Protein Folding Prediction with Genetic Algorithms [*]

Yi-Yao Huang, Chang-Biau Yang, Kuo-Tsung Tseng, and Chia-Ning Yang
cbyang@cse.nsysu.edu.tw

Key words: protein structure, folding, genetic algorithm, secondary structure

Abstract

Hydrophobic-hydrophilic model (HP model) is one of the most simplified and popular protein folding models. This model considers the hydrophobic-hydrophobic interactions of protein structures, but the results of prediction are not encouraged enough. Therefore, we suggest that some other features should be considered, such as SSEs, charges, and disulfide bonds. In this paper we propose a genetic algorithm (GA) with more possible considerations based on the lattice model to predict the 3D structure of an unknown protein, target protein, whose primary sequence and secondary structure elements (SSEs) are assumed known. Experimental results show that these additional features indeed improve the prediction accuracy by comparing our prediction results with their real structures with RMSD.

1 Introduction

Protein folding prediction, sometimes called *protein structure prediction* (PSP), is one of the most important issues for understanding living organisms. It is believed that the biological function of a protein is determined by its structure, and that is why proteins play so many different roles in cells, such as catalyzing, regulating, and transporting, and so on. Biological scientists have dedicated themselves to solving the structures of proteins by experimentations, like *X-ray crystallography* and *nuclear magnetic resonance* (NMR) spectroscopy [17, 19, 22]. Both of them, however, are time-consuming and the difficulty of crystallizing the structures is increasing due to the high degree complexity of the protein structures. As a result, many computer scientists have taken part in the research of PSPs by using computational strategies. Generally speaking, there are three commonly used strategies for solving PSPs: *ab initio*, *homology modeling*, and *threading* methods.

PSPs based on lattice models have been proved to be NP-complete problems [2, 6, 11, 13, 20]. Therefore, it is unfeasible to use the thorough search (i.e. the brute force algorithm) of one protein's conformation for these problems. Consequently, heuristic optimization methods are considered to solve PSP problems. In recent years, many heuristic methods for PSPs have been proposed. And, *genetic algorithm* (GA) is one of the most popular strategies used by many scientists [7, 10, 15, 16, 21].

Whereas the results of traditional *hydrophobic-hydrophilic models* (HP models) that consider the hydrophobicity only does not seem to be very satisfied, and many methods that only obtain improvements in how to get better folding do not upgrade the accuracy of PSPs explicitly. We suggest that some other characteristics, besides the hydrophobicity, have to be considered [3].

In this paper we propose a hybrid method of homology modeling and folding strategies. We use GA as our method's architecture and perform folding on the 3D lattice model to help us obtain a rough conformation of a target protein. We consider the hydrophobicity, the charges of the *side chains*, the specificity of some amino acids, and most important one, the *secondary structure elements* (SSEs), in our fitness function of GA to improve the prediction accuracy. In the past, some methods have ever used the SSEs to get the accuracy improvement of PSPs [5, 9].

The rest of this paper is as follows. Section 2 introduces the levels of protein structures, the HP model, and the genetic algorithm (GA). In Section 3, we propose our predicting method and in Section 4, we discuss the scoring aspects of our fitness function. Section 5 presents our experimental results on some small proteins and shows that our results, *RMSD* (*Root Mean Square Distance*) values, are better than those of the previous methods. We concludes our paper in Section 6.

[*] This research work was partially supported by the National Science Council of the Republic of China under contract NSC-92-2213-E-110-005.

2 Preliminaries

In this section, we shall give some basic concepts of proteins, the HP model we used and the genetic algorithm (GA).

There are twenty kinds of amino acids in proteins. These amino acids have different side chains, or the so-called *residues*. Table 1 lists these twenty amino acids, their abbreviations, and their hydrophobicity.

Protein structures are so complicated, so that they are categorized into three levels of structures, as shown in Figure 1. We briefly introduce these levels as follows.

Primary structure, the first level, of a protein is the amino acid sequence.

Secondary structure is the second level. The regular arrangements that can be found in proteins have repetitive hydrogen bonding between the amide N-H and the carboxyl group of the peptide backbone. The bonds in the peptide backbone are important for structures. *α helix* and *β sheet* are two major types of secondary structure. Other random parts are called *loops*.

Tertiary structure is the further folding of secondary structures and loops. The tertiary structure of a protein is stabilized by the forces of hydrophobic interactions between amino acids mainly and other forces, like disulfide bonds that can be formed by cysteines, charge attractions, and so on.

The *hydrophobic-hydrophilic model* (HP model) proposed by Dill [8] is one of the most simplified and popular of protein folding models [2, 10, 15, 16, 18, 21], where H represents hydrophobic (*non-polar* or "water-hating") and P represents hydrophilic (*polar* or "water-liking"). In the HP model, amino acids in a protein sequence are abstracted by hydrophobic (H) or by hydrophilic (P). And then, the protein sequence folds on this lattice such that each amino acid occupies at one grid point and two consecutive amino acids, called *sequential* amino acids, in the sequence are also adjacent at the grid points of the lattice model. Besides, any two amino acids can not occupy at the same grid point. Therefore, the folding of a protein sequence is restricted on the lattice space. For one protein sequence, there may be many feasible folding arrangements as long as they follow the rules mentioned above.

According to the law of thermodynamics, natural proteins or the so-called folded proteins are under the status of minimal free energy. For the above reason, on the HP model, the natural conformation of a protein is supposed to have the minimal free energy when there are maximal non-sequential hydrophobic amino acid pairs. That is, the main provider of the free energy is defined as

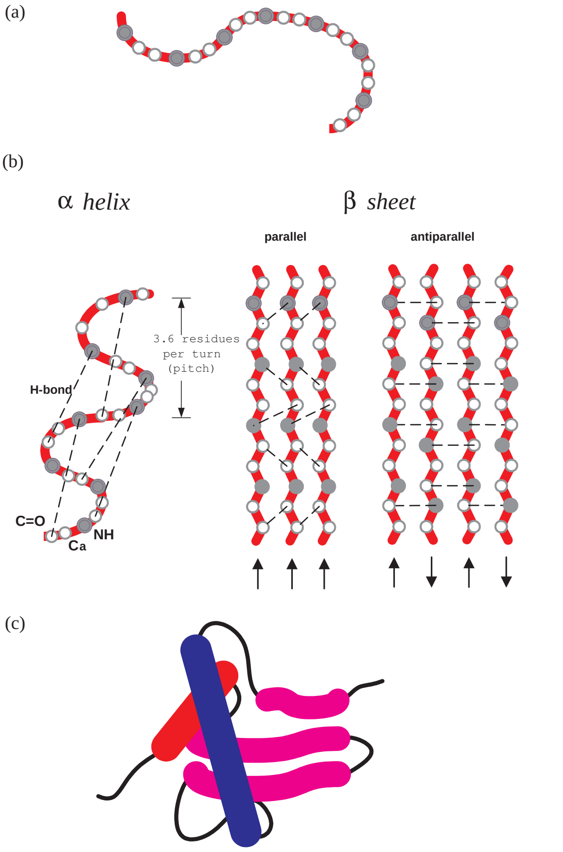

Figure 1: Levels of protein structures. (a)Primary structure. (b)Secondary structure. (c)Tertiary structure.

Table 1: Twenty amino acids. H means "hydrophobic" (water-hating) and P means "polar" (water-liking).

	One-letter Abbreviation	Three-latter Abbreviation	Name	Hydrophobicity
1	A	Ala	Alanine	H
2	C	Cys	Cysteine	P
3	D	Asp	Aspartic Acid	P
4	E	Glu	Glutamic Acid	P
5	F	Phe	Phenylalanine	H
6	G	Gly	Glycine	H or P
7	H	His	Histidine	P
8	I	IlE	Isoleucine	H
9	K	Lys	Lysine	P
10	L	Leu	Leucine	H
11	M	Met	Methionine	H
12	N	Asn	Asparagine	P
13	P	Pro	Proline	H
14	Q	Gln	Glutamine	P
15	R	Arg	Arginine	P
16	S	Ser	Serine	P
17	T	Thr	Threonine	P
18	V	Val	Valine	H
19	W	Trp	Tryptophan	H
20	Y	Tyr	Tyrosine	P

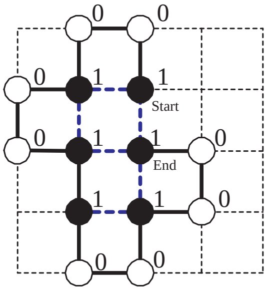

Figure 2: One folding conformation of a protein sequence on the 2D HP model with free energy 6. The sequence starts from "Start" and ends at "End", where amino acids are linked by solid lines. The solid node represents "hydrophobic" and the hollow one represents "hydrophilic". Non-sequential hydrophobic amino acid pairs are linked by dot lines, where two amino acids of each pair are not sequential in sequence (linked by solid line) and are adjacent on the grid points of the lattice model.

the interactions between hydrophobic amino acids such that the hydrophobic amino acids often form a hydrophobic core interior, and the global conformation surrounded by hydrophilic ones. Figure 2 shows one folding conformation of a protein sequence on the 2D HP model with free energy 6, where six non-sequential hydrophobic amino acid pairs (pairs linked by dot lines) are included.

The genetic algorithm (GA) was first proposed by Holland [14]. The basis of GA is "survival of the fittest", which was referred to Darwin's evolutionary theory, and the evolution of species is dominated by recombination and mutation of gene. Therefore, the main idea of GA is to simulate the process of natural evolution such that individuals of the population undergo recombination and mutation to adapt the new environment. In other words, individuals with better fitness have larger chances to survive.

The GAs are adaptive heuristic search methods for solving the optimization problems according to the fitness functions [1, 12], especially for the problems that do not have precisely-defined solving methods. The flow chart of GA is shown in Figure 3.

The key points that a GA is successful or not are (1) how the fitness function is defined, (2) which representation of sequences are used, and (3) what genetic operators are applied. We focus our attention on how to define the better fitness function to satisfy our desire. The further discussion about the fitness function will be given in Section 4.

3 A New Method Based on the Lattice Model

Actually, protein folding based on the HP model does not work very well and cannot predict the structure of a target protein successfully. When using GA to solve this problem, the fitness function should be defined well. For the HP model, the fitness function considers the hydrophobicity only. Though the hydrophobicity of amino acids does provide the energy and the force of one protein to fold its structure, there are some other characteristics that affect the conformation. And more important, from a sequence to a 3D structure, several segments of the polypeptide chain will at first form to regu-

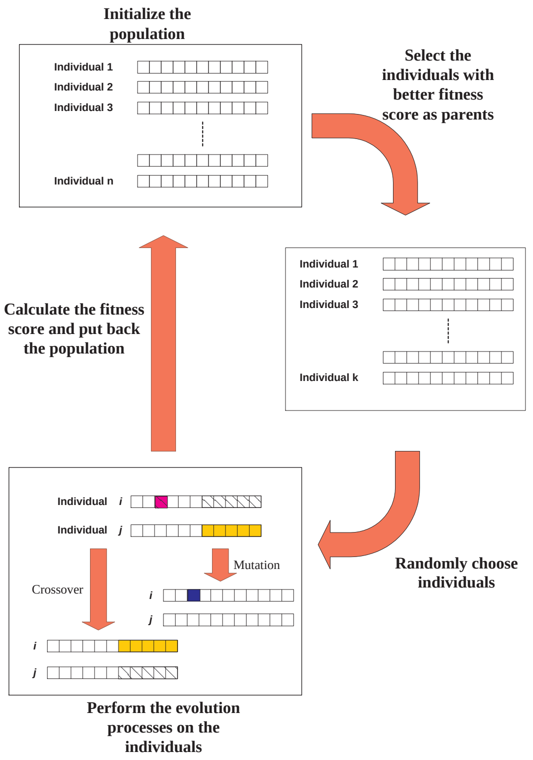
Initialize the population
Individual 1
Individual 2
Individual 3
Individual n
Select the individuals with better fitness score as parents
Individual 1
Individual 2
Individual 3
Individual k
Calculate the fitness score and put back the population
Individual i
Individual j
Mutation
Crossover
i
j
i
j
Randomly choose individuals
Perform the evolution processes on the individuals

Figure 3: The procedure of the genetic algorithm.

lar structure segments, the secondary structure elements (SSEs), and then these SSEs will fold to a 3D structure further. Besides, for those secondary structure elements, they do not change the basic conformations of secondary structure under the stable structure conformation. That is, while folding the sequence, we suggest that the secondary structure elements should be kept and considered. At the same time, some other characteristics, such as side chain charges and disulfide bonds, should also be considered.

As follows we show the overall steps of our algorithm (boldface parts) and give an example to illustrate the algorithm step by step.

Algorithm:

Homology Modeling in Folding Algorithm

Input: A target protein sequence S_1, where its secondary structure is known but its tertiary structure is unknown.

Let $\mathtt{S}_1$ = SSKCSRLKTFPQNLVQACVYHK and its secondary structure is

$SS_{\mathtt{S}_1}$ = ----SS--HHHHHHHH-SSTT-.

Output: The folding conformation of S_1.

Step 1: Perform sequence alignment on S_1 and each sequence in database, such as PDB, to find one template sequence S_2 that has the highest sequence similarity with S_1. If more than one sequence has the highest sequence similarity with S_1 at the same time, randomly choose one sequence from them as S_2.

Find $\mathtt{S}_2$ = SVYCSSLACSDHN, and the alignment result is as follows:

```
S1 = SSKCSRLKTFPQNLVQACVYHK
     |--||:|---------||--|:
S2 = SVYCSSL---------ACSDHN
```

where | represents that residues are identical, : represents that they are similar, and - represents a gap or that they are not similar.

Step 2: Find the structurally conserved regions, which have 50% or higher sequence similarity and the sequence alignment score is positive. Copy the coordinates of structurally conversed regions, except gaps, in the template structure S_2 to the target protein structure S_1.

```
S S K C S R L
| - - | | : |
S V Y C S S L
      ↑
```

Conserved Region

```
K T F P Q N L V Q
- - - - - - - - -
- - - - - - - - -
        ↑
```

Remaining Region

```
A C V Y H K
| | - - | :
A C S D H N
     ↑
```

Conserved Region

Step 3: For those remaining regions, *R-regions*, that are not structurally conserved regions, apply the folding algorithms on the 3D lattice model with GA to predict their folding conformations.

Then for the remaining region, R_1, we translate the residues to 1 or 0 according to their hydrophobicity, where 1 represents "hydrophobic" and 0 represents "hydrophilic". HP_{R_1} is the sequence of hydrophobicity of R_1.

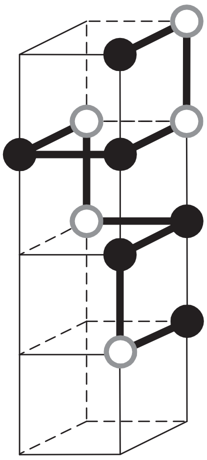

Figure 4: The folding conformation. The solid node represents “1” while the hollow one represents “0”.

$$\begin{aligned} \mathtt{R_1} &= \mathtt{LKTFPQNLVQA} \\ \mathtt{HP_{R_1}} &= \mathtt{10011001101} \end{aligned}$$

Besides, its corresponding secondary structure segment, SS_{R_1}, is shown as follows:

$$\begin{aligned} \mathtt{R_1} &= \mathtt{LKTFPQNLVQA} \\ \mathtt{SS_{R_1}} &= \mathtt{--HHHHHHHH-} \end{aligned}$$

Then, perform the GA to run the folding algorithm to predict its possible folding conformation. Figure 4 shows an example of folding conformation of R_1.

Step 4: For each R-region R_i, smooth the folding conformation such that it may be feasible as a real structure. Then search for the segments with same length as R_i in the set of proteins of known structure and apply the curve alignment between the folding structure and these segments of known structure [4]. Copy the coordinates from the most similar protein structure segment that gets the highest score.

After having the conformation, smooth the folding conformation as shown in Figure 5 and then find out the segment which is most similar to R_1 based on structure alignment with curve alignment [4].

Step 5: Construct the complete protein structure model of S_1.

Finally, we copy the coordinates of this segment and combine it with the conserved segments to form the final predicted conformation. Figure 6 shows the final predicted result.

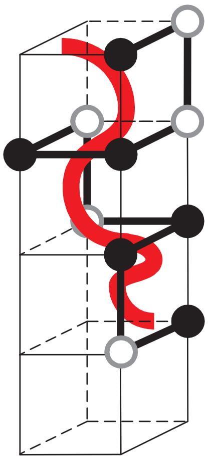

Figure 5: The smooth folding conformation.

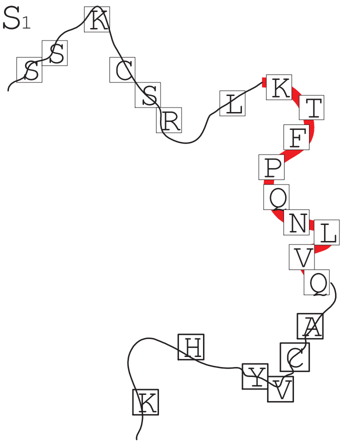

Figure 6: The final predicted conformation.

4 The Fitness Function

For one remaining region, there are one or more elements of secondary structure. Each element may be an α helix, β strand or loop. The scoring aspects are as follows.

1. Hydrophobic-hydrophobic interactions.

The hydrophobic-hydrophobic interactions are roughly the same except some conditions. For the HP model, the hydrophobic-hydrophobic interaction occurs between two amino acids that are not consecutive in sequence but are adjacent to each other at the grid points of the lattice model. However, we suggest that the interaction should be restricted to the pairs of amino acids belonging to the same secondary structure element.

The interactions between the amino acids in an α helix should occur only for those amino acids with distance of three or four amino acids in sequence. This is because that there are 3.6 residues per turn and those amino acids having such distance locate at the same side and then have chances to interact. Thus, for each residue A_j in one α helix segment, as long as its adjacent amino acids also belong to this segment, we hope they are A_{j-3}, A_{j-4}, A_{j+3}, or A_{j+4} and then the fitness score will be set higher.

On the other hand, for β sheet, the interactions should occur only between the amino acids that belong to different β strands and should not occur between those amino acids within the same β strand. Roughly, each strand of one β sheet is straight and not bended too much. Therefore, it is of rare occurrence to have interactions between amino acids belonging to one strand.

For the above reason, the interactions of these conditions are invalid and therefore will be penalized. That is, this condition will get a negative fitness score.

Then, we define the fitness function of this parameter as follows:

$$S_1 = (\ \#\ of\ H\text{–}H\ interactions\ )$$
$$+PenH \times (\ \#\ of\ invalid\ pairs\ ),$$

where $PenH$ is the penalty of invalid pairs and it is negative.

2. Charges of side chains.

For two adjacent amino acids at the grid points of the lattice model, if they have opposite charges, that is, one is positively charged residue and the other is negatively charged one, there is attraction between this pair of amino acids, and it will get a positive fitness score.

On the other hand, residues with the same charge should locate far from each other. For two adjacent amino acids at the grid points of the lattice model, if they have the same charge, there is anti-attraction between this pair of amino acids, and it will get a negative fitness score.

$$S_2 = BonC \times (\ \#\ of\ different\ charge\ pairs$$
$$-\#\ of\ same\ charge\ pairs\ ),$$

where $BonC$ is the bonus of charge pairs.

3. The segment conformation of α helix.

As described above, in α helix, there are 3.6 residues per turn. Thus the folding conformation of the α helix segment is regular, and each *pitch* of it should have the same length. Therefore, we define pitch P_k of one α helix is the distance from A_k to A_{k+4}. Next, we assume that the length of pitch P_1 is defined as the standard pitch length l and the pitch length of P_k is l_k, for k =1, 2, 3, ... Then we hope that the difference of each l_k and l is not too large. The total penalty of fitness score of this α helix is defined as follows.

$$S_3 = \begin{cases} \sum_{i=1}^{k} |l_k - l| \times PenA & if\ the\ element \\ & is\ an\ \alpha\ helix \\ 0 & otherwise \end{cases}$$

where $PenA$ is the penalty of deviations.

4. The segment conformation of β sheet.

For the reason mentioned before, we also hope that the folding conformation of each β strand tends to be straight and, in most conditions, close strands can be formed as a β sheet, parallel or antiparallel. Assume the length, number of amino acids, of one β strand is l and A_1 (A_l) is the first (last) residue of this β strand. And we define the distance from A_1 to A_l on the lattice to be str. We hope $|l - str|$ is not too large neither. Thus, the best conformation of this β strand is that $str = l$. Therefore, the penalty of fitness score of this β strand is defined as follows.

$$S_4 = \begin{cases} (l - str) \times PenB & if\ the\ element\ is \\ & a\ \beta\ strand \\ 0 & otherwise \end{cases}$$

where $PenB$ is the penalty of deviations.

5. The strong strength of disulfide bonds.

For one protein structure, the disulfide bond may be one kind of great strength to conformation. Though there are only two to three amino acids with sulfonium (S) and disulfide bonds are not very common, once that the bond forms, the strength is very strong and has quite large effect to the structure. Thus, for two adjacent amino acids at the grid points of the lattice model, if the disulfide bond can be formed between them, they will get more positive fitness score.

$$S_5 = BonS \times (\ \#\ of\ disulfide\ bonds\ ),$$

where $BonS$ is the bonus of disulfide bonds.

Table 2: Parameters used in GA.

Parameter	Value
Generation	500
Population Size	100
Crossover Rate	0.8
Mutation Rate	0.05

Table 3: Weights of features.

Feature	Fitness Value
PenH	-1
BonC	1
PenA	10
PenB	10
BonS	5

Therefore, the fitness score function given is as follows:

$$F = \sum_{i=1}^{n} R_i,$$

where R_i is the fitness score of the ith remaining region,

$$R_i = \sum_{j=1}^{m} E_{ij},$$

where E_{ij} is the fitness score of the jth element of R_i, and

$$E_{ij} = \sum_{k=1}^{5} S_{ijk},$$

where S_{ijk} is the S_k score of the jth element of R_i.

5 Experimental Results

In this section, we shall show some experimental results by comparing our model with the folding method based on the original HP model [4]. Most of the results show that the additional characteristics we use do improve the predicted structures.

Table 2 lists some parameters used in GA. And, Table 3 lists some coefficients in the fitness function.

Table 4 shows that we have fair improvements for the RMSDs even though the results are not good enough. Here, we notice that the first template protein 1CFD of target protein 1LIN is not a good template because that they have 100% sequence similarity, but the real structures of these two protein are quite different. Thus, lacking good template proteins may be one of the factors that affect the prediction results.

However, in Table 4, we also find that some prediction results of templates with lower sequence similarity are better than those with higher sequence similarity. As a result, we suggest that it is not always right to choose the templates with higher sequence similarity.

Table 5 demonstrates another case about template proteins. 1JYQ and 1JYU both have 90.4% sequence similarity with 1QG1. However, 1JYQ is a better template protein than 1JYU obviously.

Table 6 and Table 7 show another two cases with no distinguished improvements.

Table 8 shows the comparison between our method and the method of Dovier et al. [9]. The input information for their method includes the primary sequence and the secondary structure of the target protein. But, they do not use the information of other known structures in the database. For most cases, we improve the prediction results with the evaluation of RMSD values. Take 1E0M for example, the result of their method is 7.2 while the result of our method is 6.05 for the entire sequence with 37 amino acids. Furthermore, for the segment of 7th-22nd amino acids, the result of their method is 5.9 while the result of our method is 4.11. We may note, in passing, that the difference of the execution times between the two methods are very much. Actually, it is not so fair to compare them because these two methods do not apply the same strategy.

For most of these test cases, we find that the improvements are more explicit in α helix than in β sheet. The possible reason is that it is not easy for folding to make the strands close, so the conformation of β sheets are not predicted well. However, we can also make the β strands go straight possibly.

Table 4: Comparison of our method with the original method for target: 1LIN(146).

Template Protein	Sequence Similarity	RMSD(Old)	RMSD(Ours)
1CFD	100%	7.34	-
1TNW	69%	18.72	13.37
1IQ5	55%	15.15	9.18
1DTL	52.9%	10.22	7.48
5PAL	36.4%	12.18	8.43

Table 5: Comparison of our method with the original method for target: 1QG1(104).

Template Protein	Sequence Similarity	RMSD(Old)	RMSD(Ours)
1JYQ	90.4%	4.15	-
1JYU	90.4%	13.89	-
1SHA	46.7%	4.82	4.82
1SHD	45.2%	8.89	6.77
1PDR	24.4%	10.55	8.0

Table 6: Comparison of our method with the original method for target: 5CPV(108).

Template Protein	Sequence Similarity	RMSD(Old)	RMSD(Ours)
1CDP	100%	0.16	-
1BU3	82.7%	1.62	-
2PVB	76.4%	5.6	5.72
1C7V	32.5%	5.32	4.92

Table 7: Comparison of our method with the original method for target: 1SHD(101).

Template Protein	Sequence Similarity	RMSD(Old)	RMSD(Ours)
1SHA	98%	3.86	-
1JYQ	48.3%	9.98	7.43
1QG1	45.2%	9.04	7.35
1PDR	29.2%	8.52	9.68

Table 8: Comparison of our method with method of Dovier et al..

Target Protein		Method of Dovier et al.		Our Method		
Name	Length	Time(s)	RMSD	Time(s)	RMSD	Template(Similarity)
1VII	36	5460	10.0 7.5(4-32)	9.6	6.01 4.98	1E0M(11.6%)
1E0M	37	69420	7.2 5.9(7-22)	27.4	6.05 4.11	1LE3(27%)
1ED0	46	64200	7.3 3.7(7-30)	37.7	6.45 6.18	1LE3(5.7%)
1ENH	54	12240	10.0 5.4(8-52)	28.1	5.57 4.29	1CMG(36%)

6 Conclusion

We propose a method with GA based on the lattice model to predict the folding of proteins. Our algorithm is a hybrid method, consisting of homology modeling method and the folding algorithm. We consider the information of secondary structure, hydrophobicity, charges of side chains, and disulfide bonds, in the fitness function of GA for the prediction of protein folding. The results of our experiments show that these features indeed improve the prediction accuracy compared with the previous methods based on the HP model.

From the experimental results, we find that even though considering these features has improvements for the predictions, the results of predictions are still not very perfect. The possible reason is that the lattice model itself is not a very proper model except that it is easy to implement. Therefore, in the future, the folding of one protein can be based on some other lattice models. For example, folding on the 3D triangular lattice should be one more suitable model for protein structures because the folding angles of protein sequences are more and the folding pathways of proteins are more close to the real structures. Besides, more chemical characteristics of amino acids can be considered when predicting protein structures, such as the structures of residues.

Furthermore, the statistics of secondary structure elements in all proteins of one large protein database, the research of supersecondary structures and domains, and so on, also offer the information. And more and more databases of structure classification are developed. They also have large contribution for protein structure predictions. It seems reasonable to conclude that there are still some challenges about protein structure predictions.

References

[1] D. Beasley, D. Bull, and R. Martin, "An overview of genetic algorithms: Part2, research topics," *University Computing*, Vol. 15, No. 4, pp. 170–181, 1993.

[2] B. Berger and T. Leight, "Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete," *Journal of Computational Biology*, Vol. 5, No. 1, pp. 27–40, 1998.

[3] M. K. Campbell and S. O. Farrell, *Biochemistry.* Brooks Cole, fourth ed., 2002.

[4] Y. Y. Chen, C. B. Yang, and K. T. Tseng, "Prediction of protein structures based on curve alignment," *Proceedings of the 20th Workshop on Combinatorial Mathematics and Computation Theory*, Chiayi, Taiwan, pp. 34–44, 2003.

[5] R. S. Cheng, C. B. Yang, and K. T. Tseng, "Protein structure prediction based on secondary structure alignment," *Proceedings of 2004 Symposium on Digital Life and Internet Technologies(Abstract, full text in CD)*, Tainan, Taiwan, pp. 29–29, 2004.

[6] P. Crescenzi, D. Goldman, C. Capadimitriou, A. Piccolboni, and M. Yannakakis, "On the complexity of protein folding," *Journal of Computational Biology*, Vol. 5, No. 1, pp. 409–422, 1998.

[7] Y. Cui, R. S. Chen, and W. H. Wong, "Protein folding simulation with genetic algorithm and supersecondary structure constraints," *Proteins*, Vol. 31, pp. 247–257, 1998.

[8] K. A. Dill, "Theory for the folding and stability of globular proteins," *Biochemistry*, Vol. 24, pp. 1501–1509, 1985.

[9] A. Dovier, M. Burato, and F. Fogolari, "Using secondary structure information for protein folding in CLP(FD)," *Electronic Notes in Theoretical Computer Science* (M. Comini and M. Falaschi, eds.), Vol. 76, Elsevier, 2002.

[10] S. Duarte-Flores and J. Smith, "Study of fitness landscapes for the HP model of protein structure prediction," *In Proceedings of the Congress on Evolutionary Computation 2003 (CEC'2003)*, Vol. 1, Canberra, Australia, IEEE Service Center, pp. 2338–2345, 2003.

[11] A. Fraenkel, "Complexity of protein folding," *Bulletin of Mathematical Biology*, pp. 1199–1210, 1993.

[12] D. Goldberg, *Genetic Algorithms.* Addison Wesley Publishing, first ed., 1988.

[13] W. Hart and S. Istrail, "Robust proofs of NP-hardness for protein folding: general lattices and energy potentials," *Journal of Computational Biology*, Vol. 4, No. 1, pp. 1–22, 1997.

[14] J. Holland, "Adaptation in natural and artificial system." Technical Report. The University of Michigan Press, USA, 1975.

[15] T. Jiang, Q. Cui, G. Shi, and S. Ma, "Protein folding simulations of the hydrophobic-hydrophilic model by combining tabu search with genetic algorithm," *Journal of Chemical Physics*, Vol. 119, No. 8, pp. 4592–4596, 2003.

[16] N. Krasnogor, W. Hart, J. Smith, and D. Pelta, "Protein structure prediction with evolutionary algorithms," *In W. Banzhaf, J. Daida, A.E. Eiben, M.H. Garzon, V. Honavar, M. Jakaiela, and R.E. Smith, editors, GECCO-99: Proceedings of the Genetic and Evolutionary Computation Conference*, Morgan Kaufman, 1999.

[17] R. C. T. Lee, "Computational biology." http://www.csie.ncnu.edu.tw/, Department of Computer Science and Information Engineering, National Chi-Nan University, Taiwan, 2001.

[18] M. Milostan, P. Lukasiak, K. Dill, and J. Blazewicz, "A tabu search strategy for finding low energy structures of proteins in HP-model," *Proceedings of Seventh Annual International Conference on Research in Computational Molecular Biology*, Berlin, Germany, 2003.

[19] J. Setubal and J. Meidanis, *Introduction to Computational Molecular Biology.* PWS Publishing Company, Boston, second ed., 1997.

[20] R. Unger and J. Moult, "Finding the lowest free energy conformation of a protein is NP-hard problem: Proof and implications," *Bulletin of Mathematical Biology*, Vol. 55, No. 6, pp. 1183–1198, 1993.

[21] R. Unger and J. Moult, "Genetic algorithms for protein folding simulations," *Journal of Molecular Biology*, Vol. 231, No. 1, pp. 75–81, 1993.

[22] M. Waterman, *Introduction to Computational Biology: Maps, Sequences and Genomes.* Chapman and Hall, London: CRC Press, 1995.