APPLICATION OF METAHEURISTIC APPROACH TO MODEL AN ASSEMBLY OF TRANSMEMBRANE HELICAL BUNDLE IN INTEGRAL MEMBRANE PROTEINS

Mr. Kanon Sujaree

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Nanoscience and Technology
(Interdisciplinary Program)
Graduate School
Chulalongkorn University
Academic Year 2014

การประยุกต์วิธีเมต้าฮิวริติกเพื่อจำลองการรวมกลุ่มของทรานเมมเบรนเฮลิกซ์ในอินทิกรัลเมมเบรนโปรตีน

นายคณน สุจารี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาวิทยาศาสตร์นาโนและเทคโนโลยี (สหสาขาวิชา)
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2557

Thesis Title	APPLICATION OF METAHEURISTIC APPROACH TO MODEL AN ASSEMBLY OF TRANSMEMBRANE HELICAL BUNDLE IN INTEGRAL MEMBRANE PROTEINS
By	Mr. Kanon Sujaree
Field of Study	Nanoscience and Technology
Thesis Advisor	Associate Professor Pornthep Sompornpisut, Ph.D.

Accepted by the Graduate School, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctoral Degree

....Dean of the Graduate School

(Associate Professor Sunait Chutintaranond, Ph.D.)

THESIS COMMITTEE

....Chairman

(Associate Professor Vudhichai Parasuk, Ph.D.)

....Thesis Advisor

(Associate Professor Pornthep Sompornpisut, Ph.D.)

....Examiner

(Assistant Professor Somsak Pianwanit, Ph.D.)

....Examiner

(Assistant Professor Nutthita Chuankrerkkul, Ph.D.)

....External Examiner

(Kwanniti Khammuang, Ph.D.)

คณน สุจารี : การประยุกต์วิธีเมต้าฮิวริติกเพื่อจำลองการรวมกลุ่มของทรานเมมเบรนเฮลิกซ์ในอินทิกรัลเมมเบรนโปรตีน (APPLICATION OF METAHEURISTIC APPROACH TO MODEL AN ASSEMBLY OF TRANSMEMBRANE HELICAL BUNDLE IN INTEGRAL MEMBRANE PROTEINS) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: รศ. ดร.พรเทพ สมพรพิสุทธิ์, 57 หน้า.

แม้ว่าจะมีการเติบโตอย่างรวดเร็วของจำนวนโครงสร้างสามมิติของโปรตีนที่วัดได้ แต่เมื่อเปรียบเทียบจำนวนทรานสเมมเบรนโปรตีนกับจำนวนโครงสร้างโปรตีนทั้งหมดที่มีอยู่ในฐานข้อมูล Protein data bank พบว่าอัตราร้อยละอยู่ในระดับต่ำ เนื่องจากความยากเชิงเทคนิคสำหรับการวัดโครงสร้างที่มีความละเอียดสูง การศึกษานี้จึงนำเสนอ อัลกอริทึ่มใหม่ชื่อ แม็ก มิน แอนท์ ซิสเตม (Max-Min ant system) ที่ถูกออกแบบเพื่อหาการรวมตัวของทรานสเมมเบรนโปรตีนชนิดเกลียวอัลฟ่า โดยใช้การจัดวางของเฮลิกซ์ชนิดแข็งเกร็งที่บังคับโดยเงื่อนไขระยะทาง วิธีการที่นำเสนอเรียกว่า ทาร์มมัส (THAMMAS : Transmembrane Helix Assembly by Max-Min Ant System) ผลิตการวางทิศทางของกลุ่มทรานสเมมเบรนเฮลิกซ์ที่หลากหลาย และหาคำตอบที่เหมาะสมกับเงื่อนไขของระยะทาง ตามพฤติกรรมการหาอาหารของมดในการค้นหาเส้นทางที่สั้นที่สุดระหว่างรังกับแหล่งอาหาร เพื่อแสดงให้เห็นถึงประสิทธิภาพของอัลกอริทึ่มชนิดใหม่นี้ THAMMAS ถูกนำมาวัดการแพคของทรานสเมมเบรนของไอออนแชนนัล KcsA และ MscL จากข้อมูลระยะทางที่ได้มาจากโครงสร้างผลึกและการแพคของโดเมนรับรู้ศักย์ไฟฟ้า KvAP โดยใช้ชุดเงื่อนไขระยะทางที่วัดจากการทดลอง เปรียบเทียบผลการทดลองกับ คอนเวนชั่นนอล ออฟติไมเซชั่น อัลกอริทึ่ม ซึ่งได้แก่ ซิมูเลคเตด อัลเนลลิ่ง มอนติ คาร์โล และจีเนติกส์ อัลกอริทึ่ม

สาขาวิชา วิทยาศาสตร์นาโนและเทคโนโลยี ลายมือชื่อนิสิต ....

ปีการศึกษา 2557 ลายมือชื่อ อ.ที่ปรึกษาหลัก ....

5287757420 : MAJOR NANOSCIENCE AND TECHNOLOGY

KEYWORDS: MAX-MIN ANT SYSTEM / GENETIC ALGORITHM / SIMULATED ANNEALING / MONTE CARLO

KANON SUJAREE: APPLICATION OF METAHEURISTIC APPROACH TO MODEL AN ASSEMBLY OF TRANSMEMBRANE HELICAL BUNDLE IN INTEGRAL MEMBRANE PROTEINS. ADVISOR: ASSOC. PROF. PORNTHEP SOMPORNPISUT, Ph.D., 57 pp.

Despite the rapid growth of solved 3D structures proteins, a relatively low percentage of all structures of the Protein Data Bank is transmembrane proteins due to technical difficulties for high-resolution structure determination. This study proposes a novel algorithm. Max-Min Ant System, designed to find an assembly of α-helical transmembrane proteins using a rigid helix arrangement guided by distance constraints. The method called THAMMAS (Transmembrane Helix Assembly by Max-Min Ant System) generates a variety of orientations of transmembrane helix bundle and finds the solution that is matched with the provided distance constraints based on the behavior of ants to search for the shortest possible path between their nest and the food source. To demonstrate the efficiency of the novel algorithm, THAMMAS are applied to determine the transmembrane packing of KcsA and MscL ion channels from distance information extracted from the crystal structures, and the packing of KvAP voltage sensor domain using a set of experimentally-determined constraints and the results are compared with those of conventional optimization algorithms, Simulated Annealing Monte Carlo method and Genetic Algorithm.

Field of Study: Nanoscience and Technology

Student's Signature ....

Advisor's Signature ....

Academic Year: 2014

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Associate Professor Pornthep Sompornpisut, Ph.D. for his kind advisor, his consultant and helpful in my work and dissertation.

I would like to thank the thesis committee: Associate Professor Vudhichai Parasuk, Ph.D Assistant Professor Somsak Pianwanit, Ph.D. Assistant Professor Nutthita Chuankrerkkul, Ph.D and Kwanniti Khammuang, Ph.D. for their comments to my dissertation.

I would like to acknowledge the 90th Anniversary of Chulalongkorn University Fund (Ratchadaphiseksomphoj Endowment Fund), the Center for Innovative Nanoscience and Nanotechnology (CIN grant 2555), NSTDA through Nanotec grant: p-11-00990 and nanoscience and technology program for enhancing and providing financial support.

I would like to thank my friends Mr. Chirayut Supunyabut, Mr. Panisak Boonamnaj and Miss Sunan Kitjaruwankul for their suggestion in chemistry and comments in thesis.

Finally I would like to thanks my mother Mrs Patthanan Sujaree for great encouragement and her love to me all the time. This is an important force which supports me to success in this work.

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

GA	Genetic Algorithm
ACO	Ant Colony Optimization
MMAS	Max-Min Ant System
SAMC	Simulated Annealing Monte Carlo
THAMMAS	Transmembrane for Max-Min ant system
TM	Transmembrane
PDB	Protein data bank
RMSD	Root Mean Square Deviation
KcsA	potassium Channel from Streptomyces lividans
KvAP	voltage-gated potassium channel from Aeropyrum pernix
MscL	mechanosensitive channel of large conductance
DE	Differential equation
SDSL	Site-directed spin labeling
EPR	Electron paramagnetic resonance
TMH	Transmembrane helix
TMP	Transmembrane protein
Cα	Alpha carbon
DOE	Design of experiment
VSD	Voltage sensor domain

CHAPTER I

INTRODUCTION

1.1 Integral membrane proteins

The cell membrane is a selectively permeable membrane that encircles the cytoplasm of organisms protecting the intracellular organelle from its extracellular surroundings. The cell membrane's function is cell to cell communication by controlling the specific molecules that go in and out between the cells. Integral membrane proteins containing membrane-spanning domain are located within the cell membrane mediating communication between the external and the interior of the cell. Their structure shares a common property, by which parts of the protein interacts with a hydrophobic membrane environment whereas other parts are hydrated. Generally, integral membrane proteins can be divided into two distinct structural classes. Those classes are α-helical and β-barrel transmembrane (TM) segments. Transmembrane proteins play a variety of biologically significant roles such as neurotransmitter, transporter, receptor, signaling transduction and catalytic activity etc. Approximately around 20-30% of the proteins encoded by human genome are membrane proteins[1]. Knowledge of their three-dimensional (3D) structure could gain new understanding of diseases and illnesses, and lead to improved human health and disease treatment. Therefore, they are valuable for the pharmaceutical industry. Transmembrane proteins, for instance, the G-protein-coupled receptor covers ~50% of all drug targets [2-4]. Despite their biological and pharmacological importance, relatively few high-resolution structures are solved, corresponding to less than 1% of the known 3D protein structures available in protein data banks[4]. Such very little information is largely due to problems of protein expression, crystallization and stability under studied conditions, which hampers high-resolution structure determination with x-ray or nuclear magnetic resonance (NMR). In spite of the fact that crystal structures are obtained typically in detergent micelles which may alter their conformation to a non-native state[5-7]. Therefore, there is a need for computational tools as an alternative approach for structure prediction of membrane proteins.

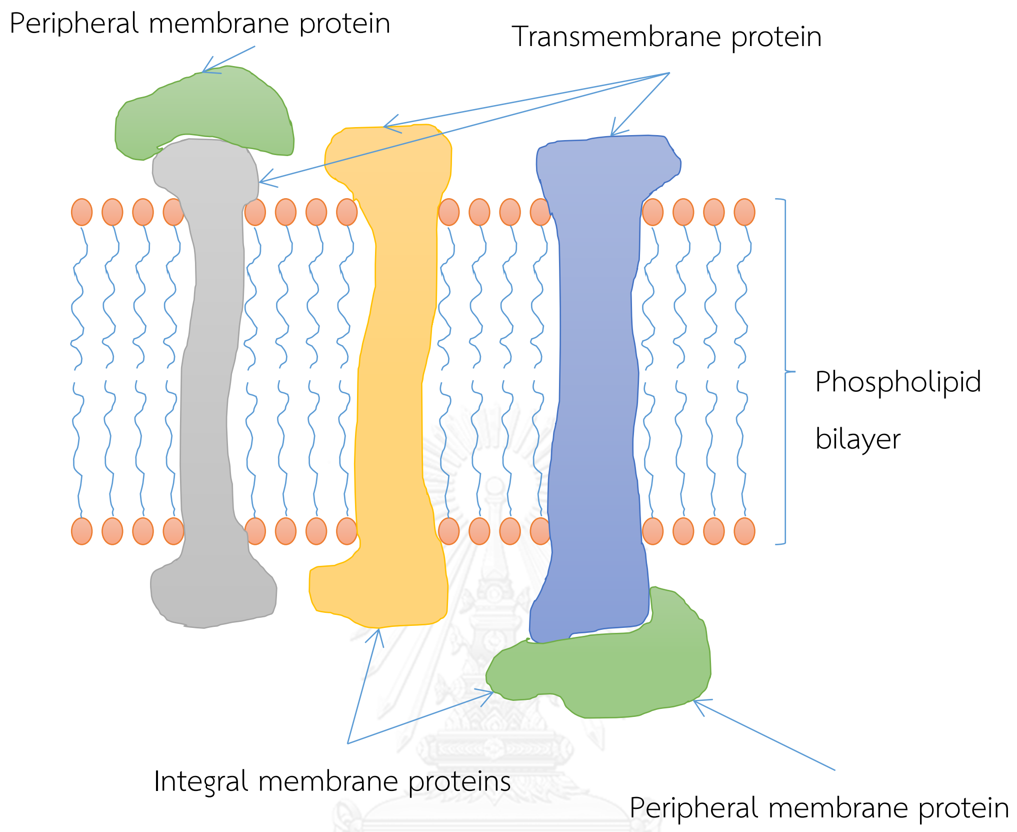
Peripheral membrane protein
Transmembrane protein
Phospholipid bilayer
Integral membrane proteins
Peripheral membrane protein

Figure 1. 1 Integral membrane protein is includes transmembrane protein and peripheral membrane protein. They are embedded in phospholipid bilayer

1.2 Optimization Algorithm

The optimization problem can be classified into two categories that are conventional optimization and approximation optimization (Figure 1.2) Conventional optimization is based on mathematics and finding the best solution. However it takes a long time when solving a large scale problem such as linear programming, goal programming dynamic programming etc. Approximation optimization is based on intelligent random searches applying large scale and complex problems. Both types are constructive approaches that use specific rules in approximating and stochastic search approaches using bio-inspire theory for each method. They reduce calculating time and may not find the best solution but close to the best solution.

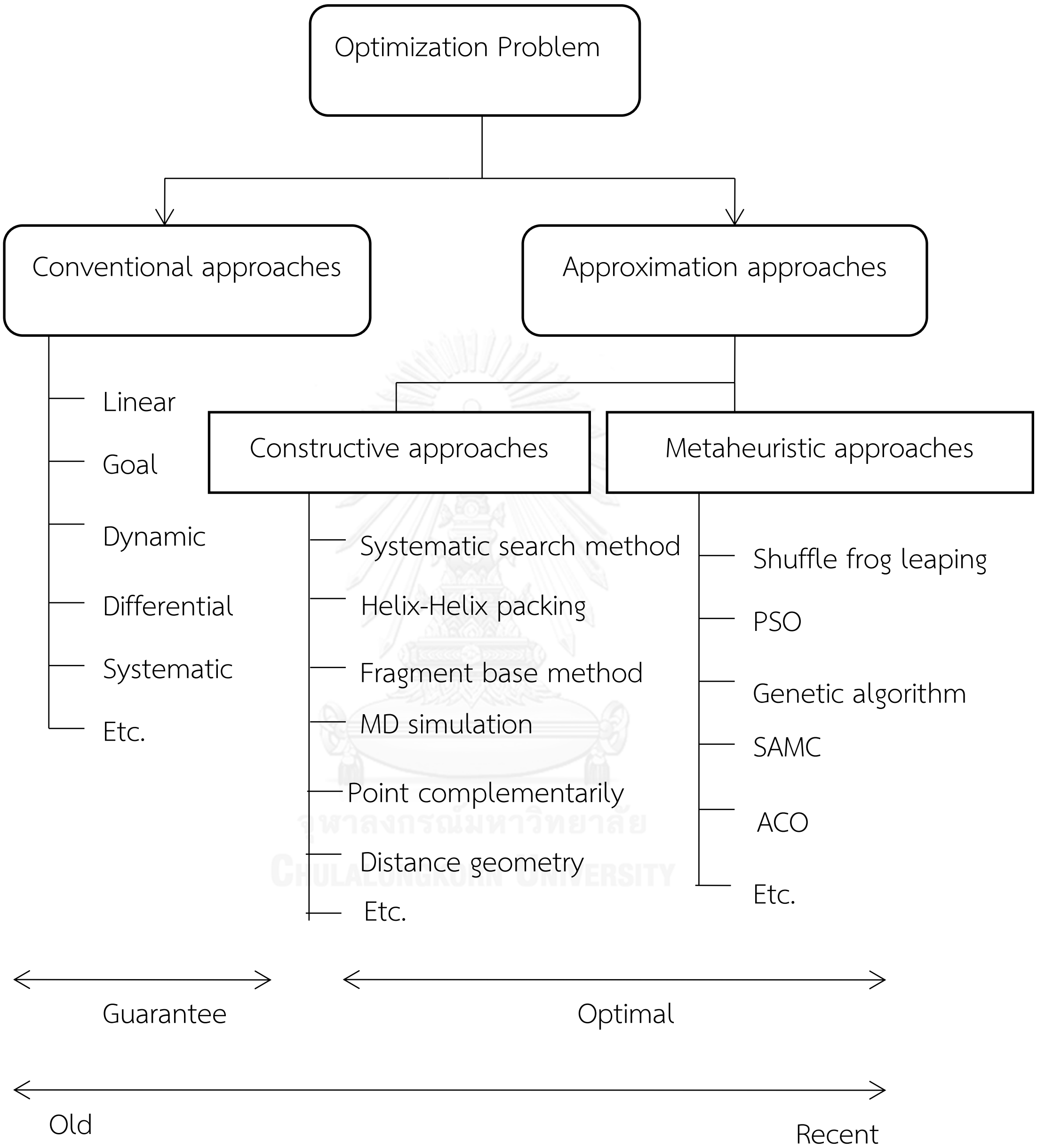

Figure 1. 2 Classification of Optimization Problem

1.2.1 Metaheuristics approach

In the last decade, a variety of recently proposed meta-heuristic search methods have attracted considerable attention for solving large-scale optimization problems. These modern global optimization techniques such as genetic algorithm simulated annealing (SA), evolutionary programming (EP), evolutionary strategy (ES), tabu search (TS), ant colony optimization (ACO), and particle swarm optimization (PSO) offer alternatives to address the difficulty in finding the global optimum solution[8-13]. Many of these algorithms are able to handle optimization problems with a good approximation to yield the global optimum in a large search space. Owing to the vast number of degree of freedom, membrane protein packing problem is considered to be one of the most significant and challenging problems for stochastic global optimization methods. Interestingly, these methods have the ability to identify a unique global minimum in a much smaller number of iterations compared to the computationally expensive Monte Carlo (MC) and molecular dynamics methods that are emphasized on the production of Boltzmann ensemble. Therefore, it is essential to develop a new efficient method combining experimentally-derived information to provide good initial guesses.

1.2.1.1 Max-Min ant system

Max-min ant system (MMAS) is an ACO algorithm, a nature-inspired meta-heuristic optimization approach[14]. The ACO algorithm imitates the behavior of ants to find the shortest routes from their colony to the food source by following and depositing a chemical substance, called pheromone on the paths that they travel[15-18]. MMAS is an improved ACO algorithm which is relied on the use of stochastic search methods for finding optimal solution based on the objective function. A notable advantage of MMAS over the standard ACO is that the pheromone intensity on all paths can be controlled within upper and lower limits to avoid premature convergence or stagnation to a local optimum [19]. As a novel algorithm with regard to computational molecular modeling in biological applications, ACO algorithms have received the considerable attention of several researchers in many fields. In computer-aided drug design, a novel docking algorithm PLANTS (Protein-Ligand Ant System) which is the ACO-based algorithm developed by Krob et al was used to predict the binding pose protein kinase A[16].

ACO algorithm was employed to identify important descriptors in QSAR (Quantitative structure-activity relationship) and to predict COX-2 inhibition activity

[20]. In both cases, the ACO has demonstrated considerable performance with satisfactory convergence rates.

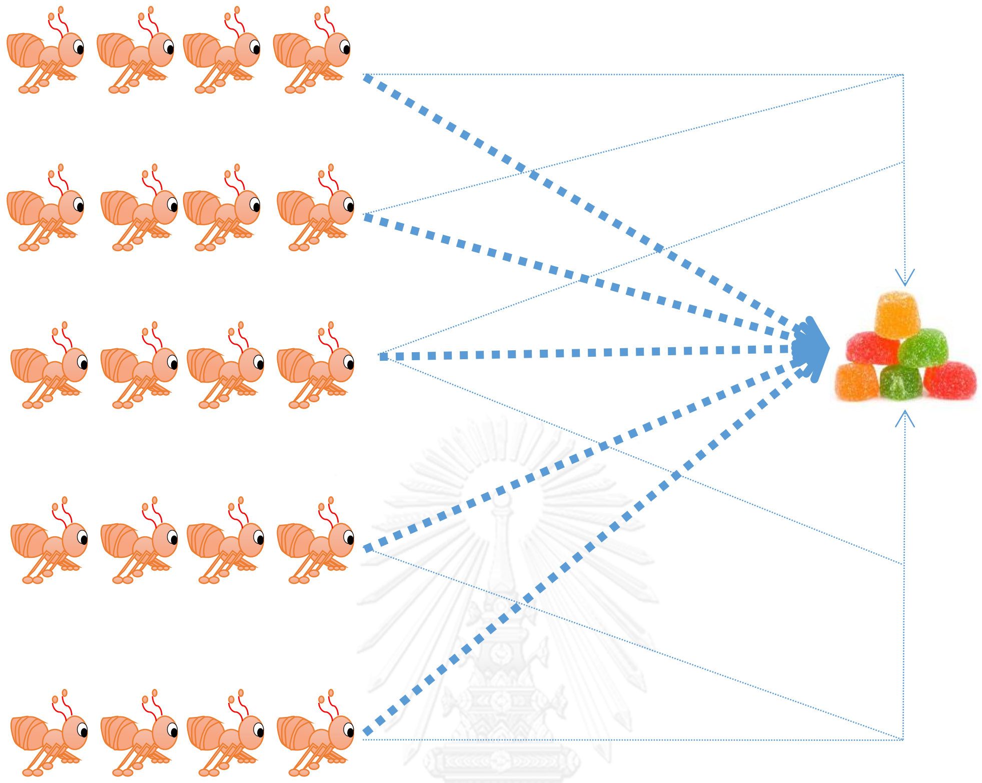

Figure 1. 3 Finding the food of the ant’s behavior that they try to find the shortest path from nest to food.

Naturally, ants are likely to find their food by randomly seeking around their nest, then drop an amount of pheromone depending on the quality and quantity of foods to mark the trail on the path[14]. The pheromone trail will lead other ants to follow the track to the food source and deposit more pheromones which in turn increase a probability of selecting that route. The ants decide to choose the path according to the pheromone concentration, the higher the pheromone intensity on the path is, the more ants are likely to choose that path. A shorter path has a higher pheromone concentration whereas the probability of choosing a long path is lower because the pheromone on the long path evaporates faster in comparison to the shorter one. Through this mechanism, it enables ants to choose with higher probability paths that are indicated by stronger pheromone concentrations. MMAS can be described using the Travelling Salesman Problem (TSP)[20]. Assuming that there are a set of *N* cities, distances between each pair of cities are known. The goal of TSP is to find the shortest tour around a set of cities.

Generally, MMAS consists of two main iterative processes: tour construction and pheromone trail update [18, 21, 22]. Initially, *m* artificial ants (a parameter specified by the user) are placed in randomly chosen cities. An amount of pheromone (τ_0) in the initialization stage is assigned according to Eq.1.1

$$\tau_0 = \frac{1}{C^m} \tag{1.1}$$

where C^m is the possible range of the tour lengths, in which this study is set to 1. In order to construct a tour, the probability (*p*) of ant *k* moving from a current city *i* to a neighboring city *j* is determined according to Eq. 1.2

$$p_{ij}^k = \frac{[\tau_{ij}]^\alpha [\eta_{ij}]^\beta}{\sum_{j \in N_i^k} [\tau_{ij}]^\alpha [\eta_{ij}]^\beta} \qquad \text{if } j \in N_i^k \tag{1.2}$$

where τ_{ij} is the amount of pheromone deposited from city *i* to city *j*;α and β are the weight of pheromone trail and of heuristic information, respectively; η_{ij} is an available heuristic information between city *i* and city *j*; N_i^k is a set of cities which ant *k* has not visited yet.

Once all ants have completed their tour construction, an update of the pheromone trails is performed. The pheromone update process consists of the pheromone evaporation and the deposition of new pheromone. The update of pheromone level τ_{ij}^{new} is computed according to Eq. 1.3

$$\tau_{ij}^{new} = (1-\rho)\tau_{ij}^{old} + \sum_{k=1}^{m} \Delta\tau_{ij}^k, \qquad \forall (\mathrm{i, j}) \in \mathrm{L} \tag{1.3}$$

Where ρ (with 0 <ρ< 1) is a fraction of pheromone evaporation and L is possible paths. Thus (1 - ρ) is a fraction of remaining trail. In the right hand side of Eq.1.3, the first term corresponds to the pheromone evaporation and the second term is the amount of pheromone deposition, which is defined according to Eq. 1.4

$$\Delta\tau_{ij}^k = \begin{cases} 1/C^k, & \text{if arc(i, j) belongs to}\,\mathrm{T}^\mathrm{k} \\ 0, & \text{otherwise} \end{cases} \tag{1.4}$$

Where T^k is the best so-far arc from the previous results and C^k is the distance of ants passed in the path. The C^k value is arbitrary for this study. We found that C^k

of 10 is an appropriate value for weighting pheromone concentration of the best solution of the current iteration used for the next iteration. For MMAS the upper (τ_{max}) and lower (τ_{min}) bound of the pheromone concentration is employed to ensure pheromone intensities lie within a given range. The upper bound and lower bound are defined according to Eq.1.5 and 1.6

$$\tau_{\max} = \frac{1}{\rho C^{bs}} \tag{1.5}$$

$$\tau_{\min} = \frac{\tau_{\max}}{2n_c} \tag{1.6}$$

Where C^{bs} is the best-so-far distance and n_c is number of cities. In this study, C^{bs} is set to 1 in order to avoid τ_{max} approaching to zero or too small value. The n_c corresponds to the total number of degrees of freedom defined for each TMH.

1.2.1.2 Genetic algorithm

Genetic algorithm (GA) is the method which depends on the concept of evolution and natural selection followed by Charles Darwin. This concept has been described as the stronger will survive for following generations, whereas the vulnerable creature will die and become extinct. It had three strategies for selection including crossover process, mutation process and cutoff offspring[8]. John Holland applied this concept to use for finding answers via optimization. The better answer is like the tribal strains which possess the strength to develop their tribe. However, the terrible answers are like the vulnerable races which will finally die and become extinct. Genetic algorithm is better than conventional AI because it is more robust. Conventional AI cannot be broken easily when the input is changed slightly. GA is based on analogy with the genetic structure and behavior of chromosomes within the population of individual followed by: first, individual in a population to complete a resource and mate. Second, those individuals will produce more offspring than the others that perform poorly. Third, genes from strong individuals increase their population and will become more suited to their environment. This GA process starts from a population of randomly generated individuals and is an iterative process with the population in each iteration called a generation. In each generation, the fitness of every individual in the population is evaluated. The fitness is usually the value of the objective function in the optimization problem being solved. Then each generation is modified to form a new generation. The new generation of solutions is then used in the next iteration of the algorithm. Generally, the algorithm finished when either a maximum number of generations have been produced. This method is composed of an initial population of solutions. Then, it improves it through repetitive application of the mutation, crossover, inversion and selective operators.

1.2.1.3 Simulated Annealing Monte Carlo method

This algorithm uses the Monte Carlo method to apply with the simulated annealing method. During each cycle of this algorithm there is a constant in temperature cycle, the legend’s position, orientation and conformation are randomly changed in case of flexibility The new position is compared to its predecessor, and new energy is less than the former. The new state is recognized. But if the new state possesses the energy more than the previous, it is also recognized probabilistically. This probability depends on the energy and cycle temperature. Mostly at high temperature, many states will be accepted. Oppositely at low temperatures, most of these probabilistic moves will be declined. The user can select the minimum energy state during a cycle and use this as the initial state for the next cycle or the last state. The best result is progressed by choosing the minimum energy state from the former cycle. The initial annealing temperature or rt0 could be the order of the average DE presented during the first cycle. This could be confirmed that the ratio of accepted to rejected part is high at the beginning. An initial annealing temperature (rt0) of a typical automated docking job is 500 which is depending on the system’s average DE. The temperature reduction factor (rtrf) is shown at 0.85 per cycle. For avoiding simulated quenching a gradual cooling should be used leading to trap systems in local minima. Conditional to the degree of complexity of the problem, a good search is showed by 50 Monte Carlo “cycles” and a maximum of 30,000 steps rejected “rejs” or 30,000 steps accepted “accs”. 10 “runs” may or may not give a range of possible binding modes and multiple runs will give relative energy. The condition of 100 runs, 50 cycles, 3000 steps accepted and

3000 rejected will show more populated clusters, hinting at the ‘density of states’ for a given conformation. A short test job is shown as 1 run, 50 cycles, 100 accepted, 100 rejected steps. The user has to particularize the maximum step a state variable can do in one step. However, this condition can be changed in Monte Carlo simulating annealing, if a reduction factor (a fraction from 0 to less than 1) for translations and rotations is shown. At the start of each cycle, the range from the former cycle is multiplied by this constant value to give the new value for translational and angular displacements. If required, the states can be sampled during a docking and output to a trajectory file. This file has all the state variables required to assign each sampled conformation, position and orientation of the ligand. The user can determine the range of cycles as a sample. This enhances the selection of the last few cycles when the docking will be close to the final docked conformation, or the selection of the total run[23].

1.3 Experimental design and analytical data

Nowadays, there are several factors impacted on the system. This thesis’s purpose is to find factors that influence with to find objective function from the system. If this work wants to develop this experiment for the best efficiency, this thesis has to use science theory to design the experiment in terms of statistic or statistical design of experiment. This is used to find the suitable data and can then be taken to statistical analyst. Finally, It enhances to get the best result and direct to its objective.[24]

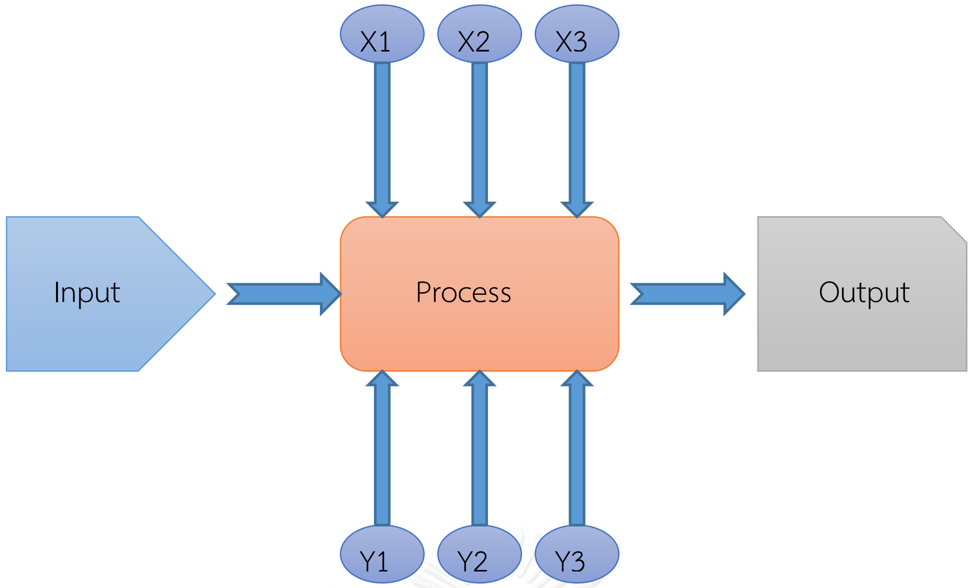

Figure 1. 4 The schematic of the experimental design

The experiment is mean testing which can have adjusted parameters in the system making it easy to observe and specify the cause of result which differs from the previous. The experiment will be a success if it is planned and designed and the method outlined step by step before working [24] Experimental designs pass procedure planning for collecting the required data. This data is analyzed via statistic approach leading the result related to its objective. Experiment design should be easy to work with the efficiency process. Moreover, it should use resource efficiency such as experimental time and funding [25]. There are many strategies of experimental design including one factor at a time and Factorial designs [25]. One factor at a time design is to start with selecting a process from chosen factors and fixes a constant value in other factors. Then test each of the factors until the factor that impacted the system can be found [25]. However it can be observed that the result of one factor depends on other factors or coordinating factors. In this case it can be called interaction of each factor. This process contains some disadvantage that cannot consider this interaction between factors. Therefore, experimental results which have to analyze on this point should not use one factor at a time method. The Factorial design is an efficiency design because there is a replication process for decreasing in variable rate and consideration in the combination result. This factor will indicate factors in term 2 or 3 factors for instance X factor is composed of x level and Y factor including y level, so in a cycle it can be explained that there are 2 impactions such as the main effect or interaction between factor. The main factor is the impaction of studied factors which is in transition when a factor is changed in level. Interaction between factors is the result of each factor which is not equilibrium when compared from level to level. It can be explained that the result of one experiment is dependent on the level of other related factors. Factorial design is popular in all fields of research. Researchers usually use this design because of decreasing time as most of the experiment is studying the impacted factor of more than 2 factors. Other than the studying of the main factor, one should also study about the interaction between related factors. Therefore factorial design is efficient and suitable for this experiment [24, 25] due to its having a variety design such as there are 2 factors and 2 levels in one experiment. So this sample can call this factorial design size 2 x 2 or 2^2 resulting in 4 times of testing.

1.3.1 2^k Factorial designs

The important design of factorial design is 2^k factorial design where k is the number of factors and 2 is the number of levels of factors. Full Factorial is composed of 2 x 2 x 2 x ... x 2 = 2^kdata, this design is given the lowest number of testing as it can study in impaction of factor all k factors perfectly. It is very useful in the beginning when there are many factors which have to be selected factorial design is popular for being used to find the influential factor. It can be said that there are several interesting factors in experiments for searching impacted factor to the system. The 2^k factorial gives a lot of testing at once. So half of 2^kfactorial have to test2^{k-1}time leads to a decrease in the number of testing and so called the one-half fraction of the 2^k design[24, 25].

1.3.2 3^k Factorial designs

This method is used when there are k factors to consider. Each of the factors consists of 3 levels including high, intermediate and low level. In one experiment composed of 3 x 3 x 3 x ... x 3 = 3^kdata which is called 3^k Factorial design. It can be replaced with a number by the first number in the level of factor X, the second number is the level of factor Y,......, and k number is the level of factor K. the design is suitable for researchers who are interested in the concave response. This process continues followed by its plan and the collected data as required. Finally, these data are statistically analyzed[24, 25].

1.3.3 Analysis of variance: ANOVA

Analysis of variance or ANOVA is efficient to test the hypothesis related to the impaction from each level factor or coordination of the average value from each of the level factors. The process of this approach is shown as below. Analysis of variance is most commonly used and advantages for considering the impaction of many factors which can respond. Not only the main factor, but it can be used to analyze the interaction between factors. To study 2 factors factor X and Y and 2 levels by denoting x as number level of factor X and y as number level of factor Y. Therefore one factorial experiment is composed of an xy test and n is the number of repeating.[25]Generally, the table of analysis of variance consists of the Source of variation, Degrees of freedom (DF),Sum of square (SS), Mean squares (MS)] and F-value.

1.3.4 Multiple comparisons

Analyses of variance when hypothesis is rejected as a result, then the experiment need to know at least one different group. It can not specify this group, then it is necessary to work in the next step to observe which factor is differently impacted to each other. This comparing such as Student-Newman-Keul (SNK), Least Significant Difference (LSD), Un can’s new multiple range test, Turkey and Scheffe’ the experiment show each method contains different limitation. In 1965 Carmer and Swanson used Monte Carlo which is the process for forming data by using randomized numbers and probability[26]. The resulting data is similar to a fact which is not stable. It presents that LSD is more efficient to find the different of average exactly if this sample display these after analysis of variance by F test at 5% significance. Moreover, Duncan’s new multiple range tests are efficient and used in statistical programs. The LSD method by Fisheries is also efficient but it should be analyzed of variance by using F-test which P-value is less than 0.05 before selecting this test.

1.4 Methods for prediction structure membrane protein

Because of the difficulty in obtaining high-resolution structures, many different approaches for predicting membrane protein structures from amino acid sequences have become an important alternative approach. The methods may be classified into three general categories. First, comparative or homology model of membrane protein structure built based on sequence similarity is the most common strategy. However, this approach requires proteins with a known 3D structure serving as the structural template .A more challenging method is template-free prediction of membrane protein structure if the structure of the homologous proteins does not exist. The *ab initio* modeling such as ROSETTA de novo structure prediction utilizes Metropolis Monte Carlo sampling approach and knowledge-based empirical energy function to successfully predict helical transmembrane proteins with various sizes and topologies at considerable accuracy. Another approach is to incorporate biophysical and biochemical data as constraints into conformational search methods (i.e. molecular dynamics, simulated annealing Monte Carlo or distance geometry) to bias sampling toward conformations that are consistent with the experimental results. The latest strategy is remarkably promising because advances in protein engineering combined with many recent techniques such as site-directed spin labeling and electron paramagnetic resonance (SDSL/EPR), chemical cross-linking and mass spectroscopy, cryo-electron microscopy (CryoEM) and fluorescence resonance energy transfer (FRET) have become routinely accessible for elucidating the distance and distance changes at selected regions within a protein or between proteins. [5-7, 27-30]. Although these approaches are limited by the low resolution and sparse distance data, incorporating the low-to-moderate resolution structural data into an efficient computational method has made it possible to calculate a 3D structure from partially unfolded structure as well as conformational changes of transmembrane proteins at moderate resolution. The obtained structure can serve as a starting point for improving a higher accuracy of the structure using additional refinement methods.

Membrane proteins can be classified into three major categories of (i) integral, (ii) peripheral and (iii) lipid-anchored membrane proteins on the basis of the nature of their association with the lipid bilayer. Integral or transmembrane protein (TMP) is the major class of membrane proteins whose structural architecture is composed of single or a bundle of membrane-spanning segments, in which two basic secondary structures, α-helices and β-barrels, have been observed. Transmembrane α-helices (TMHs) are the most common structural motif of TMP [31]. The α-helical transmembrane proteins are an ideal model system for the method development. Unlike β-barrel transmembrane proteins which are mainly found in the outer membranes, α-helical proteins found in all cellular membranes. They are responsible for numerous and diverse cellular functions, but their structures remain largely unsolved due to the lack of structural homologues. Based on observations of known membrane protein structures, each transmembrane helix (TMH) was found to possess a tilt angle of less than 40° with respect to the bilayer normal[32, 33]. This is owing to structural constraints imposed on TMHs by the hydrophobic lipid bilayer. Because of the geometric restriction, the accessible conformation space is substantially reduced, hence, reduces the complexity for sampling a large variety of the assembly search space. A significant reduction of the degrees of freedom is considered to be the major advantage in the structure prediction of transmembrane proteins in this class [34].

Comparative or homology modeling approaches is the most well-known structure prediction method. These methods require known 3D structure of homologous protein as a template. However, homology modeling is not generally used for structural modeling of membrane proteins because of the limited available number of atomic-resolution structures of transmembrane protein families. Assembly of TM helical bundle has an important role in stabilizing global-fold structure of membrane proteins. A number of experimental and computational methods have been reported to address the molecular mechanism of TM assembly[27]. One of the most well-known membrane prediction approaches is the helix-helix packing-based approach [28-30]. In this approach, Monte Carlo algorithm was employed to randomly generate structure models, of which an assessment was subsequently carried out to choose the final model. This structural evaluation relies on the scoring or probability methods which have been derived from contact propensities between inter-helical contacting residue pairs (polar and non-polar groups) in membrane proteins. A hydrophobicity-based method has successfully been introduced to model a number of transmembrane proteins. The method is based on the statistical frequency analysis of amino acid sequence in membrane proteins.

1.5 Problem statement

1.5.1 KcsA potassium channel: KcsA / TM2

Many transmembrane proteins contain two or more transmembrane segments. For instance, the bacterium *Streptomyces lividans* (KcsA) potassium channel, a homotetramer protein, comprises two transmembrane α-helices of each monomer. Each monomer contains 160 amino acid residues. The 2Å-resolution x-ray structure of KcsA potassium channel (a PDB accession code: 1k4c) is shown in Figure 1.5

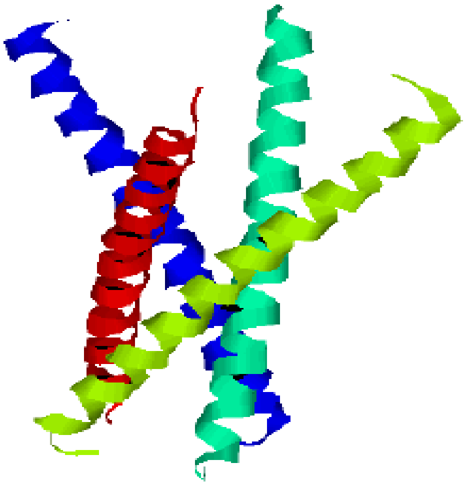

Figure 1. 5 The inner transmembrane segments of the tetramer structure of KcsA potassium channel

A possible arrangement of a single transmembrane helix with respect to the membrane normally consists of a total of five configurational parameters including a translational parameter (σ) and four rotational parameters (θ_1-θ_4) σ was employed in a range from 0 to 20Å. θ_1 θ_2 and θ_3 associated with the rotation along the x-, y- and z-axis, respectively, were subjected to rotate the helix in a range from 0° to 90°. The rotation about its helical axis defined as θ_4 was given in a range between 0° and 360° (Figure 1.6, Figure 1.7).

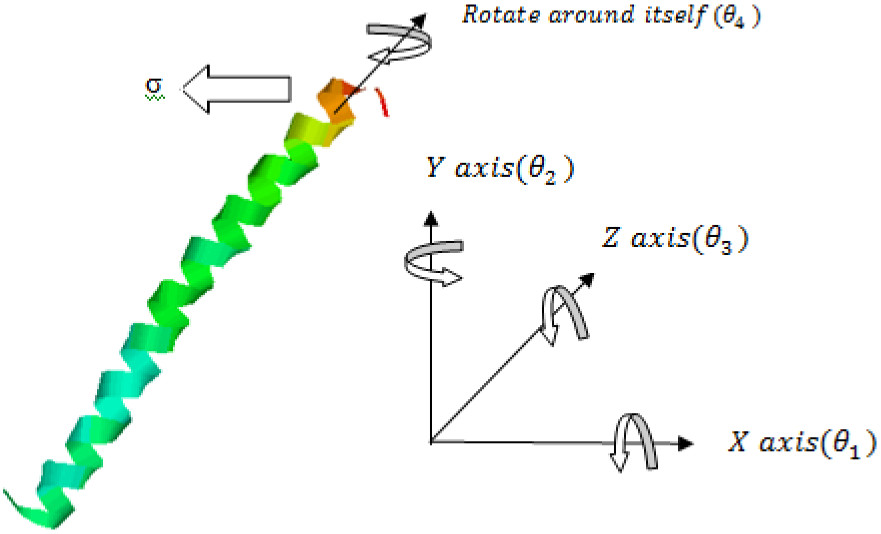

Figure 1. 6 A single transmembrane helix and the definition of configurational parameters.

A given value of σ, θ_1, θ_2, θ_3 and θ_4, 3D structure of a transmembrane helix is generated as an initial segment. Subsequently, the tetramer structure of KcsA is obtained by performing a fourfold symmetric operation along the channel axis (Figure 1.7). After structure generation, root-mean-square-deviation (RMSD) of Cα atoms relative to the reference was used to measure the difference between the computed structure and the x-ray crystal structure.

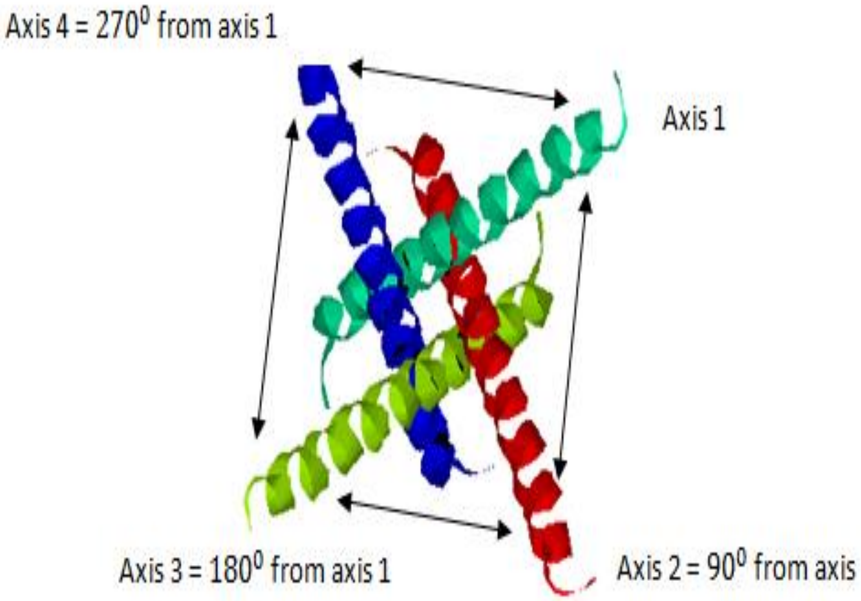

Figure 1. 7 The tetramer structure of the channel after fourfold symmetric transformation of the structure coordinates.

1.5.2 KcsA potassium channel: TM1-TM2

KcsA / TM1-TM2 contains two helixes (figure 1.8). This type has 10 values of parameters including σ_1 , σ_2, θ_1 , θ_2 , θ_3 , θ_4 , θ_5 , θ_6 , θ_7 and θ_8. For σ_1 and σ_2 were employed in a range from 0 to 20 Å. θ_1, θ_2 , θ_3 , θ_5 , θ_6 and θ_7 associated with the rotation along the x-, y- and z-axis, respectively, were subjected to rotate the helix in a range from 0° to 90°. The rotation about its helical axis defined as θ_4 and θ_8 were given in a range between 0° and 360°. (Figure 1.8)

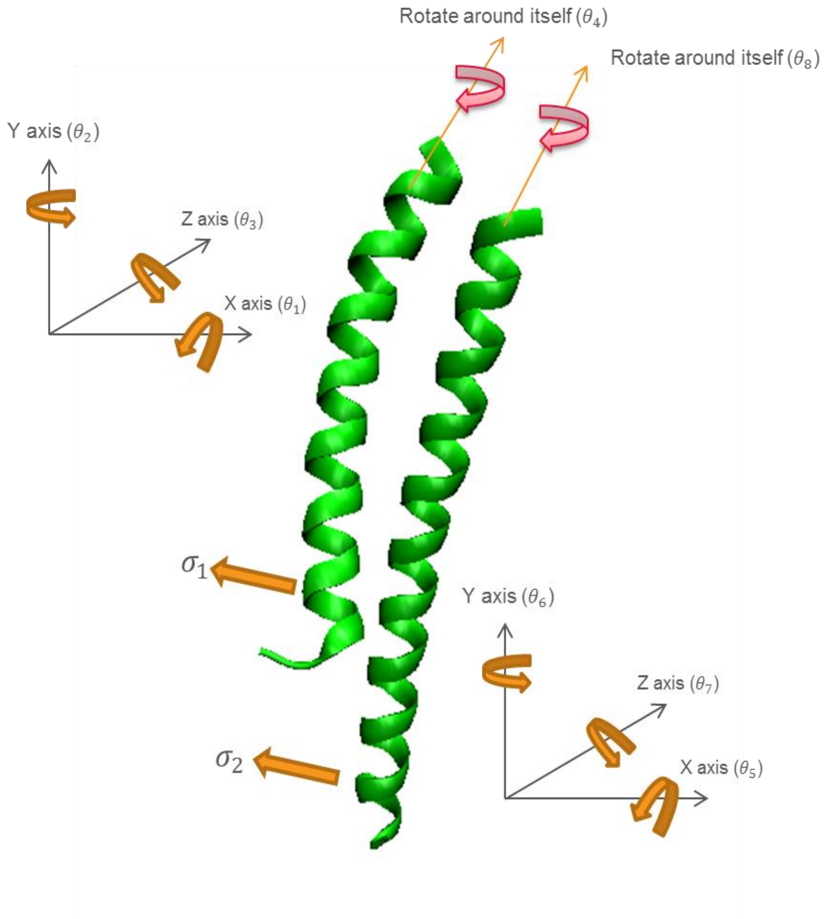

Figure 1. 8 Degree of freedom of rotational parameters for KcsA / TM1-TM2

1.5.3 KvAP / VSD

The input of KvAP/VSD file has four helixes (figure 1.9). There are denoted as A,B,C,D segments, However this study considers only on A,B,D which are associated with twelve parameters. The θ_1 , θ_2 and θ_3 in helix A are allowed to change in range from 0° to 20°. θ_4 ranges from 0° to 360°. The rotation of helix B is defined by θ_5 , θ_6 , θ_7 and θ_8. They are sampled in a range from 0° to 20° except for θ_8 (0° to 360°). The range of D helix rotation is from 0° to 40° in θ_9 , θ_{10} , θ_{11} and θ_{12} range 0° to 360° (Figure 1.9).

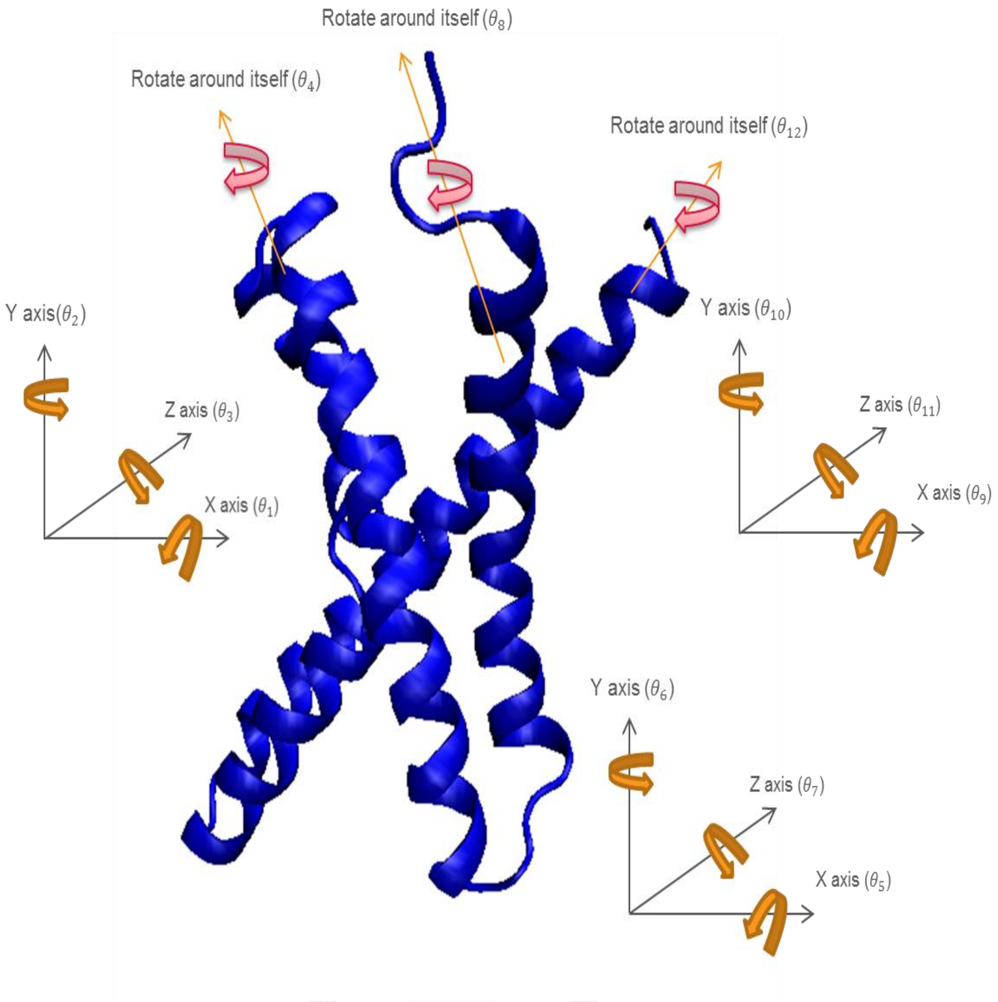

Figure 1. 9 Degree of freedom of rotational parameters for KvAP / VSD

1.5.4 MscL / TM1-TM2

MscL has two helixes: TM1 and TM2 (Figure 1.10) . The parameters of MscL are σ_1 , σ_2 θ_1 , θ_2 , θ_3 , θ_4 , θ_5 , θ_6 , θ_7 and θ_8. σ_1 , θ_1 , θ_2 , θ_3 and θ_4 are in helix one and σ_2 , θ_5 , θ_6 , θ_7 and θ_8 are in helix two.

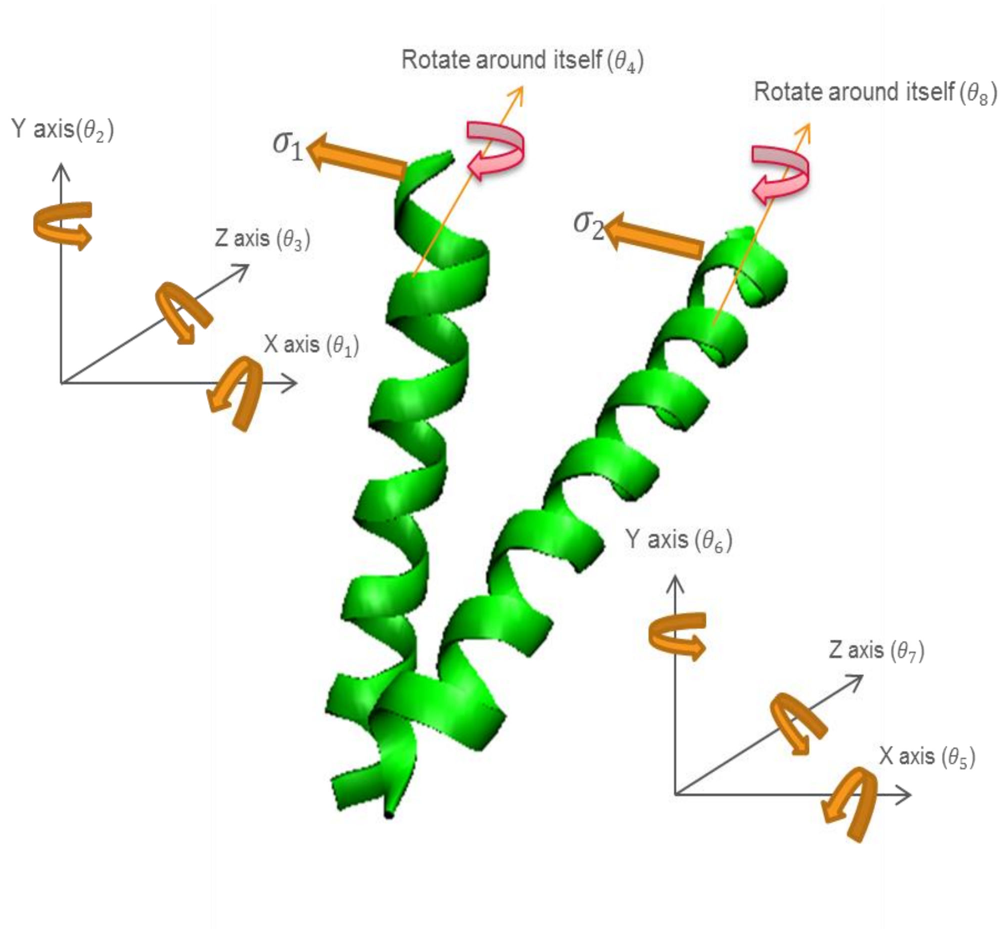

Figure 1. 10 Degree of freedom of rotational parameters for MscL / TM1-TM2

1.6 Rationale and Objectives

The framework of this thesis is to develop a novel computational approach for modeling on assembly of transmembrane proteins. The approach is based on nature-inspired metaheuristic that is Max-Min ant system algorithm. An efficiency of the purposes algorithm is compare with conventional methods such as genetic algorithm and simulated annealing monte carlo method. The testing models limit to α- helical membrane proteins

CHAPTER II

MATERIALS AND IMPLEMENTATION

2.1 Materials

2.1.1 Hardware

The desktop computers with intel core i7 processor 3.7 GHz 3770 CPU RAM 8.00 GB were used to operate the simulation in this work.

2.1.2 Software

2.1.2.1 The Visual molecular dynamics (VMD)

VMD is the graphic program for molecular use to check input and output files. It represents a trajectory file.

2.1.2.2 The visual basic (VB)

VB is used to show how to base an object oriented program and graphic user interface by Microsoft. It is a tool for developing a program on a windows operation system.

2.1.2.3. The statistical package for social science (SPSS)

SPSS is a program for statistical and analytical data. It can be represented in charts, graphs and data tables.

2.1.2.4 The Origin program

Origin is the program for data analysis and scientific graphing. It supports graph 2D and 3D

2.1.2.4 Tool command language and Toolkit (Tcl/Tk)

Tcl/Tk is to develop for dynamic programming language and easy to graphic interface by user.

2.2 Penalty Functions and distance constraints

THAMMAS is used to search for the conformational space by minimizing the objective penalty function (P) that satisfies a set of distance constraints[35]. The penalty value is determined by a sum of the square of the residual or violations (*viol*) which considers the differences between the distances calculated from the assembly model of TMHs (r^{calc}) and the distance constraints from the equivalent pairs. The penalty function is introduced as a harmonic function that is

$$P = \sum k(viol)^2$$

2.1

$$viol_{ij} = \begin{cases} r_{ij}^{calc} - r_{ij}^{upl} & ; \quad r_{ij}^{calc} > r_{ij}^{upl} \\ r_{ij}^{lol} - r_{ij}^{calc} & ; \quad r_{ij}^{calc} < r_{ij}^{lol} \\ 0 & ; \; r_{ij}^{lol} \le r_{ij}^{calc} \le r_{ij}^{upl} \end{cases}$$

2.2

Where r_{ij}^{upl} and r_{ij}^{lol} are the upper and lower distance constraints between ith and jth residues, respectively, $r_{ij}^{calc}$is the distance of the model, *k* is an arbitrary value and use as a weighting factor. In this study, a set of inter-helical Cα-Cα distances was generated based on the crystal structure of the reference proteins [36, 37]. Then, r_{ij}^{upl} and r_{ij}^{lol}were computed by an addition or subtraction of 1Å to the generated distances in the set.

2.3 Implementation

In this study, the author has developed THAMMAS (Transmembrane Helix Assembly by Max-Min Ant System, the first MMAS was designed for an assembly prediction of α-helical transmembrane proteins using a sparse set of distance constraints. THAMMAS generates a large variety with finite number of orientations of transmembrane helix bundle and finds the solution that is matched with the provided distance constraints based on the MMAS algorithm. The author demonstrates the efficiency of THAMMAS in identifying the solution for helical arrangement of transmembrane proteins, KcsA, MscL and KvAP voltage sensor domain with considerably small sampling iterations. The efficiency of the proposed method has been compared with that of two conventional optimization approaches. These are the simulated annealing monte carlo (SAMC) and genetic algorithms, which have widely been used in many applications in protein structure modeling and drug discovery such as protein-ligand docking, structure calculation, structure prediction and prediction of biological activity etc.

2.3.1 MMAS methods

The implementation of the MMAS algorithm used in the present work has been written in Visual Basic. Specifically, our algorithm, namely THAMMAS, generates a set of TMHs assembly models and the solution is based on the best distance constraint penalty. A random starting configuration of TMHs and THAMMAS parameters including the Number of iterations and ants (I/A), Evaporation rate of the pheromone (ρ), the exponent values of the pheromone trail (α) and of the heuristic measure in the random proportional rule (β) have been introduced in the process. The THAMMAS *parameters* were assigned to the values according to Dorigo and Stützle's work[10], and subsequently optimized using a full factorial design with three levels of all factors and statistical analysis. The initial amount of pheromone is placed in every dimension of all of the search areas. By using a rigid-body transformation of TMHs, each helix consists of five degrees of freedom (σ, θ_1, θ_2, θ_3 and θ_4) which parameters used are described later. The assembly search space is given by the degrees of freedom of TMHs.

For constructing a tour, THAMMAS computes are used one at a time and uses the probability of the current node according to Eq. 1.2, and selects the next node using the proportional roulette wheel method and the available values within a specified range of translation and rotation. The ant will read these pheromone information and move according to the selected paths in the tour. After finishing the tour, THAMMAS computes an assembly model of TMHs. At this step, symmetry and coordinate transformations for generating structure coordinates are performed if the tested protein is homo-oligomer. Then, the penalty value of the obtained model is computed and stored in the array of ant *k*. In this stage, the other ant is released in order to repeat another random tour until the maximum number of ants is reached. Once all ants have completed all the tours, the penalty values are sorted. The best tour as shown by the lowest penalty of the current iteration is kept as the best-so-far route and gets extra pheromone while the rest of the ants do not deposit the pheromone on their paths. The pheromone values are added to the best-so-far route depending on the quality of the calculated penalty value. After applying the pheromone evaporation to all search paths, the pheromone update is performed and controlled within the upper and lower bounds. The updated pheromone affects the decision of the ant to choose the solution in the next iteration. This iterative process is repeated until a given termination criterion is achieved. Figure 2.1 shows the corresponding flowchart of THAMMAS. The pseudo code of the basic THAMMAS algorithm is illustrated in this figure.

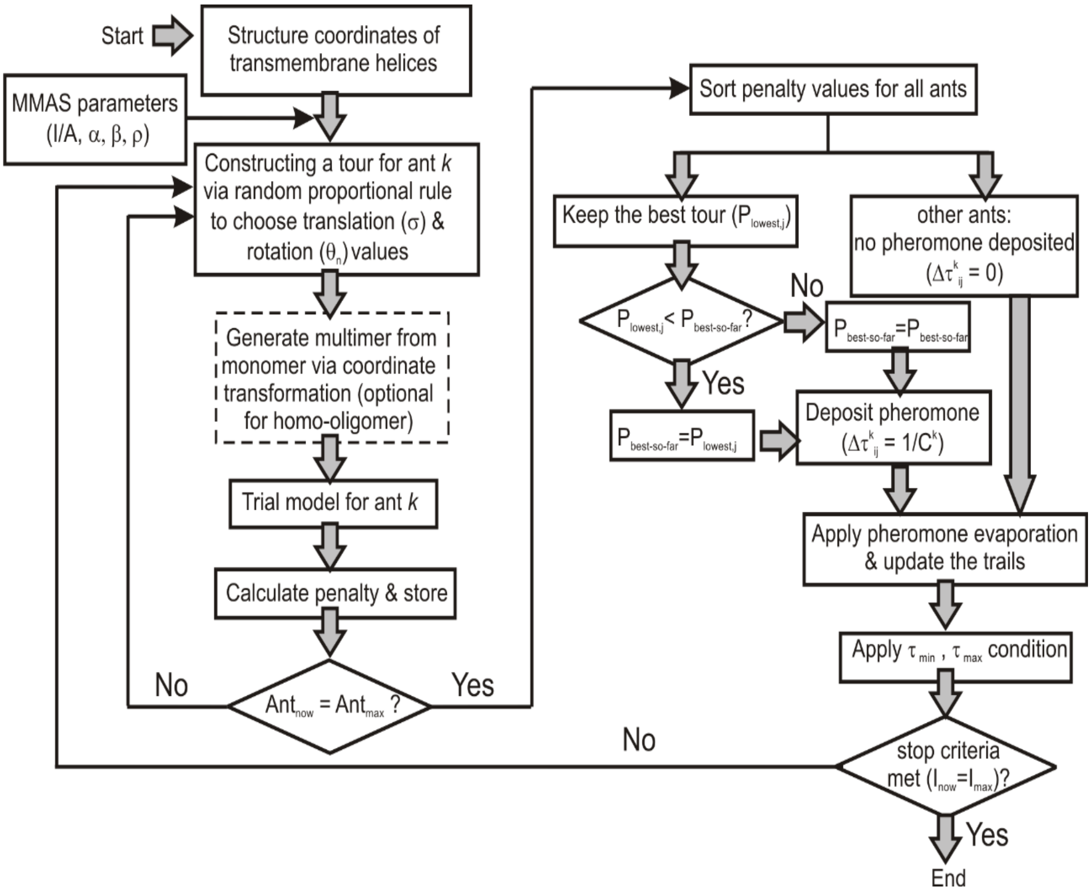
Start
Structure coordinates of transmembrane helices
MMAS parameters (I/A, α, β, ρ)
Constructing a tour for ant k via random proportional rule to choose translation (σ) & rotation (θ_n) values
Generate multimer from monomer via coordinate transformation (optional for homo-oligomer)
Trial model for ant k
Calculate penalty & store
No
$\mathrm{Ant}_{now} = \mathrm{Ant}_{max}$?
Yes
Sort penalty values for all ants
Keep the best tour ($P_{lowest,j}$)
other ants: no pheromone deposited ($\Delta\tau^k_{ij} = 0$)
$P_{lowest,j} < P_{best\text{-}so\text{-}far}$?
No
$P_{best\text{-}so\text{-}far} = P_{best\text{-}so\text{-}far}$
Yes
$P_{best\text{-}so\text{-}far} = P_{lowest,j}$
Deposit pheromone ($\Delta\tau^k_{ij} = 1/C^k$)
Apply pheromone evaporation & update the trails
Apply τ_{min} , τ_{max} condition
No
stop criteria met ($I_{now} = I_{max}$)?
Yes
End

Figure 2. 1 MMAS flow chart

2.3.2 Genetic algorithm

This algorithm begins to define value parameters including crossover percent rate, mutation percent rate, number of generation and number of population. Secondly, random degrees of freedom (σ, θ_1, θ_2, θ_3 and θ_4) into chromosome array.

Third, generation of systematic operation and calculation penalty function. Fourth, take condition for crossover and mutation by roulette wheel selection. Fifth, this step is to select and cutoff offspring to the next generation. Finally, when the stopping criteria is met the algorithm terminates.

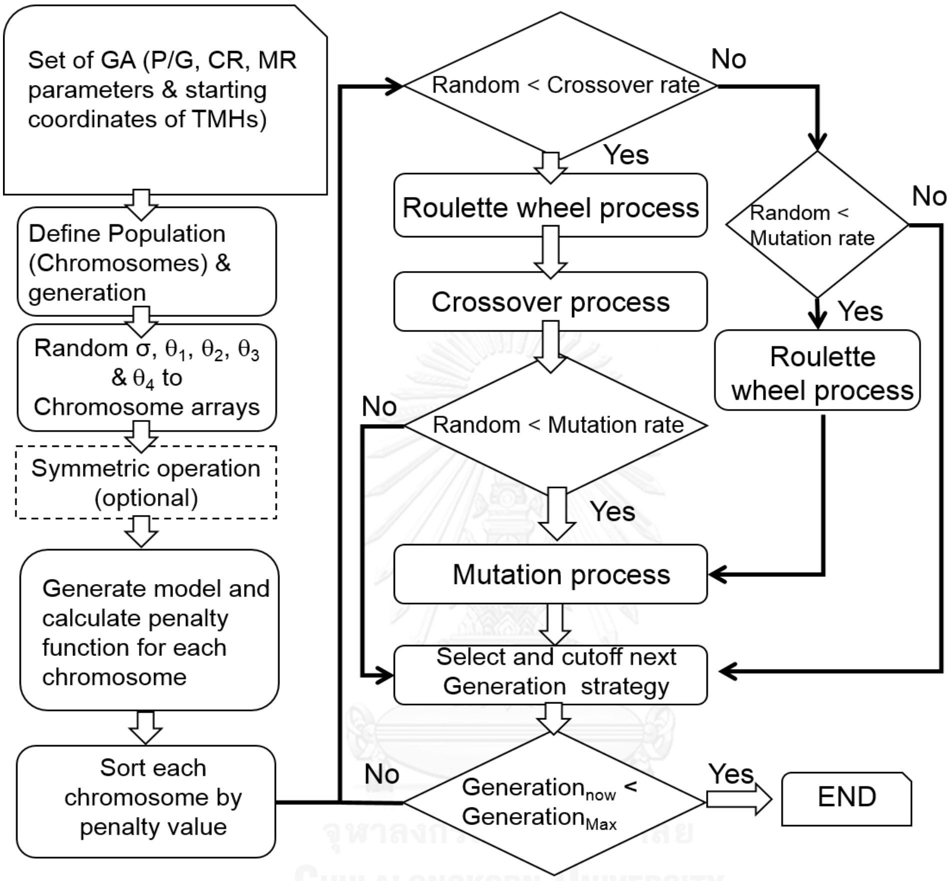
Set of GA (P/G, CR, MR parameters & starting coordinates of TMHs)
Define Population (Chromosomes) & generation
Random σ, θ1, θ2, θ3 & θ4 to Chromosome arrays
Symmetric operation (optional)
Generate model and calculate penalty function for each chromosome
Sort each chromosome by penalty value
Random < Crossover rate
No
Yes
Roulette wheel process
Crossover process
Random < Mutation rate
No
Yes
Mutation process
Select and cutoff next Generation strategy
Random < Mutation rate
No
Yes
Roulette wheel process
No
Generation_now < Generation_Max
Yes
END

Figure 2. 2 GA flow chart

2.2.3 Simulated annealing monte carlo

The scope of this method is based on annealing of metals and sampling technique of monte carlo approach. The initial step of this algorithm is to define iteration cycles.. Next random value of σ, θ_1, θ_2, θ_3 and θ_4 are generated and store into array. Afterwards, structure model is generated and evaluated by calculating penalty function. The model with the lowest penalty is kept as the best so far penalty value. The process is repeated iteratively until the stopping criterion is obtained.

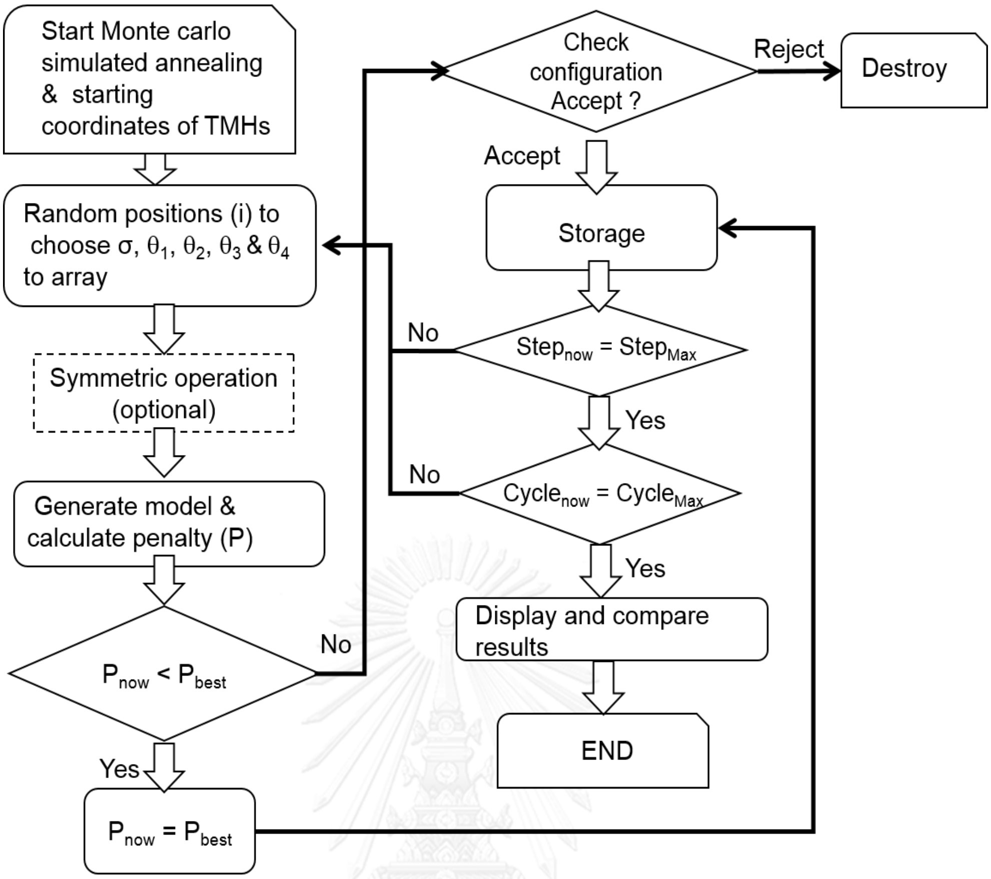
Start Monte carlo simulated annealing & starting coordinates of TMHs
Random positions (i) to choose σ, θ1, θ2, θ3 & θ4 to array
Symmetric operation (optional)
Generate model & calculate penalty (P)
Pnow < Pbest
Yes
No
Pnow = Pbest
Check configuration Accept ?
Reject
Destroy
Accept
Storage
No
Stepnow = StepMax
Yes
No
Cyclenow = CycleMax
Yes
Display and compare results
END

Figure 2. 3 MCSA flow chart

2.4 Testing models

This thesis has tested our present method with three α-helical transmembrane proteins of known structure including the pH-gate potassium channel from *Streptomyces lividans* (KcsA), the mechano sensitive channel from *Mycobacterium tuberculosis* (MscL) and the voltage sensor domain of the voltage-dependent potassium channel from *Aeropyrumpernix* (KvAP). The structure coordinates of TMHs were taken from the crystal structures corresponding to PDB code 1K4C for KcsA[38] and 2OAR for MscL[38]. Since the experimental EPR data of the bi-functional spin label is employed in the case of KvAP, The TMHs structure of voltage sensor domain was taken from the work published previously[39]. Briefly, this KvAP structure was the crystal structure(1ors)[40] modified by attaching EPX-pseudoatoms, which is covalently connected to the two-Cα of the corresponding labeled residues[41]. This work used KcsA and MscL as model systems for the assembly prediction of symmetric multimer membrane proteins while the isolated voltage sensor domain of KvAP was used to demonstrate an example of multi-helical packing within the subunit. KcsA and MscL are the pore-forming ion channels composed of two TMHs arranged in four-fold and five-fold symmetry respectively. For KvAP, this thesis focused only on the isolated voltage sensor domain which is comprised of four TMHs. Detailed information on the transmembrane segments of the proteins that were used in the study are described in the Results and Supplementary Information.

2.5 Assembly search space: degrees of freedom

The conformational search space of transmembrane helix bundles is performed aiming to optimize inter helical distance penalty over rigid-body helices [36, 42]. In addition, the structure of loops was not considered in this study. The arrangement of rigid-body helix with respect to the membrane normal is defined by five degrees of freedom including a translational parameter (σ) and four rotational parameters (θ_1-θ_4). θ_1, θ_2 and θ_3 are associated with the rotation of TMH along the Cartesian x-, y- and z-axis, respectively, whereas θ_4 defines the rotation of its helical axis. The helix axis is defined as the vector connecting between the average CA coordinates of the first and last four residues of the N- and C-terminal ends of a helix. Unless otherwise specified, the translation σ was employed in a range from -20 to 20Å at intervals of 1Å, while the available range of the rotation is from 0° to 90° for θ_1-θ_3 and 0° and 360° for θ_4 with the interval of 10°. Cartesian transformation matrices are applied to generate 3D structure coordinates.

2.6 Statistics methods

The Design of Experiment (DOE) method was employed based on the factorial design with three levels: high, medium and low. Therefore, factorial 3^k design of experiments (k=4 factors that are I/A, α, β, and ρ) were performed. To assign the values of I/A, First determined the sample size (n) which is the representative of the overall populations. To determine the appropriate sample size, this thesis used Yamane's random sampling principle[43]. For populations of a known size, the sample size is determined according to Yamane's rule as below:

$$n = \frac{N}{1 + Ne^2} \qquad 2.3$$

Where *N* is the total number of population and *e* is the level of precision presented in terms of percent of accepted error.

For a given helix, there are 40 possible values for sampling the translation position over the range of -20 to +20Å in a 1Å step size. The number of the rotational values can be considered into two sets. The first set involves a helix tilting along the x-, y- and z-axis and the second set is defined by a helix twist. There are 10 possible values for a tilting range of -0 to +90° and 36 possible values for the 360° helix rotation in a 10° step size. Therefore, the total number of population is 40×10×10×10×36 = 1.44 × 10^6. The sample size for a 99% confidence level is:

$$n = \frac{1.44 \times 10^6}{1 + (1.44 \times 10^6) \times (0.01)^2} = 9931 \qquad 2.4$$

Thus, the sampling size used in the study is set to 10000. This number corresponds to the total numbers of tour generated by the algorithm, or in the other words the total conformational space to be generated for each THAMMAS run.

In the algorithm, the sample size is given by the number of iterations multiplied by the number of ants (I/A). With a sample size of n=10000, three sets of I/A parameter were assigned to 50/200, 100/100 and 200/50 for the low, medium and high levels of the factors, respectively. Values of the α, β, and ρ parameters for the three levels are summarized in Table 3.1. The (dependent variable) penalty was computed based on the distance differences between those obtained in an assembly model and in the constraints. In this step, distance constraints generated from the distances between the Cα atoms of residue pairs belonging to the inner helical bundle of the KcsA x-ray structure were taken into consideration in THAMMAS runs to guide an arrangement of the four inner helices of TM2 of KcsA. The measured distances could potentially be derived from EPR through spin-spin dipolar coupling of the spin label sample. It should be emphasized that starting with only one subunit of TM2 of the crystal structure. THAMMAS generates the other three around the four-fold axis of symmetry to obtain an assembly model. This model system represents the simplest assembly problem to be solved by the method. The dependent and independent variables are used for statistical hypothesis testing.

This dissertation have investigated all possible combinations for all of the factors and levels of factors used in the THAMMAS algorithm to determine the most significant factors. This corresponds to a total of $3^4 \times 3$ =243 runs required to complete the DOE process. Each THAMMAS run was repeated 3 times with different random seeds for generating the initial structure. Analysis of variance (ANOVA) was performed to investigate which of the THAMMAS variables significantly affect the structure with low penalty which corresponds to a good assembly of TMHs. Optimal parameters were chosen on the basis of the statistical results that gave the minimum penalty value of predicted models.

CHAPTER III

RESULTS AND DISCUSSION

The framework of this thesis is to introduce a novel computational approach for modeling an assembly of transmembrane segments of membrane proteins. The approach is based on nature-based metaheuristics including genetic algorithms and the Max-Min ant system. The proposed approach will be compared with existing methods such as Simulated annealing monte carlo method. The testing models are limited to membrane proteins that contain transmembrane α-helical motifs. This research predicted membrane proteins called KcsA, KvAP and MscL which is in the protein data bank. A scoring function which is in direct change with the RMSD value was used to predict this KcsA. Three algorithms were presented including MMAS, GA and SAMC. It was demonstrated that the pattern of distribution due to the MMAS and GA algorithms have developed a result which led to a crowded result. As opposed to SAMC, it has a random basis. Each of the methods showed a result of 1000 points. These results were the statistical analysis for finding a parameter which was suitable for each algorithm. They use 10000 structures per running time which was the same in the three algorithms. It was presented that SAMC used the least calculating time, after which MMAS and GA were both slower in calculating time. The calculating time of GA, MMAS and MCSA was 1400-1600, 1800 and 1000 seconds respectively. However, for the number of received good answers which were less than 3 angstrom, MMAS gave the best answer followed by GA. SAMC gave the answer which was least in cutoff.

3.1 MMAS parameters

This research used 3^k Factorial designs with 3 levels: high, medium and low which are composed of 4 parameters including calculating cycle per number of ant(A/I),Weight of pheromone (WOP), Weight of heuristic information (WOH) and Evaporation rate (ER) for finding parameters which suitable for this problem. This is explained by specifying the level factor of each parameter, hypothesis testing of

ANOVA, impaction to Main factor and conclusion and comparing the answer. Their detail is as shown below.

Table 3. 1 The level and parameters of MMAS

Factors		Level	
	Low	Medium	High
Iterations/Ants	50/200	100/100	200/50
Weight of pheromone	0.5	1.5	2.5
Weight of Heuristic data	1	2.5	5
Evaporation rate	0.1	0.5	0.9

This dissertation has investigated all possible combinations for all of the factors and levels of factors used in the THAMMAS algorithm to determine the most significant factors. This corresponds to a total of $3^{4}\times 3$ =243 runs required to complete the DOE process. Each THAMMAS run was repeated 3 times with different random seeds for generating the initial structure. Analysis of variance (ANOVA) was performed to investigate which of the THAMMAS variables significantly affect the structure with low penalty which corresponds to a good assembly of TMHs. The results are shown in

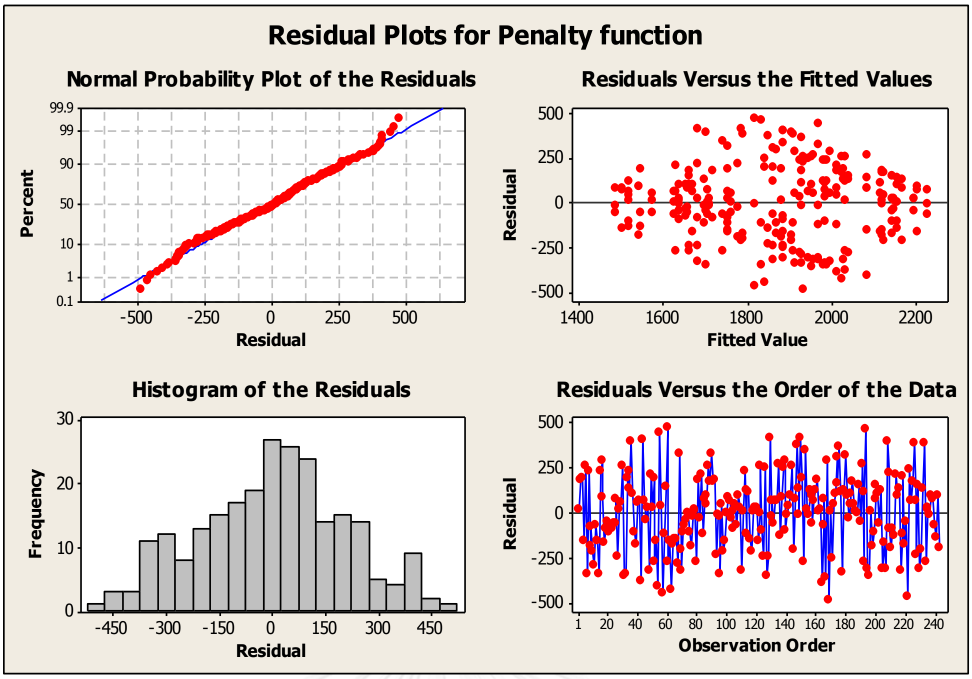

Figure 3. 1 Residual plots of MMAS

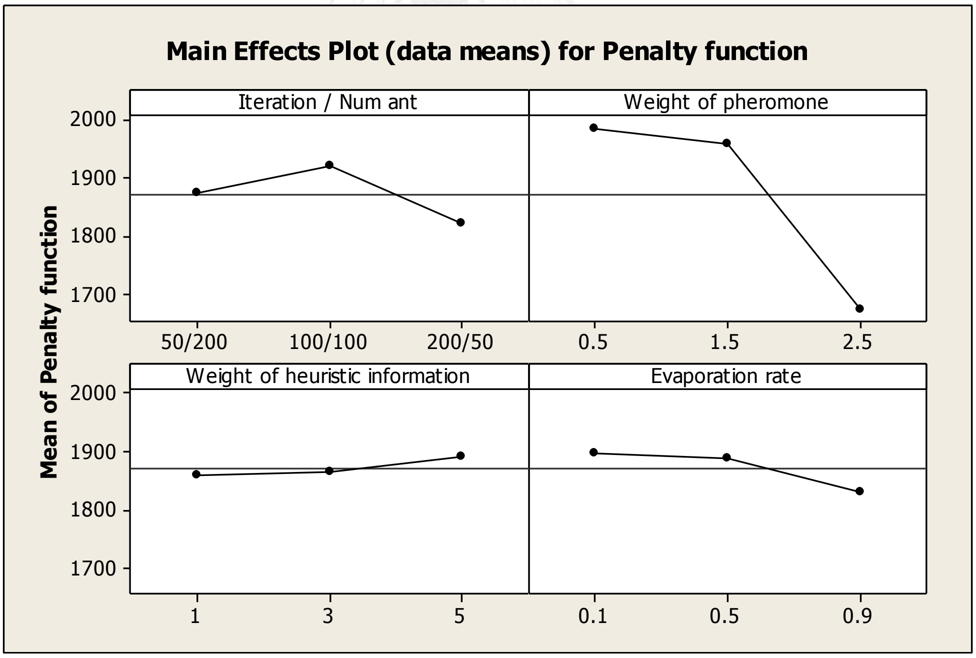

Figure 3. 2 Main effect plots of MMAS

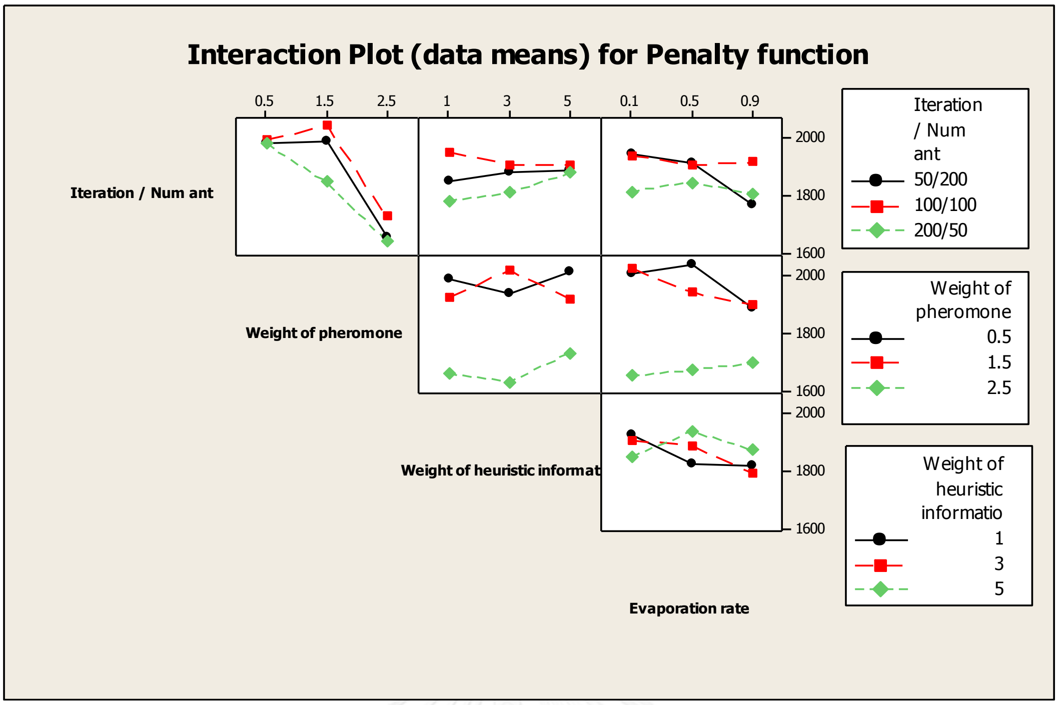

Figure 3. 3 Interaction plots for evaluating impact of MMAS parameters on the penalty function. The interaction plots were used to find the optimal set of parameters for obtaining the minimum penalty. From p-value, α is the most significant parameter, by considering the interaction plots associated with the α parameter.

Table 3. 2 Results of ANOVA for predicting transmembrane arrangement of KcsA inner helices

Factors	df	SS	MS	F	P
IA	2	388159	194080	3.04	0.05
α	2	4755031	2377515	37.27	0.00*
β	2	42215	21107	0.33	0.719
ρ	2	210407	105203	1.65	0.195
IA$\times\alpha$	4	276349	69087	1.08	0.367
IA$\times\beta$	4	151966	37992	0.6	0.666
IA$\times\rho$	4	282004	70501	1.11	0.356
$\alpha\times\beta$	4	348200	87050	1.36	0.248
$\alpha\times\rho$	4	384508	96127	1.51	0.203
$\beta\times\rho$	4	315193	78798	1.24	0.398
IA$\times\alpha\times\beta$	8	391082	48885	0.77	0.633
IA$\times\alpha\times\rho$	8	274585	34323	0.54	0.827
IA$\times\beta\times\rho$	8	241974	30247	0.47	0.873
$\alpha\times\beta\times\rho$	8	521149	65144	1.02	0.422
IA$\times\alpha\times\beta\times\rho$	16	494351	30897	0.48	0.952
Error	162	10333379	63786		
Total	242	19410552			

Df degree of freedom, SS sum of squares, MS mean square

From an analysis of the ANOVA table, it reveals the relative contribution of THAMMAS variables for constructing near native structure models with low penalty values. F-ratio, which is the ratio of MS to the error, was used to evaluate the significance of a variable on the correctness of the predicted model. In general, when F-ratio is much greater than the critical value for the F-distribution, this suggests that the variable is important for giving a good quality of model. Any terms with a P-value greater than 0.05 are not significant and can be discarded. From the ANOVA table, it appears that the weight of pheromone (α) was the most significant variable. The I/A factor is the next influencing factor, but not as strong as the α factor. The β, ρ and interactions of the factors are neglected as being very small values to contribute a significant effect on the quality of the assembly model. Based on the main effects and interaction plots (Figure 3.3), the optimal set of parameters for obtaining the best penalty values corresponds to I/A= 200/50, α = 2.5, β = 3 and ρ = 0.1.

3.2 GA parameters

The parameter optimization for the GA algorithm was also conducted using DOE 3k factorial designs with three levels; high, medium and low. The test protein was the same as DOE did for MMAS. The GA variable parameters including Generation/Population (G/P), Crossover Rate (CR) and Mutation Rate (MR) are summarized in Table 3.3 in the Supplementary Material. From an analysis of the ANOVA table (Table 3.4), the G/P factor is the most significant variable for calculating a good model. The interactions plots (Figure 3.6) indicated that the optimized parameters of GA correspond to 50/200 for G/P, 100% for CR and 10% for MR. As for SAMC, it is not necessary to perform DOE because the nature of the algorithm is based on random. The MC cycle is set to 10000 cycles per run, which is equivalent to the total number of sample size used in GA and MMAS.

Table 3. 3 The level and parameters of GA

Factors		Level	
	Low	Medium	High
Generation/Population(G/P)	50/200	100/100	200/50
Crossover Rate(CR in %)	80	90	100
Mutation Rate (MR in %)	10	20	30

This thesis has investigated all possible combinations for all of the factors and levels of factors used in the GA algorithm to determine the most significant factors. This corresponds to a total of $3^3 \times 3$ = 81 runs required to complete the DOE process. Each GA run was repeated 3 times with different random seeds for generating the initial structure. Analysis of variance (ANOVA) was performed to investigate which of the GA variables significantly affect the structure with low penalty which corresponds to a good assembly of TMHs. The results are shown in Table 3.4. Optimal parameters were chosen on the basis of the statistical results that gave the minimum penalty value of predicted models.

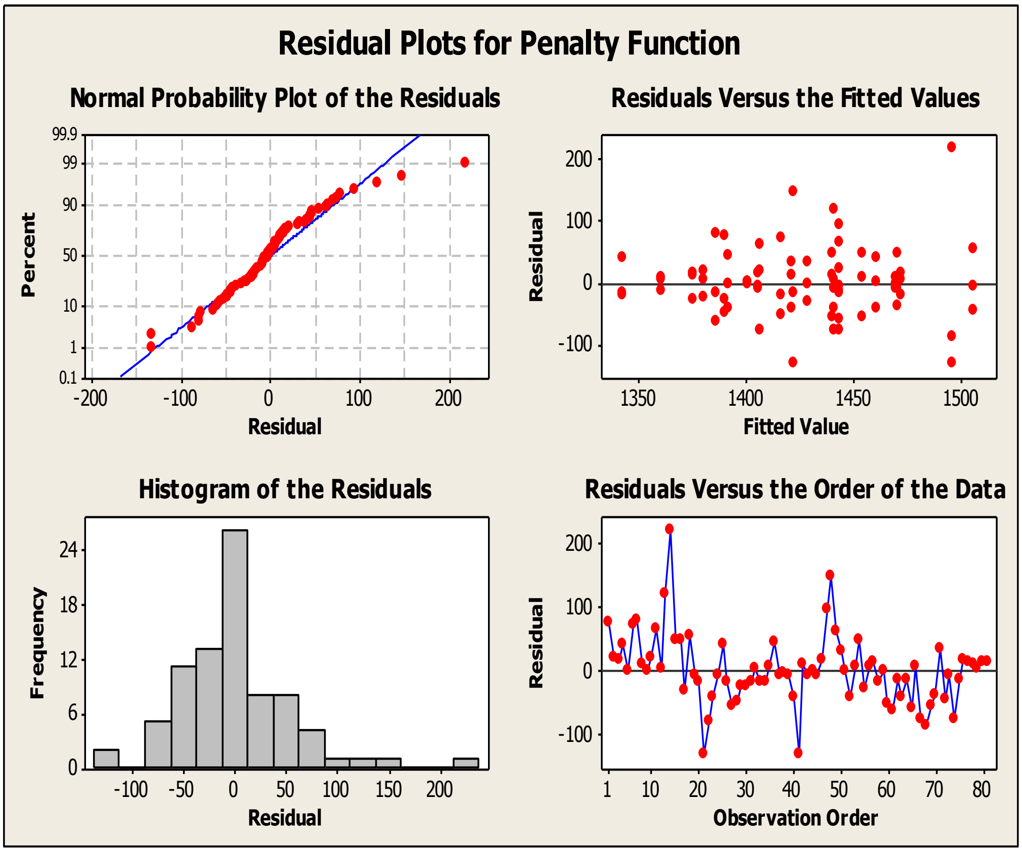

Figure 3. 4 The Residual plots of GA

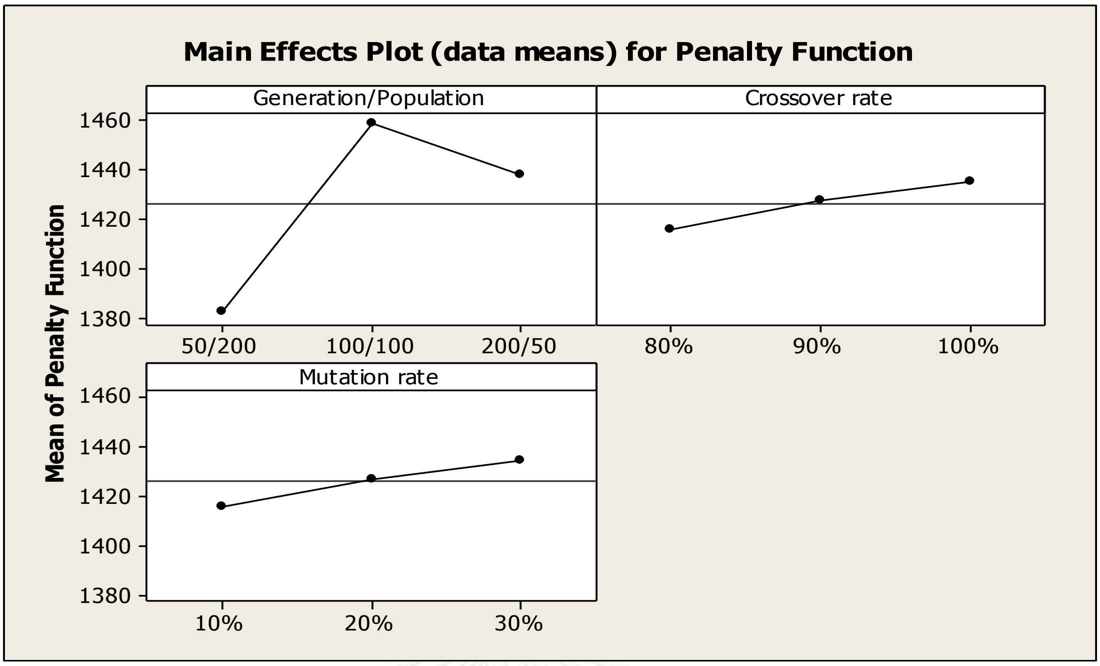

Figure 3. 5 The Main effect plots of GA

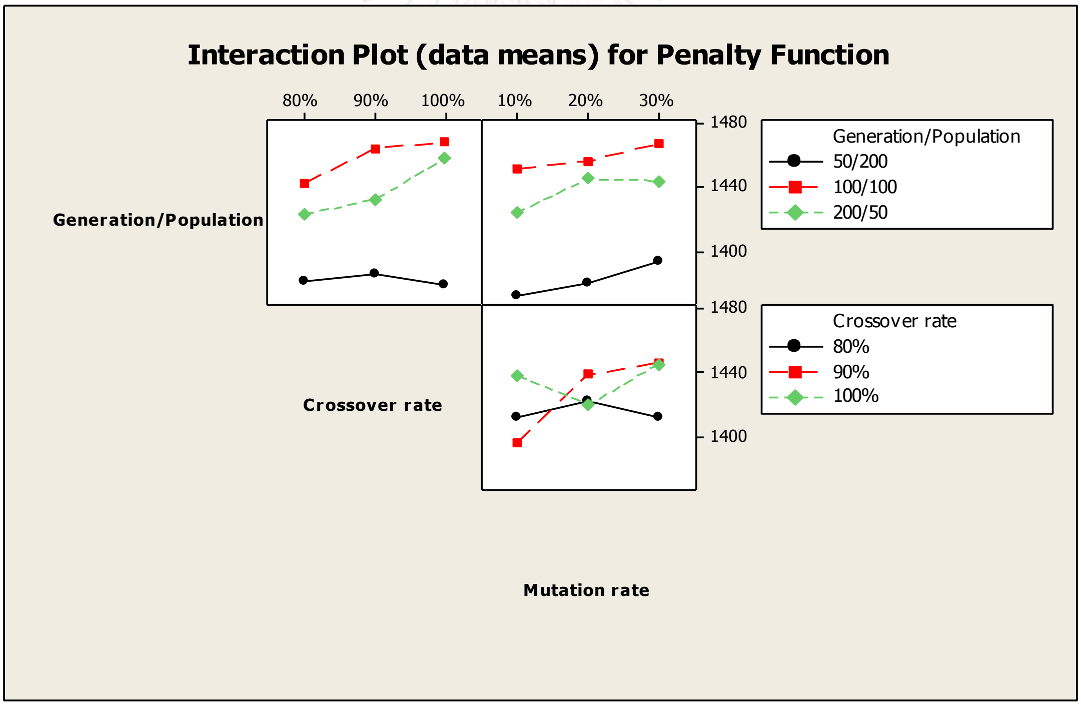

Figure 3. 6 Interaction plots for evaluating impact of GA parameters on the penalty function. The interaction plots were used to find the optimal set of parameters for obtaining the minimum penalty. From p-value, α is the most significant parameter, by considering the interaction plots associated with the G/P parameters.

Table 3. 4 Results of ANOVA for predicting transmembrane arrangement of KcsA inner helices

Factors	df	SS	MS	F	P
G/P	2	83173	41586	9.53	0
CR	2	5016	2508	0.57	0.566
MR	2	4894	2447	0.56	0.574
G/P×CR	4	4318	1080	0.25	0.91
G/P×MR	4	924	231	0.05	0.995
CR×MR	4	12130	3032	0.69	0.599
G/P×CR×MR	8	17527	2191	0.5	0.849
Error	54	235670	4364		
Total	80				

3.3 Transmembrane Assembly Scenarios

To demonstrate the ability of the proposed algorithm, the prediction test of transmembrane assembly of the three protein models has been divided into four scenarios; 1) TM assembly of KcsA inner TM helices (denoted as KcsA/TM2), 2) TM assembly of KcsA inner and outer TM helices (denoted as KcsA/TM1-TM2), 3) TM assembly of the two TMHs of MscL (denoted as MscL/TM1-TM2) and 4) TM assembly of the isolated voltage sensor domain of KvAP (denoted as KvAP/VSD). The inter- or intra-subunit Cα-Cα distance constraints were extracted from the crystal structures of the proteins. Except for KvAP, the experimental distance constraints taken from the literature were used to fit with the distances between EPX-pseudo atoms of the previously built model (34).

Table 3. 5 Total number of intra- and inter-subunit distance constraints used

Proteins	Type	Total number of constraints
KcsA/TM2	Inter	41
KcsA/TM1-TM2	Intra	11
MscL/TM1-TM2	Intra	3
	Inter	12
KvAP/VSD	Intra	10*

Note:* Experimental distances were taken from[41].

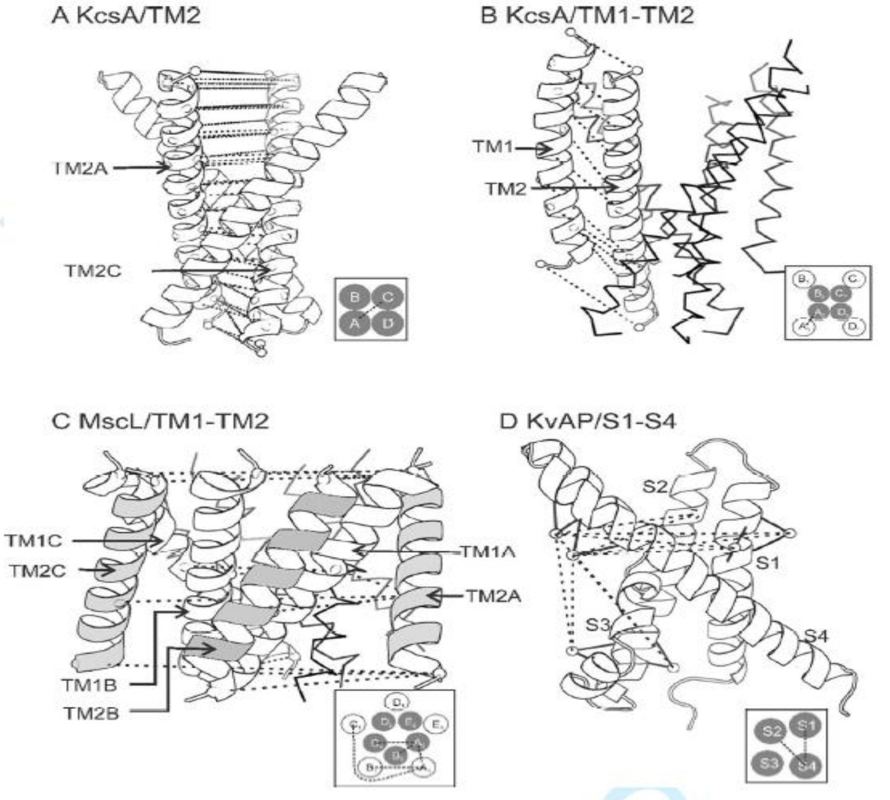

Figure 3. 7 Transmembrane assemblies of test proteins of known 3D structure.

With the KvAP/VSD scenario, the available values to search for translation and rotation of TMHs have been reassigned. Due to a lack of experimental distance data available that can only be used to constraint between S1and S4, and S2 and S4, a limit of the search range for the θ_1, θ_2 and θ_3 values was specified within 20° for S1 and S2, and 40°for S4, while S3 was kept fixed in its original position. Only for θ_4 (rotation around its helix axis), the available values remain the same (360rotation).

Table 3. 6 Inter-subunit Cα-Cα distances of the selected residues on the TM2 segment of KcsA. Distances were derived from the crystal structure with PDB code 1k4c.

Residue &chainID	Distance (Å)	Residue &chain ID	Distance (Å)	Residue &chain ID	Distance (Å)	Residue &chain ID	Distance (Å)
84A-84C	29.2	94A-94C	31.2	104A-104C	11.0	114A-114C	16.6
85A-85C	36.5	95A-95C	27.1	105A-105C	16.3	115A-115C	10.9
86A-86C	39.9	96A-96C	20.4	106A-106C	17.4	116A-116C	15.2
87A-87C	41.9	97A-97C	22.9	107A-107C	10.6	117A-117C	21.0
88A-88C	35.7	98A-98C	25.6	108A-108C	9.8	118A-118C	18.7
89A-89C	31.7	99A-99C	20.0	109A-109C	16.2	119A-119C	14.5
90A-90C	36.2	100A-100C	14.4	110A-110C	15.9	120A-120C	21.0
91A-91C	35.3	101A-101C	18.7	111A-111C	9.6	121A-121C	25.4
92A-92C	27.7	102A-102C	20.9	112A-112C	11.5	122A-122C	21.4
93A-93C	27.7	103A-103C	15.3	113A-113C	18.0	123A-123C 124A-124C	21.0 14.2

Table 3. 7 Intra-subunit Cα-Cα distances of the selected residues on the TM1 and TM2 segments of KcsA. Distances were derived from the crystal structure with PDB code 1k4c.

Residue ID	Residue ID	distance(Å)	TM segments to be constrained
22	124	20.6	TM1-TM2
25	121	20.2	TM1-TM2
28	118	21.5	TM1-TM2
31	115	21.5	TM1-TM2
34	112	20.5	TM1-TM2
37	105	14.8	TM1-TM2
40	102	12.0	TM1-TM2
43	99	11.6	TM1-TM2
46	96	14.3	TM1-TM2
49	93	15.7	TM1-TM2
52	86	12.2	TM1-TM2

Table 3. 8 Ten experimentally-determined DEER distances between bi-functional spin labels on the S1, S2, S3 and S4 of KvAP voltage sensor domain. Experimental distances were taken from Q. Li et al.[44]

Double cysteine		Distance*(Å)	TM segments to be constrained
39/43	118/121	21.7	S1-S4
	121/125	19.8	S1-S4
40/44	118/121	29.9	S1-S4
	121/125	29.9	S1-S4
57/61	118/121	26.7	S2-S4
	121/125	30.8	S2-S4
72/75	118/121	29.7	S2-S4
	121/125	29.7	S2-S4
74/77	118/121	21.8	S2-S4
	121/125	22.8	S2-S4

Table 3. 9 Inter- and Intra-subunit Cα-Cα distances of the selected residues on the TM1 and TM2 segments of MscL. Distances were derived from the crystal structure with PDB code (2oar).

a) Inter-subunit distances

Residue ID	Chain	Residue ID	Chain	Distance (Å)	Residue ID	Chain	Residue ID	Chain	Distance (Å)
14	A	14	B	9.7	71	A	71	B	18.6
14	A	14	C	15.9	71	A	71	C	30.0
28	A	28	B	12.1	83	A	83	B	20.4
28	A	28	C	19.0	83	A	83	C	33.6
40	A	40	B	20.3	92	A	92	C	42.3
40	A	40	C	32.8	92	A	92	B	25.4

b) Intra-subunit distances

Residue ID	Chain	Residue ID	Chain	distance(Å)	TM segments to be constrained
14	A	92	A	26.5	TM1-TM2
28	A	83	A	10.0	TM1-TM2
40	A	71	A	7.4	TM1-TM2

The plots of the penalty value vs. RMSD to native for a representative case of 10,000 experimental runs for each scenario by each method are shown in Figure 3.8. The general trend is that helical assemblies closer in RMSD to the native structure have lower penalty scores than those farther from the packing of the native structure. 10000 predicted models of the four assembly scenarios exhibit RMSD to native with the range between 2.0 Å and 5.0 Å for THAMMAS, revealing a quite acceptable performance in the prediction. The scattered plot produced by the GA algorithm shows a broader distribution graph with the RMSD ranging from 2.0 Å to 7Å. The experimental runs conducted by the SAMC algorithm exhibits a quite broad RMSD distribution ranging from 2 Å to more than 8 Å in the four assembly scenarios. One can observe that the scattered plots generated by the GA and SAMC algorithms have a lower density near native structure compared to THAMMAS. In addition, while THAMMAS has shown to be the superior performance for predicting the assembly of the transmembrane helix bundle, GA performs better than the SAMC algorithm. According to the results, THAMMAS shows noticeable consistency by finding native-like or near native assembly with the narrowest range of RMSD to the native conformation compared to the GA and SAMC algorithms.

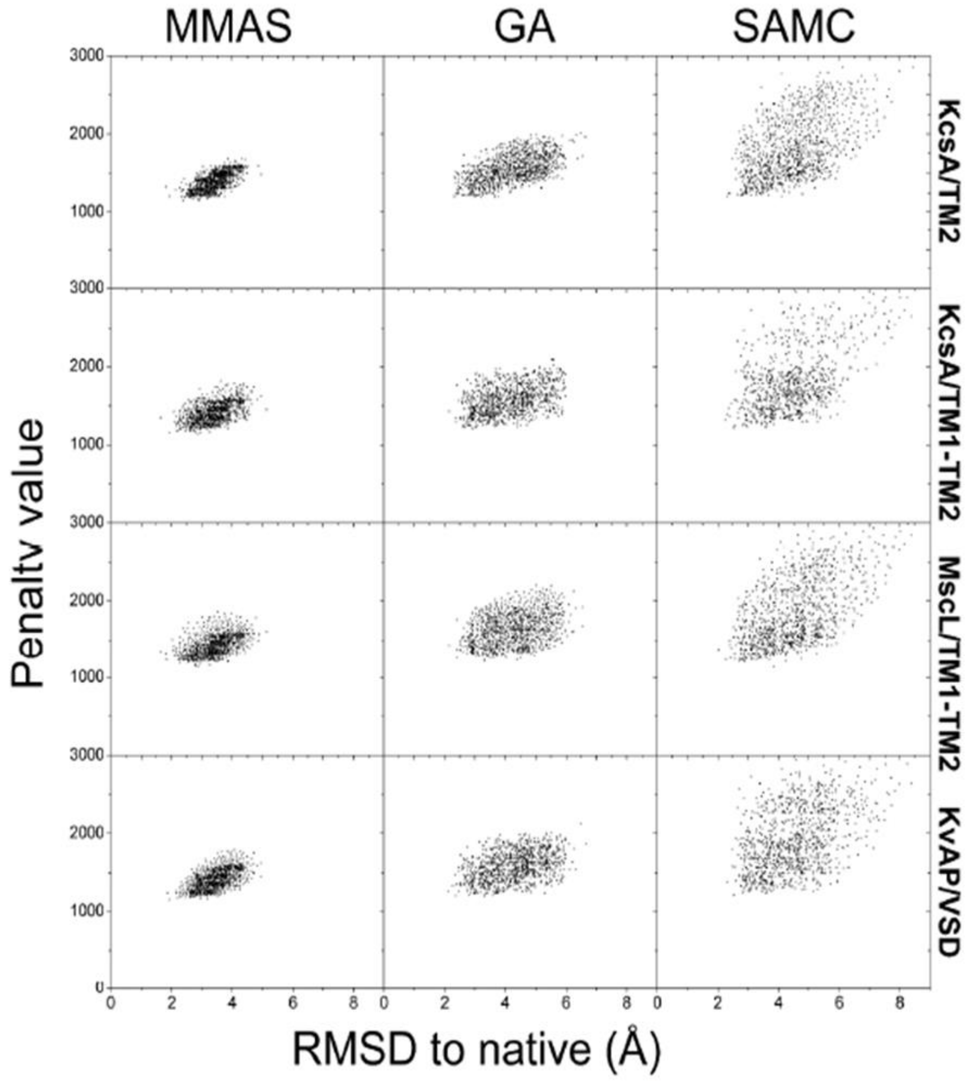

Figure 3. 8 Scattered plots comparing the RMSD distribution as a function of penalty of 10000 experimental.

Figure 3.9 presents the average percentage of predicted structures clustered by each specified RMSD range by each method. If the RMSD cutoff of 3Å is chosen for discrimination between native-like and non-native assembly, the robustness and superiority of THAMMAS has achieved with a rate of correct prediction greater than GA and SAMC. At a RMSD range of 3-4Å, the THAMMAS predictions also give the majority containing significant transmembrane assemblies that were qualitatively correct. These outstanding results suggested that THAMMAS has a greater efficiency over a well-known Monte Carlo method and genetic algorithm.

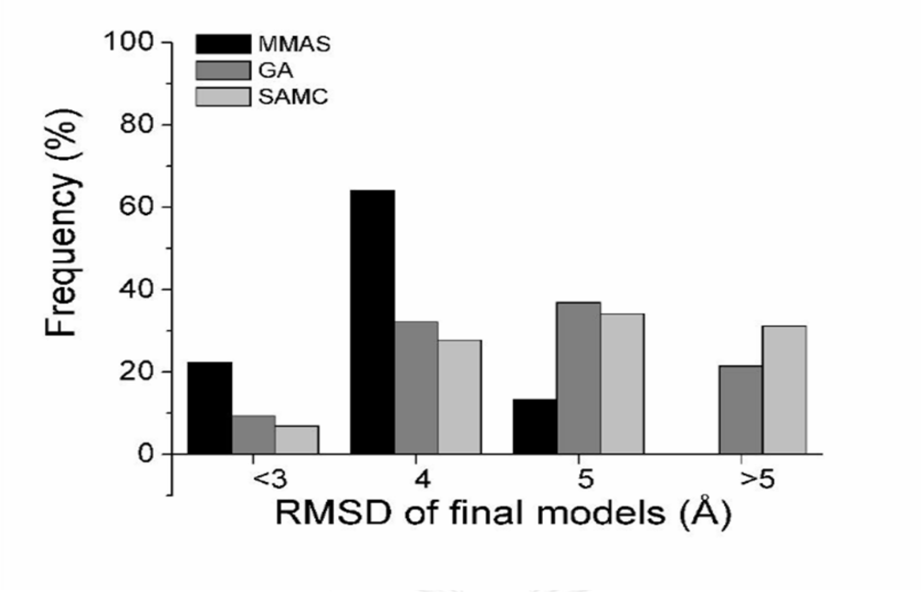

Figure 3. 9 RMSD distributions clustered by average percentage of the frequency for given RMSD range

3.4 Optimization performance

In order to better understand the performance of the proposed algorithm, the history of the optimization profile of the three algorithms has been analyzed. For this evaluation, this work randomly chose five representative models for each assembly scenario. This gives rise to a total of 20 experimental runs per algorithm. The stopping criteria of 50 iterations, (generations or MC cycles) was employed. Note that the three algorithms designed for minimization of the penalty function update iteratively during the minimization process, and then store the resulting structures, which in turn are subjected to calculate RMSD values. Figures 3.10 show plots of RMSD as a function of iteration steps for 20 experimental runs of each method. It appears from the history of the optimization profiles that in most cases THAMMAS has the ability to quickly find better structures towards the native structure with faster RMSD convergence compared to GA and SAMC. This implies that THAMMAS has strong capability in terms of speed of structure convergence. According to the computational results, SAMC has a limitation in finding native-like or near native assembly under the presented protocols. It should, however, be noted that the GA algorithm also has the capability of finding the solutions with an acceptable convergence rate. The Boltzmann fitted curves shown in Figure 3.10 (D) compare the average performance of the THAMMAS, GA and SAMC. It can be clearly seen that the convergence rate of THAMMAS is considerably better than the SAMC, but not much significantly different from GA for a long iteration. As observed in this figure, THAMMAS takes the least iterations for convergence compared with the other two algorithms. The presented results confirm the efficiency of THAMMAS in the optimization problem of transmembrane helix assembly.

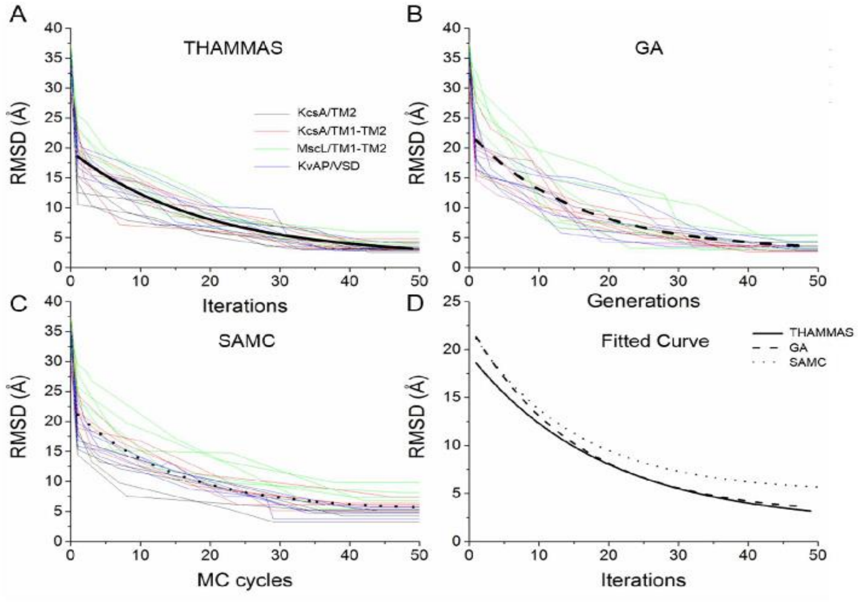

Figure 3. 10 Optimization profile comparing the performance of (A) MMAS (B) GA and (C) SAMC algorithms. RMSD profiles of 20 experimental runs are shown as light lines. Heavy lines in (A)-(C) and (D) are the curves fitted with Boltzmann’s function.

CHAPTER IV

CONCLUSION

Metaheuristic methods have become popular tools in solving large scale optimization problems for a variety of systems. The goal of this thesis is to introduce a metaheuristic approach to predict transmembrane protein structure from an assembly of membrane spanning α-helices. The MMAS method is considered as a novel computational approach for its application to membrane protein structure prediction. In this paper, The author propose Max-Min Ant System algorithm, a new and efficient approach for solving the optimization problem of transmembrane helix assembly to satisfy the distance based penalty function. The MMAS algorithm, based on an ant colony optimization method, is capable of predicting the correct assembly of transmembrane proteins with a considerable rate of success. Benchmark studies show the performance and effectiveness of the proposed approach, compared with genetic algorithm and simulated annealing Monte Carlo methods. In terms of structure convergence, the proposed algorithm can outperform the well-known algorithms by comparing the rapid search for a good solution from iteration to iteration. This thesis anticipate that the MMAS algorithm gives a promising alternative that is useful in the structural bioinformatics and computational biophysical.

REFERENCES

[1] Stevens, T.J. and Arkin, I.T. Do more complex organisms have a greater proportion of membrane proteins in their genomes? Proteins 39(4) (2000): 417-20.

[2] Bennett, D.L., and C. G. Woods. Painful and painless channelopathies. Lancet neurology 13 (2014): 587-599.

[3] Yildirim, M.A., Goh, K.-I., Cusick, M.E., Barabasi, A.-L., and Vidal, M. Drug[mdash]target network. Nat Biotech 25(10) (2007): 1119-1126.

[4] Giguere, P.M., Kroeze, W.K., and Roth, B.L. Tuning up the right signal: chemical and genetic approaches to study GPCR functions. Curr Opin Cell Biol 27 (2014): 51-5.

[5] Lacapere, J.J., Pebay-Peyroula, E., Neumann, J.M., and Etchebest, C. Determining membrane protein structures: still a challenge! Trends Biochem Sci 32(6) (2007): 259-70.

[6] Taraska, J.W. Mapping membrane protein structure with fluorescence. Current Opinion in Structural Biology 22(4) (2012): 507-513.

[7] McHaourab, H.S., Steed, P.R., and Kazmier, K. Toward the fourth dimension of membrane protein structure: insight into dynamics from spin-labeling EPR spectroscopy. Structure 19(11) (2011): 1549-61.

[8] Holland, J.H. Adaptation in Natural and Artificial Systems. University of Michigan Press, 1975.

[9] Neumaier, A. Molecular Modeling of Proteins and Mathematical Prediction of Protein Structure. SIAM Review 39(3) (1997): 407-460.

[10] Talbi, E.G. A Taxonomy of Hybrid Metaheuristics. Journal of Heuristics 8(5) (2002): 541-564.

[11] Gendreau, M., and J.-Y. Potvin. Handbook of metaheuristics. New York: Springer, 2012.

[12] Alba, E. Optimization techniques for solving complex problems. N.J.: Wiley, 2009.

[13] Cotta, C., Fernández, A. J., Gallardo, J. E., Luque, G. and Alba, E. Metaheuristics in Bioinformatics: DNA Sequencing and Reconstruction, in Optimization techniques for solving complex problems. Hoboken ed. N.J.: Wiley, 2009.

[14] Bonabeau, E., M. Dorigo, and G. Theraulaz. Swarm Intelligence. Oxford University Press, 1999.

[15] Dorigo, M., Birattari, M., and Stu?tzle, T. Ant colony optimization artificial ants as a computational intelligence technique. IEEE Comput. Intell. Mag. 1(4) (2006): 28-39.

[16] Korb, O., Stützle, T., and Exner, T. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. in Dorigo, M., Gambardella, L., Birattari, M., Martinoli, A., Poli, R., and Stützle, T. (eds.),Ant Colony Optimization and Swarm Intelligence, pp. 247-258: Springer Berlin Heidelberg, 2006.

[17] Dorigo, M. and Blum, C. Ant colony optimization theory: A survey. Theoretical Computer Science 344(2–3) (2005): 243-278.

[18] Dorigo, M. and Stützle, T. Ant Colony Optimization. MIT Press, Bradford Books, 2004.

[19] St, T., #252, tzle, and Hoos, H.H. <italic>MAX-MIN</italic> Ant system. Future Gener. Comput. Syst. 16(9) (2000): 889-914.

[20] Shi, W.M., Shen, Q., Kong, W., and Ye, B.X. QSAR analysis of tyrosine kinase inhibitor using modified ant colony optimization and multiple linear regression. Eur J Med Chem 42(1) (2007): 81-6.

[21] Dorigo, M. and Gambardella, L.M. Ant colony system: a cooperative learning approach to the traveling salesman problem. Trans. Evol. Comp 1(1) (1997): 53-66.

[22] Dorigo, M., Di Caro, G., and Gambardella, L.M. Ant algorithms for discrete optimization. Artif Life 5(2) (1999): 137-72.

[23] Vanderbilt, D. and Louie, S.G. A Monte carlo simulated annealing approach to optimization over continuous variables. Journal of Computational Physics 56(2) (1984): 259-271.

[24] Box, G.E.P., J. S. Hunter, and W. G. Hunter. Statistics for experimenters : design, innovation, and discovery. Hoboken ed. N.J.: Wiley-Interscience, 2005.

[25] Montgomery, D.C. Design and analysis of experiments. Hoboken ed. N.J.: ohn Wiley & Sons, 2005.

[26] Swanson, R.N. and Cramer, H.E. A Study of Lateral and Longitudinal Intensities of Turbulence. Journal of Applied Meteorology 4(3) (1965): 409-417.

[27] Schiemann, O. and Prisner, T.F. Long-range distance determinations in biomacromolecules by EPR spectroscopy. Q Rev Biophys 40(1) (2007): 1-53.

[28] Altenbach, C., Flitsch, S.L., Khorana, H.G., and Hubbell, W.L. Structural studies on transmembrane proteins. 2. Spin labeling of bacteriorhodopsin mutants at unique cysteines. Biochemistry 28(19) (1989): 7806-7812.

[29] Grohmann, D., Klose, D., Klare, J.P., Kay, C.W., Steinhoff, H.J., and Werner, F. RNA-binding to archaeal RNA polymerase subunits F/E: a DEER and FRET study. J Am Chem Soc 132(17) (2010): 5954-5.

[30] Leitner, A., et al. Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. Mol Cell Proteomics 9(8) (2010): 1634-49.

[31] Berg, J.M., J. L. Tymoczko, and L. Stryer. Biochemistry. New York: W.H. Freeman, 2012.

[32] Bowie, J.U. Helix packing in membrane proteins. J Mol Biol 272(5) (1997): 780-9.

[33] Bowie, J.U. Helix packing angle preferences. Nat Struct Biol 4(11) (1997): 915-7.

[34] Gerken, U., et al. Membrane Environment Reduces the Accessible Conformational Space Available to an Integral Membrane Protein. The Journal of Physical Chemistry B 107(1) (2003): 338-343.

[35] Jiang, Y., et al. X-ray structure of a voltage-dependent K+ channel. Nature 423(6935) (2003): 33-41.

[36] Sompornpisut, P., Liu, Y.S., and Perozo, E. Calculation of rigid-body conformational changes using restraint-driven Cartesian transformations. Biophys J 81(5) (2001): 2530-46.

[37] Chen, K.Y., Sun, J., Salvo, J.S., Baker, D., and Barth, P. High-resolution modeling of transmembrane helical protein structures from distant homologues. PLoS Comput Biol 10(5) (2014): e1003636.

[38] Liu, Y.S., Sompornpisut, P., and Perozo, E. Structure of the KcsA channel intracellular gate in the open state. Nat Struct Biol 8(10) (2001): 883-7.

[39] Zhou, Y., Morais-Cabral, J.H., Kaufman, A., and MacKinnon, R. Chemistry of ion coordination and hydration revealed by a K+ channel-Fab complex at 2.0 A resolution. Nature 414(6859) (2001): 43-8.

[40] Jiang, Y.X., et al. X-ray structure of a voltage-dependent K+ channel. Nature 423(6935) (2003): 33-41.

[41] Li, Q.F., Wanderling, S., Sompornpisut, P., and Perozo, E. Structural basis of lipid-driven conformational transitions in the KvAP voltage-sensing domain. Nature Structural & Molecular Biology 21(2) (2014): 160-+.

[42] Chen, K.Y.M., Sun, J.M., Salvo, J.S., Baker, D., and Barth, P. High-Resolution Modeling of Transmembrane Helical Protein Structures from Distant Homologues. Plos Computational Biology 10(5) (2014).

[43] Yamane, T. Statistics, an introductory analysis. New York: Harper, 1973.

[44] Li, Q.F., Wanderling, S., Sompornpisut, P., and Perozo, E. Structural basis of lipid-driven conformational transitions in the KvAP voltage-sensing domain. Nature Structural & Molecular Biology 21(2) (2014): 160-6.

APPENDIX

VITA

Mr. Kanon Sujaree was born on December 23, 1981 in Chainat Thailand. He received a Bachelor Degree of Engineering (B.ENG Industrial Engineering) at Naresuan University in 2004 and Master Degree of Engineering (M.Eng Engineering Management) at Naresuan University in 2009.During his studies towards the PhD. degree, He was awarded the 90 th Anniversary of Chulalongkorn University Fund. His present address is 55/33 Moo 14, Klongneung, Klongleung, Pathumthanee 12120