Protein 3D Features and surface modeling research

Huayong Yang[1] and Xiaoli Lin[2, 3]

[1]Information and Engineering Department of City College, Wuhan University of Science and Technology, Wuhan, 430083, China

[2]Hubei Key Laboratory of Intelligent Information Processing and Real-time Industrial System, School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, 430065, China

[3]Information and Engineering Department of City College, Wuhan University of Science and Technology, Wuhan, 430083, China

Yang_welcome@163.com; 693716397@qq.com

Keywords: protein molecular;intelligent computing;3D features

Abstract: In the twenty-first century, the life science has been studied from the genome research into the structure and function of the structure, which can further reveal the mystery of life. In this paper, it takes the relationship between protein structure and its function as the breakthrough point, with the aid of introducing the protein molecular structure database, analyzing the application of intelligent computing method in the prediction of protein structure. Moreover, by setting up the 3D model of protein molecules, so as to realize the detection of protein 3D features.

Introduction

As human beings begin to carry out functional gene research, the status of proteomics has been raised to an unprecedented level, therefore, proteomics will become one of the most strategic points in the future- one of the top strategical points to conquer human beings gene war. Checking the data of gene sequencing, analyzing the gene expression -the structure of protein, function and the relationship, which is the important component of the plan.[1] In the study, it found out that the function of protein was determined by the spatial structure of protein. Therefore, the study on the spatial structure of protein can play a very important role in the whole proteomics project.

Protein molecules are the main functional units of biological macromolecules, moreover, the function and structure are closely related, that is, the various functions of protein are realized by different three-dimensional structures. The structure of protein is very complex, so far, the basic hypothesis of the study is that the structure of protein is determined by the sequence of protein, which means the structural information of protein is hidden in the protein sequence. Therefore, the research on the spatial structure of protein should focus on studying the hidden relationship between the amino acid sequence and the spatial structure of the protein.

Relationship between Protein Structure and Its Function

Proteins are molecular machinery, building materials, and offensive and defensive weapons in living cells, which can catalyze chemical reactions, control gene expression, maintain cell and tissue structure and so on.

Protein molecules are in tree structure. A molecule can be composed by one or more peptide chains, a peptide chain can contain one or more amino acids, while an amino acid is composed by many atoms, among them, each atom contains the sequence number of atom, atom name, name of amino acid, the belonging peptide, three-dimensional atomic coordinates and other information. The structure of protein molecules can be shown in Fig. 1.

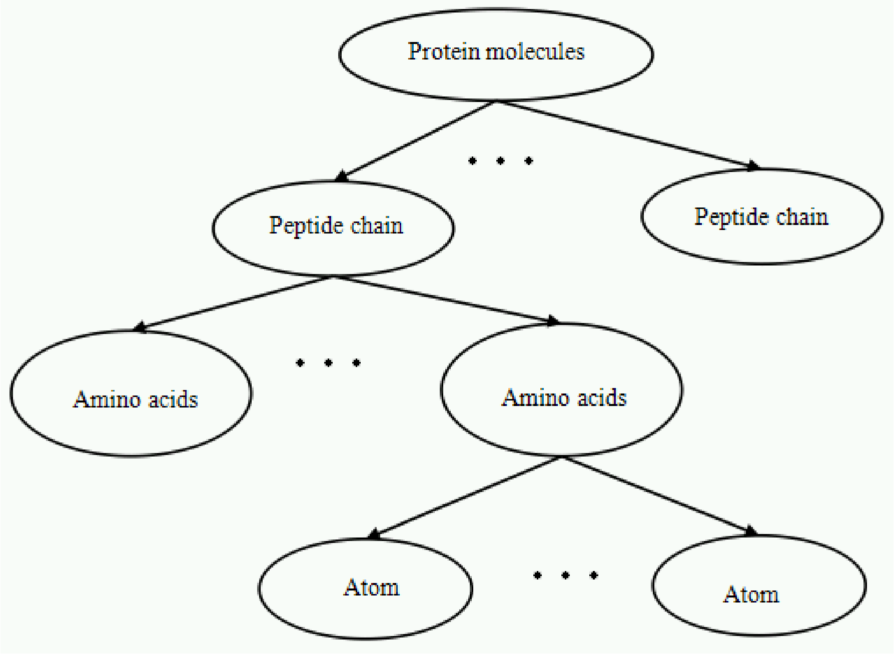

Fig. 1 The Tree Structure of Protein Molecules

A variety of functions of proteins have a very close relationship with their spatial structure. In particular, the surface structure of protein can play an important role in the study of protein function, such as the surface structure of protein can play a very important role in studying the interaction between proteins and the connection of proteins as well as other research. Therefore, it is very important to determine the molecular surfaces of proteins, which can play a very important role in analyzing the protein molecular and setting up reasonable models, so as to provide a theoretical guidance for the design of drugs based on protein structure.

The function of protein is very complex, but as a kind of material, in the function exercise process, it has some common function way, which has the strict structural requirement of the three dimensional level. Only by obtaining a special three-dimensional space structure through the special interaction, can protein molecules have specific biological activities, so the research on the three-dimensional structure of proteins is necessary to understand how proteins perform its function.

Studies on protein sequences have to study DNA sequences obtained by sequencing. This is because the genetic information of living is transmitted from the sequence of DNA to the sequence of protein , that is to say, the sequence of protein is obtained by the sequencing of the DNA sequence. In biology, the genetic information from the sequence of DNA to the sequence of protein sequence can be called as the first genetic code, while the relationship that the protein structure is determined by the sequence of the protein can be called as the second genetic code. Biological molecular data and their relationships can be summarized in Fig.2

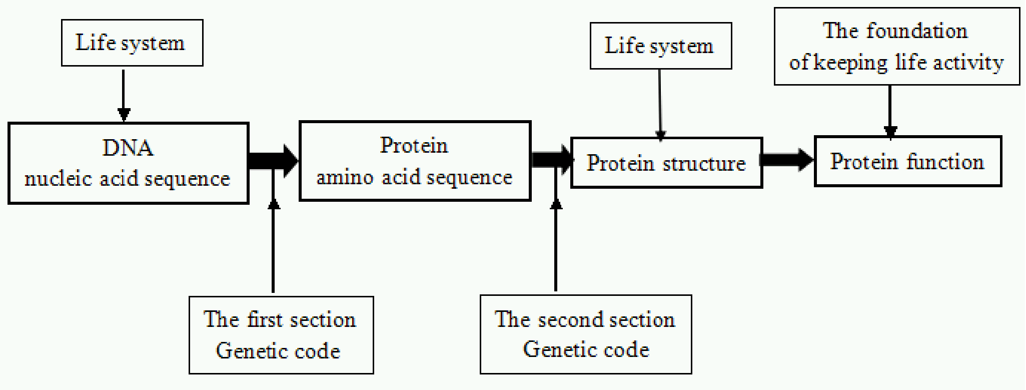

Fig.2 The Data Relationship of Biological Molecular

Carson

This kind of method is widely used as one of the methods to draw protein ribbon model , the key point that is associated with protein chemistry is peptide plane, which is used to be as a basis for geometric modeling, the key point of this method that is associated with the computer graphics is B spline curve , which is used to construct as the smooth normal curve of the ribbon model. The specific algorithm of the ribbon geometric modeling of Carson are as follows:

Determining the plane of peptide. Inputting a series of sequenced three-dimensional coordinates of Cα and alpha carboxyl oxygen atoms. The methods to define the plane of peptide are as follows: (1) calculating vector $A = CA(i +1) -CA(i)$; (2) calculating vector $B = O (i) - CA (i)$; (3) generating normal vector of petide plane $C = A \times B$; (4) generating the vector that is paralleled to petide plane (which is perpendicular to A) $D = C\times A$; (5) unified vector C, D.

Determining the value of smooth curve. According to the ribbon width and the direction of vector, taking Cα atom as the center to calculate the sampling points on both sides of the ribbon. Because CA, C and N is in the same plane, which can be even close to the same line, therefore, it can generate a curve curve that is not too smooth, moreover, the data will cause the increase of calculation workload , thus taking backbone of each of the two peptide plane at the junction of the midpoint of the M of $Ca (i + 1)$ and $Ca (i)$ as the data points, then using cubic B spline curve to generate ribbon. The algorithm of generating the value point of the model are as follows: (1) vector $M = (CA(i +1) +CA(i)) / 2$; (2) vector $E = (W/ 2)\times D$ (where D is the D after being unified, W is the width of the ribbon, take 1.5 as the value); (3) calculating point $M_1 = M - E;$ (4) calculating point $M_2 = M + E;$ (5) taking all points from M_1 and M_2 as the peptide chain type value points to generate cubic B spline curve. Forming the line $M1\text{->}M2$, its width is the width of the ribbon and plane which is paralleled to the plane of peptide.

Using the cubic B spline curve to generate function so as to generate a smooth curve. According to the type value point to calculate the cubic B spline curve, obtaining the interpolation points through the the curve fitting, then according to the interpolation points to draw the cubic B spline curve, so as to generate smooth curve.

In the end, having reasonable construction for the scene of ribbon model, so as to realize the visualization of ribbon model of protein molecules.

Evaluation Method for the Prediction of Protein Structure

In this paper, it adopts a typical method , which is used to predict the percentage of all two-level structural residues rightly through evaluating the number of total residues, which can be calculated out as the following formula:

$$Q=(P_H+P_E+P_C)/N$$

Among them, the number of residues of H, E, C, can be represented by P_H、P_E、P_C, respectively, N is the total number of all residues. In addition, the predicting accuracy of these two-level structures can be given in three types, respectively:

$$Q_i=P_i/N_i$$

Among them, $i \in\{H,E,C\}$; Ni are the total number of three residues of H, E, C in protein training set or test set. While Pi is the number of residues of the three state H, E, and C that have been correctly predicted.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (61502356)

Reference

[1] C.G. Anfinsen. 1973. Principles that govern folding chains. Science. vol.181, pp: 223-230.

[2] Bairoch A, Apweiler R. 2000. The SWISS-PROT protein sequence database and its supplement Tr EMBL in 2000. Nucleic Acids Research. vol.28, pp:45-48.

[3] Peter B, Mc Garvey, Hongzhan Huang, el al. 2000. PIR: a new resource for bioinformatics. Bioinformatics Applications Note. vol.16, pp:290-291.

[4]RCSB Protein Data Bank. http://www.rcsb.org.

[5] Homik K M, Stinchcombe M, White H. 1989. Multilayer Feed Forward Networks Are Universal Approximators. Neural Networks. vol.2, pp:359-366.