Protein Structure Prediction Based on Improved Genetic Algorithm

Jiaxi Liu

***Abstract*—The prediction of protein three-dimensional structure from amino acid sequence has been a challenge problem in bioinformatics, owing to the many potential applications for robust protein structure prediction methods. Protein structure prediction is essential to bioscience, and its research results are important for other research areas. Methods for the prediction and design of protein structures have advanced dramatically. The prediction of protein structure based on average hydrophobic values is discussed and an improved genetic algorithm is proposed to solve the optimization problem of hydrophobic protein structure prediction. An adjustment operator is designed with the average hydrophobic value to prevent the overlapping of amino acid positions. Finally, some numerical experiments are conducted to verify the feasibility and effectiveness of the proposed algorithm by comparing with the traditional HNN algorithm.**

***Index Terms*—Protein structure prediction, genetic algorithm, amino acid hydrophobicity.**

I. Introduction

With the development of science and technology, pure biology gradually fails to meet further research demands. A burgeoning cross discipline, bioinformatics therefore was born: on the basis of biodata, bioinformatics combines statistics, computer science, bioscience, and several subjects together, efficiently extracting and organizing a huge number of biodata. Proteins are indispensable to any organism, their roles are different but important in vivo, for example, transportation and storage, immunity, catalysis, memory identification, etc. Thus, exploring the pattern of their functions is momentous in life sciences. Protein structure prediction can not only promote the progress of bioscience, but also apply its research result into medicine, food, husbandry, industrial manufacture, environmental protection and so on.

Understanding the latent relationship between amino acid sequence and structure of protein is an essential problem in structural bioinformatics [1]. Protein secondary structure is the local special conformation of amino acid residue in polypeptide chains; it has eight different types: α-helix (H), β-bridge (B), folding (F), helix -3(G), helix -5(I), Turn (T), strand (S) and Loop (L). Each one of the secondary structural types is affected by the interaction of both locality and long range of amino acid residue in protein chains. The main task of protein secondary structure prediction is to make an amino acid sequence that contains twenty types of amino acids A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y be mapped to its corresponding secondary structural sequence. Protein secondary structure prediction is also related to protein folding prediction and protein tertiary structure prediction. Particularly, in addition to the information of protein secondary structure prediction being beneficial to determine protein tertiary structure prediction, it can be exerted in protein function prediction for interacted prediction.

The earlier research on protein secondary structure prediction primarily focused on the three coarse-grained types of secondary structure prediction, i.e., roughly categorize the eight types of structures into three types: helix, folding, and strand; the representative algorithms include PHD [2], PSIPRED [3], Jpred [4], etc. Document [5] advances SCPRED method which uses linear sequence of second order structural elements to describe characteristics of protein structure. Acquiring nine main protein characteristics through feature selection algorithm, SCPRED method substantially increases the accuracy of prediction of sequence structure with low similarity, reaching approximately 80% [6]. On the top of B-RNN, document [7] proposes prediction algorithm SSpro8; however, considering SSpro8 cannot model the dependency between the secondary structure types of adjacent residue, document [8] proposes to adopt conditioned neural field in order to establish a pattern of secondary structure prediction. Document [9] combines position-specific scoring matrix, amino acid sequence, and secondary structure sequence to extract characteristics; the accuracy of its prediction are 94.5%, 91.1%, and 83.0% respectively by testing data as well as FC699, 1189, 640, and 25PDB.

The majority of documents listed above predict protein structure from the similarity of its sequence rather than the chemical and physical properties of protein itself. This article will take factors like the reciprocity of hydrophobic amino acid, covalent bonds, and Van der Waals' force into account, which impact the stability of protein structure. And on the basis of the potential energy of VDW, modified genetic algorithm will be used to calculate protein secondary structure.

II. Related Work

One of the key challenges in protein science is determining three dimensional structure from amino acid sequence. Although experimental methods for determining protein structures are providing high resolution structures, they cannot keep the pace at which amino acid sequences are resolved on the scale of entire genomes.

Early methods of secondary structure prediction, introduced in the 1960s and early 1970s, focused on identifying likely alpha helices based on helix-coil transition

Manuscript received March 21, 2020; revised July 13, 2020.

Jiaxi Liu is with the Masters School, 49 Clinton Avenue, Dobbs Ferry, NY 10522, USA (e-mail: jiaxi.liu@mastersny.org).

doi: 10.18178/ijesd.2020.11.9.1289

models. More accurate predictions were introduced in the 1970s such as beta sheets, relying on statistical assessments with probability parameters derived from known solved structures. These methods, are typically at most about 60-65% accurate. The evolutionary conservation of secondary structures can be exploited by processing many homologous sequences in a multiple sequence alignment, by calculating the net secondary structure propensity of amino acids. Larger databases of known protein structures have been established and modern machine learning methods such as neural nets and support vector machines can achieve up to 80% overall accuracy based on the databases. The theoretical upper limit of accuracy is around 90%, due to idiosyncrasies in DSSP assignment near the ends of secondary structures. Limitations are also imposed by secondary structure prediction's inability to account for tertiary structure. Dramatic conformational changes related to the protein's function or environment can also alter local secondary structure.

Several computational tools have been developed to predict different levels of protein structural hierarchy. APSSP2 server predicts secondary structure of protein's from their amino acid sequence with high accuracy. It uses the multiple alignment, neural network and MBR techniques [10]. PSA server allow user to analysis of protein sequence and present the analysis in Graphical and Textual format. This allows property plots of 36 parameters. BetatTPred2 uses Neural Network and multiple alignment techniques. This is highly accurate method for beta turn prediction [11]. TBB pred I predicts the whether a protein is outer membrane betat-barrel protein or not. It also predicts transmembrane Beta barrel regions in a given protein sequence [12]. PEP str predicts the tertiary structure of small peptides with sequence length varying between 7 to 25 residues. The prediction strategy is based on the realization that ?-turn is an important and consistent feature of small peptides in addition to regular structures [13]. AR NH Pred, a web server for predicting the aromatic backbone NH interaction in a given amino acid sequence where the pi ring of aromatic residues interact with the backbone NH groups. The method is based on the neural network training on PSI-BLAST generated position specific matrices and PSIPRED predicted secondary structure [14]. CH predict predicts two types of interactions: C-H...O and C-H...PI interactions. For C-H...O interaction, the server predicts the residues whose backbone Calpha atoms are involved in interaction with backbone oxygen atoms and for C-H...PI interactions, it predicts the residues whose backbone Calpha atoms are involved in interaction with PI ring system of side chain aromatic moieties [15].

III. Mathematical Model of Amino Acid Hydrophobicity

Let a binary digit represent an amino acid residue: hydrophobic residue is 1, and hydrophilic residue is 0. There are eight digits in total; a number between 0 and 255 can be used to model the hydrophobicity and hydrophilicity of the eight residue fragments. The numerical values that associate with the characteristic pattern of α-helix are 9, 12, 13, 17, … , 201, 205, 217, 219, 237, while the characteristic pattern of β-fold is consist of either successive 1 or alternating 01. In the process of secondary structural prediction, according to the amino acid fragment count point model, if the value of point model is the characteristic number of α-helix, then this fragment prediction demonstrates α-helix; if it is the characteristic number of β-fold, then this fragment prediction instead presents β-fold. All the other circumstances except α-helix and β-fold are random coil.

The computing method of hydrophobicity of sequence fragment depends on the numerical value of hydrophobicity of each amino acid residue. For any of a protein sequence, a sliding window is applied to scan this sequence, and it is possible to calculate the average hydrophobicity and the hydrophobic moment of each amino acid under the sliding window. The width of the window could be adjusted; in order to acquire more information and reduce noise interference, the window width of 9~15 residues is commonly selected. The equation of the average hydrophobicity is as below:

$$\overline{H} = \frac{1}{n} \bullet \sum_{i=1}^{n} H_i \tag{1}$$

where H_i is the hydrophobicity of the i_{th} residue of the fragment. The hydrophobic matrix is:

$$\overrightarrow{H} = \sum_{i=1}^{n} (H_i \bullet S_i) \tag{2}$$

where S_i is the unit vector of α carbon atom to the center of the side chain. The hydrophobic moment graph is corresponding to the protein hydrophobic graph. Analyzing these graphs can effectively help to predict protein secondary structure.

IV. Protein Structure Prediction Method Based on Improved Genetic Algorithm

A. Mathematical Model of Genetic Algorithm

At the beginning, m ants are randomly placed in n cities with an initial pheromone intensity value of $\tau_{ij}(0)$ on each side of the city. The first element of k_{th} ant's taboo table is the beginning city. Then each ant goes from the i_{th} to j_{th} city, according to the probability function:

$$p_{ij}^{k}(t) = \begin{cases} \dfrac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}]^{\beta}}{\sum_{s \in allowed_k} [\tau_{is}(t)]^{\alpha}[\eta_{is}]^{\beta}}, & j \in allowed_k \\ 0, & \end{cases} \tag{3}$$

The choice of the city depends on the distance between the cities and the strength of the pheromone. Which $\tau_{ij}(t)$ represents the strength of the pheromone on the side (i, j), and η_{ij} represents the inter-city distance factor, usually taken as the reciprocal of the distance, the set $allowed_k = \{1,2,\cdots,n\} - \{Tabu_k\}$. α and β are the parameters that control the relative importance of the pheromone and the visibility. The visible transition probability is a trade-off between visibility and pheromone

strength at time *t*.

After *n* cycles, all the ant's taboo tables are filled, and calculate the length of each ant's path, then find the shortest path to save, and record the path and change the pheromone on the path. The formula for pheromone update is:

$$\tau_{ij}(t+1) = \rho.\tau_{ij}(t) + \Delta\tau_{ij} \tag{4}$$

$$\Delta\tau_{ij} = \sum_{k=1}^{l} \Delta\tau_{ij}^{k} \tag{5}$$

where $\Delta\tau_{ij}$ represents the sum of the newly added pheromone on a certain edge, ρ represents the level of pheromone dissipation, and $\Delta\tau_{ij}^{k}$ indicates the pheromone left by the *k*th ant at the edge (*i, j*) between time *t* and *t* + *n* Quantity. If the *k*th ant passes the edge (*i, j*) between time *t* and *t*+*n*, then

$$\Delta\tau_{ij}^{k} = \begin{cases} \dfrac{Q}{L_k}, ant_k \in edge(i,j) \\ 0 \end{cases} \tag{6}$$

B. Application of Improved Genetic Algorithm in Protein Structure Prediction

After the amino acid sequence is crossed and mutated, the progeny may have a circulating state, that is, the same position is occupied by two amino acids at the same time, so we need to adjust the operator of the original genetic algorithm. The basic principle of genetic algorithm is shown as follows: firstly, the coordinates of each amino acid are determined according to the initial point and coding, and then the first amino acid in the sequence is tested. If the repeated amino acid in the sequence is encountered, then begin at the point of the repeated amino acid, adjusting backwards until the last point has no repeat coordinates.

Using the roulette selection method in the genetic algorithm, the probability that each amino acid sequence is selected is directly proportional to the average hydrophobic value. But the probability of intersecting individual is selected is inversely proportional to the hydrophobic value:

$$P_i = \frac{E_i}{\sum_{j=1}^{n} E_j} \tag{7}$$

In the protein structure prediction, three crossover sequences are selected for each cycle, and two sequences for mutation are selected.

When performing cross-calculation, the crossover probability P_C is 0.7, and then a random probability *r*, *0<r<1*. When *r*< P_C, cross-calculation is performed, then the two parts of each two are exchanged to form a new individual. After completing the cross, six new individuals are created.

When performing the mutation calculation, set the mutation probability P_m to 0.05. If the random probability is less than the mutation probability, conduct the mutation operation, and randomly mutate a mutation point, and change the coding code of the mutation point. The variation rule is as follows: 1→2, 2→3, 3→1, and two new amino acid coding sequences are generated.

After completing a loop operation, the average individual's hydrophobic value is recalculated for the new individual in the genetic algorithm. Compared with the previous generation population, the fitness of the next generation is greater than that of the previous generation, and the next generation of individuals is retained. The individuals of the previous generation are eliminated, and the crossover and mutation operations are repeated until the 6000th generation.

V. Numerical Simulation

In order to verify the proposed algorithm, the protein sequence used is >>gi|390408302|gb|AFL70986.1| gag protein, partial [Human immunodeficiency virus], and the predicted results are shown in Fig. 1.

From Fig. 1, we can see the state of each amino acid in the sequence, so the predicted secondary structure of the sequence is as follows: 1 to 9 amino acids are random coils. 10 to 33 amino acids are α-helices. 34 to 37 are β-folded. 38 to 45 are random coils. 46 to 49 are α-helices. 50 to 53 are random coils. 54 to 65 are α-helices. 66 to 72 are random coils. 73 to 95 The position is α-helices. 96 to 101 are random coils. 102 to 108 are β-sheets. 109 to 115 are random coils, and 117 are β-sheets.

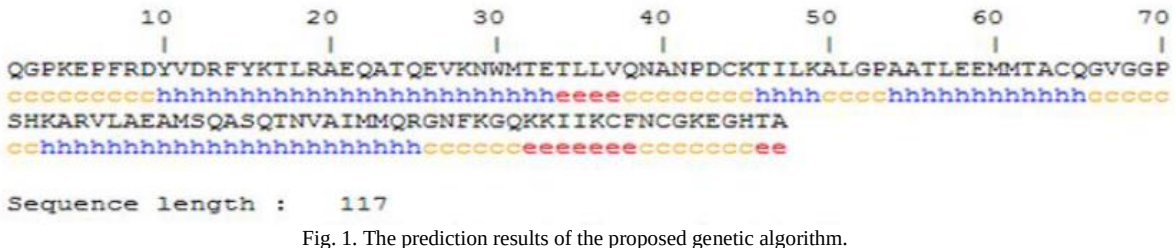

Fig. 1. The prediction results of the proposed genetic algorithm.

Fig. 2. Prediction results of the HNN algorithm.

The predicted results of the traditional HNN algorithm are shown in Fig. 2. The amino acids 1 to 6 are random coils. 7 to 34 are α-helices. 35 to 37 positions are β-folded. 38th is α-helx. From 39 to 44 are random coiling. 45 to 49 are α-helices. 50 to 55 are random coils. 56 to 65 are α-helices. 66 to 71 are random coils. 72 to 83 are α-helices. 84 to 86 are random coils. 87 to 95 are α-helices. 96 to 102 are random coils. 103 to 108 are β-folded, and 108 to 117 are random coils.

Fig. 3 (a) and Fig. 3 (b) show the proportion of the structures in the sequence using the proposed algorithm and HNN algorithm. The Alpha helix predicted by this algorithm accounts for 53.85%, and the extended strand accounts for 11.11%. Random coil Accounted for 35.04%, no other secondary structure.

Alpha helix	(Hh)	:	63 is	53.85%
3_{10} helix	(Gg)	:	0 is	0.00%
Pi helix	(Ii)	:	0 is	0.00%
Beta bridge	(Bb)	:	0 is	0.00%
Extended strand	(Ee)	:	13 is	11.11%
Beta turn	(Tt)	:	0 is	0.00%
Bend region	(Ss)	:	0 is	0.00%
Random coil	(Cc)	:	41 is	35.04%
Ambigous states	(?)	:	0 is	0.00%
Other states		:	0 is	0.00%

Fig. 3. (a) The proposed method.

Alpha helix	(Hh)	:	65 is	55.56%
3_{10} helix	(Gg)	:	0 is	0.00%
Pi helix	(Ii)	:	0 is	0.00%
Beta bridge	(Bb)	:	0 is	0.00%
Extended strand	(Ee)	:	9 is	7.69%
Beta turn	(Tt)	:	0 is	0.00%
Bend region	(Ss)	:	0 is	0.00%
Random coil	(Cc)	:	43 is	36.75%
Ambigous states	(?)	:	0 is	0.00%
Other states		:	0 is	0.00%

Fig. 3. (b) HNN algorithm.

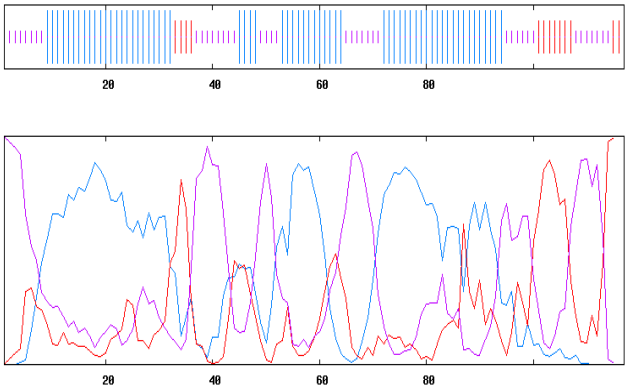

Fig. 4. (a) The secondary structure distribution of the algorithm in this paper.

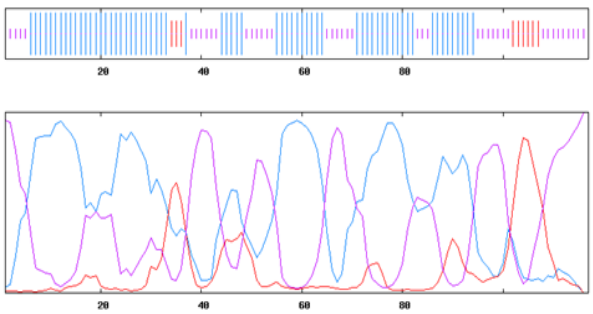

Fig. 4. (b) The secondary structure distribution of HNN algorithm.

Fig. 4 (a) and Fig. 4 (b) show the state of each amino acid in the sequence obtained by the proposed algorithm and the HNN algorithm, and the distribution of the secondary structure of the secondary structure in the whole sequence.

From Fig. 3(a) and Fig. 3(b), it can be seen that the prediction results of the genetic algorithm in this paper are similar to the HNN prediction results. However, as shown in Fig. 1 and Fig. 2, there are differences in the conclusions given by the prediction. Because there are differences in the state of the amino acids in the boundary regions of different secondary structures. For example, in the prediction results of this algorithm, 1 to 9 amino acids are random coils, and 10 to 33 are α-helices, but in HNN prediction results, 1 to 6 are random coils and7 to 34 is an α-helix, showing a large difference in predictions at different secondary structure boundaries, which is shown in the Fig. 4(a) and 4(b).

VI. Conclusion

This paper discusses the prediction of protein structure based on average hydrophobic values. The average hydrophobic value is taken as a mathematical optimization model, and the variable is the hydrophobicity of any amino acid. The genetic algorithm is used to solve the optimization problem of hydrophobic protein structure prediction. The average hydrophobic value is proposed as the adjustment operator to improve the genetic algorithm, and the problem of preventing the overlapping of amino acid positions is solved, so that the amino acid sequence can be accorded to the true extent to the maximum extent. In numerical experiments, the improved genetic algorithm obtained similar prediction results as the traditional HNN algorithm. The feasibility and effectiveness of the proposed algorithm were proved in predicting the state of the amino acids in the boundary regions of different secondary structures. An adjustment operator is designed with the average hydrophobic value to prevent the overlapping of amino acid positions. Some numerical experiments are conducted to verify the feasibility and effectiveness of the proposed algorithm by comparing with the traditional HNN algorithm.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Jiaxi Liu wrote the paper and had approved the final version.

Acknowledgment

Jiaxi Liu thanks the help from Dr. Zhao.

References

[1] J. Cheng, A. N. Tegge, and P. Baldi, "Machine learning methods for protein structure prediction," *IEEE Reviews in Biomedical Engineering*, vol. 1, pp. 41-49, 2008.

[2] B. Rost and C. Sander, "Prediction of protein secondary structure at better than 70% accuracy," *Journal of Molecular Biology*, vol. 232, no. 2, pp. 0-599, 1993.

[3] D. W. A. Buchan, S. M. Ward, A. E. Lobley *et al.*, "Protein annotation and modelling servers at University College London," *Nucleic Acids Research*, vol. 38(Web Server), pp. 563-568, 2010.

[4] A. Drozdetskiy, C. Cole, J. Procter *et al.*, "JPred4: A protein secondary structure prediction server," *Nucleic Acids Research*, vol. 43, no. W1, pp. W389-W394, 2015.
[5] L. Kurgan, K. Cios, and K. Chen, "SCPRED: Accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences," *BMC Bioinformatics*, vol. 9, no. 1, pp. 0-226, 2008.
[6] M. J. Mizianty and L. Kurgan, "Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences," *Bmc Bioinformatics*, vol. 10, no. 1, pp. 414-0, 2009.
[7] G. Pollastri, D. Przybylski, B. Rost *et al.*, "Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles," *Proteins-Structure Function & Bioinformatics*, vol. 47, no. 2, pp. 228-235, 2010.
[8] Z. Wang, F. Zhao, J. Peng *et al.*, "Protein 8-class secondary structure prediction using conditional neural fields," *Proteomics*, vol. 11, no. 19, pp. 3786-3792, 2011.
[9] L. Nanni, S. Brahnam, and A. Lumini, "Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition," *Journal of Theoretical Biology*, vol. 360, pp. 109-116, 2014.
[10] H. Lesso and R. A. Li, "Helical secondary structure of the external S3-S4 linker of pacemaker (HCN) channels revealed by site-dependent perturbations of activation phenotype," *Journal of Biological Chemistry*, vol. 278, no. 25, pp. 22290-7, 2003.
[11] B. Issac and G. P. S. Raghava, "FASTA servers for sequence similarity search," *Proteomics Protocols Handbook*, pp. 503-525, 2007.
[12] N. K. Natt, H. Kaur, and G. P. Raghava, "Prediction of transmembrane regions of beta-barrel proteins using ANN- and SVM-based methods," *Proteins Structure Function & Bioinformatics*, vol. 56, no. 1, pp. 11-18, 2004.
[13] H. Kaur, "PEPstr: A de novo method for tertiary structure prediction of small bioactive peptides," *Protein & Peptide Letters*, vol. 14, no. 7, 2007.
[14] H. Kaur and G. P. Raghava, "Role of evolutionary information in prediction of aromatic-backbone NH interactions in proteins," *Febs Letters*, vol. 564, no. 1, pp. 47-57, 2004.
[15] H. N. Chua, W. K. Sung, and L. Wong, "Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions," *Bioinformatics*, vol. 22, no. 13, pp. 1623-1630, 2006.

Jiaxi Liu is a junior student of the Masters School, 49 Clinton Avenue, Dobbs Ferry, NY 10522, USA. She has extensive research experience, and very passionate at molecular biology. She is also doing research on red blood cell, stem cell and traumatic brain injury. She attended 2019 9th International Conference on Environment Science and Biotechnology (ICESB 2019) and did presentation during the conference.