Prediction of Protein Backbone Based on the Sliced Lattice Model *

Chia-Chang Wang, Chang-Biau Yang†, Hsing-Yen Ann and Hsiao-Yen Chang
Department of Computer Science and Engineering
National Sun Yat-sen University, Kaohsiung, Taiwan 80424
†cbyang@cse.nsysu.edu.tw

April 25, 2008

Abstract

In the past decades, a significant number of studies on the prediction of protein 3D tertiary structures have been extensively made. However, the folding rules, the core issue of protein structure prediction, still stay unsolved. Given a target protein with its primary amino acid sequence, the protein backbone structure prediction (PSP) problem is to construct the 3D coordinates of α-carbon atoms on the backbone. We propose a hybrid method by combining the homology model and the folding approach to solve the PSP problem. Our idea of protein folding is performed on the combined sliced cubic lattice, which mixes coarse lattices with fine lattices. Our computation is based on the HP (Hydrophobic-Polar) model, combined with the constraint of disulfide bonds. The folding is optimized by using the ant colony optimization (ACO) algorithm. Our experimental results show that our prediction accuracy is better than previous methods by the measurement of RMSD.

Key words: bioinformatics, protein backbone, folding, ant colony optimization, sliced lattice

1 Introduction

The driving force behind the structural researches is the basic tenet of molecular biology: to understand the biological reactions requires understanding the structure of the participating molecules. The conformations of proteins can be affected by mechanical supports, such as the enzymatic catalysis, the immune protection and generation and transmission of nerve impulses among many others [22]. Those are protein engineering, where the gene of the existing protein is mutated in order to change its function to design new proteins from combinational factors. These endeavors would have benefits if the structure of a protein can be predicted from its amino acid sequence. The primary sequence of a potential protein may now be determined from its DNA sequence. Thus the structure prediction methods would be extremely useful if the protein *tertiary structure* could be predicted directly from its *primary structure*.

A structure prediction method should be first tested by using sequences with known tertiary structures. While the known protein structure plays a role in folding pathway and in correcting the misfolds, it is generally assumed that the native structure is of the minimum free energy. For a pair of naturally related proteins with more than 33There is also evidence that the sequence identity of pairs of homologous proteins with similar remote structures is mostly less than 25% [18]. These remote homologous protein sequences cannot be found by general sequence alignment methods. Thus, searching for remote homologous is never an easy task.

There is a tradeoff for deciding how many factors should be considered in a structure prediction method. In this paper, we focus on the prediction of protein backbone, i.e. the 3D coordinates of α-carbon atoms. In our approach, we use a lattice model of protein folding based on the biological simplification [8], which divides the amino acids into two kinds, *hydrophilic* and *hydrophobic*. The lattice model provides valuable perception of the general complexity of protein backbone structure prediction problem. Protein backbone prediction has been proved to be NP-hard on the lattice model [2, 12].

Our approach is a hybrid method which combines the homology model with the folding approach. At the beginning, we use the relationship between similar sequences and structures to find a homologous protein structure as the template by sequence alignment. According to the sequence align-

*This research work was partially supported by the National Science Council of the Republic of China under contract NSC-93-2213-E-110-004.

ment, we copy the structure of the regions whose similarities are larger than the predefined threshold from the template protein. For those misaligned regions, the folding approach is invoked for constructing these regions. We solve the folding problem based on the HP model with the ant colony optimization (ACO) and make some modifications on the original lattice model. The main modification is that we partition each grid unit into more units, such as 5 or 10 units, which is called the *sliced lattice model*. We also consider the connection of disulfide bonds, which provide a significant influence on the framework [3, 5]. Our folding approach is more global to prevent falling into the local optimal than the original approach that folds only on the misaligned regions.

The organization of this paper is as follows. In Section 2, we will introduce some preliminaries of proteins and the previous works in the related research. Next, Section 3 presents our method for predicting protein backbone structures. There are some discussions on the sliced lattice model in Section 4. Section 5 shows the experimental results of our methods, which gets the better results compared to the previous works. The conclusion is given in Section 6.

2 Preliminaries

Amino acids are composed together to form a protein (polypeptide) chain. The repeating elements are amide planes containing peptide bonds. The central carbon (α-carbon) atoms of these amide planes can be writhed to create the three-dimensional conformations of the proteins. The protein structures can be described in four levels. In this paper, we focus on the prediction of the 3D tertiary structure of a protein backbone (the 3D coordinates of α-carbon atoms) from its primary amino acid sequence. *Primary structure* is the order in which the amino acids are covalently linked together by peptide bonds. The primary structure is the first step to specify the 3D structure of a protein. *Tertiary structure* includes the arrangement of side chains and the prosthetic groups (groups of atoms other than amino acids). The interaction between the side chains has large influence on the folding of proteins.

2.1 Folding Problem

Minimizing the total free energy is one approach for protein prediction to determine the positions of protein atoms. However, it is hard to find the accurate energy function in practice. Therefore, simplified models are used to approximate the energy function, even if the simplified models are complex. Here we consider a folding approach on the lattice model [8]. Amino acids can be *hydrophilic*, which means that they do well in water, or *hydrophobic*, which means that they do not. The adjacency in this model could be regarded as a grid in space. Roughly, conformations tend to have the hydrophobic amino acid residues inside and they are surrounded by hydrophilic amino acid residues.

Folding on the lattice model has been proved to be NP-hard [2, 12] on the *square lattice* as is the problem on the *cubic lattice*. Many complexity analyses and algorithms focused on Dill's hydrophobic-polar model [8], which is called the *HP model*, where P represents "polar" because of the nature of hydrophilic amino acid, and H means the hydrophobic amino acid. In this model, we reduce the twenty amino acids into two-alphabet {H, P} system as input. Some heuristics for fingding optimal HP structures have been proposed, especially evolutionary algorithms [6, 23, 17]. None of these methods for solving folding problem on the HP model with either heuristics or exact methods are able to robustly solve the problem with scale about 100 amino acids. Numerous studies focus on drawing near optimal solution, which could be approximate with $2h$ where h is the number of hydrophobic residues. Mauri et al. gave a 1/4 approximation on the square lattice [13]. Newman[15] presented a 1/3 approximation algorithm and showed the upper bound cannot be higher than 1/2. On cubic lattice model, Hart and Istrail[10] gave a 3/8 approximation algorithm and it can be 3/5 approximation with some modifications.

Figure 1 shows a grid for 9 residues. The hollow circles represent hydrophilic residues (P) and the solid circles represent hydrophobic ones (H). In Figure 1(a), the number of hydrophobic contacts is three, where (3, 4), (4, 5) and (8, 9) are hydrophobic neighbors. Figure 1(b) displays a fold that adds one more hydrophobic contact represented by the gray line connecting them.

2.2 The Evolutionary Algorithms

Many heuristic optimization methods have been proposed to solve the protein folding problem. *genetic algorithm* (GA) and *ant colony optimization* (ACO) algorithm are two of the mostly used methods on this problem.

GA is a randomly global search method which mimics the conditions of natural biological evolution [7]. GA operates on a population of possible solutions applying the principle of survival of the fittest one to produce better and better approximations to the solution. A new set of populations is created by selecting individuals according to their degree of fitness and breeding them with operators simulating the natural genetics. This process results in the individuals that are more suitable under the particular environment than the original ones.

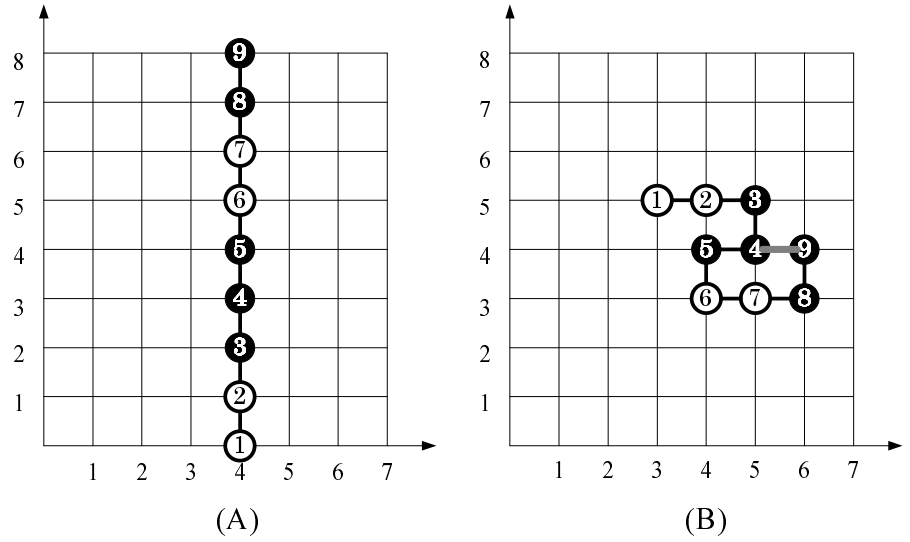

Figure 1: A protein folding on a 9 × 9 HP lattice model. (A) Zero H-bond. (B) One H-bond. The solid circles represent the H atoms and the hollow circles represents the P atoms. The gray lines represent the hydrogen bonds between two H atoms.

ACO is a population-based approach, which simulates the behavior of real ant colonies. This algorithm was proposed by Dorigo et al. [9] to solve various combinatorial optimization problems. The fundamental idea underlying ACO is the information (pheromone trails) exchange among ants. The algorithm is an iterative process, which is done through simple agents ("ants") to construct possible solutions. The solution is judged by the pheromone trails left by ants. Each of the ants chooses a path depending upon the smell, pheromone, left by other ants. While an ant moves at random and encounters a previously left trail, it can detect and select the path with high probability. It shows an autocatalytic behavior where the more the ants follow a trail, the more attractive the trail becomes for being followed. The pheromone trail shows a kind of information passing on the colonies. Previous researches showed that the ACO algorithms applied on the HP folding problem are more efficient than other heuristic algorithms[19, 20, 21].

3 A Prediction Method on the Sliced Lattice Model

A number of studies focused on the homology modeling in protein structure prediction (PSP). Though those approaches used in homology modeling seem more reasonable than ab initio methods, the prediction performance can be improved by considering more information. In this paper, we propose a hybrid method combining homology modeling and folding approach. Our method reduces the errors that result from taking a poor template for homology modeling. Our algorithm is as follows.

Algorithm:

Input: A target protein sequence S_1.

Output: The coordinates of the backbone conformation (α-carbon atoms) of S_1.

Step 1: Use the sequence alignment approach to find a template protein sequence S_2 which is similar to S_1.

Step 2: Based on the result of sequence alignment, find the structurally conserved regions in S_2 with more than 50% similarity. Copy the coordinates of those regions, except gaps, in S_2 to be the resultant structure of S_1.

Step 3: Apply the sliced HP folding lattice model on the whole sequence of S_1, then take the folding results as their conformations for the remaining regions U_i's that are not structurally conserved regions.

Step 4: For each U-region U_i found in Step 3, copy the coordinates of the folding conformation directly.

Step 5: Merge the regions obtained by the folding approach and the structurally conserved regions got from sequence alignment. Construct the coordinates of the protein structure backbone of S_1.

In the Step 3, we use ACO algorithm to approximate the folding problem. The details are given as follows.

Step 3.1 Initialize the pheromone trails of ACO and other related parameters. The folding path is encoded into related directions, straight (S), left (L), right (R), up (U) and down (D).

Step 3.2 All ants construct the probabilities of all solutions by the pheromone.

Step 3.3 Calculate the score of each solution with Equation (1) and Equation (2), explained later.

Step 3.4 Based on the scores of the solutions, update the pheromone trails.

Step 3.5 If the best solution does not change for several iterations, terminate the algorithm. Otherwise, go to Step 3.2.

Suppose the target protein is a sequence of length L. In the initial stage, $L-2$ local structure motifs, each of length 3, are obtained by sliding the sequence with window size 3. In the construction phase of ACO, each of the ants randomly selects a starting point by choosing a sequence position between 1 and $L-1$ with the uniform distribution. From the starting position, the target sequence is folded until the ending position. When extending a conformation $p_k...p_i$ to p_{i+1} during the construction phase, the relative direction is determined by the pheromone and heuristic values according to the following probability:

$$p_{i,d} = \frac{[\rho_{i,d}]^{\alpha}[\varphi_{i,d}]^{\beta}}{\sum_{e\in\{S,L,R,U,D\}}[\rho_{i,e}]^{\alpha}[\varphi_{i,e}]^{\beta}},$$

where $\rho_{i,d}$ is the pheromone value of the position i with direction d and denotes the previous experience, $\varphi_{i,d}$ is the heuristic value of the position i with direction d and denotes to the current energy of the relative direction. After the construction phase, the ants update the pheromone values in the following way:

$$\rho_{i,d} \leftarrow (1-\mu)\rho_{i,d} + \Delta_{i,d,c},$$

where $0 \le \mu \le 1$ is the pheromone persistent rate and $\Delta_{i,d,c}$ is the amount of pheromone gathered from the ant c in the previous iteration.

The scoring formula used in the cubic lattice model is not suitable in the sliced lattice model. From the statistical data of the protein data in PDB, the distance between two consecutive residues is about $3\mathring{A}$ to $6\mathring{A}$. Therefore, when the distances between two residues that are not consecutive are in the range, we consider that there should be a bond between them. The function for evaluating the score of the new folding model is as follows:

$$Score = (\sum_{a=1}^{n}\sum_{b=1}^{a}\min\{\frac{D_{DLL}^2}{D_{a,b}^2}, \frac{D_{a,b}^2}{D_{DLL}^2}\}) \times \rho - \mu \quad (1)$$

where D_{DLL} is the user-defined lattice length of the folding model, $D_{a,b}$ is the distance between the residues a and b, ρ is the user-defined score for every bond and μ is the penalty function if the distance between two residues is less than one predefined distance τ. The definition of μ is as follows:

$$\mu = \begin{cases} \phi_p, & \text{if } D_{a,b} \le \tau, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

where ϕ_p is the user-defined penalty and τ is the user-defined distance. The function μ denotes that the residues of the result folding with impractical distance in the native protein structures.

4 Implementation Details

In the following, some examples are given to illustrate the details of our algorithm. Let us consider two highly similar protein sequences S_1 and S_2, where S_1 is the target sequence and S_2 is the template sequence. After sequence alignment is performed on S_1 and S_2, the target sequence S_1 is divided into structurally conserved regions and U-regions. Then we copy the coordinates of the structurally conserved regions from S_2 to S_1. Users may adjust the threshold of the score for choosing the structurally conserved regions here to get better results.

Next, convert S_1 into a 0/1 sequence, which maps to the hydrophobic and polar residues in the HP model. And then apply the folding approach to position the residues. In this phase, the *cubic lattice* and the *triangular, lattice* (*face centered cubic*, FCC) models [11, 1] are usually used to approximate the native conformation of proteins. The real structures rarely locate on the positions that are exactly on the grid, and more positions could be chosen on the triangular lattice model than the cubic lattice model. As the experimental results shown in Table 1, the prediction accuracy on the triangular lattice model is about 10% to 40% better than the cubic lattice model. However, the RMSD of the triangular model is still not good enough. Therefore, we modify the lattice model to the sliced lattice model.

Our basic concept is to slice the original lattice into multiple little lattices. On the sliced lattice model, we have more choices to put the residues. Thus, the conformation is closer to the real structure. Here, we only consider the sliced cubic lattice model. Because the directions of putting the next residue in the triangular lattice are more than

Table 1: Comparison of the folding results obtained by various models, measured by RMSD. "Cubic": cubic lattice; "Tri.": triangular lattice; "*L*": basic sliced cubic lattice; "*OL*": combined sliced cubic lattice; "S. tri.": sliced triangular lattice. ACO is used in all folding optimizations, except that "*OL GA*" uses GA.

	ID	Length	Cubic	Tri.	*L*	*OL*	S. tri.	*OL GA*
1	1B3A	67	10.29	9.27	9.13	8.26	8.87	8.79
2	1AZG	72	10.10	9.18	8.65	8.07	7.85	8.85
3	1BRH	108	13.21	10.23	9.97	9.87	9.62	10.50
4	5CPV	108	14.55	8.98	8.23	8.01	7.86	8.61
5	1B1U	117	15.71	11.44	10.30	9.81	9.77	10.39
6	1DTL	149	17.60	11.23	9.77	9.51	10.99	10.35
7	1A6N	151	15.48	12.12	10.48	10.44	11.74	10.48
8	118L	162	14.80	12.75	11.63	11.20	11.90	10.94
9	1FW9	164	17.36	12.79	11.70	11.49	11.70	12.18
10	1B8K	180	17.80	15.00	14.40	13.49	13.25	14.13
11	1BGS	199	18.53	13.81	11.54	11.34	11.93	11.90
12	12CA	255	21.29	15.41	13.70	13.12	13.50	13.24
13	1ACD	260	20.54	14.26	13.16	12.77	14.64	13.77
14	1BZM	260	21.76	15.71	14.37	14.11	14.09	14.71
15	1BC1	318	21.83	15.52	14.74	13.64	13.97	14.52
16	1BOS	345	24.80	17.74	15.62	14.56	14.90	15.09
17	1B6A	355	24.13	19.56	15.23	14.07	15.12	15.46
18	1C16	359	23.74	19.34	17.55	16.69	16.22	16.69
19	1G9Q	407	29.53	18.20	16.43	16.08	16.32	17.31
20	1IM8	436	24.61	19.99	17.14	16.74	16.79	16.19

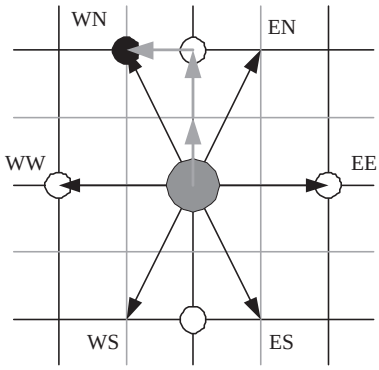

Figure 2: An example of converting the directions of the 2D triangular lattice to the 2D sliced cubic lattice model. In this example, we slice the original lattice into two little lattices, and N, N, W (gray arrows), may simulate the direction WN of the 2D triangular lattice model.

those in the cubic lattice, the process becomes more time-consuming to reach the convergent condition when we solve the folding problem by evolutionary algorithms, and the situation get worse when the lattice is sliced more. Furthermore, as shown in Figure 2, the directions of the triangular lattice can be represented approximately when the cubic lattice is sliced appropriately. As shown in Table 1, the folding results obtained on the sliced model are about 10% better than those obtained on the triangular lattice model.

Based on the idea of slicing the lattice, there is a further modified model. In origin, the idea of slicing the lattice is used to improve the unsuitable layout of the original cubic lattice model. We combine the original lattice and the sliced lattice, called the *combined sliced lattice*. The way of representing a position that alters into one original lattice plus multiple little lattices is shown in Figure 3. From the viewpoint of algorithms, the combined sliced lattice model can be regarded as giving small amount of shift to overcome the limitation of the original lattice model.

Though the percentage of the improvement between the basic and combined sliced lattice models is less than those compared with others, this is especially noteworthy in the case of using only the hydrophobicity of the amino acid sequence. That might be a bottleneck of the folding approach.

As we can see, the sliced cubic lattice model outperforms the sliced triangular lattice model. Therefore, considering the cost of time and performance, we take the sliced cubic lattice model as our folding model. Comparing to GA, the ACO algorithm has better results and is more efficient based on the score function of the HP model [19, 20, 21].

Table 2 shows the experimental results of folding with and without disulfide bond constraint. All of the protein data in Table 2 contain disulfide bonds in their real tertiary structures. The improvement is about 10% to 50% when the constraint of disulfide bonds is involved. The results represent that, if the disulfide bonds are ignored, some structural information may lose. Thus, consideration for both hydrogen bonds and disulfide bonds seems to be a more reasonable approach.

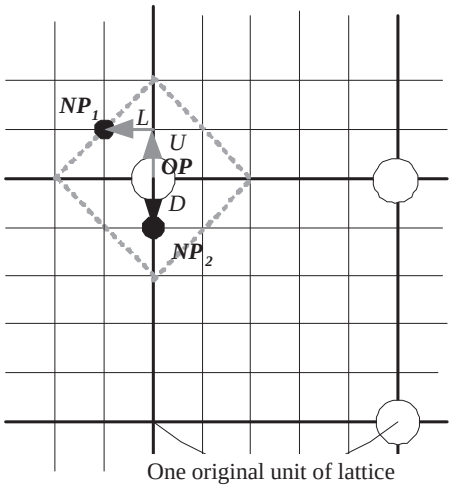

Figure 3: An example of the combined sliced lattice model. "OP" denotes the original position. "NP_1" denotes the position that moves one "U" and one "L" from "OP".

Table 2: Comparison of the folding with and without the constraint of disulfide bonds. The experiments are simulated on the sliced cubic lattice model. "No disulfide" and "Disulfide" denote the folding results without and with the constraint of disulfide bonds, respectively.

	Num.	Length	No disulfide	Disulfide	Im.(%)
1	1BXJ	29	5.5429	4.7539	16.60
2	1E0F	36	10.3608	9.0476	14.51
3	1CR8	42	7.0111	5.9700	17.44
4	1BMR	67	8.9200	7.7000	15.84
5	1DOY	96	9.7956	8.0148	22.22
6	1TGJ	112	16.0995	14.4165	11.67
7	1B7S	130	12.7071	9.7681	30.01
8	1EXT	162	20.8476	17.9490	16.15
9	1BNL	191	14.5605	11.0170	32.16
10	1YNA	193	16.4267	13.0962	25.43
11	1B6U	199	17.0831	14.3846	18.76
12	1SAC	204	14.8066	12.1942	21.42
13	1CFB	205	16.5892	14.9328	11.09
14	1BOY	211	18.9438	16.8339	12.53
15	1CHG	227	15.6382	12.3747	26.37
16	2CHA	236	16.1777	13.2871	21.76
17	1JT1	262	18.3408	12.6064	45.49
18	1H8D	288	19.7429	17.1597	15.05
19	1D3D	289	19.7625	14.9631	32.08
20	1BP3	383	21.8088	16.9998	28.29

After folding, we have the folding conformation to compare with the real protein structures of PDB. First, the B-spline curve technique is applied to transform the folding conformation to smooth curves. Then, we take the folding conformation after applying B-spline curves as the coordinates of those U-regions.

Finally, we merge the structures from the homology modeling and the conformation from the folding approach together, then the prediction of the the target sequence is complete.

5 Experimental Results

In this section, we will show our experimental results and discuss their performance. We use the protein data published on PDB published on April 17, 2005 as our testing database. There are 20380 proteins in the database. All the simulations are executed under a PC with AMD Athlon 1.67GHz processor and 512 MB RAM. There are some redundant data in PDB, which means that some identical sequences map to different PDB codes, may influence the accuracy of our method. Therefore, we take those similar protein data away from our template pool. Only the sequences that sequence identities compared with the target sequence are lower than 90% are hold before we apply our method on each protein. All identities are calculated by the program Blastp [14].

We use the corresponding protein sequences of CASP6 in PDB as our testing data set to compare with the Chen's method[4]. The comparison results are shown in Table 3. The codes of CASP6 are from T0282 to T0196, except T0265, T0245, T0244, T0243, T0241, T0222, T0221, T0218, T0206, since there is no corresponding protein data in the PDB. And there is no "T0225" code in CASP6. Totally, there are 77 proteins in our test. Though we have the presumption that the sequence identity of the template sequences should be over 30%, however, the selected templates of 28 proteins are below 30% identity, because there does not exist similar protein sequence for these 28 proteins in PDB. The proteins with the highest similarity are selected to be our template. Even the other 47 template proteins that fit the presumption of 30% identity, the identities are still low mostly. Since we select templates based on the result of Blastp, we take no template for the other two proteins T0272 and T0257 because there is no output from Blastp. The conformations of these two protein sequences are predicted with the folding approach only.

In Table 3, it is marked by an underline if our result is inferior to Chen's result. We win in 48 proteins and lose in 29 proteins. In summary, we have some improvement compared to Chen's results.

Table 3: Comparison with Chen's[4] method on protein data of CASP6. "L" denotes the length of the protein. "F" denotes the folding length. "I" and "S" denote the sequence identities and the sequence similarities between the target protein and the template protein, respectively. "R_{Chen}"and "R_{ours}" denote the RMSD of Chen's method and our method, respectively. "pH" and "T" denote the acid value and the temperature in Kelvin of environment, respectively.

Protein		Experimental data							Environ.	
CASP6	PDB	Temp.	L	F	I (%)	S (%)	R_{Chen}	R_{ours}	pH	T
T0282	1XFK	5CEV	332	125	7	6	16.28	12.08	8.50	113
T0281	1WHZ	1WG8	67	67	11	19	31.82	6.67	8.00	100
T0280	1WD5	1FW8	205	205	35	9	17.06	13.70	6.10	120
T0279	1WD7	1SRY	254	254	40	7	22.60	13.09	–	–
T0278	1WIY	1N97	389	0	82	97	0.91	**13.46**	4.20	160
T0277	1WTY	1JOG	115	23	33	23	7.22	**8.34**	8.00	100
T0276	1WKC	1SOU	166	11	41	34	4.13	**9.44**	6.50	298
T0275	1WJG	1MJH	135	15	26	20	7.27	**11.28**	7.00	100
T0274	1WGB	1I0S	156	56	26	12	15.60	**16.18**	8.40	100
T0273	1WDJ	1G59	186	186	28	6	28.45	10.59	6.70	100
T0272	1WJ9	–	188	188	–	–	13.24	**16.42**	–	–
T0271	1VGG	1RLH	159	42	49	50	10.35	**13.49**	5.00	150
T0270	1VDH	1T0T	248	9	54	52	3.79	**13.23**	–	100
T0269	1VGS	1QQ2	237	161	39	24	10.74	**12.08**	6.00	98
T0268	1WG8	1KA9	281	281	34	5	13.01	12.73	7.50	290
T0267	1WK4	1VHS	173	91	37	8	10.48	10.01	–	–
T0266	1WDV	1DBX	148	21	21	15	7.88	4.06	–	100
T0264	1WDE	1VHV	284	171	30	12	30.76	14.57	–	–
T0263	1WD6	1CI9	96	96	34	11	14.49	10.36	7.50	110
T0262	1WFX	1KO7	177	177	33	7	16.08	**17.44**	7.50	100
T0261	1WDI	1VKY	294	102	42	51	30.12	15.97	–	100
T0260	1T33	1PB6	220	124	30	10	12.20	12.16	7.00	200
T0259	1T57	1VP8	178	35	50	49	5.52	**18.08**	7.50	100
T0258	1T5B	1V4B	199	77	81	91	8.62	**12.73**	–	–
T0257	1Y12	–	154	154	–	–	15.81	**15.93**	–	–
T0256	1XQA	1NPB	223	204	26	13	27.61	9.13	6.50	113
T0255	1U69	1U7I	152	75	33	18	8.95	**11.98**	7.00	100
T0254	1U9D	1SF9	120	46	61	15	11.09	**11.73**	4.60	100
T0253	1Y0B	1P4A	191	10	25	14	12.93	10.50	7.50	100
T0252	1U60	1MKI	309	140	29	24	18.03	10.88	7.50	100
T0251	1XG8	1O1H	108	108	47	9	18.39	7.76	7.00	100
T0250	1XV2	1X8M	227	18	24	6	13.75	**15.23**	7.50	100
T0249	1T6S	1AJ4	162	52	32	8	17.96	**20.17**	6.70	303
T0248	1TD6	7REQ	286	286	25	4	24.59	10.90	7.50	95
T0247	1VLO	1WOS	354	80	39	28	7.37	7.21	4.25	100
T0246	1VLC	1CM7	362	30	54	56	2.91	**6.30**	7.50	277
T0242	2BLK	1X9N	113	113	30	10	22.48	9.23	4.80	100
T0240	1U07	1T3C	178	178	32	13	18.09	10.13	5.60	100
T0239	1RKI	1AWC	101	101	55	12	12.62	8.33	9.00	100
T0238	1W33	1W3Z	181	79	88	99	1.61	**17.97**	6.00	100
T0237	1W81	1HN6	364	364	27	23	19.47	15.17	3.40	303
T0236	1W53	1T80	85	85	26	14	18.49	7.66	7.50	100
T0235	1VJV	1NBF	359	112	24	7	23.62	14.47	5.50	100
T0234	1VL7	1WNE	136	136	31	9	16.78	11.98	6.00	100
T0233	1VQU	1V8G	332	65	38	33	13.70	12.24	5.80	100
T0232	1ZGM	1R2T	232	58	21	5	14.92	12.89	8.50	85
T0231	1VKK	1V6F	138	58	80	81	1.88	**10.49**	6.00	298
T0230	1UWD	1QQP	102	102	26	12	30.16	8.69	7.50	290
T0229	1VLA	1ML8	135	50	37	25	8.44	**11.90**	4.60	100
T0228	1VLP	1YIR	409	143	33	22	23.22	14.43	7.50	100
T0227	1WK2	1SRY	82	82	29	15	28.56	7.98	–	–
T0226	1WIW	1MZJ	278	33	38	5	12.81	**14.98**	7.40	100
T0224	1RHX	1W63	87	87	36	13	44.45	7.82	7.00	100
T0223	1VKW	1IK0	213	204	31	7	48.51	10.81	6.00	298
T0220	1VLL	1X7D	321	321	31	22	24.55	13.02	6.25	100
T0219	1VRB	1H2N	297	48	28	5	21.93	15.70	7.50	100
T0217	1VPQ	1VPY	260	119	30	19	8.22	**11.00**	4.50	100
T0216	1VL4	1VPB	435	106	25	17	17.43	8.77	7.00	100
T0215	1X9B	1P4J	53	53	30	19	26.24	3.69	6.70	293
T0214	1S04	1XNE	110	15	31	25	5.86	**8.09**	6.50	298
T0213	1TE7	1UN7	103	103	26	12	15.59	8.73	6.50	120
T0212	1TZA	1XVS	257	207	52	51	43.14	13.36	6.50	100
T0211	1XPW	1GQP	143	0	26	9	8.73	**10.52**	4.60	100
T0210	1YUD	1XE7	152	26	24	13	9.09	**15.17**	5.60	100
T0209	1XQB	1W1Z	183	0	37	8	13.98	13.23	8.50	100
T0208	1TZ9	1KMZ	334	334	35	4	63.52	13.23	6.00	100
T0207	1TTZ	1WJK	74	14	52	14	9.51	7.42	7.00	298
T0205	1VM0	1S0G	188	188	26	12	31.41	9.80	6.00	100
T0204	1VKV	1HXQ	298	75	21	9	12.05	11.74	–	113
T0203	1VKP	1XKN	358	119	33	29	15.45	13.26	5.60	100
T0202	1SUW	1U0T	249	42	28	20	11.42	**13.03**	6.00	100
T0201	1S12	1DY6	94	94	42	13	10.06	9.18	8.50	291
T0200	1T70	1T71	255	35	35	28	11.73	10.72	6.50	90
T0199	1STZ	1MX0	323	0	25	5	18.99	**19.50**	5.50	100
T0198	1SUM	1T8B	225	42	26	25	3.87	**12.88**	7.50	100
T0197	1YEM	1MLW	163	163	42	7	11.74	10.62	6.00	100
T0196	1XE1	1WB3	89	89	38	17	40.69	8.36	–	100

Table 4 shows the experimental results of the data set, proposed by Palu et al.[16], and illustrates the comparison. Palu's method uses molecular dynamics simulation on the FCC lattice model. The simulations of Palu are done under a 1.533 GHz AMD Athlon processor. The results are also measured by RMSD. Our results do not only obtain better results, but also spend much less time. In Table 4, we give two results for "1YPA". The first result that takes "2CI2" as the template returns a much better result than theirs. However, the homology of the target sequence and the template sequence may affect much. Therefore, the second result that takes "1TIN" as the template represents an example with a large amount of folding. Though the accuracy decreases a little than the first result, the RMSD still remains much closer to the real structure than theirs. In our experiments, we always select a protein with the highest identity to the target protein as our template. However, protein homology was not considered in Palu's method, which may be the reason of getting low accuracy. Homology modeling seems very important in the protein structure prediction.

6 Conclusion

Protein structure prediction can be roughly divided into two approaches, *ab initio* methods and homology modeling. Generally, homology modeling provides a better prediction accuracy because of the increasing protein data. However, it still depends deeply on the choice of a suitable template. In the paper, we propose a hybrid method, the combination of homology modeling and protein folding optimization. Our method decreases the influence of unsuitable templates. Our folding optimization is performed on the combined sliced cubic lattice, which mixes coarse lattices with fine lattices. The original lattice models cannot express the protein structures well; our combined sliced lattice model is shown closer to the real structures. So that we can use the folding approach on those dissimilar regions stand independently and bypass the curve fitting techniques. At the same time, the benefits of the homology modeling and folding approach still hold.

Though we decrease the influence of selecting an unsuitable template, our results are still affected by that the sequence similarity does not reflect structure homology. Therefore, how to find out whether our results are affected by unsuitable template or not is still a problem. The structural properties that we consider are the hydrophobicity and the links of disulfide bonds. Other stereochemistry or biochemistry properties may be considered to adjust the errors after our predictions finish.

In addition to the problems we have mentioned, some problems still have to be considered, such as the measurement of accuracy. RMSD is usually used to measure the accuracy of protein structure prediction. The predictions of the secondary structures are also one point of view from their functional expressions. It is believed that there are still many challenges about protein structure prediction.

References

[1] R. Agarwala, S. Batzoglou, and V. Dancik, "Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model," *Journal of Computational Biology*, Vol. 4, No. 3, pp. 275–296, 1997.

[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] S. F. Betz, "Disulfide bonds and the stability of globular proteins.," *Protein Science*, Vol. 2, pp. 1551–1558, 1993.

[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] C. C. Chuang, C. Y. Chen, J. M. Yang, P. C. Lyu, and J. K. Hwang, "Relationship between protein structures and disulfide-bonding patterns.," *Proteins: Structure, Function, and Bioinformatics*, Vol. 53, pp. 1–5, 2003.

[6] T. Dansdekar and P. Argos, "Folding the mainchain of small proteins with the genetic algorithm," *Journal of Molecular Biology*, Vol. 236, No. 3, pp. 844–861, 1994.

[7] L. Davis, *Genetic Algorithms and Simulated Annealing*. Morgan Kaufmann Publishers Inc., 1987.

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

[9] M. Dorigo, V. Maniezzo, and A. Colorni, "The ant system: Optimization by a colony of cooperating agents," *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, Vol. 26, No. 1, pp. 29–42, 1996.

[10] W. Hart and S. Istrail, "Fast protein folding in the hydrophobic-hydrophilic model within three-eights of optimal," *Journal of Computational Biology*, Vol. 3, No. 1, pp. 53–96, 1996.

Table 4: Comparison of our method with Palu[16]. "Temp." denotes the template used in our prediction. "L" denotes the length of proteins. "RMSD" denote the results of prediction with templates. "Iden." and "Simi." denote the sequence identities and similarities, respectively.

Target		Method$_{Palu}$		Our Method				
Name	L	Time(s)	RMSD	Time(s)	RMSD	Temp.	Iden.	Simi.
1ZDD	34	1045	4.0	2.703	3.12	1Q2N	0.66	0.61
1VII	36	14280	7.4	3.047	**12.59**	1UNC	0.74	0.70
1EOM	37	36000	3.4	3.093	**17.41**	1I5H	0.47	0.49
1EDO	46	36000	7.2	3.656	**11.54**	1NBL	0.55	0.56
2IGD	61	174960	11.5	7.469	8.01	1MVK	0.79	0.74
1YPA	64	420840	9.4	6.687	0.34	2CI2	0.79	0.97
				6.687	1.83	1TIN	0.31	0.40

[11] W. Hart and S. Istrail, "Lattice and off-lattice side chain models of protein folding: Linear time structure prediction better than 86% of optimal," *Journal of Computational Biology*, Vol. 4, No. 3, pp. 241–259, 1997.

[12] 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.

[13] G. Mauri, A. Piccolboni, and G. Pavesi, "Approximation algorithms for protein folding prediction.," *Proceedings of the 10th Annual Symposium on Discrete Algorithms (SODA)*, San Antonio, USA, pp. 945–946, 1999.

[14] N. I. H. (National Institutes of Health), "NCBI (National Center for Biotechnology Information)." http://www.ncbi.nlm.nih.gov/.

[15] A. Newman, "A new algorithm for protein folding in the HP model," *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms*, San Francisco, California, pp. 876–884, 2002.

[16] A. D. Palu, A. Dovier, and F. Fogolari, "Constraint logic programming approach to protein structure prediction.," *BMC Bioinformatics*, Vol. 5, pp. 186–198, 2004.

[17] A. A. Rabow and H. A. Scheraga, "Improved genetic algorithm for the protein folding problem by use of a cartesian combination operator.," *Protein Science*, Vol. 5, No. 9, pp. 1800–1815, 1996.

[18] B. Rost, "Protein structures sustain evolutionary drift.," *Folding & Design*, Vol. 2, pp. 19–24, 1997.

[19] A. Shmygelska, R. Hernandez, and H. H. Hoos, "An ant colony optimization algorithm for the 2d hp protein folding problem.," *Proceedings of the 3rd International Workshop on Ant Algorithms*, pp. 40–52, 2002.

[20] A. Shmygelska and H. H. Hoos, "An improved ant colony optimisation algorithm for the 2d hp protein folding problem.," *Proceedings of the 16th Canadian Conference on Artificial Intelligence*, pp. 400–417, 2003.

[21] A. Shmygelska and H. H. Hoos, "An ant colony optimisation algorithm for the 2d and 3d hydrophobic polar protein folding problem," *BMC Bioinformatics*, Vol. 6, No. 30, 2005.

[22] M. L. Teodoro, G. N. P. Jr, and L. E. Kavraki, "A dimensional reduction approach to modeling protein flexibility.," *Proceedings of the sixth annual international conference on Computational biology*, Washington, DC, USA, pp. 299–308, 2002.

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