**MULTI-DIMENSIONAL SCALING AND MODELLER BASED EVOLUATIONARY ALGORITHMS FOR PROTEIN MODEL REFINEMENT**

A Thesis

Presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Yan Chen

Dr. Yi Shang, Thesis Supervisor

December 2013

The undersigned, appointed by the dean of the Graduate School, have examined the thesis entitled

MULTI-DIMENSIONAL SCALING AND MODELLER BASED EVOLUATIONARY ALGORITHMS FOR PROTEIN MODEL REFINEMENT

Presented by Yan Chen,

A candidate for the degree of Master of Computer Science,

And hereby certify that, in their opinion, it is worthy of acceptance.

Professor Yi Shang

Professor Dong Xu

Professor Ioan Kosztin

### Acknowledgement

Firstly, I would like to thank my Holy Father who loves me and gives me strength to overcome every difficulty.

I am sincerely presenting deep appreciation to my advisor, Dr. Yi Shang, who continuously guides me in every stage of my M.S. study with great patience. His spirit of hard working and strictness in academy impresses me much and this will be a great treasure in all my life.

I would also like to thank the committee members for their guidance and participation during my M.S. study. Great gratitude is also given to my parents and my husband who love and encourage me to overcome difficulties in study and life. Many thanks are also given to my officemates, Son P. Nguyen, Jingfen Zhang, Zhiquan He, Hongbo Li, Jiong Zhang, Junlin Wang and Wei Xiong for their help and support in the completion of this thesis.

**Table of contents**

Acknowledgement II

List of figures V

List of tables VII

Abstract VIII

Chapter 1: INTRODUCTION 1

Chapter 2: RELATED WORK 5

2.1 Evolutionary Algorithms 5

2.2 Protein Quality Evaluation 6

2.2.1 Similarity Consensus 7

2.2.2 Energy or scoring functions 9

2.3 MDS 11

2.4 MODELLER 12

Chapter 3: THREE NEW EVOLUTIONARY ALGORITHMS FOR PROTEIN MODEL REFINEMNT 13

3.1 Initial Population 15

3.2 Evaluation 16

3.3 Selection 18

3.4 Crossover 20

3.4.1 MDS-based method 20

3.4.2 MODELLER-based method 23

3.4.3 Hybrid method 23

Chapter 4: EXPERIMENTAL Results 25

4.1 Data Set 25

4.2 Computation Time 27

4.3 Results 30

4.3.1 Refinement results using ProQ2 as quality assessment method 30

4.3.2 Refinement results using CGDT-TS as quality assessment method 35

Chapter 5: CONCLUSIONS AND FUTURE WORK 41

REFERENCES 43

### List of figures

[Figure 2. 1 Implementation of evolutionary algorithm 6](#_Toc374529224)

Figure 3. 1 The evolutionary algorithm of the whole process 14

Figure 3. 2 ProQ2 and GDT-TS of initial population for T0654 17

Figure 3. 3 CGDT-TS and GDT-TS of initial population for T0654 17

Figure 3. 4 Selection probability of ProQ2 for T0654 19

Figure 3. 5 Selection probability of CGDT-TS for T0654 19

Figure 3. 6 The WMDS-based crossover (WMDS-C) algorithm 20

Figure 3. 7 The MODELLER-based crossover (MODELLER-C) algorithm 23

Figure 3. 8 The Hybrid crossover (Hybrid-C) algorithm 24

[Figure 4. 1 The all computation time for different algorithms 28](#_Toc374529235)

[Figure 4. 2 The average iteration time in each step for different algorithms 29](#_Toc374529236)

[Figure 4. 3 The computation time for steps in WMDS-based EA of each iteration 30](#_Toc374529237)

[Figure 4. 4 The GDT-TS of best model using ProQ2 evaluation 33](#_Toc374529238)

[Figure 4. 5 The RMSD of best model using ProQ2 evaluation 33](#_Toc374529239)

[Figure 4. 6 The average GDT-TS of best 10 models using ProQ2 evaluation 34](#_Toc374529240)

[Figure 4. 7 The average GDT-TS of all models using ProQ2 evaluation 35](#_Toc374529241)

[Figure 4. 8 The GDT-TS of best model using CGDT-TS evaluation 37](#_Toc374529242)

[Figure 4. 9 The RMSD of best model using CGDT-TS evaluation 38](#_Toc374529243)

[Figure 4. 10 The average GDT-TS of best 10 models using CGDT-TS evaluation 38](#_Toc374529244)

[Figure 4. 11 The average GDT-TS of all models using CGDT-TS evaluation 39](#_Toc374529245)

[Figure 4. 12 The average improved GDT-TS for different algorithms 40](#_Toc374529246)

### List of tables

[Table 4. 1 Summary of CASP10 refinement targets 26](#_Toc374529252)

[Table 4. 2 Summary of refinement results on CASP10 targets using ProQ2 32](#_Toc374529253)

[Table 4. 3 Summary of refinement results on CASP10 targets using CGDT-TS 36](#_Toc374529254)

### Abstract

To computationally obtain an accurate prediction of the three-dimensional structure of a protein from its primary sequence is one of the most important problems in bioinformatics and has been actively researched for many years. Although a number of software packages have been developed and they sometimes perform well on template-based modeling, further improvement is needed for practical use. Model refinement is a step in the prediction process, in which improved structures are constructed based on a pool of initially generated models. Since the refinement category being added to the Critical Assessment of Structure Prediction (CASP) competition in 2008, CASP results show that it is a challenge for existing model refinement methods to improve model quality consistently.

This project focuses on evolutionary algorithms for protein model refinement. Three new algorithms have been developed, in which multidimensional scaling (MDS), MODELLER, and a hybrid of both are used as crossover operators, respectively. The MDS-based method takes a purely geometrical approach and generates a child model by combining the contact maps of multiple parents. The MODELLER-based method takes a statistical and energy minimization approach and uses the remodeling module in MODELLER program to generate new models from multiple parents. The hybrid method first generates models using the MDS-based method and then run them through the MODELLER-based method, aiming at combining the strength of both. Promising results have been obtained in experiments using CASP datasets. The MDS-based method improved the best of a pool of predicted models in terms of the global distance test score (GDT-TS) in 9 out of 16 test targets. For instance, for target T0680, the GDT-TS of a refined model is 0.833, much better than 0.763, the value of the best model in the initial pool.

**Keywords**: MDS, MODELLER, Evolutionary Algorithm

### Chapter 1: INTRODUCTION

Proteins are essential biochemical compounds that contribute to many processes in life. Functional properties of cells are dependent on the correct protein structure [1]. Misfolded proteins may lead to some diseases, such as Alzheimer’s, Parkinson’s, Type II Diabetes, even certain cancers [2] [3]. The knowledge of the protein tertiary structure can help in basic research on protein functions as well as in drug development. Both experimental methods and computational methods can be used in protein structure acquirement. X-ray crystallography is the most wildly used method to finding the protein structures in experimental methods. Quaternary structure of large proteins can be determined by electron microscopes (cryoEM) or nuclear magnetic resonance (NMR). However, it is costly, slow and difficult to find the protein tertiary structures through experimental technologies [4]. So better computational techniques are quit desirable for predicting protein structure from primary sequences information. Therefore obtaining an accurate prediction of the three-dimensional structure of a protein by automatic prediction is one of the most important problems in bioinformatics and has been actively researched for many years.

The CASP is a biannual world wild contest in the structure prediction community to assess the current protein modeling techniques and identify their quality. There are different groups trying to predict structure of a protein whose structure is unreleased to the outside world or refine an existing model close to native structure. Evaluating the models’ quality and refining the existing models are the two main problems of predicting the three dimensional protein models.

Model refinement is a step in the prediction process, in which improved structures are constructed based on a pool of initially generated models. The refinement category has been added to the CASP competition since 2008 to evaluate the current technology for further improvement to the best predicted models. Participants can focus on the best submitted protein models, sometimes the certain part of a protein needs to refinement or the accuracy of the protein models are provided. The aim of protein refinement is to move the backbone or the side chain conformations closer to native. Several methods were explored and software packages were developed to improve the quality of protein structure models. i3Drefine software is the only fully-automated server which can improve the global and local structure quality in CASP10. This new technology minimizes the energy iteratively with physics and knowledge-based force fields, and hydrogen bonding (HB) network optimization technique [5]. GalaxyRefine has the best performance in CASP10 by improving the local structure quality. It rebuilt the side chains, repacked them, and relaxed the structure by molecular dynamics simulation [6]. KoBaMIN is another method based on minimization of a knowledge-based potential of mean force [7]. I-TASSER is an automated pipeline for predicting protein 3D structure by multiple threading alignments and iterative structure assembly simulations [8]. Methods mentioned in references [9] and [10] also can improve the initial model structure.

Although a number of software packages have been developed and they sometimes perform well on template-based modeling, further improvement is needed for practical use. During CASP8 and CASP9 [11] [12], only a few groups were able to improve the protein model quality consistently. In CASP10, seven groups were better than naïve (null) method, only two or three groups were significantly better. Most groups are still unable to improve the quality of the starting models; even high-performing groups were only able to make modest improvements; and no predicted models were more similar to the native structure than to the starting model. The maximum improvement in high accuracy of GDT-TS (GDT-HA) is only about 0.1 [13].

In this project, three new evolutionary algorithms which are MDS-based, MODELLER-based and a hybrid of MDS-based and MODELLER-based crossover have been developed. The MDS-based method is a purely geometrical approach that combines the contact maps of multiple parents to breed a new offspring. The MODELLER-based method takes a statistical and energy minimization approach and uses the remodeling module in MODELLER program to generate new models from multiple parents. The hybrid method first generates models using the MDS-based method and then runs them through the MODELLER-based method, aiming at combining the strength of both.

This thesis is started with some related work for this project of protein tertiary structure refinement in Chapter 2. Chapter 3 presents the three evolutionary algorithms to refine protein models and Chapter 4 describes the refinement results. Some conclusions and future work are shown in Chapter 5.

### Chapter 2: RELATED WORK

### 2.1 Evolutionary Algorithms

Evolutionary Algorithms simulate the natural evolution progress [14]. This computational method is based on evolutionary and biological principles to achieve optimal results. The size of the population is invariable by filtering the reproduction and selecting the better survivals. This evolutionary algorithm uses a fitness function that represents the quality of the population as a number. Several evolutionary algorithms are used for solving the protein model predication problems [15] [16] [17] [18]. Figure 2.1 shows the general process of an evolutionary algorithm.

Step 1. Initial [Population](http://en.wikipedia.org/wiki/Population)

Generate the initial population randomly

Step 2. Fitness Function Evaluation

Evaluate the fitness of each individual in the pool

Step 3. Repeat Generation until Termination

Step 3.1. Selection

Select the best-fit individuals as parents

Step 3.2. Reproduction

Breed new generation by crossover or mutation

Step 3.3. Evaluation

Evaluate the individual’s fitness of new generation

Step 3.4. Replacement

Knock out least-fit individuals

**Figure 2. 1 Implementation of evolutionary algorithm**

### 2.2 Protein Quality Evaluation

Picking the most suitable candidates from the template is essential to model refinement. The basic hypothesis of protein models is that the native structure has the minimum free energy in general [1]. Using energy or scoring functions is a most adoptable approach for evaluating the models quality. An accurate energy function could help to find the best prediction protein models. There are two general categories of energy functions: one is physics-based potential and the other is knowledge-based potential [19]. Physics-based potential function uses physics law to evaluate the models’ quality. The knowledge-based statistical method is based on the thermodynamic equilibrium and known information of protein models. The second approach is clustering based. This theory assumes that good models have more structural neighbors. The third approach is using consensus information. The major selection is based on pairwise similarity between structures [20].

### 2.2.1 Similarity Consensus

The pairwise similarity could be retrieved from the root-mean-squared deviation (RMSD), the template Modeling Score (TM-score) and the total score of global distance test (GDT-TS) between each protein model pairs etc.[2].

RMSD is the measure of average distance between the C-α atoms of two protein models. RMSD is commonly used in the study of protein conformations; it measures the similarity in three-dimensional structure of the C-α atomic coordinates. Low RMSD score indicates two protein models are similar. Given two sets of n points v and w, the RMSD is defined as follows by Eq. (1):

(1)

TM-score [21] [22] is an algorithm to measure the similarity of topologies between two protein models. Because TM-score considers the close matches more than the distant matches, TM-score is more accurate than RMSD. The range of TM-score is (0,1] for each pairs comparison, where 1 means a perfect match between two protein models.

GDT-TS [23] is a global quality measure of the correct positioning of amino acid sequences between two protein models. It is the most commonly parameter used to evaluate the quality of a predicted protein model with the experimentally determined structure. It is also widely accepted as a quality assessment of protein models prediction tool in CASP competition. The GDT-TS score is more sensitive to outlier regions of a protein model, so it is more accurate than RMSD score. The GDT-TS score is calculated by averaging percentage of residues with C-α atom distance in the model structure within certain distance cutoff of their positions. GDT-TS ranges from 0 to 1 with higher value indicating better accuracy. GDT-HA is a high accuracy version of the GDT score [24]. It is more rigorous and uses smaller cutoff distance, only half the size of GDT-TS.

GDT-TS is used as major assessment criteria in CASP competition to compare the predicted model with native structure. The GDT is defined as follow:

(2)

Where and are two protein 3D structures and is the percentage of amino acid residues' alpha carbon atoms from that can be superimposed with corresponding residues from within a defined distance cutoff , [20].

However, in the refinement procedure, the knowledge of native structure is not exposed. For predicted model’s evaluation, most groups used consensus GDT-TS (CGDT-TS) ，which considers the similarity between each pair of models. Given a set of prediction structures set *S* and a reference structure set *R*, the CGDT-TS score for each prediction structure *Si* is defined as:

 (3)

The reference structure set *R* can be the same structure set as *S*. The Eq. (2) and (3) indicates that the range of CGDT-TS is [0, 1]. The larger the CGDT score, the more similarity between two protein models. In this thesis, TM-score software developed by Yang Zhang’s group is used to compute the CGDT-TS [21].

### 2.2.2 Energy or scoring functions

OPUS\_Cα [25] is a knowledge-based potential function, only using the information of C-α position. This software is based on seven major representative molecular interactions in proteins: distance-dependent pairwise energy with orientation preference, hydrogen bonding energy, short-range energy, packing energy, tri-peptide packing energy, three-body energy, and salvation energy. dDFIRE [26] program treats each polar atom as a dipole and is based on the orientation angles in dipoles interactions and distance between two atoms dipoles. This approach considers the hydrogen bonding interaction as the physical dipole-dipole interaction. The possible orientation-dependent interactions between polar and nonpolar atoms and interactions between non-hydrogen-bonded polar atoms are treated fairly. This method produces an all-atom parameter-free statistical energy function. The calRW was developed by Zhang and Zhang’s group [27] by involving two functions: the first function is a pairwise distance-dependent atomic statistical potential function using an ideal random walk chain as reference state; the second function is using a side chain orientation-dependent energy function. GOAP [28] is a generalized orientation and distance-dependent all-atom statistical potential that is determined by the relative orientation of the planes, which relate on each heavy atom in interacting pairs. This algorithm only considers the distance and angle dependent between representative atoms or blocks of side-chain or polar atoms.

ProQ2 as the best single-model method in CSAP10 is a model quality assessment program, which could be used to evaluate the model’s quality or as a scoring function for sampling-based refinement. This model quality assessment algorithm uses support vector machines to validate each residue quality and the global quality of protein models. It combines previously used features with updated structural and predicted features to evaluate the predicted models [29].

For each individual residue, ProQ2 uses S-score which is the transformation of the normal RMSD to predict its quality. S-score is described by the following formula:

 (4)

where *RMSDi* is the local RMSD deviation for residue *i* based on a global superposition trying to maximize essentially the sum of S-score over the whole model.

### 2.3 MDS

Proteins’ three dimensional structures determine the functions of a protein and amino acid sequence. A protein model can be presented as a result of spatial contacts between amino acids. Neighbors in sequence are not corresponding to the nearest neighbors in space. The contact map of a protein is a two dimensional matrix in which each value represents the distance between all amino acid residue pairs of the 3D protein models. For a protein with n residues, the contact map is n × n symmetrical matrix, the distance of ( *i* , *j* ) is range from 0 to 1. “1” means to residue *i* and *j* are close and “0” means far. In certain conditions the contact map can reconstruct the 3D coordinates of a protein [30]. Contact map predictions have been used in the modeling of protein 3D structures [31] [32] [33] [34].

Metric multi-dimensional scaling (MDS) [35] [36] [37], also called Principal Coordinates Analysis (PCO), is a dimensional reduction technology which aims to identify patterns in a distance matrix. Using MDS, a 3D protein model could be constructed. Multi-dimensional scaling tries to find the low dimensional space to represent the distances between points in space that match the dissimilarities [38].

MDS can be used to compare orthologous sequence sets [39] and predict the protein models binding [40], achieve better, clash-free placement of loops obtained from a database of protein models [41].

### 2.4 MODELLER

MODELLER is a computer program for comparing protein models modeling by satisfaction of spatial restraints which are deduced from homology to template structures and energy objective functions [42] [43].

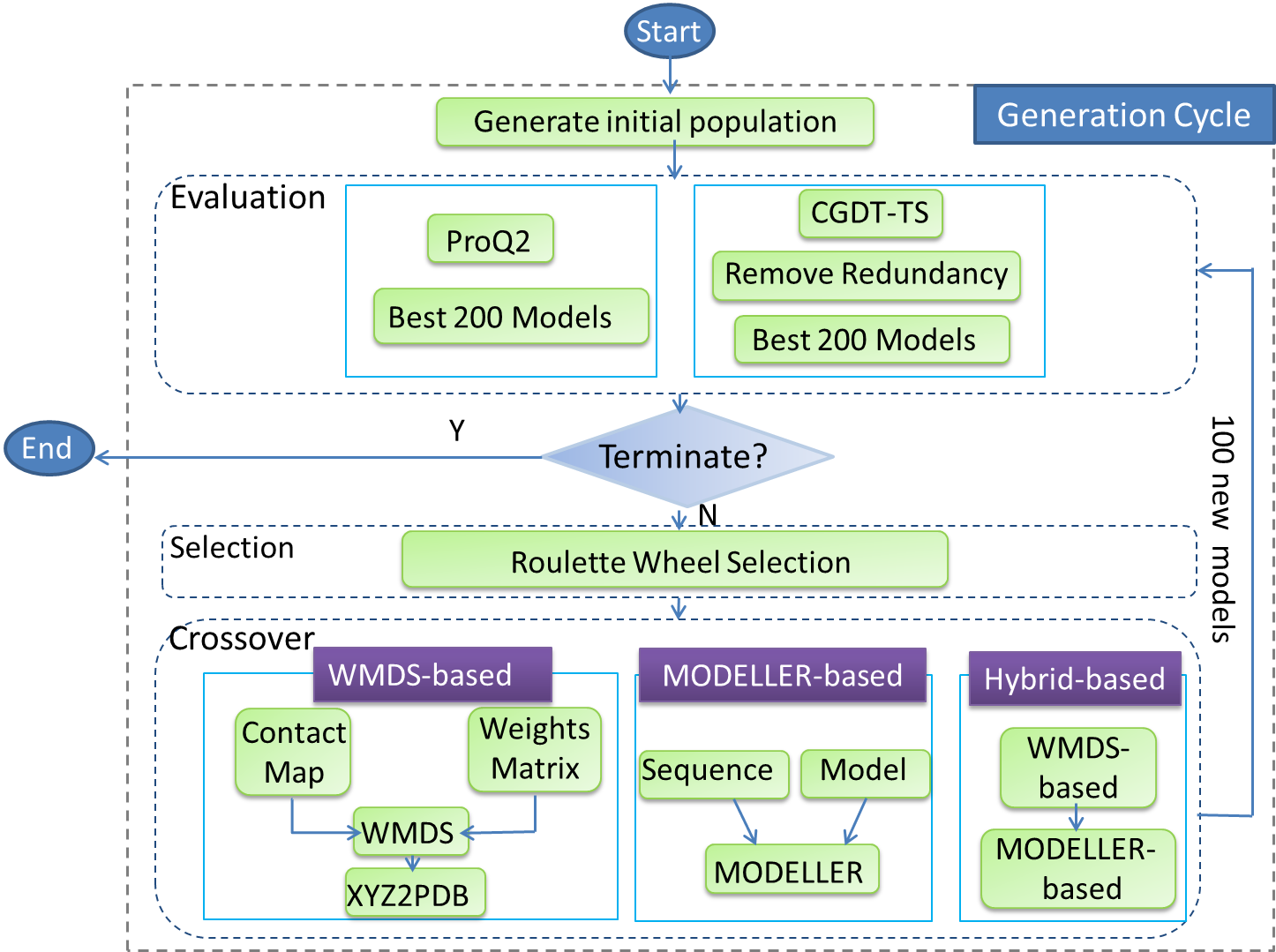
The basic inputs include an alignment of a sequence, the atomic coordinates of the templates and a script file. MODELLER calculates many distances and angle restraints from the alignment with the template tertiary structure and generates a model containing all non-hydrogen atoms within minutes. Except model building, MODELLER could be used in fold assignment, multiple alignments of protein sequences and calculation of phylogenetic trees, etc. [44].

The spatial restraints [44] include homology-derived restraints, stereo chemical restrains, statistical preferences for dihedral angles and non-bonded inter-atomic distances, and optional manually curated restraints. Those restrains presented as probability density functions are optimized by a combination of conjugate gradients and molecular dynamics with simulated annealing considering them altogether as a function.

### Chapter 3: THREE NEW EVOLUTIONARY ALGORITHMS FOR PROTEIN MODEL REFINEMNT

The evolutionary algorithm begins by creating a population of individuals, where the user defines the number of individuals in the population. The creation of the initial population is followed by loop over generations, repeating until the maximum number of generations is attained. A generation consists of three stages: protein model quality evaluation, selection, and crossover.

The following Fig. 3.1 shows the flow chart for the whole process.



**Figure 3. 1 The evolutionary algorithm of the whole process**

### 3.1 Initial Population

More than 100 research groups participated in CASP competition all over the world. First of all, the host sends out the identified sequences to each group; then each group uses different algorithms to generate the 3D structure partially or fully. Finally, at most, five decoys can be submitted to the CASP center. There are about 250 prediction structures for each protein target.

Some predicted structures generated from the same group or the similar algorithms are nearly identical to each other. For the diversity of the original pool, those similar models will be removed before calculating. In the previous study of our research group [45], if the similarity between two models is equal to 1 by computing the consensus GDT-TS scores, they will be considered to be similar and will discard one of them from the initial pool.

The original population pool contains protein models with different sequence lengths. Decoys with the same sequence length will generate the same size of the contact map and weights matrixes. For this experiment, we only consider protein models with the same length as native sequence and remove the too long or too short sequence for each target.

After removing redundant structures and partial length structures, we pick the best 200 models as our initial population. But actually, after rigorous screening, for some targets, the size of initial pool may be less than 200. The smallest population for this experiment is 164 for target T0711. After the first iteration, howover, the size of population would be kept as 200.

### 3.2 Evaluation

Good model quality assessment methods are essential and crucial for the end users of a model. They could contribute to the improvement of predicted models as well as the selection of the most likely model in a population. ProQ2 as a new single scoring function method has good performance in CASP10 competition, which takes one single structure as input and assigns a score. Consensus GDT-TS was widely used in protein models prediction, which assigns a score to each model as its average structural similarity (GDT score) to all other models in the set. If most predicted protein models have good quality for one protein model, the consensus GDT-TS could find the best predicted model for this target. We use TM-score program to get the pairwise similarity value between pairs. It can be used to calculate the GDT-TS score and RMSD value and evaluate the quality of predicted protein models comparing to native one.

In this experiment, single scoring function ProQ2 and consensus based GDT-TS involved as the protein quality assessment methods.

Figures 3.2 and 3.3 show examples of initial population for T0654.

**Figure 3. 2 ProQ2 and GDT-TS of initial population for T0654**

**Figure 3. 3 CGDT-TS and GDT-TS of initial population for T0654**

### 3.3 Selection

In the evolutionary algorithm, roulette wheel selection is the most widely used selection method. This selection is based on the probability of individual, which is proportional to their fitness values. The probability segments of individuals constituted a roulette wheel; the larger the segment size is, the more chances to be chosen. The selection probability *Pi* for individual *i* defined as follow:

(5)

where *f1*, *f2*,…,*fn* are the fitness values of individual 1, 2,…, n.

The roulette wheel selection gives a chance to all individuals. Therefore, it can keep the diversity in the population [46]. In this experiment, we chose 100 pairs as parents set. Figures 3.4 and 3.5 show the selection probability for T0654.

**Figure 3. 4 Selection probability of ProQ2 for T0654**

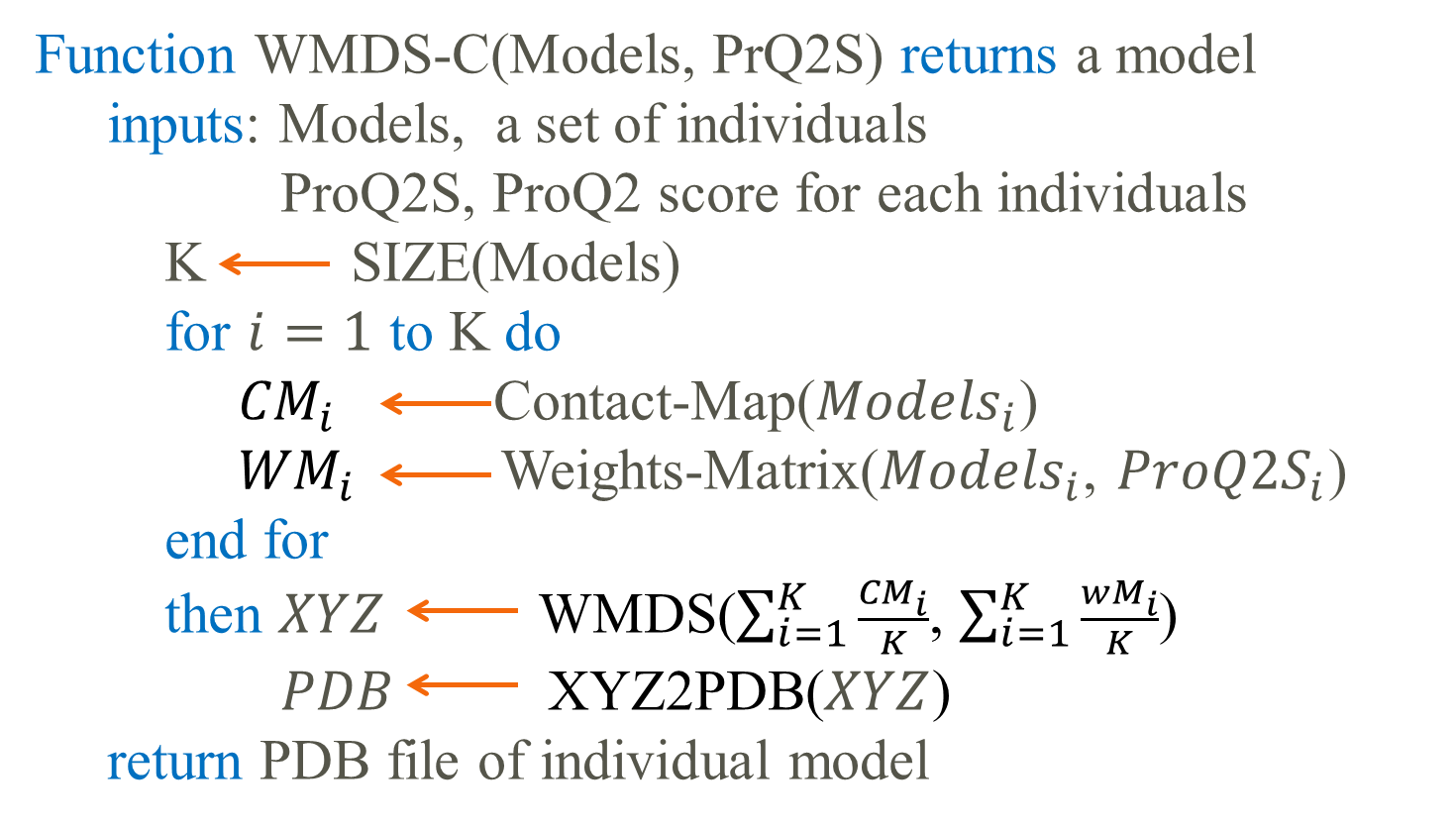
**Figure 3. 5 Selection probability of CGDT-TS for T0654**

### 3.4 Crossover

The parents are chosen by employing roulette wheel selection. Then the following methods are used to combine them together.

### 3.4.1 MDS-based method

The contact map of protein models reflects the distance relationship of each residue. Because the protein models are not always accurate in some segments, we need to combine parts of contact map with added the weights on the contact map to breed new generation. The WMDS-based crossover algorithm is shown in Figure 3.6.



**Figure 3. 6** **The WMDS-based crossover (WMDS-C) algorithm**

The distance between amino acids of a protein may be confident in some part, but not all of them. ProQ2 program can evaluate each residue of a protein. Those quality values are used to construct the weights matrix, and added to the distance matrix to refine the protein 3D structure. In the following part, four different weight matrices are constructed with *qi* representing the quality of residue *i* in a protein using ProQ2.

Weights matrix *w1* shown by [Eq.](app:ds:equation) (6) is the multiply quality of each residue pairs.

(6)

Except for weight matrix *w1*, the following weight matrices are also tested, but the results are worse for those matrices which are unable to receive a ProQ2 score. That is because the formula of weight matrix *w1* works as probability of distance between each pairs of amino acid in a protein models.

Weight matrix *w2* shown as in Eq. (7) is the sum of quality of each residue pairs.

(7)

Weights matrix *w*3 as in equation (9) is the multiply of quality of each residue pairs firstly, divided by the sum of the quality of all residue pairs.

(8)

(9)

Weights matrix *w4* as in Eq. (11) is the sum the quality of each residue pairs, divided by the sum of the quality of all residue pairs.

(10)

(11)

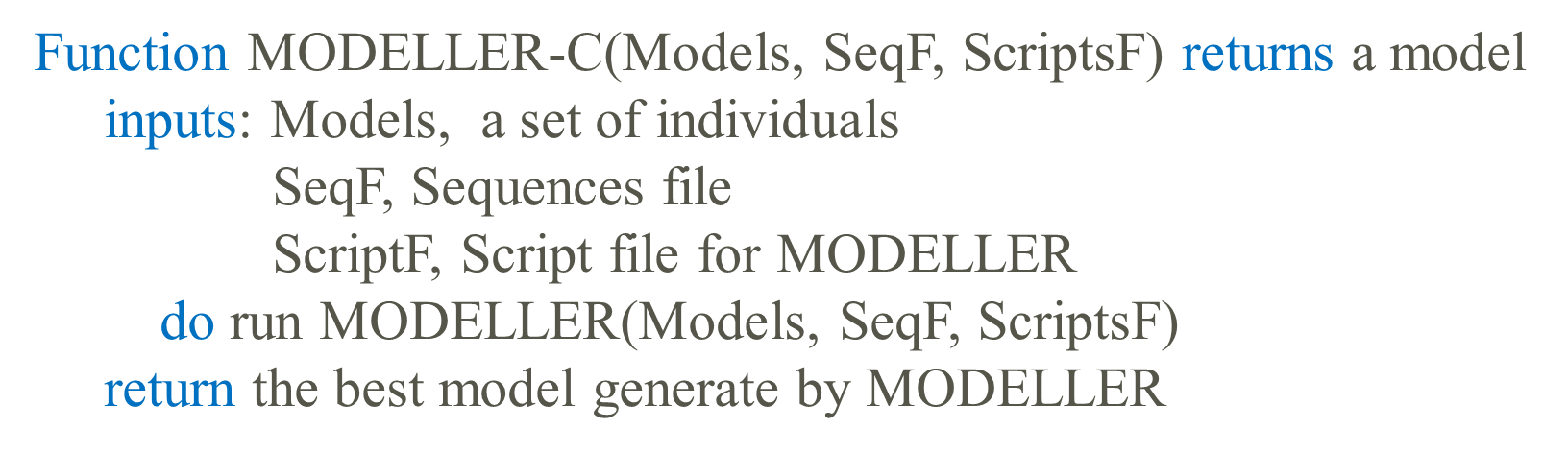
After selecting the parents, crossover is applied to generate the next generation. There are many methods to combine two different 3D structures. The better feature would be kept, and the bad ones are abandoned. The “Average” combine method is very simple and it averages the contact maps and weight matrix of parents. Average merges two contact maps into one.

By using MATLAB\_mdscale (non-classical multidimensional scaling) to implement the weight metric multidimensional scaling (WMDS), we allow those weights to maximize the fit to the original distances. Points with large weights will have a stronger influence on the results. The MATLAB function ‘MDSCALE’ is used to generate a three dimensional matrix M. Non-classical multidimensional scaling determines the type of scaling, non-metric or metric. Since the RMSD formula is used to get the contact map for each 3D protein model, it should be metric. The corresponding parameters are “metricstress”, “metricsstress”, “sammon”, and “strain”. After testing the parameters, “metricstress” is stable and has better performance, so “metricstress” is chosen as the default parameter when running the WMDS by MATLAB. In metric scaling, “MDSCALE” wants to configure points whose pairwise Euclidean distances is equal to the dissimilarities approximately.

After that, a Perl script xyz2pdb is used to reconstruct the PBD file of a prediction model from the coordinates which are derived by MDS scaling.

### 3.4.2 MODELLER-based method

In the MODELLER-based refinement algorithm, the roulette wheel selection is employed to select three parents as the template structure firstly, and then we extracted their alignments file that has the input sequences aligned fully with the template sequences and 3D coordinates of templates. They are combined together with a script as the input of MODELLER, and the default “automodel” modeling protocol in MODELLER is used. In this way, MODELLER will refold the models based on the input and give birth to a baby. The default number of templates is five, but three parents are chosen as a pair template in the calculation. If the templates are not very good models in the prediction pool, the next generation are also not good models. Better remodeling models could be obtained if the models in the pool are good. The MODELLER-based crossover algorithm is shown in Figure 3.7.

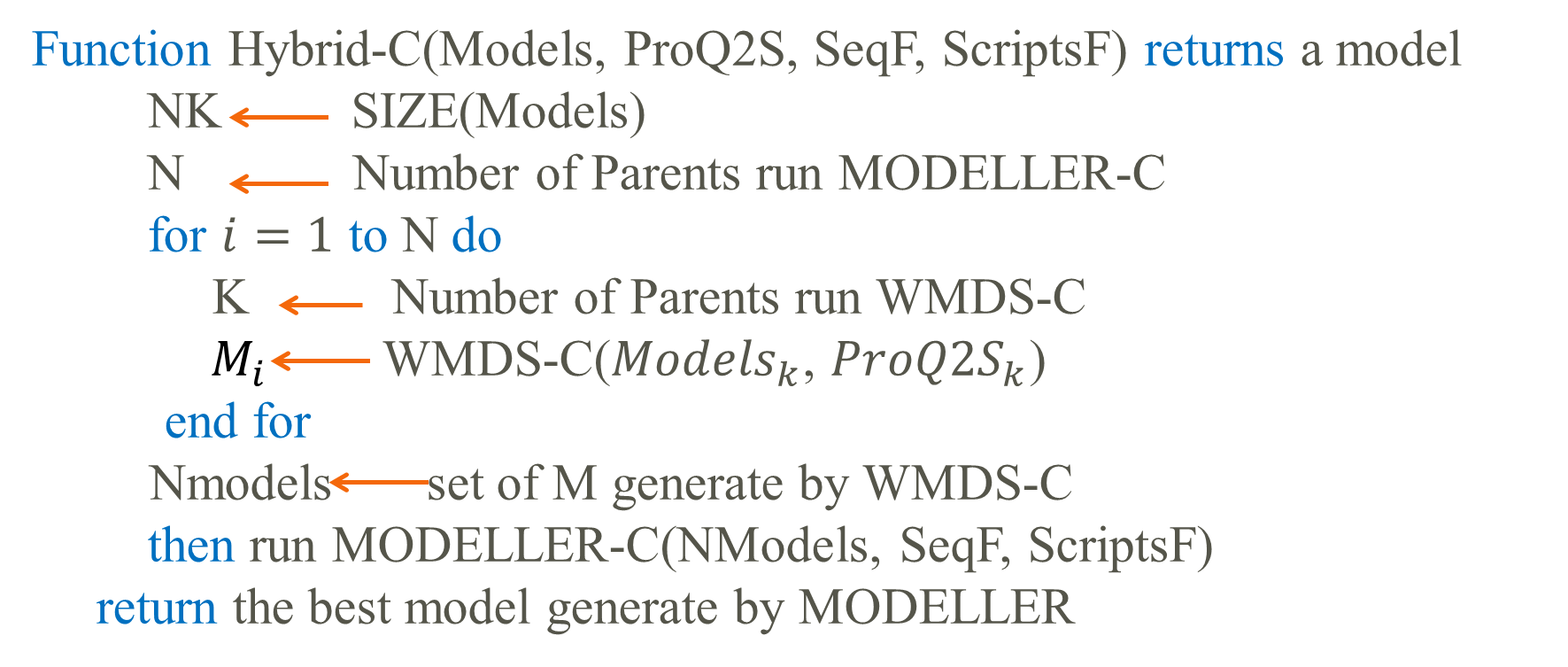


**Figure 3. 7 The MODELLER-based crossover (MODELLER-C) algorithm**

### 3.4.3 Hybrid method

In this algorithm, in order to compare this three crossover algorithms, the same number of protein models in each generation are generated. Firstly, the above MSD- based algorithm is repeated three times to generate enough parents, and then 100 sets of three models will run MODELLER and refold as refinement.

After generating the new breeds, the new models will compete with the original pool of their parents, then using the selection methods to keep the size of the pool and prepare for the next generation. The steps of selection, crossover and evaluation will be run iteratively until the end. Figure 3.8 shows the hybrid algorithm.



**Figure 3. 8 The Hybrid crossover (Hybrid-C) algorithm**

### Chapter 4: EXPERIMENTAL Results

This chapter presents the results from the algorithms that have been implemented in accordance with the design in Chapter 3.

This project was written in Perl, MATLAB, C/C++ and ran on Red Hat Enterprise Linux Server 5.4 with 24 Intel Xeon(R) CPU X5660 processors running @ 2.80GHz, total 24GB memory installed, and Linux kernel at version 2.6.32-220.el6.x86\_64. The MATLAB version is the newest MATLAB 8.2 with release name of R2013b.

### 4.1 Data Set

The whole system is using evolutionary algorithm to refine models iteratively. In this paper, the targets sequence information and templates are selected from CASP competition as test samples. Sixteen CSAP10 targets were used as experiment objects to generate the initial population for each target. At each time, an individual’s ProQ2 score or Consensus GDT-TS score is calculated, and the best predicted protein models was chosen in certain size of each target. After this, a model will be selected to reproduce by roulette wheel selection. Crossover is performed on two or three members of the population. The optional crossover refinement algorithm includes WMDS-based, MODELLER-based, and hybrid of WMDS-based and MODELLER-based methods. The evolutionary algorithm iterates over generations until the termination.

**Table 4. 1 Summary of CASP10 refinement targets**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Target | Residues | Average CGDT-TS | Best GDT-TS | Best RMSD |
| T0648 | 102 | 0.5663 | 0.8227 | 3.494 |
| T0654 | 166 | 0.5822 | 0.7682 | 2.542 |
| T0657 | 154 | 0.6325 | 0.8703 | 2.254 |
| T0659 | 85 | 0.6347 | 0.9696 | 0.861 |
| T0662 | 79 | 0.7473 | 0.8882 | 1.481 |
| T0665 | 67 | 0.2395 | 0.9948 | 0.626 |
| T0668 | 86 | 0.6501 | 0.4295 | 7.532 |
| T0669 | 109 | 0.4613 | 0.683 | 2.908 |
| T0673 | 88 | 0.2566 | 0.6532 | 5.349 |
| T0675 | 75 | 0.4423 | 0.6096 | 4.4 |
| T0678 | 161 | 0.2163 | 0.4172 | 6.28 |
| T0680 | 119 | 0.2730 | 0.763 | 3.298 |
| T0696 | 111 | 0.5571 | 0.7075 | 3.453 |
| T0698 | 119 | 0.5391 | 0.6471 | 3.916 |
| T0700 | 86 | 0.3450 | 0.9643 | 0.802 |
| T0709 | 33 | 0.6516 | 0.9896 | 0.685 |

The 3rd, 4th and 5th columns are the CGDT-TS of whole decoys for each target, the GDT-TS and RMSD of the best prediction model.

### 4.2 Computation Time

Figure 4.1 compares the computation time for different algorithms. In the figure, P and C stand for ProQ2 and Consensus GDT-TS, respectively, in the quality evaluation part. 2P and 3P stand for two or three parents in crossover, respectively. In crossover, there are three methods: WMD-based method, MODERLLER-based method, and combined WMDS and MODELLER-based method. For example, P-2P-WMDS means using ProQ2 program to evaluate the protein quality, selecting two models as parents and using WMDS-based crossover. C-2P-WMDS uses consensus GDT-TS as the evaluation method compared to P-2P-WMDS.

As shown in Figure 4.1, the computation time for WMDS-based refinement is shorter than MODELLER-based refinement. Two or three parents’ combination within WMDS-based refinement did not make a significant difference in the crossover section because the geometry computation runs very fast on the Linux system. The overall computation time for hybrid refinement method is not equal to the sum of three times computation time for running WMDS-based and MODELLER-based refinement.

**Figure 4. 1 The all computation time for different algorithms**

Figure 4.2 shows a summary result for the computation time in each step for different algorithms. The computation time for MODELLER-based refinement using ProQ2 as QA method over WMDS-based refinement is 5.9, the ratio is 2.5 when using CGDT-TS as QA method. CGDT-TS selection method spent more than 8 minutes to assess the whole data set and prepared for the next generation in iteration, while ProQ2 selection method takes less than 2 minutes. ProQ2 as one single model method was running very fast which the predicted models had the same sequence length and used the same base name file. Because CGDT-TS compared each pairs of prediction models in the population, the computation time higher than Proq2 evaluation method. For the same quality assessment method, because the size of population is the same, the computation times for predicted protein models evaluation are very similar. In the step of selection, the implement time is very short.

**Figure 4. 2 The average iteration time in each step for different algorithms**

Figure 4.3 shows the computation time for WMDS-based refinement. In the WMDS-based crossover method, most computation time was spent on running the WMDS in MATLAB program which spent more than 1 minute to get the coordinates of 100 models.

**Figure 4. 3 The computation time for steps in WMDS-based EA of each iteration**

### 4.3 Results

### 4.3.1 Refinement results using ProQ2 as quality assessment method

In Table 4.2, the best final models produced by evolutionary algorithms were selected based on the true GDT-TS against the native experimental structure using ProQ2 as quality assessment method. Figures 4.4 and 4.5 show the GDT-TS and RMSD of the best models obtained by ProQ2 evaluation, respectively.

P-2P-WMDS refinement algorithm is using ProQ2 as protein model quality assessment method and WMDS-based crossover method. It could get better models with higher true GDT-TS in 9 out of 16 and best 5 target models didn’t change during the refinement. For target T0680, the true GDT-TS was even increased from 0.763 to 0.8333 and its RMSD was decreased from 3.298 to 2.502. P-3P-WMDS chooses three models as parents and uses WMDS-based crossover. It didn’t improve the best model quality and even got worse model for target T0673. In MODELLER-based refinement, the maximum improvement occurred on target T0662 and its true GDTS-TS was improved 0.046. 8 out of 16 GST-TS of the best models was improved but RMSD didn’t change with 2 of 16 decreased. P-2P-WMDS-MODELLER combined WMSD-based and MODELLER-based refinement and it could also improve the best model’s quality. The biggest improvement happened to target T0669 and its GDT-TS increased by 0.0412. The RMSD of the best models had been improved by crossover evolutionary algorithm but their GDT-TS were not increased. The predicted model quality didn’t relate on its geometry similarity with the native structure.

**Table 4. 2 Summary of refinement results on CASP10 targets using ProQ2**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Target | Initial Best | | P-2P-WMDS | | P-3P-WMDS | | P-3P-MODELLER | | P-2P-WMDS-  MODELLER | |
| GDT-TS | RMSD | GDT-TS | RMSD | GDT-TS | RMSD | GDT-TS | RMSD | GDT-TS | RMSD |
| T0648 | 0.8227 | 3.494 | 0.8343 | 3.494 | 0.8227 | 3.494 | 0.8314 | 3.465 | 0.843 | 2.936 |
| T0654 | 0.7682 | 2.542 | 0.7682 | 2.542 | 0.7682 | 2.542 | 0.7646 | 3.064 | 0.7518 | 2.817 |
| T0657 | 0.8703 | 2.254 | 0.8741 | 2.218 | 0.8703 | 2.254 | 0.8759 | 2.194 | 0.8759 | 2.229 |
| T0659 | 0.9696 | 0.861 | 0.973 | 0.8 | 0.9696 | 0.861 | 0.9696 | 0.797 | 0.9696 | 0.818 |
| T0662 | 0.8882 | 1.481 | 0.9178 | 1.433 | 0.8882 | 1.481 | 0.9342 | 1.316 | 0.9013 | 1.453 |
| T0665 | 0.9948 | 0.626 | 0.9948 | 0.626 | 0.9948 | 0.626 | 0.9948 | 0.671 | 0.9948 | 0.626 |
| T0668 | 0.4295 | 7.532 | 0.4423 | 6.404 | 0.4295 | 7.532 | 0.4295 | 7.513 | 0.4423 | 6.277 |
| T0669 | 0.683 | 2.908 | 0.7036 | 2.876 | 0.683 | 2.908 | 0.7165 | 2.85 | 0.7242 | 2.908 |
| T0673 | 0.6532 | 5.349 | 0.6532 | 5.349 | 0.4395 | 11.06 | 0.6532 | 5.302 | 0.6532 | 5.349 |
| T0675 | 0.6096 | 4.4 | 0.5965 | 4.509 | 0.6096 | 4.4 | 0.5789 | 5.048 | 0.5965 | 4.509 |
| T0678 | 0.4172 | 6.28 | 0.4318 | 6.139 | 0.4172 | 6.28 | 0.4221 | 7.054 | 0.4172 | 6.28 |
| T0680 | 0.763 | 3.298 | 0.8333 | 2.502 | 0.763 | 3.298 | 0.763 | 3.944 | 0.763 | 3.298 |
| T0696 | 0.7075 | 3.453 | 0.705 | 3.453 | 0.7075 | 3.453 | 0.7475 | 3.892 | 0.705 | 4.127 |
| T0698 | 0.6471 | 3.916 | 0.6702 | 3.982 | 0.6471 | 3.916 | 0.6723 | 4.019 | 0.6618 | 3.982 |
| T0700 | 0.9643 | 0.802 | 0.9643 | 0.802 | 0.9643 | 0.802 | 0.8714 | 1.313 | 0.8429 | 1.641 |
| T0709 | 0.9896 | 0.685 | 0.9896 | 0.694 | 0.9896 | 0.685 | 1 | 0.51 | 0.9792 | 0.768 |

**Figure 4. 4 The GDT-TS of best model using ProQ2 evaluation**

**Figure 4. 5 The RMSD of best model using ProQ2 evaluation**

Figures 4.6 and 4.7 show the average GDT-TS of best 10 and all models using ProQ2 evaluation method, respectively. Only 3 of best 10 models’ average GDT-TS were not improved in this experiment by P-2P-WMDS evolutionary algorithm. For target T0698, P-2P-WMDS, P-3P-MODELLER, P-2P-WMDS-MODELLER methods raised the average GDT-TS of best 10 models to 0.0427, 0.0501 and 0.0358, respectively. The average quality of total generated models was improved except P-3P-WMDS EA method. From Fig. 4.7, the average GDT-TS of all models went up 0.1474, 0.1645 and 0.1534, respectively for target T0657.

**Figure 4. 6 The average GDT-TS of best 10 models using ProQ2 evaluation**

P-3P-WMDS method couldn’t achieve better models sometimes and even got worse models. This may be due to involvement of too many parents that make a mess of the constructed contact maps and weight matrixes. P-2P-WMDS utilized the geometry benefits of both parents and added the appropriate probability weights which are generated by ProQ2 program to improve the predicted model quality.

**Figure 4. 7 The average GDT-TS of all models using ProQ2 evaluation**

### 4.3.2 Refinement results using CGDT-TS as quality assessment method

Table 4.3 describes the results of EAs on CASP10 refinement target by CGDT-TS evaluation.

According to the Table 4.3, MDS-based and MODELLER-based refinement algorithms using CGDT-TS evaluation method failed to achieve better predicted models. The reason for this is that CGDT-TS compared all individual models and similar models could get higher score, using CGDT-TS evaluation the number of generated offspring is too big and most of them are similar and has poor quality, so it can’t keep the good predicted protein models in iteration of the pool.

**Table 4. 3 Summary of refinement results on CASP10 targets using CGDT-TS**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Target | Initial Best | | C-2P-WMDS | | C-3P-WMDS | | C-3P-MODELLER | | C-2P-WMDS-  MODELLER | |
| GDT-TS | RMSD | GDT-TS | RMSD | GDT-TS | RMSD | GDT-TS | RMSD | GDT-TS | RMSD |
| T0648 | 0.8227 | 3.494 | 0.811 | 4.474 | 0.314 | 7.408 | 0.8169 | 4.544 | 0.7965 | 5.017 |
| T0654 | 0.7682 | 2.542 | 0.7427 | 3.259 | 0.2682 | 7.25 | 0.7518 | 3.383 | 0.7445 | 3.46 |
| T0657 | 0.8703 | 2.254 | 0.8722 | 2.303 | 0.2556 | 7.339 | 0.8722 | 2.225 | 0.8609 | 2.319 |
| T0659 | 0.9696 | 0.861 | 0.9561 | 0.855 | 0.3581 | 5.292 | 0.9561 | 0.923 | 0.9527 | 0.941 |
| T0662 | 0.8882 | 1.481 | 0.8257 | 1.79 | 0.3388 | 5.848 | 0.8059 | 1.86 | 0.8224 | 1.788 |
| T0665 | 0.9948 | 0.626 | 0.9583 | 0.837 | 0.4479 | 4.343 | 0.9635 | 0.827 | 0.9635 | 0.8 |
| T0668 | 0.4295 | 7.532 | 0.391 | 7.049 | 0.266 | 9.055 | 0.4199 | 7.308 | 0.375 | 7.767 |
| T0669 | 0.683 | 2.908 | 0.6649 | 4.2 | 0.2784 | 7.468 | 0.6495 | 4.154 | 0.6418 | 4.757 |
| T0673 | 0.6532 | 5.349 | 0.3508 | 9.259 | 0.2177 | 13.395 | 0.4758 | 11.139 | 0.3508 | 10.399 |
| T0675 | 0.6096 | 4.4 | 0.5 | 7.233 | 0.2588 | 11.147 | 0.5044 | 7.058 | 0.5132 | 7.07 |
| T0678 | 0.4172 | 6.28 | 0.4026 | 6.479 | 0.0828 | 16.182 | 0.3782 | 7.72 | 0.3653 | 7.31 |
| T0680 | 0.763 | 3.298 | 0.4766 | 12.546 | 0.763 | 3.298 | 0.5677 | 12.413 | 0.4688 | 11.485 |
| T0696 | 0.7075 | 3.453 | 0.5025 | 12.041 | 0.2175 | 14.913 | 0.5075 | 11.946 | 0.5025 | 11.937 |
| T0698 | 0.6471 | 3.916 | 0.6176 | 4.702 | 0.2773 | 7.494 | 0.5987 | 4.941 | 0.6324 | 4.722 |
| T0700 | 0.9643 | 0.802 | 0.8786 | 1.448 | 0.3214 | 7.492 | 0.9786 | 0.793 | 0.8429 | 1.531 |
| T0709 | 0.9896 | 0.685 | 0.9792 | 0.636 | 0.5521 | 3.162 | 0.9896 | 0.679 | 0.9688 | 0.852 |

**Table 4.3 Summary of refinement results on CASP10 targets using CGDT-TS**

Figures 4.8 and 4.9 are showing the GDT-TS and RMSD of the best models using CGDT-TS evaluation, respectively. Almost all the generated best protein model for each target with CGDT-TS evaluation and running for 10 iterations were worse than the initial best model in the template pool. The similar result shows in the Fig. 4.10 which is the average GDT-TS for best 10 models using CGDT-TS evaluation.

**Figure 4. 8 The GDT-TS of best model using CGDT-TS evaluation**

**Figure 4. 9 The RMSD of best model using CGDT-TS evaluation**

**Figure 4. 10 The average GDT-TS of best 10 models using CGDT-TS evaluation**

While it can be seen from Fig. 4.11 that the average GDT-TS for all models using CGDT-TS evaluation method is improved except C-3P-MODELLER evolutionary algorithm.

**Figure 4. 11 The average GDT-TS of all models using CGDT-TS evaluation**

Figure 4.12 shows the average improved GDT-TS for different algorithms. Except choosing three models as parents with WMDS-based crossover, overall predicted models in the population are improved. Many evidence shows that more than two parents can increase the performance of EAs [47]. But studies on different methods and different types of fitness should be considered, in this experiment, choosing three models as parents failed to reach good results. For the top 1 and top10 models, the average improved GDT-TS were not significant.

**Figure 4. 12 The average improved GDT-TS for different algorithms**

Despite not being conclusive, the study indicated that using MDS-based crossover method to refold existing predicted models can be a promising approach to improve the best predicted protein models’ quality.

### Chapter 5: CONCLUSIONS AND FUTURE WORK

The current project applied evolutionary algorithm framework and implemented WMDS-based, MODELLER-based, hybrid of MDS-based and MODELLER-based crossover algorithms to refine the predicted protein model structures. The performance of the evolutionary algorithm is compared based on the ProQ2 and CGDT-TS evaluation.

In WMDS-based crossover algorithm, choosing three models as parents is hard to get better models. Using CGDT-TS as QA method replaced the last population too fast with bad models that couldn’t improve the overall model quality. In this method, using ProQ2 as QA method and choosing two models as parents could achieve better predicted models.

In MODELLER-based crossover algorithm, the average computation time for each target is more than 3.5 hours. It could achieve better models in some cases.

The hybrid method could attain few better models sometimes, but it can’t improve the GDT-TS for each target on top1 model and the average GDT-TS of top10 models.

Comparing those three EAs, the computation time for MODELLER-based method are slow than WMDS-based method; using ProQ2 method to evaluate the models’ quality is fast than CGDT-TS method; except choosing three models as parents, all methods could improve the overall quality of the population; only P-2P-WMDS method could improve the GDT-TS of each target on top1, the average GDT-TS of top 10 and all population.

The performance of WMDS-based method with three parents is very bad and the reason are still not clear, so multi-parents for WMDS-based crossover algorithm needs more research on it.

In the CASP refinement problem, the best starting model will be provided, but this experiment only consider the whole population for each target, more experiment should be done for the next CASP competition.

In this study, the global ProQ2 score was used as protein quality assessment method and the local ProQ2 score was used to construct the weight matrices. For further study, other promising program such as IDDT [48] could be used in this evolutionary algorithm to see whether better results could be obtained.

Another problem is how to pick the best model from the finial population using better quality assessment method. Evolutionary algorithm as a stochastic searching process could get the similar performance for different runs. The final results may be different and the exactly same results couldn’t be obtained in this experiment.

# REFERENCES

|  |  |
| --- | --- |
| [1] | C. Anfinsen, "Principles that govern the folding of protein chains," *Science,* vol. 181, pp. 223-230, 1973. |
| [2] | C. Anjum A, "Protein Tertiary Model Assessment Using Granular," 2012. |
| [3] | E. Reynaud, "Protein Misfolding and Degenerative Diseases," *Nature Education,* vol. 3, no. 9, p. 28, 2010. |
| [4] | M.S. Johnson, N. Srinivasan, R.Sowdhamini, TL.Blundell, "Knowledge-based protein modeling," *Crit Rev Biochem Mol Biol,* vol. 29, pp. 1-68, 1994. |
| [5] | D. Bhattacharya, J.Cheng, "i3Drefine Software for Protein 3D Structure Refinement and Its Assessment in CASP10," *PLoS ONE,* vol. 8, no. 7, p. e69648, 2013. |
| [6] | L. Heo, H. Park and C. Seok, "GalaxyRefine: protein structure refinement driven by side-chain repacking," *Nucleic Acids Research,* vol. 41, pp. 384-388, 2013. |
| [7] | João P. G. L. M.Rodrigues,M.Levitt and G.Chopra, "KoBaMIN: a knowledge-based minimization web server for protein structure refinement," *Nucleic Acids Research,* vol. 40, pp. 323-328, 2012. |
| [8] | D. Xu, J. Zhang, A. Roy, and Y. Zhang, "Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement," *Proteins,* vol. 79(Suppl 10), pp. 147-160, 2011. |
| [9] | H. Park and C. Seok, "Refinement of unreliable local regions in template-based protein models," *Proteins,* vol. 80, pp. 1974-1986, 2012. |
| [10] | H.Park,J.Ko, K.Joo, J.Lee, C. Seok, and J. Lee, "Refinement of protein termini in template-based modeling using conformational space annealing," *Proteins,* vol. 79, pp. 2725-2734, 2011. |
| [11] | J.L.MacCallum,L. Hua,M.J. Schnieders,V.S. Pande,M.P.Jacobson and K.A.Dill, "Assessment of the protein structure refinement category in CASP8," *Protein,* vol. 77(Suppl.9), pp. 66-80, 2009. |
| [12] | J.L.MacCallum,A.Perez, M.J.Schnieders,L.Hua,M.P.Jacobson, and K.A.Dill, "Assessment of protein structure refinement in CASP9," *Proteins,* vol. 79(Suppl.10), pp. 74-90, 2011. |
| [13] | T.Nugent,D.Cozzetto and D.T.Jones, "Evaluation of predictions in the CASP10 Model Refinement Category(Accepted Article)," *Proteins:Structure,Function,and Bioinformatics,* 2013. |
| [14] | J. Holland, "Adaptation in Natural and Artificial Systems," in *University of Michigan Press*, Michigan, 1975. |
| [15] | N.Mansour, F.Kanj,H.Khachfe, "Evolutionary Algorithm for Protein Structure Prediction," in *Natural Computation (ICNC), 2010 Sixth International Conference*, Yantai,ShanDong, 2010. |
| [16] | A. Warmflash, P. Francois and E.D. Siggia, "Pareto evolution of gene networks: an algorithm to optimize multiple fitness objectives," *Physucal Biology,* vol. 9, no. 5, 2012. |
| [17] | Y.Zhang, L.Wu, "Artificial Bee Colony for Two Dimensional Protein Folding," *Advances in Electrical Engineering Systems,* vol. 1, no. 1, pp. 19-23, 2012. |
| [18] | MT.Oakley, EG.Richardson, H.Carr, RL.Johnston, "Protein Structure Optimisation With a "Lamarckian" Ant Colony Algorithm," *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM,* 2013. |
| [19] | J. Skolnick, "In quest of an empirical potential for protein structure," *Curr Opin Struct Biol,* vol. 16, pp. 166-171, 2006. |
| [20] | Q. Wang, Y.Shang, and D. Xu, "Improving a Consensus Approach for Protein Structure Selection by Removing Redundancy," *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 8, no. 6, pp. 1708-1715, 2011. |
| [21] | Y.Zhang and J.Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins,* vol. 57, no. 4, pp. 702-710, 2004. |
| [22] | Y.Zhang and J.Skolnick, "TM-align: a protein structure alignment algorithm based on the TM-score," *Nucleic Acids Res,* vol. 33, no. 7, pp. 2302-2309, 2005. |
| [23] | A. Zemla, "LGA: a method for finding 3D similarities in protein structures," *Nucleic acids research,* vol. 31, pp. 3370-3374, 2003. |
| [24] | R.J. Randy and G. Chavali, "Assessment of CASP7 predictions in the high accuracy template-based modeling category," *Proteins,* vol. 69, no. s8, pp. 27-37, 2007. |
| [25] | Y. Wu,M. Lu,M. Chen,J. Li and J. Ma, "OPUS-Ca: A knowledge-based potential function requiring only Ca positions," *Protein Science,* vol. 16, no. 7, pp. 1449-1463, 2007. |
| [26] | Y. Yang, Y. Zhou, "Specific interactions for ab initio folding of protein terminal regions with secondary structures," *Proteins,* vol. 72, pp. 793-803, 2008. |
| [27] | J.Zhang,Y.Zhang, "A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction," *PLoS ONE,* vol. 5, no. 10, pp. 1-13, 2010. |
| [28] | H. Zhou and J. Skolnick, "GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction," *Biophysical Joural,* vol. 101, pp. 2043-2052, 2011. |
| [29] | A. Ray, E. Lindahl and B. Wallner, "Improved model quality assessment using ProQ2," *BMC Bioinformatics,* vol. 13, p. 224, 2012. |
| [30] | M.Vassura,L. Margara,P. Di Lena, F.Medri,P.Fariselli and R.Casadio, "Reconstruction of 3D Structures From Protein Contact Maps," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 5, no. 3, pp. 357-367, 2008. |
| [31] | A.N. Tegge, Z. Wang, J. Eickholt, and J. Cheng, "NNcon: Improved Protein Contact Map Prediction Using 2D-Recursive Neural Networks," *Nucleic Acids Research,* vol. 37, pp. w515-w518, 2009. |
| [32] | J. Bacardit, P. Widera,A. Márquez-Chamorro,F.Divina,J. S. Aguilar-Ruiz,and N.Krasnogor, "Contact map prediction using a large-scale ensemble of rule sets and the fusion of multiple predicted structural features," *Bioinformatics,* vol. 28, no. 19, pp. 2441-2448, 2012. |
| [33] | Z.Wang and J. Xu, "Predicting protein contact map using evolutionary and physical constraints by integer programming (extended version)," *Bioinformatics,* vol. 29, no. 13, pp. 266-273, 2013. |
| [34] | W.Ding, J.Xie,D.Dai,H.Zhang,H.Xie and W.Zhang, "CNNcon: Improved Protein Contact Maps Prediction Using Cascaded Neural Networks," *PLoS ONE,* vol. 8, no. 4, p. e61533, 2013. |
| [35] | W. Togerson, Theory and methods of scaling, New York: Wiley, 1958. |
| [36] | H. Abdi, "Metric multidimensional scaling. In: Salkind NJ, editor," in *Encyclopedia of Measurement and Statistics*, Thousand Oaks (CA), Sage, 2007, pp. 598-605. |
| [37] | Y.Takane,S.Jung and Y.Oshima-Takane, "Multidimensional scaling. In: Millsap R, Maydeu-Olivares A, editors," in *Handbook of quantitative methods in psychology*, London, Sage Publications, 2009, pp. 219-242. |
| [38] | F.Cox Trevor and A.A.Cox Michael, Multidimensional Scaling(Second Edition), Chapman and Hall/CRC, 2000. |
| [39] | J.Pelé,H.Abdi,M.Moreau,D.Thybert and M.Chabbert, "Multidimensional Scaling Reveals the Main Evolutionary Pathways of Class A G-Protein-Coupled Receptors," *PLoS ONE,* vol. 6, no. 4, p. e19094, 2011. |
| [40] | H.Mogi and Y-h.Taguchi, "Protein binding prediction using non-metric multidimensional scaling method," in *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference*, Atlanta, GA, 2011. |
| [41] | D.Holtby, S.Cheng Li and M. Li, "LoopWeaver: Loop Modeling by the Weighted Scaling of Verified Proteins," *Journal of Computational Bioligy,* vol. 20, pp. 212-223, 2013. |
| [42] | A.Sali and TL. Blundell, "Comparative Protein Modeling by Satisfaction of Spatial Restraints," *Journal of Molecular Biology,* vol. 234, no. 3, pp. 779-815, 1993. |
| [43] | A. Fiser, R.K. Do, and A. Sali, "Modeling of loops in protein structures," *Protein Science,* vol. 9, pp. 1753-1773, 2000. |
| [44] | N. Eswar, M. A. Marti-Renom, B. Webb, M. S. Madhusudhan, D. Eramian, M. Shen, U. Pieper, A. Sali, "Comparative Protein Structure Modeling With MODELLER," *Current Protocols in Bioinformatics, John Wiley & Sons, Inc.,* vol. Supplement 15, pp. 5.6.1-5.6.30, 2006. |
| [45] | Q.Wang, K. Vantasin,D.Xu,and Y. Shang, "MUFOLD-WQA: A new selective consensus method for quality assessment in protein structure prediction," *Proteins,* no. 79(Suppl 10), pp. 185-195, 2011. |
| [46] | M.R. Norainii,J. Geraghty, "Genetic Algorithm Performance with Different Selection Strategies in Solving TSP," 2011. |
| [47] | L. D. Chambers, Practical Handbook of Genetic Algorithms: Complex Coding Systems, Volume 3, CRC Press, 1998. |
| [48] | V. Mariani,M. Biasini,A.Barbato and T.Schwede, "IDDT:a local superposition-free score for comparing protein structures and models using distance difference tests," *Bioinformatics,* vol. 29, no. 21, pp. 2722-2728, 2013. |