Research on Protein Structure Prediction Based on Sequence Pattern

JuanjuanYin, Guojian Cheng, Fenggang Ma
School of Intelligence Science and Information Engineering
Xi'an Peihua University, Xi'an, China
294539515@qq.com, 1683939013@qq.com, 1103364369@qq.com

Keywords: Protein structure prediction; Hydrophobicity; Apriori all algorithm; KMP pattern matching

Abstract. In order to find meaningful data for human beings in a large number of biological information and then reveal the nature of life, this paper focuses on analyzing the characteristics of the hydrophobicity of amino acids and adjacent information, using sequential pattern mining Apriori All algorithm to extract amino acid sequences, using KMP pattern matching query algorithm to predict amino acid sequences, and then obtaining the secondary structure of protein. The prediction results have high reliability.

1 Introduction

In biological macromolecules, proteins dominate all activities of life, so that the major research topics that reveal the mysteries of life are closely related to them. The function of a protein depends on its complex structure. Therefore, the study of protein structure has not only been a theoretical problem, but has produced new technology of protein engineering, which uses modern biotechnology to modify the protein branch structure to make it more suitable for human needs, such as improving enzyme activity and increasing the stability.

In this paper, the protein amino acid sequence of the drug synthesis database is selected for the prediction of protein secondary structure and the method of biological sequence mining is used to analyze and predict the secondary structure of protein. At the beginning of the excavation, the hydrophobicity of amino acids and the characteristics of adjacent information are emphatically analyzed, so that the selected information can better map the characteristics of protein sequence itself. The sequential pattern mining Apriori All algorithm is used to extract the amino acid sequence, which makes the sequence pattern mining have higher and better functions. The KMP pattern matching query algorithm is used in the prediction phase of amino acid sequence, which makes the final prediction conclusion more reliable, makes the protein secondary structure prediction on the basis of sequential pattern mining method has more accurate analysis results, and then lays a solid foundation for further analysis and research on the significance of protein function to its structure [1].

2 Protein Analyses

2.1 Basic Components of Protein

Amino acids are the basic units of protein.Amino acids are organic acids with amino acids, which consists of an amino group, a carboxyl group, a hydrogen atom and an R group.Several amino acids form a peptide chain, and many peptides form polypeptides, which are then aggregated into protein structures by polypeptides.Table 1 contains the abbreviations and meanings of the 20 most common amino acids that make up proteins.

2.2 Protein Structure

Any amino acid has a unique physicochemical property, depending on the structure and nature of the side chain. All of the amino acid side chain compounds ultimately ensure its three-dimensional structure and chemical reaction. The structure of protein molecules is generally divided into one-dimensional and multi-dimensional.

Table 1 Abbreviations and Meanings of 20 Amino Acids

Symbol	Abbreviation	Meaning	Symbol	Abbreviation	Meaning
Aln	A	alanine	Pro	P	proline
Cys	C	cysteine	Gln	R	arginine
Asp	D	Aspartic acid	Ser	S	serine
Glu	E	glutamate	Thr	T	threonine
Phe	F	phenylalanine	Trp	W	tryptophan
Gly	G	glycine	Tyr	Y	tyrosine
Hls	H	histidine	Agx	B	Aspartic acid or asparagine
Ile	I	isoleucine	Glx	Z	Glutamic acid or glutamine
Lys	K	lysine	Xaa	X	Unknown or otherwise
Leu	L	leucine	Met	M	methionine

The primary structure of the protein is its basic structure, which is the arrangement of amino acids in the peptide chain. And the order of the amino acids in the protein primary structure ensures the protein secondary structure, thus finally ensures its advanced structure, for example, tertiary structure, and finally decides the function of protein in biological cells. The secondary structure of proteins depends on the stable structure of hydrogen bonds between different amino acids between c=o and n=h groups, mainly spiraling and folding. As more and more protein structures are resolved, more secondary structures are found, such as, corners, and other less-common secondary structural elements. The tertiary structure of the protein refers tothe protein polypeptide chain which is further folded on the basis of various secondary structures and the three-dimensional conformation of each protein is formed by assembling the elements in the secondary structure. The quaternary structure of the protein refers to that the interaction of several peptides forms a functional protein (that is, the interaction between protein and protein) so this intermolecular interaction is a quaternary protein structure.

2.3 Hydrophilicity and Hydrophobicity of Amino Acids

The secondary structure of the protein and the hydrophobicity of amino acids have a very large relationship, and then the secondary structure is predicted by using the hydrophobicity of amino acid residues.From previous studies, it can be concluded that the majority of chemical substances can be easily divided into two parts: the ones having the ability to react with water and the ones which do not have the ability to react with water.Hydrophobic amino acids include Ala, Val, Leu, Ile, Phe, Pro and Met[3].

To a certain extent, hydrophobic substances and hydrophilic substances are the products of uniform and balanced life. The amino acids of 20 different types of side chains make up the protein. Some of them are polar, very easy to react with aquatic products, or make up o-h, or very easily dissolved in water. However, there is another part of the side chain which is non-polar, which does not show the intention and ability to interact with other polar genes or H_2O.

3 Protein Secondary Structure Prediction Research

With the continuous accumulation of the protein structure data, the prediction method of protein secondary structure with the chemical properties of amino acid hydrophobicity was developed[1]. To find out whether the secondary structure of protein and amino acids are hydrophobic, the really specific goal is to find the frequent links between the amino acids that appear to each other. For this purpose, this paper uses the data mining method of sequential pattern to achieve the goal.

3.1 Hydrophobic Grade Conversion of Amino Acids

To use the chemical properties of amino acid hydrophobicity to analyze data, the sequence of amino acids must be converted into data that can show hydrophobicity. The hydrophobic parameters of 20 amino acids are shown in table 2. It is positive and the larger the value indicates that it has more

hydrophobic characteristics. Similarly, the value is negative and the smaller the value is, it has higher hydrophilic characteristics.

Table 2 Hydrophobic Parameters of 20 Common Amino Acids

Serial Number	Amino Acid	Hydrophobic Value	Hydrophobic Grade
1	A	1.8	3
2	R	-4.5	1
3	N	-3.5	1
4	D	-3.5	1
5	C	2.5	3
6	Q	-3.5	1
7	E	-3.5	1
8	G	-0.4	2
9	H	-3.2	1
10	I	4.5	4
11	L	3.8	4
12	K	-3.9	1
13	M	1.9	3
14	F	2.8	3
15	P	-1.6	2
16	S	-0.8	2
17	T	-0.7	2
18	W	-0.9	2
19	Y	-1.3	2
20	V	4.2	4

The data of 20 hydrophobic values are divided into intervals, and the range is as follows: [-4.5,-3], （-3,0],（0,3],（3,4.5], and then we use the decimal number: 1,2,3,4. It can be seen from the above table that the level of amino acids is obvious, and the larger the value indicates that it has more hydrophobic characteristics. Similarly, the smaller the number is , the higher hydrophilic characteristics is. From this, it can be seen that the level values obtained by this transformation have the ability to reflect the characteristics of the original amino acid.With the above table, the characters can be reflected in the collection,for example, sequence data s1=NDCQE, which is mapped to s1 '=11311.Thus, four adjacency relations can also be represented by numbers, for example, s1 'has 11,13,31,11. Then, through the search of a large number of known protein sequences, it is found that which link relationship has high probability relation to the composition of a secondary structure.

3.2 The adjacency item Conversion of Amino Acids

In this paper, according to different structure situations, the 487 protein sequences provided in the comprehensive database of drugs are classified into: H type (4 Angle spiral（αspiral).The shortest length of four residue), E (parallel beta fold, and/or the parallel folding form chain (extension), the shortest length of two residues), C (to represent the curly) structure.Each structure is divided into a sequence of 5.The results are as follows: 11653 sequence segments are obtained through the same method, and the H, C and E are 2348, 3368 and 5937 respectively.Table 3 shows the contiguous set of the sequence fragments of h-type secondary structure converted from table 2. In total, N-1 adjacency items can be generated for sequence fragment length N. To facilitate the mining of sequence patterns in table 3 and make it easier to identify and record data, a hexadecimal number is used instead of a set of adjacency combination data. Table 3 is the last list of adjacent items that are replaced.

4 Explore the Sequence Pattern of Hydrophilic Amino Acids

The prediction of the secondary structure of local protein is based on the hydrophobicity of the amino acids in sequence fragments. Namely, in order to find the correlation patterns of different hydrophilic residues in protein sequences, the method of searching for the separation of the hydrophobic residues with distinct characteristics is used in the sequence.

Table 3 Some Sequence Adjacency Items forming H-type Secondary Structure

Protein Name	Secondary Structure	Starting Location	Termination Location	Sequence Fragment	Adjacency Items	Substitution code for Adjacency Items
3tima.all	H	47	51	LAMTK	{(43),(33),(32),(21)}	{(E),(A),(9),(4)}
3tima.all	H	130	134	LQERE	{(41),(11),(11),(11)}	{(C),(0),(0),(0)}
3tima.all	H	218	222	ARTLY	{(31),(12),(24),(42)}	{(8),(1),(7),(D)}
4rhv3.all	H	142	146	RREAM	{(11),(11),(13),(33)}	{(0),(0),(2),(A)}
4rhv1.all	H	51	55	VECFL	{(41),(13),(33),(34)}	{(C),(2),(A),(B)}
4gr1.all	H	440	444	KADFD	{(13),(31),(13),(31)}	{(2),(8),(2),(8)}
6rlxc-1-DOMAK.all	H	17	21	KRSLA	{(11),(12),(24),(43)}	{(0),(1),(7),(E)}
6cts.all	H	38	42	VDMSY	{(41),(13),(32),(22)}	{(C),(2),(9),(5)}
6tmne.all	H	225	229	QDNGG	{(11),(11),(12),(22)}	{(0),(0),(1),(5)}
1ecpf-1-AUTO.1.all	H	115	119	VNRIR	{(41),(11),(14),(41)}	{(C),(0),(3),(C)}
5sici-1-DOMAK.all	H	94	98	ECEMN	{(13),(31),(13),(31)}	{(2),(8),(2),(8)}
1ecl-1-AS.all	H	135	139	IDRVN	{(41),(11),(14),(41)}	{(C),(0),(3),(C)}
1eceb-1-AUTO.1.all	H	268	272	GYLFN	{(22),(24),(43),(31)}	{(5),(7),(E),(8)}
6cpp.all	H	29	33	VQEAW	{(41),(11),(13),(32)}	{(C),(0),(2),(9)}
2trt-1-AUTO.1.all	H	95	99	GAKVH	{(23),(31),(14),(41)}	{(6),(8),(3),(C)}
2tsca.all	H	70	74	AYLHE	{(32),(24),(41),(11)}	{(9),(7),(C),(0)}
1gal-3-AS.all	H	27	31	FNETF	{(31),(11),(12),(23)}	{(8),(0),(1),(6)}
2sil-1-AS.all	H	374	378	LPVIK	{(42),(24),(44),(41)}	{(D),(7),(F),(C)}

4.1 Apriori All algorithm

The purpose of the analysis of amino acid sequence pattern is to find the continuous sequence without interval. Since there is no time constraint, the most basic sequence model can be used to meet the requirements. In this paper, a sequential pattern mining method based on Apriori features is selected, mainly Apriori All algorithm[9]. The characteristics of frequent sequences are: if the sequences with length M are not frequent, then the sequences larger than M cannot be non-frequent. According to this feature, this algorithm can effectively cut the sequence set of verification in the process of searching longer sequences according to the frequent sequence of the generation. In the process of algorithm operation, the original sequence database is repeatedly traversed. After a traversal, the frequent sequence is selected. Each time a frequent sequence is made, the sequence database removes the iterated sequence from the collection. As the number of scans increases, the frequency in the collection becomes smaller and smaller. The algorithm keeps being executed until there are no candidate sequences or no frequent sequences.

In each traversal process, a new large sequence can be generated by starting with a subset of the large sequence and using a subset. Through calculating the support degree of the selected sequence, the sequence is determined and the next subset is formed. The Apriori All algorithm is used to accumulate all large sequences, including non-maximal sequences, and these non-maximal sequences must be deleted when looking for the maximum phase.

Through sequential pattern mining, the following sequence patterns are obtained:

(1) 003#=>H, confidence level 0.81;

(2) *7D*=>E, confidence level 0.86;

(3) *D7*=>E, confidence level 0.83; … …

The results of pattern (1) show that the amino acid sequence of a length of 5, such as 1, 2, three level of hydrophobic amino acid residues is 1, the fourth is 3, the corresponding formation of the secondary structure of H. Similar pattern (2), (3) indicate that if a sequence of protein contains 242 or 424 hydrophobic grade conjoined fragments, the secondary structure is E.

4.2 KMP Pattern Matching Query Algorithm

After the rich sequence pattern in the database, the sequence prediction of the protein secondary structure is carried out through the retrieval and matching method, and the highest confidence

sequence pattern was used as the prediction result. Therefore, the efficiency of pattern matching query algorithm is critical to the prediction speed. In general string pattern matching search, KMP algorithm is the most widely used and most efficient pattern matching algorithm.

Compared with ordinary string pattern, the amino acid hydrophobic adjacency sequence pattern of biological protein has its own particularity. It introduces the wildcard "#" and "*", so it is necessary to make some improvements to KMP from two aspects: Firstly, if you encounter a "#" in the pattern string, because the wildcard is just a placeholder, any character and "#" are matched; Secondly, if the "*" is encountered in the pattern string, any number and type of characters are matched before the next matching character appears in the pattern string.

In the matching sequence pattern, the highest sequence pattern is selected as the prediction result. If there is a given sequence seq = "RTDCYGNVNRIDTTGASCKTAKPEGLSYCG", according to the mining sequential patterns in the library take the sequence's length sliding window for 5, the matching queries is carried out by using respectively the existing pattern and the sequence in the window. If the match records the serial number of the pattern, when all modes in the sequence pattern library are compared, the secondary structure corresponding to the highest confidence mode in the matching mode is selected as the prediction result. Seq sequence pattern matching "CCCCCCCCCCCCCCCCCHHHHCCCCCCCCC" was the final secondary structure sequence.

5 Conclusions

In this paper, we use data mining to discover the sequence pattern method to predict the protein secondary structure of amino acid hydrophobicity. In the study, we try to restore the mechanism of the interaction between the chemical molecules, which makes the prediction have sufficient basis. We focus on the properties of hydrophobic adjacency relations, rather than simply treating the amino acids as ordinary characters. Because the protein sequences of known structures in the research process are not rich enough, the sequence patterns obtained by mining are not comprehensive enough to affect the accuracy of sequence analysis. In this paper, due to the limitation of the condition, only the hydrophobicity is considered, and the overall forecast accuracy is only 56.8%, of which the prediction accuracy of h-type structure is better than 63.2%. We believe that with the continuous improvement of data information and mining algorithm, the prediction accuracy of protein secondary structure based on sequence pattern will be further improved.

References

[1] Liu Maojuan. The construction of the mining model of gene expression profile data [D]. China Ocean University,2015.

[2] Sang Yongsheng. Research on biological sequences based on data mining [D]. University of Electronic Science and Technology,2006.

[3] Long Haixia. Evolutionary algorithm and its application in biological information [D]. Jiangnan University,2010.

[4] XueJie. Two types of biocomputing problems and their application in data mining [D]. Shandong Normal University,2015.

[5] Zhang Hongyan. Research on clustering algorithm based on DNA calculation [D]. Shandong Normal University,2011.

[6] Ma Meng. Association rules mining algorithm for biological data and its application research [D]. China University of Science and Technology,2008.

[7] Miao Yuqing. Association rule mining and its application in gene expression data [D]. China University of Science and Technology,2007.

[8] Li Rong. Research and application of a number of key issues in biological information data mining [D]. Fudan University,2004.

[9] Gu Zhaohui. The establishment of new gene prediction platform for biological information mining and the preliminary study on cloning and function of tseg-3 [D]. Huazhong University of Science and Technology,2010.