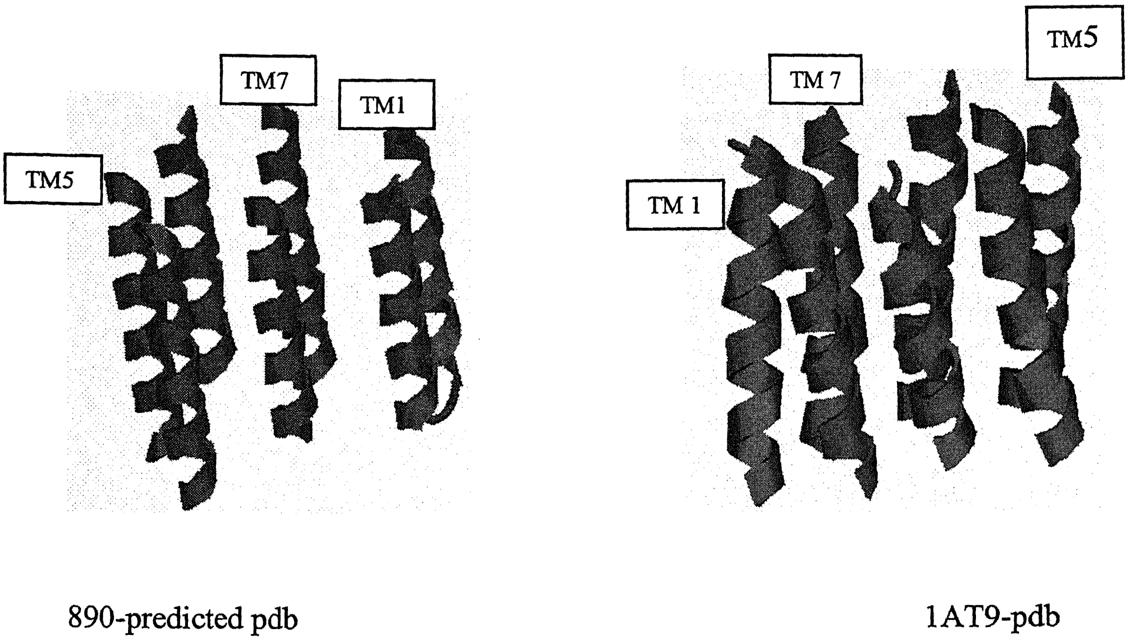
pdb 1AT9 pdb

Figure 7.5 Comparison between predicted structure and known structure.

Figure 7.5 shows the arrangement of the 7 TM regions in the predicted structure with lowest free energy (-817 KJ/mol) and that of the native structure (1AT9). The approach produced different distances between the alpha carbon backbones of packed transmembrane regions in the predicted structure. This allowed the generation of a structure with high stability and low free energy, but the distances between TM regions are a poor match to those of the native structure.

This indicated that the GA method generated a viable structure with varying distances but comparable stability. This structure is a mismatch with the native structure but its new features could be used in genetic engineering.

This comparison suggests that the performance of the GA technique could be improved even further by considering the helix tilt for each specified TM region in order to optimise the predicted 3D structure of TM protein.



890-predicted pdb

1AT9-pdb

Figure 7.6 Comparison between predicted structure and known structure.

The comparison between another predicted structure and the known structure is shown in the figure 7.6. This result indicated that the GA is also able to predict a new structure with high similarity RMSD 2.5Å but relatively low stability with potential energy 890 KJ/mol. This result suggests that the performance of the GA technique in TMGA system was quite impressive, but that it requires refinement in order to generate structures of both high structural reliability and high stability. This improvement could be performed by evaluating the interaction between all residue pairs in their environment for each structure in order to find more accurate structure in terms of TM region adjacency and end-on rotation also for residues appearing at the correct transmembrane helix depth on adjacent helices.

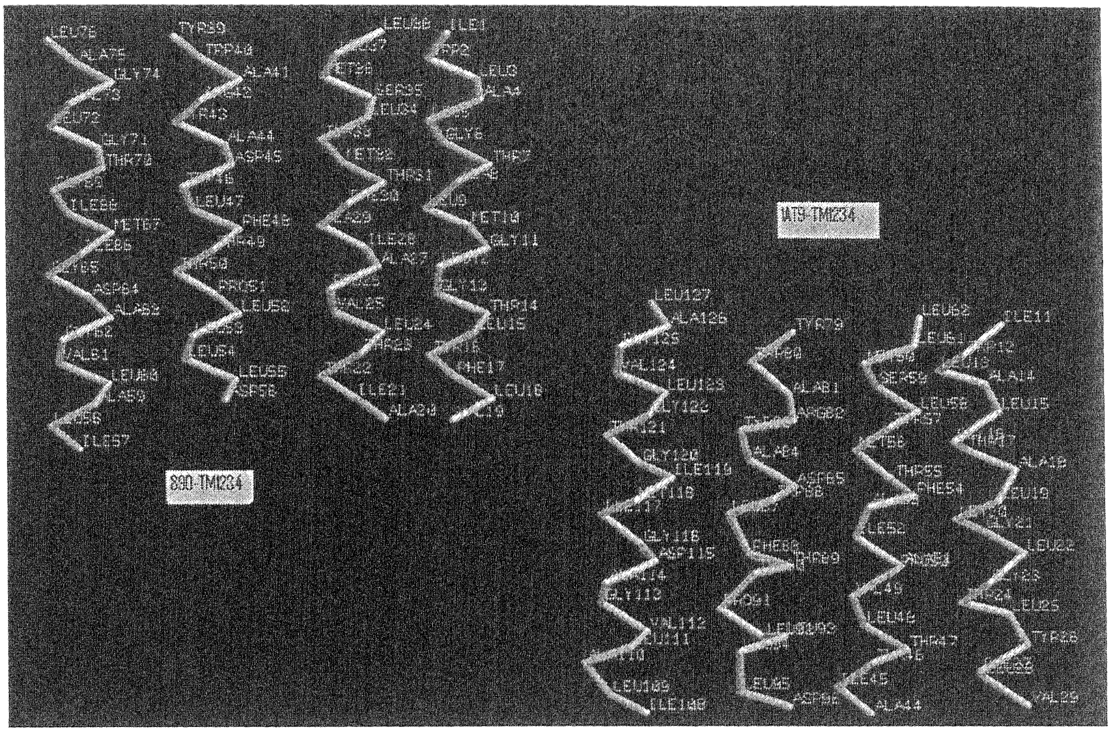


Figure 7.7 The comparison between TM1,TM2,TM3 and TM4 (8907KJ/mol) with the same TM regions in 1AT9

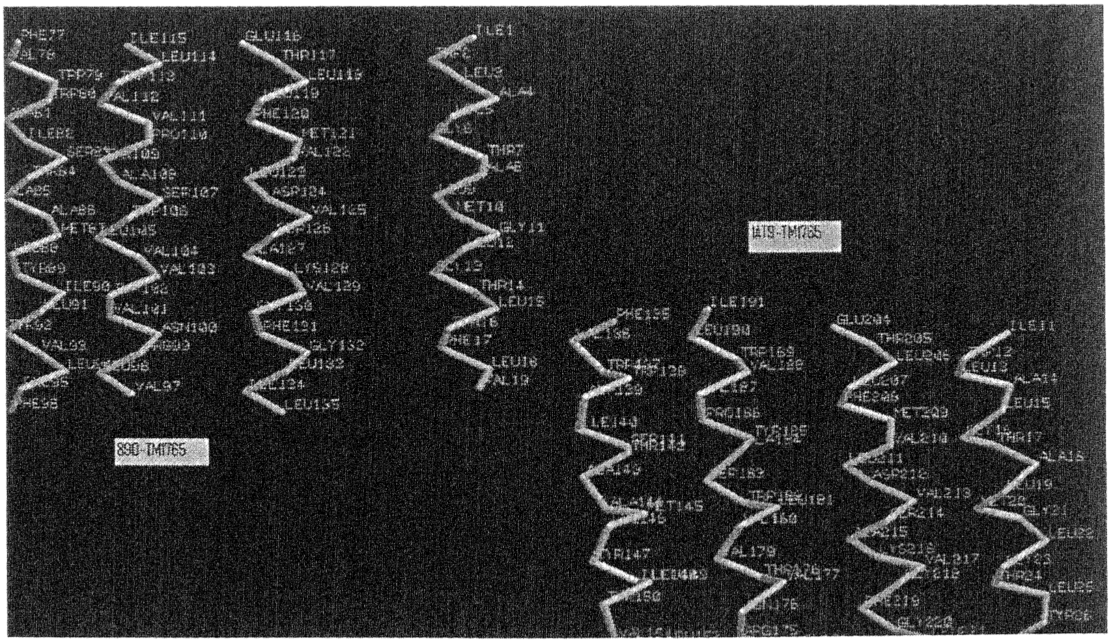


Figure 7.8 The comparison between TM1,TM7,TM6 and TM5 (890KJ/mol) with the same TM regions in1AT9

Figure 7.7 and 7.8 show the arrangement of TM regions in the native bacteriorhodopsin structure (1AT9) and a predicted structure with energy 890 KJ/mol. The distance between TM2 and TM3 and also TM1 and TM7 are reduced compared with a predicted structure with energy -817 KJ/mol. The distance could have an effect on the interaction between amino acids on adjacent helices and cause a less stable structure although its similarity is increased.

The comparison between the arrangement and orientation of TM regions in two predicted structures (with free energy of 890 KJ/mol and -817 KJ/mol respectively) suggested that reliable distances between adjacent TM regions and reliable end-on rotation could bring about correct interaction between the residue pairs and their environment which leads to increased stability of predicted structures.

As the results indicate, the TMGA system is able to solve the problem of prediction of TM region proximity and orientation by application of the GA technique that is evidently an effective and fast approach for problems where an optimal solution must be found in an enormous conformational search space.

7.2 Improving GA Technique in this application

The first set of experiments reported in chapter 6 show the effectiveness of GA as a search technique for predicting TM protein structure. The results show that in order to use the GA in this application some modifications are required.

As the conformation space for membrane protein is large (Faulon et al., 2003) this approach proved that with using a population of 100, it is possible to optimise the prediction of the 3D structure of a TM protein. Figure 6.3 shows the improved results obtained by using three input variables (x,z, and r) and selecting a population of 100.

Jones et al. (1998) suggested that small populations may find a solution quickly but large populations allow the search space to be sampled more thoroughly as shown in figure 6.4, where using a population of 50 with two input variables was shown to be unable to generate a structure with low energy. These results show that the selection of input variables in the GA technique is crucial, as well as population size, in the prediction of membrane protein structure.

As shown, when using a population of 50 with three input variables the GA is able to generate structures with low energy but the use of a small population limits the search space to modify the new structure. These experiments proved that a population of 100 for three input variables is able to successfully modify the new structures, resulting in predicted structures that are near to native structure.

In this study, the selection operator is based on the Roulette Wheel technique in order to invoke a random selection, as for a number slot from a roulette wheel. Fitter individuals of a population are represented by a higher number of slots compared to other individuals with lower fitness.

The GA process has been designed to make sure that in each generation parents with the same energy value are not selected together and also to discard those solutions with higher energy value from the population. By improving the selection process, the chance of finding solutions with lower energy values will be increased for the next subsequent generation.

The GA operators are improved in this approach in order to develop the conformation space for membrane protein structure and be able to optimise the 3D structure. Three point mutation is implemented randomly on three different variables (x,y, and r) which could be from the same alpha carbon atom starting positions, the starting atom of the initial residue of each helix or selected from different atoms. Each point mutation was considered for one of the variables that is able to change in a fixed range in order to prevent generating unexpected structure. The mutation probability is very low to prevent convergence to a local optimum (Jones *et al.*,1998). In the TMGA system, 0.02 is selected for the mutation rate.

Khimasia *et al.* (1997) designed a GA for prediction of protein structure by using the using HP model. They found that simple genetic algorithms (SGA) are not promising in themselves and for improving SGA they suggested multi-points crossovers.

In order to predict new membrane protein structures, the TMGA system has been designed to implement two point crossovers, with each point located on each of two TM regions. The crossover probability controls the rate at which solutions are subjected to crossover. In this approach, it is 0.8 as higher values for crossover probability can introduce new solutions to the population but can also be disrupted faster (Srinivas and Patnaik, 1994).

Won et al. (2005) used a GA for prediction of protein secondary structure. They utilised crossover by choosing two parent strings at random. Each parent is represented as a sequence of block structure on HMM topology. Some of the numbers of each block are swapped randomly to create two children. The position of the block does not carry any meaning so there is no constraint imposed on which block is swapped.

In this work, each parent was represented as a sequence of the alpha helical TM regions and is specified by the position {x, y and z} of the atoms in a protein that is defined by a Protein Data Bank (PDB) file. Each position holds a specific meaning that allows the crossover operator to swap those positions of axis coordinates that belong to certain TM regions.

The evaluation function in this project is based on free energy which is calculated by SwissPDBViewer. Swiss-PdbViewer includes a version of the GROMOS 43B1 force field (Gunsteren et al., 1996). This force field allows the evaluation of the energy of a structure as well as repairing distorted geometries through energy minimization.

The TMGA system has been designed to cooperate with the SwissPDBViewer in order to calculate the energy of each predicted structure. This cooperation is effective and less time consuming in use, and allows faster generation of results.

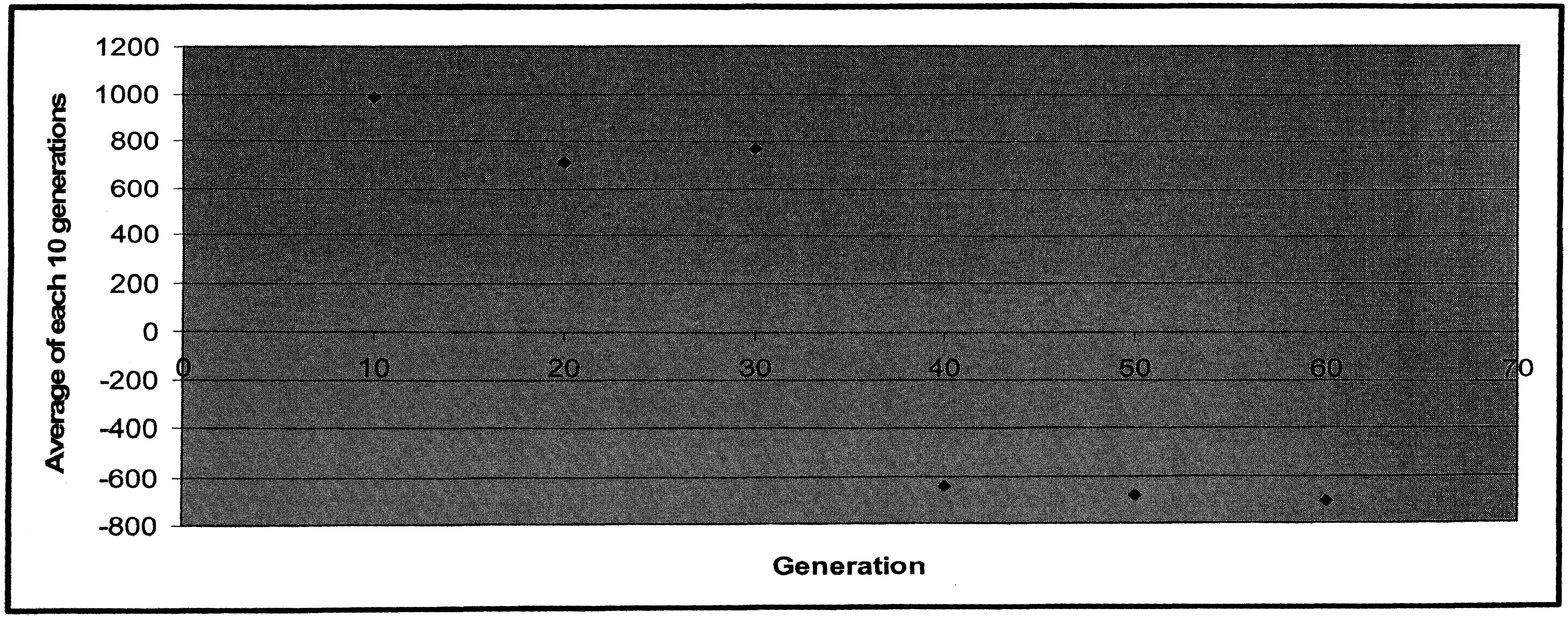


Figure 7.9 The performance of GA after each 10 generations against average energy

Figure 7.9 illustrates the performance of the GA in this application. The average energy has been calculated for each ten generations in order to show the general improvement of results with successive generations using the GA technique. This graph indicates that the average energy is reduced gradually after 50 generations after which there is no significant change in the next 10 generations and this is taken as the termination point of the GA.

7.3 Accuracy of the TMGA system

The TMGA system was tested in order to compare the list of residue pairs on adjacent helices found within a given distance, obtained for the native structure against the generated structures. This comparison was performed at three different distances, namely: 5Å, 8Å, and 10Å, counting the number of coincident pairs of residues found comparable distances apart in the predicted and native structures.

As figure 6.9 shows, at 5Å distance the percentage of coincident pairs of residues between predicted and native structures remains at a similar level throughout the generations, but variability between consecutive generations decreases as the run progresses i.e. with increased number of generations, the results become less variable.

In general, the accuracy of predicted structures indicated the performance of the GA after 30 generations in being able to effectively explore the search space of a membrane protein structure to generate structures with high similarity in which the average number of coincident residue pairs at close range between different helices is increased.

In figure 6.10, the percentage of coincident residue pairs within a greater distance range is shown. By increasing the distance limit to 8Å, generally the percentage accuracy is increased and the variation between consecutive generations is less than at 5Å.

At this distance, the comparison is performed for a long list obtained for the native structure against the generated structures and indicates higher accuracy, especially after 28 generations.

The distance range of 8 Å was selected as this is close to the average distance between the alpha carbon backbones of tightly packed transmembrane regions in determined membrane protein structures. The results in the region of 20% are indicative of predicted structures that consist of transmembrane regions which are placed at feasible distances apart, as well as possessing residues clearly subjected to accurate end-on helical rotation and fixing of membrane depth.

The last comparison was performed at 10Å distance and this produced the largest set of comparable residue pairs. As figure 6.11 shows, the percentage of coincident residues for each new structure and the native structure are high as well as the accuracy of the GA. The 10 Å range is a more generous limit, based on the upper limit of interhelical distances between alpha carbon backbones found in determined membrane protein structures, and allows for all residue pairs that are found in both structures with comparable TM region adjacency, comparable end-on rotation and at comparable membrane depth.

These results show that the GA is a promising technique for the prediction of the 3D structure of membrane protein with high stability and reasonable similarity and could obtain better accuracy in terms of similarity by further improvement such as adding extra feature for each TM region in order to estimate the angles between TM regions as well as distances and end-on rotation. The second improvement will be considered the interaction between TM regions with each other and its environment in order to use as an evaluation function instead of calculating the free energy.

Generally, the TMGA system proved successful in utilising an evolutionary approach by utilising a GA in order to predict a membrane protein structure which is near to native structure. In this research, the effectiveness of evolutionary algorithms to predicting membrane protein structure is demonstrated.

Figure 6.12 shows the percentage of correctly predicted residue pairs for the predicted structure file with the lowest energy in each generation. The results indicate that there are certain predicted structures with less stability and high percentages of coincident residue pairs, and as shown in the figure, the highest percentage accuracy is associated with the relatively high free energy of 890 KJ/mol.

The GA proved that it is able to create these predicted structures after 27 generations by selecting the fittest structures with comparable TM region adjacency, comparable end-on rotation and at comparable membrane depth. This structure contains the combination of its parents features that caused to be more similar to the native structure but possess less stability.

This indicates the advantage of the GA in exploring the search space of the predicted structures and is able to terminate the process after 56 generations by selecting the more fit structure and eliminating unfit structures.

7.4 Summary

A number of approaches are described in this chapter, and the outcomes suggest that a GA has been successfully applied to predicting membrane protein structure for the first time. In this application, the GA is able to optimise the distance between helices as well as the rotation of each helix in order to generate solution structures with reasonable interhelical distances and high stability.

Comparison between newly predicted structures and the native structure indicate that the developed GA technique represents an efficient and fast method for prediction of TM protein structure.

A number of modifications have been made to the GA method in the TMGA system for prediction of TM protein structure. These enhancements allow the prediction of new structures of reasonable similarity to native structure and of acceptable energetic stability.

The performance of the GA technique could be improved even further if more features of each TM region are added to further refine structural predictions

CHAPTER 8

8 Conclusion and further work

This chapter details the main conclusions that may be drawn from the experiments performed using the TMGA system. The future work intended to further enhance the design and application of the genetic algorithm for the prediction of membrane protein structure is discussed.

8.1 Conclusion

This thesis investigates a new approach for prediction of transmembrane protein structure by using the TMGA system. The developed TMGA system applies a modified genetic algorithm technique to generating new structures that are near to native structure in terms of energetic stability.

The main findings in this research is the impact of the choice of what would compose an individual, upon which operations and selections are made and how the solutions were represented as an approximate 3D template of TM protein structure. The results show that the selection of input variables in the GA technique is crucial, as well as the population size.

Likewise, the method of predicting the 3D structure of TM proteins by using a GA technique to manipulate the distance between helices, and the rotation of each TM region, and using an energy function to evaluate each new structure is novel.

The wide search space was explored efficiently by means of manipulating the wide range of possible distances between TM helices and also the rotation of each TM region, since comparative results were found to change drastically for different distance and rotation values.

The successful generation of feasible TM protein structures illustrates the potential of the GA method for solving the membrane protein structure prediction problem. This approach is simple, fast and efficient and is able to cooperate with other software to reduce computational time. In addition, the generation of new TM protein structures with new features prove the effectiveness of evolutionary approaches in this work. The approach has potential application to the field of membrane protein engineering.

As a result of these experiments, it can be concluded that using a genetic algorithm method proved to be reliable and robust approach. It is evident that the power of genetic algorithms is greatly enhanced by automated combination with structural bioinformatics applications.

These approaches should prove most valuable to the field of membrane protein structural bioinformatics, where fewer than 1% of protein structures are known, as the vast majority of proteins, the structure of which have not been determined, will now be open to improved and evolving prediction. This technique can be used to predict the unknown membrane protein structure based on the energetic stability and transmembrane region arrangement and orientation.

8.2 Further work

Following the successful prediction of a TM protein structure near to native structure, the next step is to be to enhance the TMGA system to predict structures with consistently high similarity and stability.

The most fundamental improvement to be made is to consider and manipulate the angles of helix tilt as in these experiments each template contains of 7 helices that are assumed to pass in parallel, straight through the membrane.

The next step would be to increase the search space by adding the variable of helix breaks and kinks. The possibility of helix breakage and kinking in TM regions are ignored in the current template. In the future, these variables will be investigated for each helix in order to generate protein structures with higher potential accuracy.

Additionally, the TMGA system may require modification to allow a wider evaluation of candidate structures, such as the incorporation of weighted scores for favoured structural motifs and arrangements, rather than being based solely on evaluation by calculation of free energy.

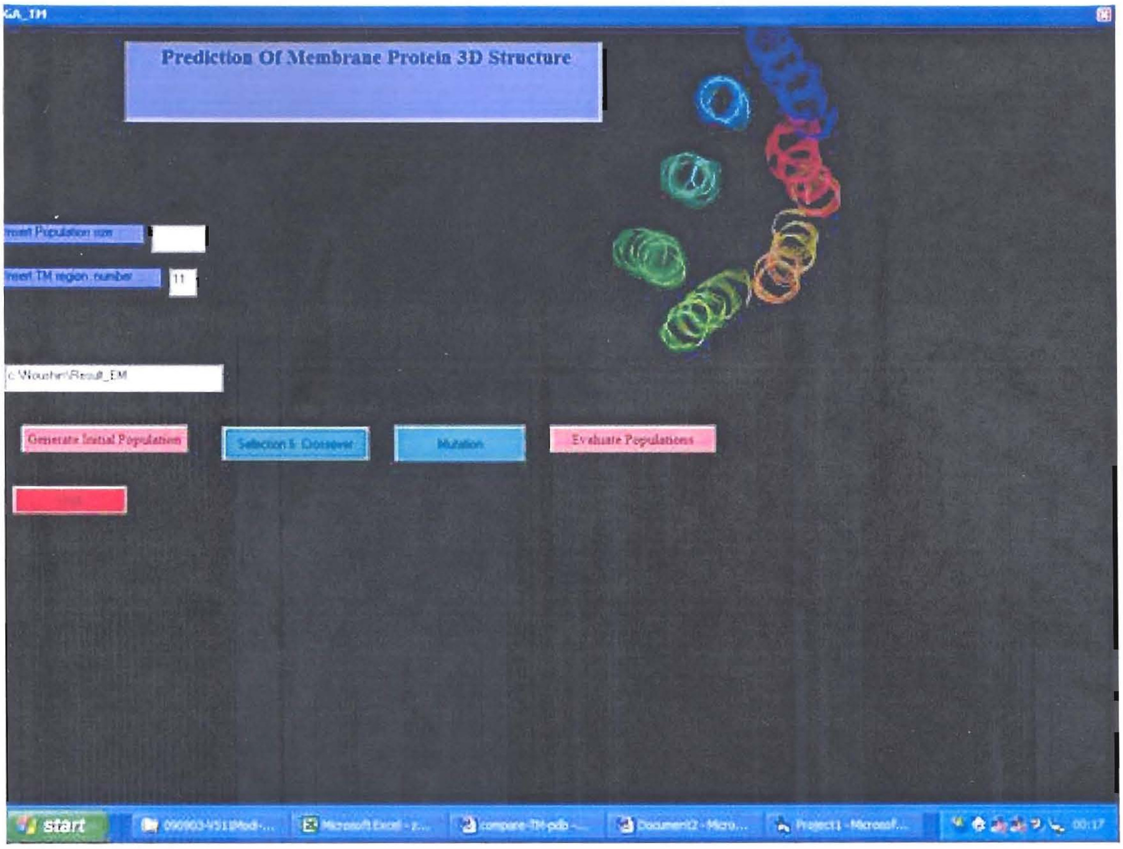
Similarly, it may be worthwhile to account for known stabilising interactions with the lipid environment of the membrane, rather than relying solely on the stability of peptide / peptide interactions. The current work is based on the determined protein structure of bacteriorhodopsin, which was chosen in order to evaluate the approach. The future application of this work will be to proteins of hitherto undetermined 3D structure.

In the longer term, these approaches may find valuable application in the fields of protein engineering and medicine.

Appendix (A)

TMGA System Screens Shot

This section contains a number of screen –captures to demonstrate the operation of the developed·TMGA system.



Prediction Of Membrane Protein 3D Structure

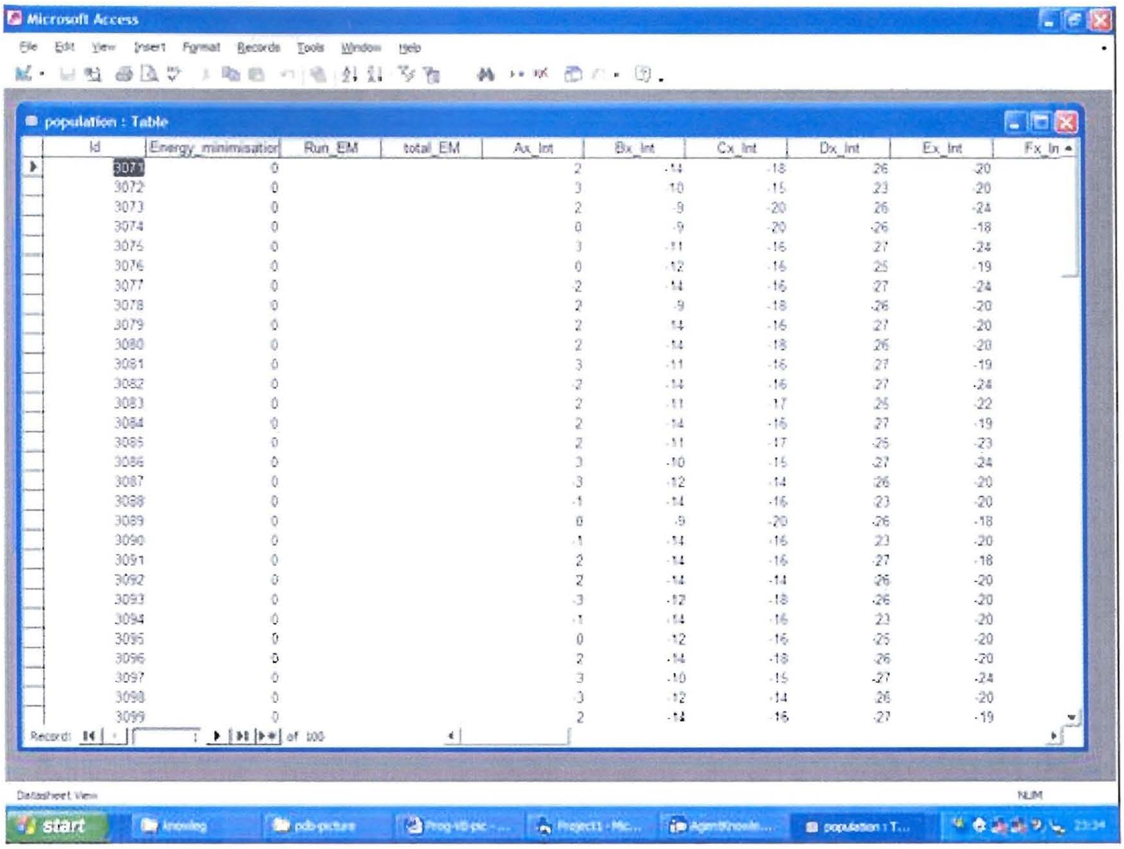
Insert Population size

Insert TM region number 11

c:\Noushin\Result_EM

Figure A.1 TMGA system interface

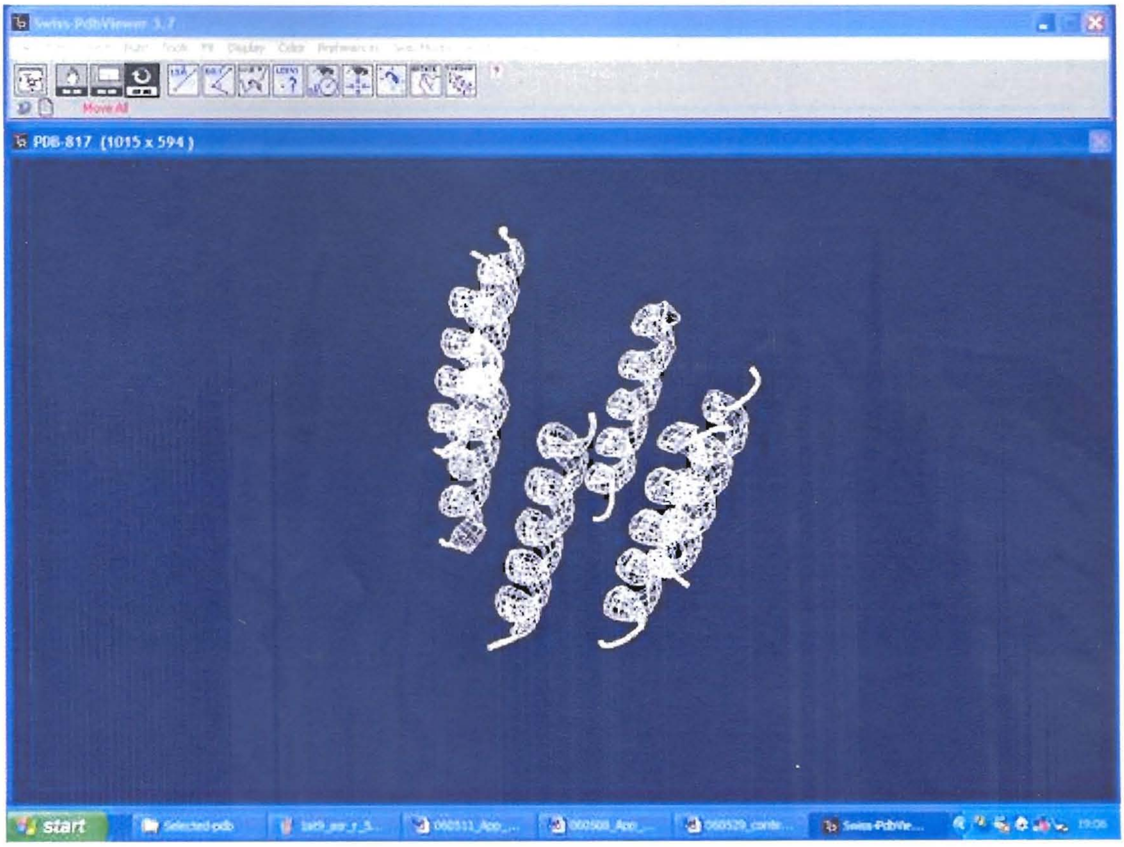
Figure A.1 shows the main interface of the TMGA system. This interface includes system menus which contains the different operations used during the implementation of GA method.



Id	Energy_minimisation	Run_EM	total_EM	Ax_Int	Bx_Int	Cx_Int	Dx_Int	Ex_Int	Fx_In
3071	0			2	-14	-18	26	-20	
3072	0			3	-10	-15	23	-20	
3073	0			2	-9	-20	26	-24	
3074	0			0	-9	-20	-26	-18	
3075	0			3	-11	-16	27	-24	
3076	0			0	-12	-16	25	-19	
3077	0			-2	-14	-16	27	-24	
3078	0			2	-9	-18	-26	-20	
3079	0			2	14	-16	27	-20	
3080	0			2	-14	-18	26	-20	
3081	0			3	-11	-16	27	-19	
3082	0			-2	-14	-16	27	-24	
3083	0			2	-11	17	25	-22	
3084	0			2	-14	-16	27	-19	
3085	0			2	-11	-17	-25	-23	
3086	0			3	-10	-15	-27	-24	
3087	0			-3	-12	-14	26	-20	
3088	0			-1	-14	-16	23	-20	
3089	0			0	-9	-20	-26	-18	
3090	0			-1	-14	-16	23	-20	
3091	0			2	-14	-16	-27	-18	
3092	0			2	-14	-14	26	-20	
3093	0			-3	-12	-18	-26	-20	
3094	0			-1	-14	-16	23	-20	
3095	0			0	-12	-16	-25	-20	
3096	0			2	-14	-18	-26	-20	
3097	0			3	-10	-15	-27	-24	
3098	0			-3	-12	-14	26	-20	
3099	0			2	-14	-16	-27	-19	

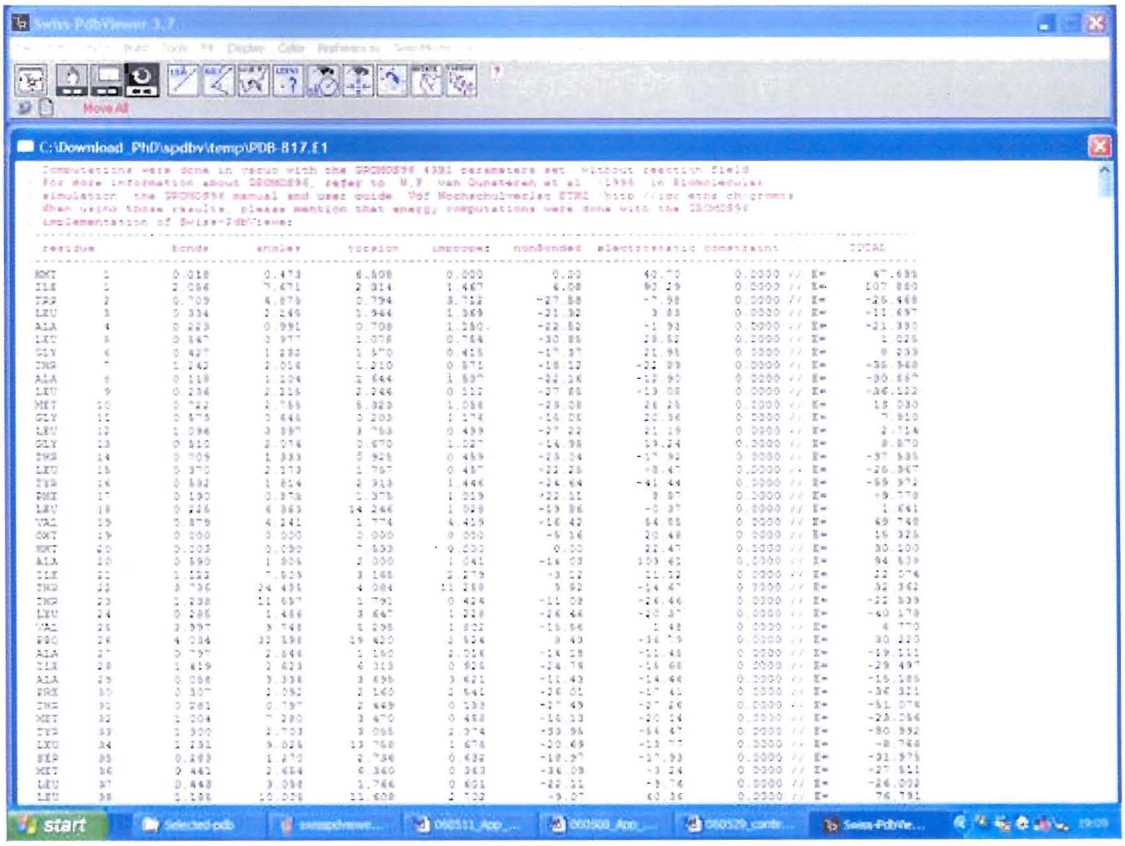
Figure A.2 Population of predicted structures in a table from Microsoft access

Figure A.2 illustrates a population of predicted structures in Microsoft access. Each row represents an arrangement of 7 alpha carbon atom starting positions {x, y, z}. The free energy of each predicted structure is added to the table when is calculated.



FigureA.3 The representation of predicted structure in SwissPDBViewer

This interface represents the predicted structure in SwissPDBViewer software that has been generated by TMGA system to be evaluated.

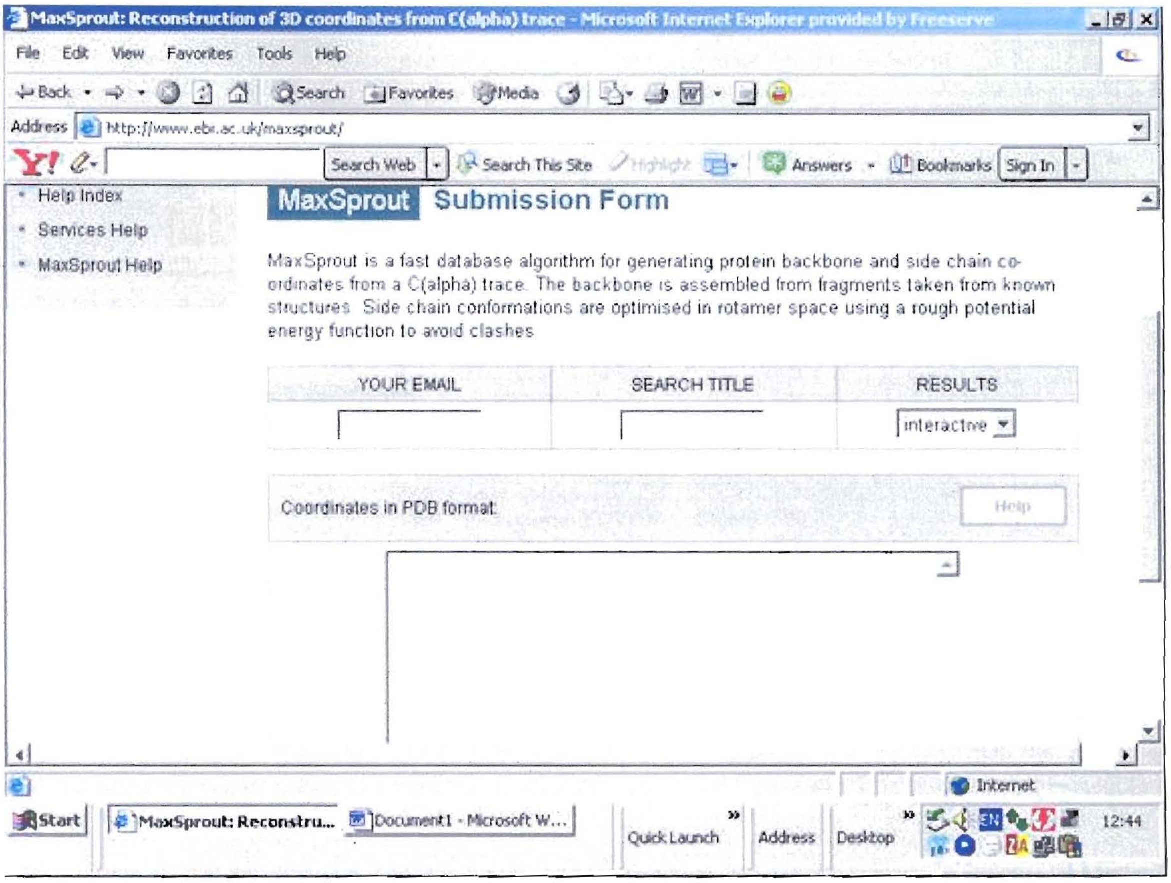


C:\Download_PhD\spdbv\temp\PDB-817.E1

```
Computations were done in vacuo with the GROMOS96 43B1 parameters set, without reaction field
For more information about GROMOS96, refer to W.F. van Gunsteren et al. (1996) in Biomolecular
simulation: the GROMOS96 manual and user guide. Vdf Hochschulverlag ETHZ (http://igc.ethz.ch/gromos)
When using those results, please mention that energy computations were done with the GROMOS96
implementation of Swiss-PdbViewer
-------------------------------------------------------------------------------------------------------------
residue          bonds     angles    torsion   improper  nonbonded electrostatic constraint      TOTAL
-------------------------------------------------------------------------------------------------------------
NMT     1        0.018     0.473     6.608     0.000      0.02       60.70      0.0000 //  E=   67.695
ILE     1        2.056     7.671     2.314     1.467      4.08       92.29      0.0000 //  E=  107.880
TRP     2        0.709     4.879     0.794     3.712    -27.88       -7.98      0.0000 //  E=  -26.468
LEU     3        0.334     2.149     1.944     1.369    -21.32        3.83      0.0000 //  E=  -11.697
ALA     4        0.223     0.991     0.708     1.250    -22.52       -1.93      0.0000 //  E=  -21.330
LEU     5        0.147     0.977     1.078     0.744    -30.85       29.62      0.0000 //  E=    1.026
GLY     6        0.427     1.232     1.570     0.415    -17.37       21.95      0.0000 //  E=    8.239
THR     7        1.242     2.016     1.210     0.571    -18.12      -22.89      0.0000 //  E=  -35.960
ALA     8        0.119     1.104     2.644     1.537    -22.16      -12.90      0.0000 //  E=  -30.667
LEU     9        0.236     2.215     2.244     0.112    -27.85      -13.08      0.0000 //  E=  -36.122
MET    10        0.722     2.755     5.023     1.056    -29.08       26.25      0.0000 //  E=   18.030
GLY    11        0.573     0.644     0.203     1.174    -16.05       20.36      0.0000 //  E=    7.910
LEU    12        1.096     3.397     3.753     0.499    -27.22       21.19      0.0000 //  E=    2.714
GLY    13        0.810     2.074     0.670     1.027    -14.98       19.24      0.0000 //  E=    8.870
THR    14        0.709     1.333     5.925     0.459    -23.34      -17.92      0.0000 //  E=  -37.535
LEU    15        0.370     2.173     1.757     0.457    -22.25       -8.47      0.0000 //  E=  -26.367
TYR    16        0.592     1.614     2.313     1.446    -24.64      -41.44      0.0000 //  E=  -59.371
PHE    17        0.190     0.878     1.378     1.019    -22.11        8.87      0.0000 //  E=   -9.770
LEU    18        0.226     4.363    14.246     1.028    -19.96       -0.37      0.0000 //  E=    1.661
VAL    19        0.879     4.241     1.774     4.419    -16.42       54.85      0.0000 //  E=   69.748
OXT    19        0.000     0.000     0.000     0.000     -5.16       20.48      0.0000 //  E=   16.326
NMT    20        0.003     0.090     7.693     0.200      0.00       22.47      0.0000 //  E=   30.100
ALA    20        0.590     1.906     2.000     1.041    -14.03      109.42      0.0000 //  E=   94.639
ILE    21        1.122     7.829     3.148     2.279     -3.22       11.32      0.0000 //  E=   22.076
THR    22        3.736    24.435     4.084    11.259      5.92      -14.67      0.0000 //  E=   32.362
TRP    23        1.238    11.657     1.791     0.424    -11.09      -26.46      0.0000 //  E=  -22.539
LEU    24        0.285     1.486     3.647     1.228    -26.44      -20.37      0.0000 //  E=  -40.179
VAL    25        2.997     3.748     5.295     1.802    -15.56        1.42      0.0000 //  E=    4.770
PRO    26        4.034    32.598    19.620     1.524      3.43      -16.79      0.0000 //  E=   30.320
ALA    27        0.797     2.844     1.160     2.016    -14.28      -11.48      0.0000 //  E=  -19.111
ILE    28        1.419     2.623     6.313     0.925    -24.78      -15.68      0.0000 //  E=  -29.497
ALA    29        0.058     3.338     3.695     3.621    -11.43      -14.64      0.0000 //  E=  -15.185
PHE    30        0.307     2.052     3.160     1.641    -26.01      -17.41      0.0000 //  E=  -36.321
THR    31        0.281     0.797     2.469     0.133    -27.49      -27.26      0.0000 //  E=  -51.076
MET    32        1.204     7.280     3.470     0.458    -16.13      -20.14      0.0000 //  E=  -23.256
TYR    33        1.300     2.703     3.065     2.374    -53.95      -54.47      0.0000 //  E=  -80.992
LYS    34        1.231     9.026    13.758     1.678    -20.69      -13.77      0.0000 //  E=   -8.768
SER    35        0.283     1.272     2.734     0.632    -18.97      -17.93      0.0000 //  E=  -31.975
MET    36        0.441     2.654     6.360     0.363    -34.09       -3.24      0.0000 //  E=  -27.511
LEU    37        0.448     3.058     1.766     0.601    -22.11       -9.76      0.0000 //  E=  -26.002
LEU    38        1.186    10.028    11.608     2.722     -9.07       60.36      0.0000 //  E=   76.791
```

Figure A.4 Interface for calculated free energy by SwissPDBViewer for the predicted structure

The Figure A.4 shows the results of the predicted structure after calculating the minimum free energy.



MaxSprout: Reconstruction of 3D coordinates from C(alpha) trace - Microsoft Internet Explorer provided by Freeserve

Address http://www.ebi.ac.uk/maxsprout/

MaxSprout Submission Form

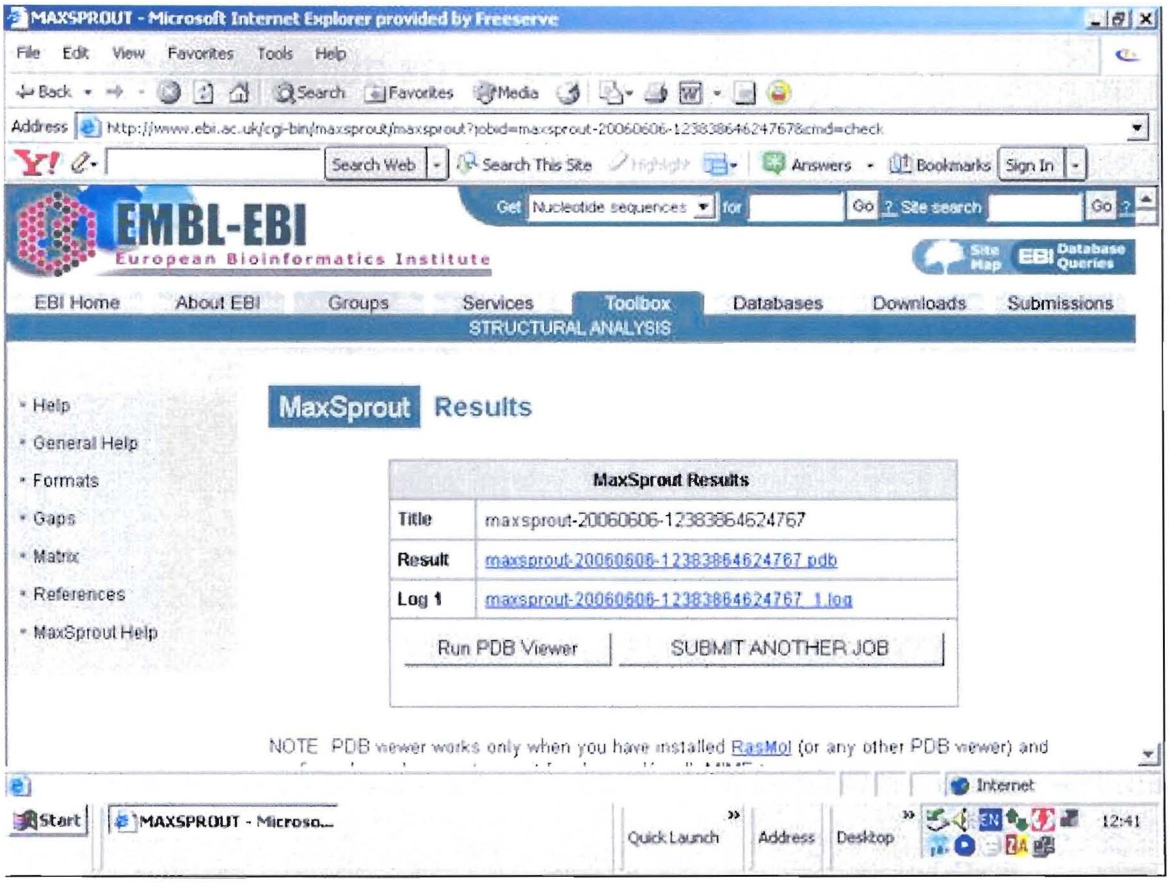
MaxSprout is a fast database algorithm for generating protein backbone and side chain co-ordinates from a C(alpha) trace. The backbone is assembled from fragments taken from known structures. Side chain conformations are optimised in rotamer space using a rough potential energy function to avoid clashes.

YOUR EMAIL	SEARCH TITLE	RESULTS
		interactive

Coordinates in PDB format:

Figure A. 5 Interface for MaxSprout submission form

Figure A5 shows MaxSprout web site for submitting PDB file which is generated in TMGA system. The MaxSprout program generates the side-chain for predicted PDB file.


MAXSPROUT - Microsoft Internet Explorer provided by Freeserve
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.ebi.ac.uk/cgi-bin/maxsprout/maxsprout?jobid=maxsprout-20060606-12383864624767&cmd=check
Search Web Search This Site Highlight Answers Bookmarks Sign In
EMBL-EBI
European Bioinformatics Institute
Get Nucleotide sequences for Go ? Site search Go
Site Map EBI Database Queries
EBI Home About EBI Groups Services Toolbox Databases Downloads Submissions
STRUCTURAL ANALYSIS
Help
General Help
Formats
Gaps
Matrix
References
MaxSprout Help
MaxSprout Results
Run PDB Viewer
SUBMIT ANOTHER JOB
Internet
Start
MAXSPROUT - Microso...
Quick Launch
Address
Desktop
12:41

MaxSprout Results	
Title	maxsprout-20060606-12383864624767
Result	maxsprout-20060606-12383864624767.pdb
Log 1	maxsprout-20060606-12383864624767_1.log

NOTE PDB viewer works only when you have installed RasMol (or any other PDB viewer) and

Figure A.6 Interace for MaxSprout results

Figure A6 shows the interface is the result file that contains the PDB file with the side-chain. The TMGA system uses this result in order to evaluate it and calculate its free energy.

Appendix (B)

Bacteriorhodopsin structure

Structure of bacteriorhodopsin at 3.0Å resolution is represented in figure C1. These crystals have been analysed by electron cryo-microscopy.

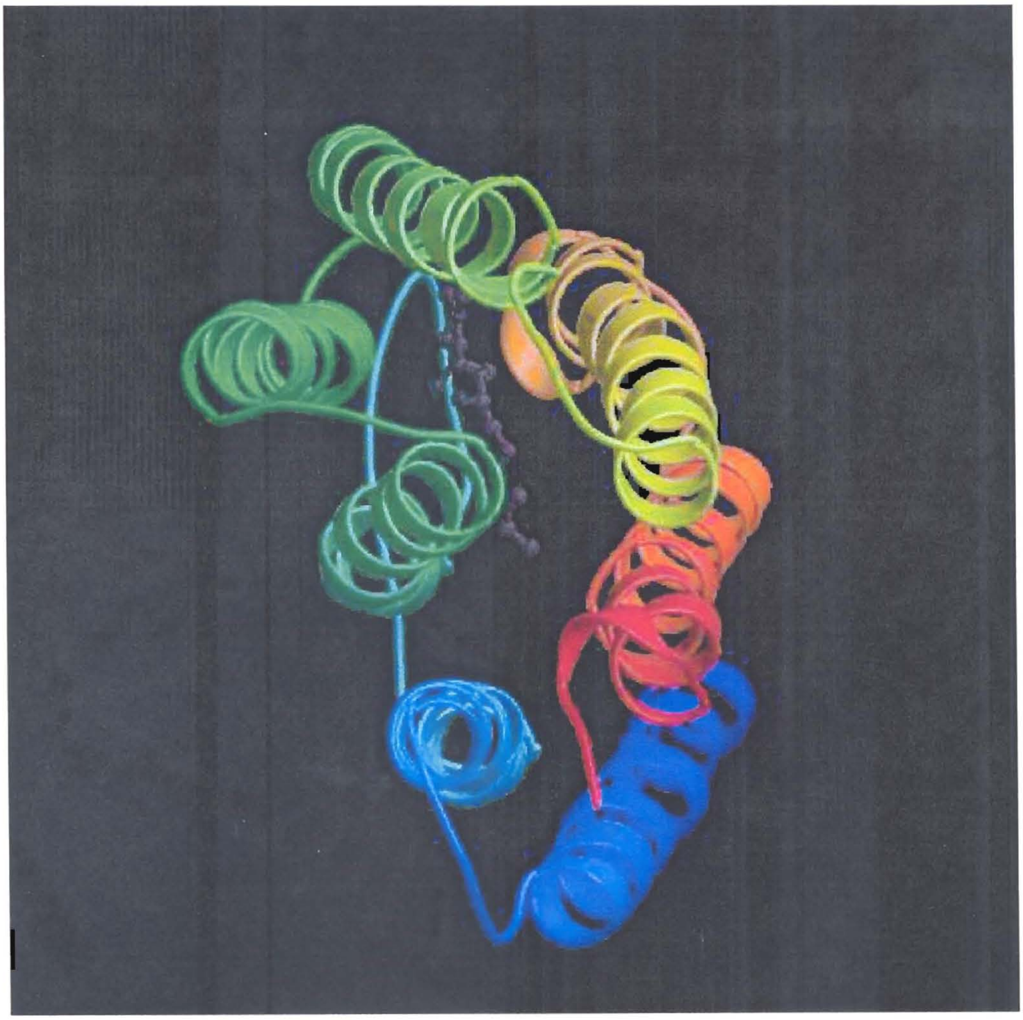


Figure B.1 Structure of Bacteriorhodopsin

Appendix (C)

Predicting 3D structure of membrane protein from its primary sequence

R.C. Togawa [1], J.F. Antoniw [2] and J.G.L. Mullins [3,*]

[1]Bioinformatics Laboratory, Embrapa Genetic Resources and Biotechnology, Parque estação Biológico Final W5 norte. Caixa Postal: 02372 CEP: 70770-900 Brasília-DF, Brazil.. [2]Plant Pathogen Interactions Division, Rothamsted Research, Harpenden, Hertfordshire, AL5 2JQ, UK and [3]Membrane Protein Group – Swansea Clinical School, University of Wales Swansea, Singleton Park, Swansea SA2 8PP, UK.

1 Abstract

We have developed a knowledge based *ab initio* method for predicting the 3D structure of membrane proteins directly from the primary sequence. Based on analysis of existing membrane protein structures, the software creates an association matrix based on the associations between the amino acids that compose the alpha helices of the membrane protein.

[*] To whom correspondence should be addressed.

Using this matrix and a propensity scale called the "kPROT scale" (Pilpel *et al.*, 1999) the software predicts the 3D structure of a given protein, based on the amino acid sequence composition of its' transmembrane regions.

Several programs were developed to analyse determined structures, predict associations between transmembrane regions, and generate low-resolution 3D structural prediction. The programs were integrated as a single module at the end. This approach gives the opportunity to evaluate each component individually, while at the same time, allowing fast prediction of structure from sequence using a single tool. The results achieved provide strong motivation to continue to improve the developed algorithm to obtain better structural predictions, and to this end, structural predictions are currently being refined using evolutionary computing approaches.

Keywords: membrane proteins, 3D structure prediction, bioinformatics.

2 Introduction

Due to the known difficulties of obtaining the 3D structures of membrane proteins from experimental data, many different computational methods of predicting the 3D structure from primary sequence have been instigated in the last decade. Many efficient methods are available to predict general topology from the primary sequence such as: DAS (Cserzö *et al.*, 1997), HMMTOP (Tusnády & Simon, 1998), PSIPRED (McGuffin *et al.*, 2000), PHDhtm (Rost *et al.*, 1996), SOSUI (Hirokawa *et al.*, 1998), TMHMM (Sonnhammer *et al.*, 1998), TMPred (Hofmann & Stoffel, 1993), TopPred (Claros & von Heijne, 1994). These methods are used to predict the transmembrane (TM) regions.

From the increasing number of completed and ongoing genome projects, more and more sequences from different organisms are elucidated and catalogued every day. Using the predictive tools cited above the TM regions are easily predicted. Laboratory methods are also used to define TM regions and the resulting TM region annotations are deposited in the protein sequence repository for public scientific use.

Following the prediction of general topology of TM regions, prediction of their general positions with respect to each other in terms of an end on configuration is the next problem to be solved. In solving this problem, the possible helix packing involved in assembly of the membrane protein 3D structure must be considered. Helix-helix packing plays a critical role in maintaining the tertiary structures of helical membrane proteins (Adamian & Liang, 2001). Studying helix-helix packing is essential in defining the final structure of an α-helical membrane protein.

Using the association matrix derived from known 3D structures of membrane proteins and a propensity scale called the "kPROT scale" (Pilpel *et al.*, 1999), an *ab initio* method for prediction of the associations between TM regions is presented here. This method results in a prediction of the general configuration and helix packing directly from the primary sequence. Also, is described the rotational method to predict the most likely angular position for each TM region and its 3D structure.

3 Materials and Methods

Databases

The primary sequence used as an input for the predictive tool are the Swiss-Prot (Bairoch & Apweiler, 2000) files. The important annotations for the development are the `'FT TRANSMEM'` that contains the transmembrane regions and the `'SQ'` that contains the protein sequence.

PDB (Berman *et al.*, 2000) files are used for the preparation of the association matrix and for the evaluation process. The important annotations are: `'DFREF'` (Contains an equivalence between the PDB file and protein sequence databases – cross-reference between SwissProt and PDB files), `'HELIX'` (Contains an α-helix region defined experimentally by X-ray crystallography or NMR.) and `'ATOM'` (Contains information about each residue in the structure in three-letter code, and contains the atom name, residue number, and XYZ co-ordinates.).

The association matrix

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

For the evaluation process, a 20x20 association matrix was created based on the following membrane protein PDB files:

Table 1 – PDB files used to create the 20x20 association matrix

PDB code	4 5 Description	Correspondent SwissProt code	Number of used TM regions for the calculation
1AP9	Photoreceptor - structure of bacteriorhodopsin from microcrystals grown in lipidic cubic phases.	P02945 (7)	7
1AR1	Complex (oxidoreductase/antibody) - structure of the paracoccus denitrificans two-subunit cytochrome c oxidase complexed with an antibody fv fragment.	P98002 (12), P08306 (2)	14
1E12	Ion pump - halorhodopsin, a light-driven chloride pump	CAB37866 (7)	7
1EUL	Hydrolase - crystal structure of calcium atpase with two bound calcium ions.	P11719 (10)	10
1F88	Signaling protein - crystal structure of bovine rhodopsin.	P02699 (Chain A 7) Chain B 7)	14
1FX8	E. Coli glycerol facilitator (glpf) with substrate glycerol.	P11244 (8)	8
1H2S	Molecular basis of transmembrane signalling by sensory rhodopsin ii-transducer complex.	P42196 (7) P42259 (2)	9
1H68	Photoreceptor - sensory rhodopsin ii.	P42196 (7)	7

1IH5	Crystal structure of aquaporin-1.	P29972 (8)	8
1IWG	Crystal structure of bacterial multidrug efflux transporter acrb.	P31224 (12)	12
1IWO	Hydrolase - crystal structure of the sr ca2+-atpase in the absence of ca2+.	P04191 (Chain A 10) (Chain B 10)	20
1JB0	Photosynthesis - crystal structure of photosystem i: a photosynthetic - reaction center and core antenna system from cyanobacteria	P25896 (11) P25897 (11) P25900 (1) P25901 (1) P20453 (2) P25902 (2)	28
1JGJ	Signaling protein - crystal structure of sensory rhodopsin ii. Insights into color tuning and transducer interaction.	P42196 (7)	7
1KQF	Oxidoreductase - formate dehydrogenase n from e. Coli	P24184 (1) P24185 (4)	5
1KZU	Light-harvesting protein - integral membrane peripheral light harvesting complex from rhodopseudomonas acidophila strain.	P26789 (3) P26790 (3)	6

1L0V	Oxidoreductase - quinol-fumarate reductase with menaquinol molecules.	P03805 (Chain C 3) (Chain O 3) P03806 (Chain D 3) (Chain P 3)	12
1L7V	Transport protein/hydrolase - bacterial abc transporter involved in b12 uptake.	P06609 (Chain A 10) (Chain B 10)	20
1OCC	Oxidoreductase (cytochrome(c)-oxygen) - structure of bovine heart cytochrome c oxidase at the fully oxidized state.	P00396 (Chain A 12) (Chain N 12) P00404 (Chain B 2)(Chain O 2) P00415 (Chain C 7)(Chain P 7) P00423 (Chain D 1) (Chain Q 1) P07471 (Chain G 1)(Chain T 1) P04038 (Chain I 1) (Chain V 1) P07470 (Chain J 1) (Chain W 1) P13183 (Chain K 1) (Chain X 1) P00430 (Chain L 1) (Chain Y 1) P10175 (Chain M 1) (Chain Z 1)	56

1OKC	Carrier protein - structure of mitochondrial adp/atp carrier in complex with carboxyatractyloside	P02722 (6)	6
1Q90	Photosynthesis - structure of the cytochrome b6f (plastohydroquinone : plastocyanin oxidoreductase) from chlamydomonas reinhardtii.	P23577 (Chain A 1), Q00471 (Chain B 4), Q08362 (Chain G 1), P50369 (Chain L 1), Q42496 (Chain M 1)	8
1QLA	Oxidoreductase - respiratory complex ii-like fumarate reductase from wolinella succinogenes.	P17413 (Chain C 5) (Chain F 5)	10
1RC2	Structure of aquaporin z.	P48838 (Chain A 6) (Chain B 6)	12
1RHZ	Protein transport - the structure of a protein conducting channel.	Q60175 (Chain A 11) Q57817 (Chain B 1)	12
1RWT	Photosynthesis - crystal structure of spinach major light-harvesting complex	P12333 (Chain A 3) (Chain B 3) (Chain C 3) (Chain D 3)	30

		(Chain E 3) (Chain F 3) (Chain G 3) (Chain H 3) (Chain I 3) (Chain J 3)	
1S7B	Transport protein - structure of the multidrug resistance efflux transporter - EMRE from escherichia coli.	P23895 (Chain A 3) (Chain B 3) (Chain C 3) (Chain D 3) (Chain E 3) (Chain F 3) (Chain G 3) (Chain H 3)	24

The kPROT scale:

The kPROT scale (Pilpel *et al.*, 1999) was used in the *TMRelate_K* development. The kPROT scale available at http://bioinformatics.weizmann.ac.il/kPROT/kPROTScales uses the knowledge-based propensities for residue orientation in TM segments, showing the value for each amino acid. The kPROT value < 0 indicates that the residue is more prevalent in the TM segments of multiple span proteins and thus assigned a higher propensity to be buried in the TM bundle. On the other hand, residues with positive kPROT values are assigned with a higher tendency to be exposed to the lipid. In the table 2 is shown the general kPROT scale for each amino acid that was used in the development.

Table 2 - The used kPROT scale

kPROT Scale			
Residue	Value	Residue	Value
A	0.0193	M	-0.3120
C	0.2672	N	-0.6757
D	-0.8658	P	-0.5092
E	-0.8551	Q	-0.5367
F	-0.1126	R	0.1782
G	-0.1247	S	-0.2141
H	-0.3423	T	-0.0162
I	0.1248	V	0.2281
K	0.2451	W	-0.1157
L	0.1908	Y	-0.1175

TMDistance algorithm

The input of the *TMDistance* program is the PDB file. The user can load more than one PDB file for the processing.

For each PDB file, the algorithm searches for the "DBREF" tag entry. Once the tag is found, it searches for "SWS" string to find the appropriate Swiss-Prot accession number. If it is necessary the program downloads and saves the corresponding Swiss-Prot file in the working directory. If it is necessary *TMDistance* converts the amino acid sequence numbers between PDB and Swiss-Prot files using the information contained in the DBREF tag (this procedure set the correct TM region based on the 'TRANSMEM' annotation for the PDB file). Then program reads the spatial co-ordinates for the atom in each TM region. With each residue pair in different TM regions, if the distance between the two residues is less than a user-selected distance (3.0Å, 3.5Å 4.0Å, 4.5Å or 5.0Å), the relevant residue-pair is added to the internal bi-dimensional array (matrix counter). After all the PDB file(s) are read, *TMDistance* creates the 20x20 association matrix output with the average distances represented in the internal bi-dimensional array.

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

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

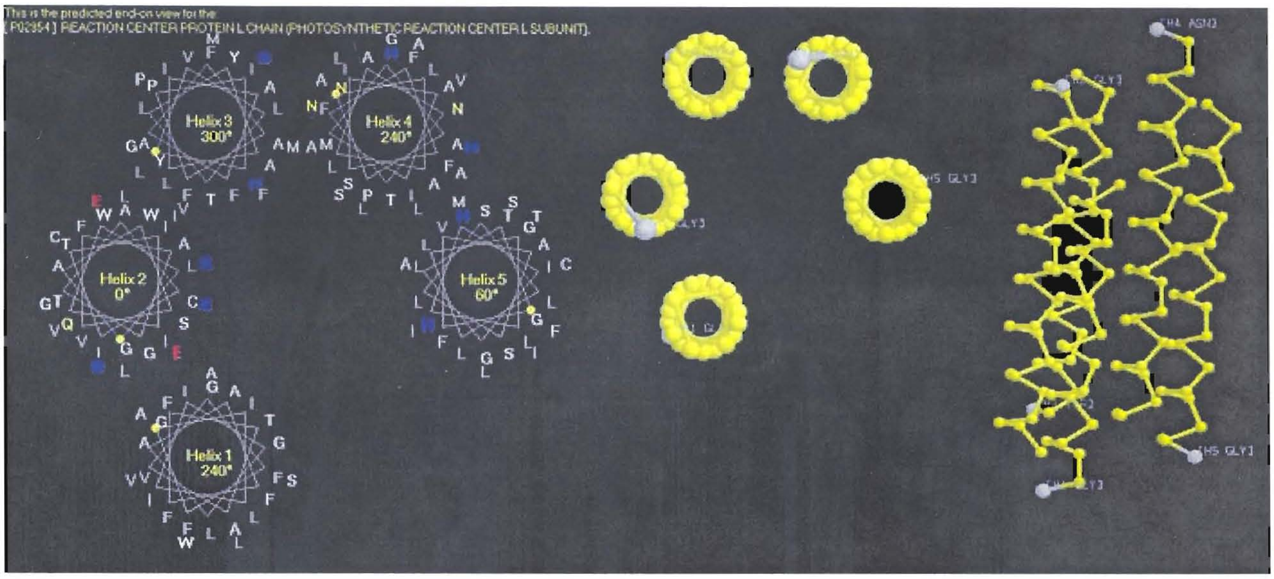
3.1 TMRelate algorithm development

The inputs for the *TMRelate* software are the created 20x20 association matrix and the membrane protein sequence in the SwissProt format.

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

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

TMRelate creates a helix wheel representation using the chosen configuration. For this step the algorithm simulates the rotation of each of the TM regions by 60° at a time, and for each rotation a score is calculated. The rotation works like an odometer, in which each TM region performs a complete rotation. After this rotation, the next TM region is rotated in turn by 60° until all the TM regions have completed one whole rotation. For each rotational position, the aggregate association scores for all the TM Regions are calculated. Again the score calculation is based on the 20x20 association matrix. In the calculation of the score for each pair of TM regions, 2 parameters are considered for inclusion of a matrix score toward the aggregate score. Firstly the depth of the 2 residues in question needs to be less than 1.5 Å apart; and secondly the angle range between the 2 residues to be computed needs to be equal or less than 60°. If the 2 conditions above are satisfied, then a score for these two residues is added to the cumulative score. At the end of all rotations, a helix wheel representation of the arrangement with the highest score is shown (figure 2).



This is the predicted end-on view for the
[P02954] REACTION CENTER PROTEIN L CHAIN (PHOTOSYNTHETIC REACTION CENTER L SUBUNIT).

This figure shows the predicted model in the form of a helix wheel representation (A). The yellow dot of each wheel represents the first residue of each TM region. The rotational orientation is anti-clockwise, rotating 60° each time. (B) Shows the created 3D structure based on the helix wheel representation "end-on" view. (C) Shows the lateral view of the same predicted 3D structure.

For the creation of a 3D structure based on the helix wheel representation, the optimal angle obtained from the helix wheel rotation is used. Using this information an atomic co-ordinate for the alpha carbon of each residue is calculated. The algorithm builds each α-helix in the corresponding position to that illustrated in the helix wheel representation. A database of the average distances between TM regions for specific protein families is in development and will provide more accurate structural predictions than that generated with the current algorithm.

A complementary development aimed at improving the quality of resulting predictions in terms of which TM region should be buried or facing out was made. This development resulted in the *TMRelate_K* version. This version differs in terms of the indices used to determine the predicted packing of TM regions and the angular orientations of TM regions with respect of the other TM regions. To this end, a knowledge-based scale called kPROT (Pilpel, *et al.*, 1999) was used (see materials and methods). This scale gives the propensities for residue orientation in transmembrane segments. It was derived from more than 5,000 non-redundant Swiss-Prot membrane protein sequences. The kPROT value for each residue is defined as the logarithm of the ratio of its proportions in single and multiple TM spans.

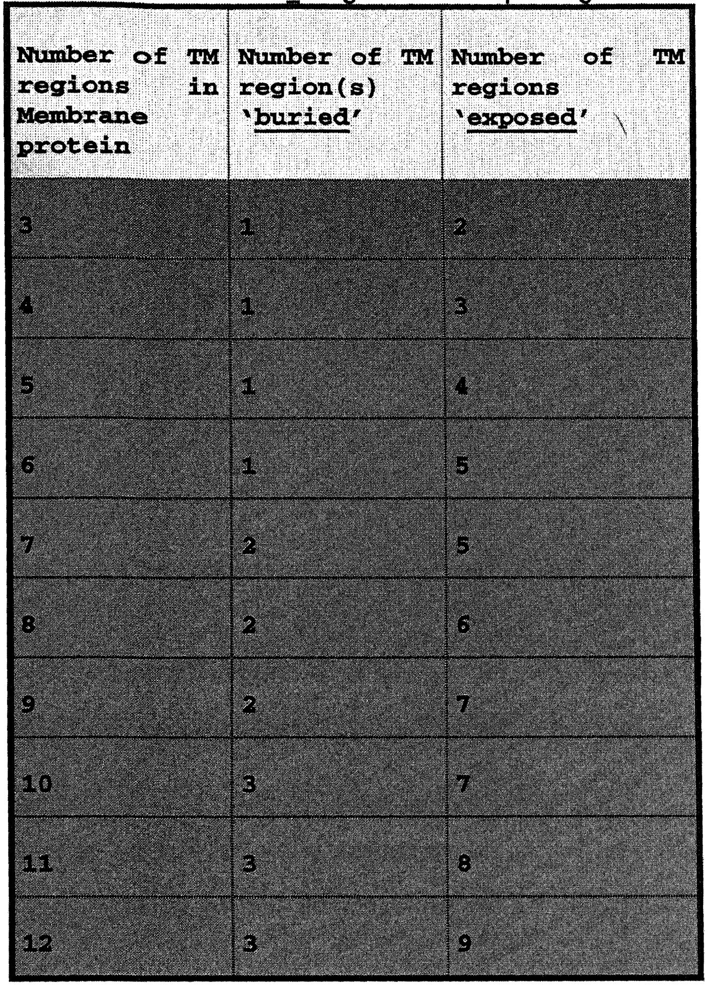
Using this scale, *TMRelate_K* calculates and predicts which TM regions are buried and which ones are facing out toward the lipid bilayer. Adding the amino acid score for each residue (Table 2) that composes the TM region, a final value is calculated for the TM region, and an overall score for each possible configuration of the whole protein is calculated. With the kPROT scale, the lower the score, the more buried the TM region is.

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

TMRelate_K algorithm

To define the helix packing for the predicted membrane protein using kPROT scale, the algorithm identifies how many TM regions are buried (TM region that is in the interior of the membrane protein) and exposed (TM region that is exposed to the lipid) depending on the number of TM regions in the protein. For example, for the *Bacteriorhodopsin precursor,* protein with seven TM regions, the algorithm considers two TM regions buried and five exposed. Table 3 shows the numbers (buried and exposed) used by the algorithm.

Table 3 – *TMRelate_K* algorithm: helix packing definition

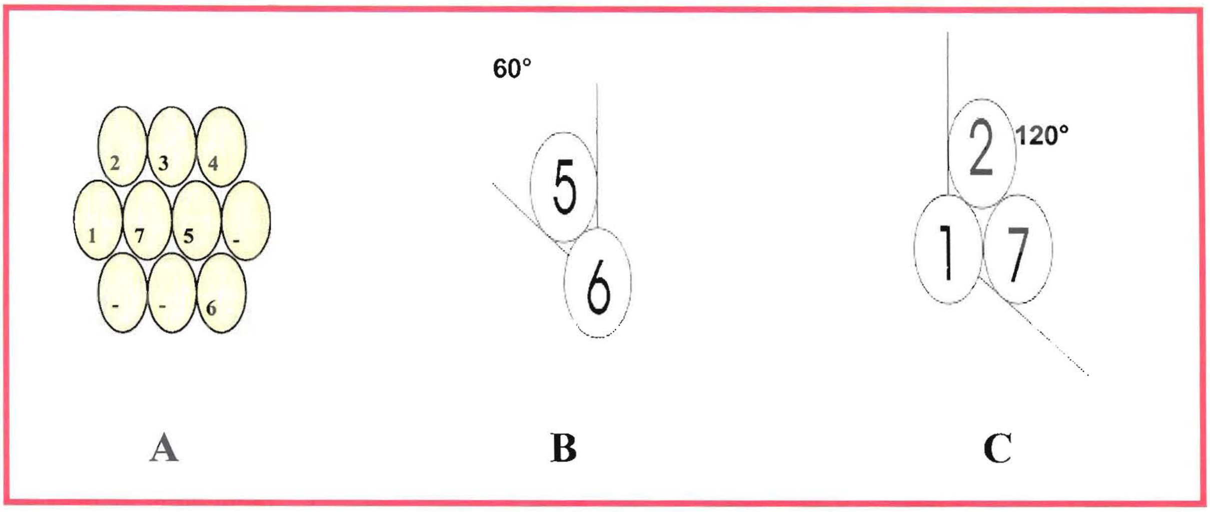


Number of TM regions in Membrane protein	Number of TM region(s) 'buried'	Number of TM regions 'exposed'
3	1	2
4	1	3
5	1	4
6	1	5
7	2	5
8	2	6
9	2	7
10	3	7
11	3	8
12	3	9

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

To predict the helix packing the algorithm calculates a score using the kPROT scale and gives a weighting based on the number of associations for each TM region. Each association between TM regions contributes 60° to the extent of "buriedness". Looking at figure 3, the TM region 6 has one association with TM 5, and the algorithm considers it as 60° buried. The TM region 1 has 2 associations, and 120° buried and so on.

Figure 3 – buried angle



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

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

Table 4 - The buried angle

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

The buried angle depending on the number

The angles shown in the table 4 would be used for the score calculation. The rationale is to use a buried angle range depending on the number of possible associations each TM region can have. The buried angle provides a higher weighting for TM regions that have more associations, leading to a higher weighted contribution from the kPROT aggregate scores. The buried angle is used as in the formula below.

```
kPROTHelixScore := (Buried angle/360) * kPROT Score For This TM region
```

6 Results

The "end on" model evaluation

An evaluation of the accuracy of the developed piece of software has been made using a set of 12 (twelve) different membrane proteins with a differing number of TM regions, as assigned in their Swiss-Prot files. To obtain the percentage of correctly predicted associations, the corresponding known high-resolution 3D structures were used (corresponding PDB files with the best resolution) for comparison. For this process, an additional program called *"TMEvaluation"* was created, while the modules *TMCompare* (Togawa *et al.,* 2001) and *TMDistance (*unpublished*)* were also used in order to prepare the data set.

TMCompare was used to select the PDB files with corresponding Swiss-Prot files. It is important to note that for the purpose of evaluating the data set it was not only necessary to match the files (PDB and Swiss-Prot), but also to find the correct arrangement associations between TM regions. In addition, *TMCompare* was used to visualize and analyse the TM regions in the real structure, which was essential in the analysis of structures like Cytochome C oxidase (PDB code 1AR1) with 12 TM regions. Analysis of such structures using a molecular rendering program like Rasmol (Sayle & Milner-White,1995) and CHIME (MDLI reference) is a difficult task due to the nature of the structure i.e., 4 different sub-units and no visual information as to where the TM region starts and where it ends. *TMCompare* facilitates this analysis by reading the TM regions from Swiss-Prot file and applying it to those amino acids in the structure, selecting and showing only the TM regions.

Using the association matrix, *TMRelate* was executed loading the membrane protein sequence in Swiss-Prot format, and the results were compared with the corresponding known structure. This process resulted in a percentage of correct associations between TM regions, calculated in an automated way by *TMEvaluation.*

TMEvaluation reads the output arrangement from *TMRelate*, counts the number of associations between each TM region and compares it with the correspondent native structure (known 3D structure), calculating the percentage of correctly predicted associations between TM regions. The results after running *TMRelate* and the *TMRelate_K* program are listed in the summary shown in table 5:

Table 5 – The results of the predicted associations obtained by *TMRelate* and *TMRelate_K*

Protein	# Of TM regions	Average percentage using *TMRelate*	Average percentage using *TMRelate_K*
Bacteriorhodopsin precursor (BR) **P02945 x 1AP9**	7	66.35%	93.94%
Rhodopsin **P02699 x 1BOK**	7	63.64%	90.91%
Photosystem Q(B) protein precursor **P02955 x 1DOP**	5	73.33%	100.00%
Rhodopsin **P02699 x 1F88**	7	63.64%	68.69%
Aquaporin **P29972 x 1FQY**	8	64.74%	67.58%
Fumarate reductase **P03805 x 1KFY**	3	100.00%	100.00%
Glycerol uptake facilitator protein (*Aquaglyceroporin*) **P11244 x 1FX8**	8	59.61%	66.66%
Adenosine A2a receptor **P29274 x 1MMH**	7	73.35%	74.75%
Fumarate reductase cytochrome B subunit **P17413 x 1QLA**	5	82.14%	95.23%
Cytochrome c *oxidase* - *Paracoccus denitrificans* **P98002 x 1AR1**	12	59.78%	60.87%
Cytochrome c *oxidase* polypeptide I - *Bos taurus* **P00396 x 1OCC**	12	58.15%	65.22%
Sensory *Rhodopsin* – Photoreceptor **P42196 x 1GU8**	7	68.62%	90.91%
Overall percentage		69.45%	81.23%

After analysing the results from *TMRelate*, an average of higher than 69% of correct predicted associations between TM regions was observed, giving very promising indications for the developed piece of software. Furthermore, the execution of the version of *TMRelate* that uses the kPROT scale (*TMRelate_K*) resulted in an even better average of 81% correctly predicted associations. The use of the kPROT scale made the software more accurate in terms of predicting the correct associations between TM regions.

It predicts the buried and exposed sides of each TM region with better accuracy, which is fundamental to the algorithm that identifies the most associated (normally the most buried) TM region, making the prediction more precise than using the association matrix.

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

Assessing the DALI (http://www.ebi.ac.uk/dali/) (Holm & Sander, 1998) evaluation for the models structures of Bacteriorhodopsin refined by the genetic algorithm (PDB-817.pdb and PDB-40.pdb), the Z score is higher for the refined models. This indicates a way that the system may be improved in the near future by this approach. The expectation is that, after aggregating the helix tilt and helix kink database into the predictive algorithm, the obtained models will be even better.

7 Discussion

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

TMRelate uses statistical information based on the associations between TM regions from known membrane protein structures (the created association matrix). Statistically, the more structures of membrane proteins that become available, the better the prediction will become, since the created matrix will be more populated giving better basis to the *ab initio* prediction.

TMRelate_K uses another knowledge-based scale, the kPROT scale (Pilpel *et al.*, 1999) that is derived from more than 5000 known membrane protein sequences deposited in the Swiss-Prot databank (Bairoch & Apweiler, 2000). The use of this scale in addition to the developed algorithm provides a useful approach for identifying TM regions, which are buried, and those that are likely to be exposed, making the prediction more accurate. This algorithm is unique in terms of combining knowledge-based approaches with statistics and mathematics for the prediction of membrane protein associations and the 3D structure α-carbon backbone.

Furthermore, the use of a permutation approach gives confidence in testing for optimum arrangements and that all TM regions have been placed in all possible positions for a chosen configuration. However, with this approach, there is a disadvantage in terms of processing time, particularly when the number of TM regions is higher than 12. This is due to the fact that the permutation is based on a factorial and one more the TM region, requires a manifold increase in the number of calculations needed.

The use of the kPROT scale combined with the algorithm using the buried angle table (table 4) is an advantage of the program, giving a very accurate result in terms of prediction of the most buried TM region. For example, visual analysis of the 3D structure with PDB code 1QLA (chain C) (Fumarate reductase cytochrome) using *TMCompare* program shows that the most buried TM region is TM 4. Running *TMRelate_K,* it is observed that TM 4 is predicted to be the most buried (*TMRelate_K* gives 95.23% of correctly predicted associations against 80.95% using *TMRelate*). The same is observed with the PDB file 1AP9 (Bacteriorhodopsin precursor (BR), where TM 3 is the most buried and predicted correctly by *TMRelate_K (TMRelate_K* gives 93.94% of correctly predicted associations against 65.66%).

The helix wheel representation created by *TMRelate* gives a graphical output, showing the rotational angle and a position for each amino acid that facilitates structural analysis.

The only similar output is found in SOSUI (Hirokawa *et. al.*, 1998) a secondary structure prediction tool available on the Internet

(http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0E.html)..

This tool gives the predicted TM region in a helix wheel representation, but without any relational associations between the predicted TM segments, showing only a picture of the helix wheels side by side.

The optimal rotational angle is another feature advanced in the developed piece of software. The method developed by Pilpel and colleagues (1999) for automatic helix orientation prediction using the kPROT scale (http://bioinfo.weizmann.ac.il/kPROT), gives a predicted angle for each TM region. In their study, it was observed that using the kPROT scale to predict the angular orientation of each TM segment is better than hydrophobic moments (Eisenberg *et al.*, 1982, 1984; Rees *et al.*, 1989) and methods based on the statistics of known high-resolution structures of integral membrane proteins to derive lipid exposure propensities of the different residues (Cronet *et al.*, 1993; Donnelly *et al.*, 1993).

However, the kPROT system does not predict the associations between TM regions; it just builds the rotational angle for each TM region considering the known configuration like *Bacteriorhodopsin* and *glycophorin* family. By contrast, *TMRelate* uses 2 stages to obtain a full structural prediction; the optimal (highest scoring) configuration based on the association score between TM regions, and the optimal rotational angle for each TM region in relation to all other TM regions, building 3D structure α-carbon backbone for each TM region.

References

Adaminan L. and Liang J. (2001). Helix-Helix packing and interfacial pairwise interactions of residues in membrane proteins. *J. Mol. Biol.*, **311**:891-907.

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**:45-48.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N. Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Research*, **28**:235-242.

Claros, M.G. and von Heijne, G. (1994) TopPred II: An Improved Software For Membrane Protein Structure Predictions. *CABIOS* **10**:685-686.

Cronet P., Sander C. and Vriend G. (1993). Modelling the transmembrane seven helix bundle. *Protein Eng.,* **6**:59-64.

Cserzö M., Wallin E., Simon I. and von Heijne G., Elofsson A. (1997). Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. *Protein Engin,* **10:**673–6.

Donnelly D., Overington J. P., Ruffle S. V., Nugent J. H. A. and Blundel T. L. (1983). Modelling α-helical transmembrane domains: the calculation and use of substitution tables for lipid facing residues. *Protein Sci.,* **2**:55-70.

Eisenberg D. Weiss R. M. and Terwilliger T. C. (1982). The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature*, **299**:371-374.

Hirokawa T.S., Boon-Chieng S.S. and Mitaku S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14:**378–9.

Hofmann K. and Stoffel W. (1993). TMBASE – a database of membrane spanning protein structure and topology. *J Magn Reson*, **144:**150–5.

Holm L. and Sander C. (1998). Touring protein fold space with Dali/FSSP. *Nucl. Acids Res.* **26**:316-319.

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

Pilpel Y., Ben-tal N. and Lancet D. (1999). KPROT: A knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane proteins structure prediction. *J. Mol. Biol.* **294:**921-935.

Ress D. C., De Antonio L. and Eisenberg D. (1989). Hydrophobic organisation of membrane proteins. *Science,* **245**:510-513.

Rost B., Fariselli P. and Casadio R. (1996). Topology prediction for helical transmembrane proteins at 86% accuracy. *Prot. Science,* **7:** 1704-1718.

Russ W. P. and Engelman D. M. (2000). The GxxxG Motif: A framework for transmembrane helix-helix association. *J. Mol. Biol.* **296**:911-919.

Sayle, R.A. and Milner-White, E.J. (1995) RasMol: Biomolecular graphics for all, *Trends in Biochemical Science (TIBS)*, (20)**9**:374.

Senes A., Gerstein M. and Engelman D.M. (2000). Statistical Analysis of Amino Acid Patterns in Transmembrane Helices: The GxxxG motif occurs frequently and associations with β-branched residues at neighbouring positions. *J Mol Biol*, **296**:921-936.

Sonnhammer E.L.L., von Heijne G. and Krogh A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. In Glasgow J., Littlejohn T., Major F., Lathrop R., Sankoff D., Sensen C., eds. Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB-98). 1998 Jun 28–Jul 1; Montreal, Canada. Menlo Park, CA: AAAI Pr. p 175–82.

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

Tusnády G.E. and Simon I. (1998). Principles governing amino acid composition of integral membrane proteins: application to topology prediction. *J Mol Biol,* **283:**489–506.

Appendix (D)

PDB files and Rasmol Software

The predicted PDB files with different free energy (-817KJ/mol, -752 KJ/mol and 890 KJ/mol) and PDB file for 1AT9 with Rasmol software in order to allow the user to display each PDB file as the 3D structure are available on CD which has been attached to this thesis. There is a ReadMe file to instruct the user to open each PDB file in Rasmol and compare each predicted structure with Native structure (1AT9).

References

Adamian, L., Jackals Jr, R., Binkowski, T.A. and Liang, J., (2003),"Higher-order interhelical spatial interactions in membrane proteins", *J.Mol.Biol*, 327, 251-272.

Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P., (2004), "Essential Cell Biology", Second edition, *Garland science*, ISBN 0-8153-3481-8.

Arkin, I.T., Brunger, A.T. and Engelman, D.M., (1997), "Are there dominant membrane protein families with a given number of helices?", *Proteins: Structure, Function and Genetics*, 28, 465-466.

Aude, J.C. and Comet, J.P., (1996), "Construction of protein sequences families", 24th Aharon Katzir-Katchalsky, conference Bioinformatic-Structure, Jerusalam, Israel, November 17-21.

Bairoch, A., and Apweiler, R., (2000), 'The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000', *Nucleic Acids Research*, 28, 45-48.

Berg, J.M., Tymoczko, J.L. and Stryer, L., (2002), "Biochemistry", Fifth edition, ISBN 0-7167-3051-0

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, J. and Bourne, P.E., (2000) "The protein data bank", *Nucleic Acids Research*, Oxford, 28, 235-242.

Bevilacqua, V., Mastronardi, G., Colaninno, A. and D'Addabbo, A., (2004), "Retina Images processing using genetic algorithm and maximum likelihood method", *Advances in computer science and technology*, 431.

Bonneau R., Baker D., (2001), "Ab initio protein structure prediction: progress and prospects", *Annu Rev Biophys Biomol Struct*, 30,173-189.

Boyd, J.C. and Savory, J., (2001), "Genetic algorithm for scheduling of laboratory personnel", *clinical chemistry*,47, 118-123.

Braden, K., (2002), "A simple approach to protein structure prediction using genetic algorithms", Stanford University, Unpublished Report. www.genetic-programming.org/sp2002/braden.pdf (Accessed May 2006)

Brande, C. and Tooze, J., (1999), "Introduction to protein structure", Garland, USA, ISBN 0-8153-2305-0.

Brunett, T.J. and Brock, O., (2005), "Improving protein structure prediction with model based search", Oxford Journal, Life Scienece, Bioinformatics, 21, 66-74

Bui, T.N. and Sundarraj, G., (2005), " An efficient genetic algorithm for prediction protein structure in the 2D HP model", Proceedings of the 2005 conference on genetic and evolutionary computation, Wshington, DC,USA, 383-392, ISBN:1-595993-010-8.

Butter, T., Rothlauf, F., Grahl, J., Hildenbrand, T. and Arndt, J., (2006), "Developing GA and mixed integer linear programs for finding optimal strategies for a student's "sports" activity", Poster, In GECCO: Proceeding of the 2006 conference on genetic and evolution computation. Wifol.bwl.uni-mannheim.de/fileadmin/files/publications/workingpapers.pdf (Accessed May 2006)

Carugo, O. and Pongor, S., (2001), "A normalized root-mean-square distance for comparing protein three dimension structure", *protein Science*, 10:1470-1473.

Chen, Z. and Xu, Y., (2006), "Structure prediction of helical transmembrane proteins at two length scales", *Journal of Bioinformatics and Computational Biology(by invitation)*

Calabretta, R., Nolfi, S. and Parisi, D., (1995), "An artificial life model for predicting the tertiary structure of unknown proteins that emulate the folding process", *Advances in Artificial life*, 862-875.

Cui, Y., Chen, R.S. and Wong, W.H.,(1998),"Protein folding simulation with genetic algorithm and super secondary structure constraints", *Protein: Structure, Function and Genetics*,31:247-257.

Davis, L.J., (1991), "Handbook of genetic algorithm", Van Nostrand Reinhold. New York, ISBN 0442001738.

De brevern, A.G., Wang, H., Tournamille, C. and Cartran, J.P., (2005), "Structural model of seven transmembrane helices, Duffy Antigan/Receptor for chemokines (DARC*)", Biochem Biophys Acta*, 1724, 288-306.

Domingues, F.S., Lackner, P. and Andreeva, A., (2000), "Structure-based evaluation of sequence comparison and fold recognition alignment accuracy", *J.Mol.Biol.*, 297, 1003-1013.

Durant, E. A., (2002), " Hearing aid fitting with genetic algorithms", University of Michigan, Unpublished thesis

Edman, M.,(2001), "Detection of sequence pattern in membrane proteins: Astory of many dimentions", Umea university,Sweden, Unpublished PhD thesis.

Eilers, M., Patel, A. B., Liu, W. and Smith, S. O., (2002), "Comparison of helix interactions in membrane and soluble α-bundle proteins", *Biophysical Journal*, 82, 2720-2736

Elliott, W.H., and Elliott, D.C., (2001), "Biochemistry and molecular biology", Imprint: Oxford: Oxford university press, second edition, ISBN: 019870048.

Engelhardt, M.O., Savic, D.A. and Walters, G.A., (2000),"Using genetic algorithms in the UK water industry", Joint conference on water resource engineering and water resources planning and management, p 204, July30-august2, Minneapolis, Minnesota, USA

Faulon, J.L., Sale, K. and Young, M., (2003), "Exploring the conformational space of membrane protein folds matching distance constrant", Sandia National Laboratories, Livermore, California 94551, USA.

Feldman, H. J.,(2003), "Computational protein structure prediction", University of Toronto, Unpublished thesis.

Fleishman, S.J. and Ben-Tal, N., (2006), "Progress in structure prediction of α-helical membrane proteins", Elsevier Ltd, Current opinion in structural biology, 16, 496-504

Fogel, L.J., Owens,A.J., Walsh, M.J.,(1966)," Artificial intelligence through simulated evolution", Wiley, New York.

Gao, P.F. and Cross, T.A., (2006), "Recent developments in membrane protein structural genomics", Genome Biology, 6, 244.

Glover, F., (1994), "Tabu search for nonlinear and parametric optimisation", *Discrete Applied Mathematics*, 49, 231-255.

Gnanadass, R., Venkatesh, P., Palanivelue, T.G. and Manivannan, K., (2004), "Evolutionary programming solution of economic load dispatch with cambined cyclo co-generation effect", *IEI Journals*, 85, 124-128.

Goldberg, D.E., (1989), "Genetic Algorithms in Search, Optimisation & Machine Learning", Addison-Wesley, New York.

Grisshammer, R. and Buchanan, S.K., (2006), "Structural biology of membrane proteins", *Springer Verlag*, ISBN 0854043616.

Guex, N. and Peitsch, M.C., (1997), "Swiss-model and the swiss-PDBViewer: An environment for comparative protein modelling" *electrophoresis*, 18, 2714-2723. www.expasy.org/spdbv (Accessed May 2006)

Gunsteren van, W.F., Billeter, S.R., Eising, A.A., Hunenberger, P.H., Kruger, P., Mark, A.E., Scott, W.R.P. and Tirani, I.G., (1996), "Biomolecular simulation: the GROMOS96 manual and user guide", ISBN 3-7281-2422-2.

Hardin, C., Pogorelov, T.V. and Luthey-Schulten, Z., (2002), "Ab initio protein structure prediction", *Elsevier Science*, 12, 176-181.

Haupt, S.E. and Haupt, R.L., (2003), "Genetic algorithms and their applications in environmental science", 3rd conference in artificial intelligence applications to the environment science.

Hess, B., (2002), "Stochastic concepts in molecular simulation", University press, Veenendaal, ISBN 9090154574

Hinds, D.A. and Levitt, M., (1992), "A lattice model for protein structure prediction at low resolution", Proc Natl Acad Sci, Usa, 89, 2536-2540.

Holland, J.H.,(1975), '*Adaptation in Natural and Artificial Systems*', The University of Michigan Press, Ann Arbor.

Holm, L., and Sander, C.,(1991), "Database algorithm for generating protein backbone and side-chain co-ordinates from a Cα trace application to model building and detection of co-ordinate errors", *J. Mol. Biol.*,218, 183-194.

Hofacker, I. and Schulten, K., (1998), "Oxygen and proton pathways in cytochrome c oxidase", University of Illinois at Urbana-Champaign, *Protein: structure, function and genetics*, Wiley-Liss, 30, 100-107.

Jones,B.F., Eyres, D.E., and Sthamer, H.H., (1998), 'A Strategy for using Genetic Algorithms to Automate Branch and Fault-based Testing', *The computer Journal*, 41(2), pp98-106.

Kendrew, J., (1994), "The encyclopedia of molecular biology", *Blackwell Science*, Cambridge mass.

Khimasia, M. and Coveney, P., (1997),"Protein structure prediction as a hard optimization problem", The genetic algorithm approach. In Molecular Simulation, 19, 205-226.

Kimball, J.W., (1994), "Biology", *Addison-Wesley*, ISBN:0201102463.

Kimura, Y., Vassylyev, D.G., Miyazawa, A., Kidera, A., Matsushima, M., Mitsuoka, K., Murata, K., Hirai, T. and Fujiyoshi, Y., (1997), "Structure of Bacteriorhodopsin at 3.0 Angstrom determined by electron crystalloghraphy", *Nature*, 389, 206-211. www.rsb.org/pdb/pubmed.do?structureId=1AT9 (Accessed May 2006)

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M. P., (1983), "Opti,isation by simulated annealing", *Science*, 220, 671-680

Kleinsmith, L.J. and Kish, V.M., (1995), "Principles of cell biology", Harper-Collins College, New York, 155-193.

Knight, M.I. and Nason, G.P., (2005), "Improving prediction of hydrophobic segments along a transmembrane protein sequence using adaptive multiscale lifting", University of Bristol

Kokubo, H. and Okamoto, Y., (2004), "Prediction of membrane protein structures by replica-exchange monte carlo simulations: Case of two helices", *J Chem Phys* **120**:10837-10847.

Krasnogor, N., Hart, W.E., Smith, J., and Pelta, D., (1999), "Protein structure with evolutionary algorithms", Proceedings of the 1999 international genetic and evolutionary computation, 1596-1601.

Krogh, A., Larsson, B., Van Heijne, G. and Sonnhameer, E.L., "Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes", *J.Mol.Biol.*, 305, 567-580.

Lesk, M., (2004), "Introduction to protein science: architecture, function and genoms", Oxford U.P., ISBN 978-0-19-926511-4.

Lodish, H., Berk, A., Zipursky, L.S., Matsudaira, P., Baltimore, D. and Darnell, J., (2000), "Molecular cell biology", *W.H. Freeman*, NewYork.

Marczyk, A., (2004), "genetic algorithm and evolutionary computation", The TalkOrigins Archive, Exploring the creation/evolution controversy, www.talkorigins.org/fags/genalg/genalg.html.

Mardle, S., and Pascoe, S., (1999), "An overview of genetic algorithms for the solution of optimisation problems", University of Portsmouth, Cheer, V13, n1, 1-8.

Matthews, H.R..Freedland, R. and Miesfeld, R. L.,(1997), "Biochemistry a short course", ISBN 0-471-02205-5

Muller, D.J., Fotiadis, D., Scheuring, S., Muller, S. and Engel, A., (1999), "Electrostatically balanced sub nanometre imaging of biological specimes by atomic force microscope", *Biophysical Journal*, 76, 1101-1111.

Opella, S.J., (1997), "NMR and membrane proteins", Nature Structure Biol, 4, 845-848.

Orlandini, E., Seno, F., Banavar, J.R., Laio, A. and Mritan, A., (2000), "Deciphering the folding kinetics of transmembrane helical proteins", *PNAS*, 97, 14229-14234.

Park, BH. And Levitt, H., (1995), "The complexity and accuracy of discrete state models of protein structure", *J Mol Biol*, 249, 493-507.

Patel, A., Davis, D, Guthrie, C. and Tuk, D., Nguyen, T. and Williams, J., (2005), "Optmisation cyclic-steam oil production with genetic algorithms", Chevron Texaco North America Upstream, Prepared for the 2005 SPE Western Regional meeting, 30 march-1 April

Patton, A.L., punch III, W.F. and Goodman, E.D.,(1995)," A standard GA approach to native protein conformation prediction", *In Eshelman*, .Proceeding of the sixth international conference on genetic algorithm, 574, 581.

Popot, J., and Engelman, D.M.,(2000), "Helical membrane protein folding, stability, and evolution", *Ann.Rev.Biochem.*,69,881-922.

Rajapakse, M., Wyse, L. Schmidt, B and Brusic, V., (2005), “Deriving Matrix of Peptide-MHC Interactions in Diabetic Mouse by Genetic Algorithm”, Intelligent Data Engineering and Automated Learning 6th International Conference, Brisbane, Australia, Lecture Notes in Computer Science, 3578, 440, ISBN: 3-540-26972-X.

Rechenberg, I, (1973), “Evolutionsstrategie: omtimierung technischer system nach prinzipien der biologischen evolution”, Frommann Hotzbog verlag, Stuttgart.

Reeves C.R., (1995), "Modern heuristic techniques for combinatorial problem", Mcgraw-Hill.

Rost, B., (1997), "Learning from evolution to predict protein structure", *Bio-computing and Emergent Computation*: 87-101.

Schwaiger, M., Lebendiker, M., Yerushalmi, H., Coles, M., Gröger, A., Schwarz, C., Schuldiner, S., and Kessler, H., (1998), “NMR investigation of the multidrug transporter EmrE, an integral membrane protein”, Eur.J.Biochem., 254, 610-619.

Schwefel, H.P., (1981), “Numerical optimisation of computer models”, Wiley, Chichester.

Segonne, F., Grimson, E. and Fischl, B., (2005), “A genetic algorithm for topology correction of cortical surfaces”, Springer-Verlag, Berlin, 3565, 393-405.

Smith, M., (1989), “Evolutionary Genetic”, Oxford university, Oxford, ISBN0198502311.

Srinivas, M., and Patnaik, L.M., (1994), 'Genetic Algorithms: A Survay', *IEEE*, 0018-9162, pp17-26.

Szustakowski, J.D. and Weng, Z.,(2000),"Protein structure alignment using a genetic algorithm", *Protein: Structure, Function and Genetics*, 38:428-440.

Tamm, L.K., Abildgaard, F, Arora, A., Blad, H. and Bushweller, H., (2003), "Structure, dynamics and function of the outer membrane protein A(OmpA) and influenza hemagglutinin fusion domain in detergent micelles by solution NMR", FEBS Letters, 555, 139-143.

Tia-Li, Y., Yassin, A. and Goldberg, D.E., (2005), "An information theoretic method for developing modular architechtures using genetic algorithm", University of illionios at urbana-champaign, Unpublished report 2005014.

Togawa R.C., Antoniw J.F., Mullins J.G.L., (2006), "Prediction of Transmembrane Region Adjacencies in Membrane Proteins", Intelligent Systems for Molecular Biology, LB-36

Tusnady. G.E. and Simon, I., (1998), "Principles governing amino acid composition of integral membrane proteins: Application to topology prediction", *J,Mol.Biol.*, 283, 489-506.

Unger, R. and Moult, J.,(1993)," A genetic algorithm for 3D protein folding simulations", Proceeding of the fifth annual international conference on genetic algorithms, PP:581-588.

Unger, R., (2004), "The Genetic Algorithm Approach to Protein Structure Prediction", DOI 10.1007/b13936, *Structure and Bonding*, 110, 153–175.

Vail, D., (2001), "Genetic algorithm as a search strategy and a novel means of potential function discovery in the protein folding problem", Unpublished report for the department of computer science, Bowdoin College.

Vorac, J., Vondrak, I. and Vicek, K., (2002), "School timetable generating using genetic algorithm", VSB-Technical University of Ostrava, Czech republic. www-e2.ijs.si/phdworkshop/2002/papers/vorac.pdf (Accessed May 2006)

White, S.H., and Wimley, W.C.,(1999), "Membrane protein folding and stability: physical principles",*Ann.Rev.Biomol.Struct.*,28,319-65.

Whitley, D., (1993), "A genetic algorithm tutorial", *Statistic and computation*, 4, 65-85.

Won, K. J., Hamelryck, T., Prügel-Bennett, A. and Krogh, A., (2005), "Evolving Hidden Markov models for protein secondary structure prediction, In proceedings of IEEE congress on evolutionary computation, pp 33-40, Edinbrgh.

Wood, T.C. and Pearson, W.R., (1999), "Evolution of protein sequences and structures", *J.Mol.Biol.*, 291, 977-995.

Zhang, Z., (2003), "An overview of protein structure prediction: from homology to Ab Initio", Unpublished Report.

Zhou, Y., and Abagyan, R. (1999). Efficient stochastic global optimization for protein structure prediction. Rigidity Theory and Application (M.F. Thorpe & P.M. Duxbury eds.), 345-35

Zoonens, M., Catoire, L.J., Giusti, F. and Popot, J.L., (2005), "NMR study of a membrane protein in detergent-free aqueous solution", Yale university, New Haven, PNAS, 102, 8893-8898.