VOL. 3 NO. 2 September 2014 - ISSN 2233 – 1859

Operations Research Society in Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: www.scjournal.com.ba

IUS Soft Computing Research Group

Predicting the Secondary Structure of Proteins by the Use of Hamming Distances and Alignment Scores

Mehmet Can, Betul Akcesme, and Faruk B. Akcesme
International University of Sarajevo, Faculty of Engineering and Natural Sciences, Hrasnicka Cesta 15, Ilidža 71210 Sarajevo, Bosnia and Herzegovina
mcan@ius.edu.ba; betul.cicek@yahoo.com; fakcesme@ius.edu.ba

Article Info

Article history:
Article received on Sep. 2014
Received in revised form Nov 2014

Keywords:
Secondary structure; Conformation of proteins; Statistical methods

Abstract

Researchers are confident about the validity of the basic hypothesis that the secondary and tertiary structures of a protein are uniquely determined by its sequence of amino acids, that is its primary structure. In this article we use a database of 200 proteins. To find the secondary structure of a new protein, the first thirteen residues of this protein are taken as a substring. Then the conformations of the central amino acids of thirteen residue substrings of the proteins in the database, whose hamming distances are less than a given threshold or alignment scores exceed a given limit are collected in a basin. The commonest conformation in this basin is attached as the conformation of the central amino acid of the substring of the unknown protein. Using this technique, for MHsim threshold 3.0, a correct estimation rate of 53.4% is obtained with 4.74% indecisives and for MHsim threshold 5.0, the success was 56.93% with76.59% indecisives. When the half of the proteins, whose secondary structure estimations are higher, subjected to same calculation the following results are obtained; for MHsim threshold 3.0, correct estimation rate is 79.52% with 58.87% indecisives and for MHsim threshold 5.0, correct estimation rate is 65.52% with 5.02% indecisives. Average correct estimation rate for the alignment scores was %54.

1. INTRODUCTION

For more than four decades, the protein folding problem has been among the most challenging problems in the biological sciences. In 1994, a protein structure prediction contest was organized with the aim of assessing the real virtues and defects of several well known methodologies. Analysis of the structures predicted by the contestants (Moult et al., 1995) has generally shown that even the most promising techniques need considerable improvement, and that the protein folding problem should still be considered unresolved. Briefly, preliminary calculations, although promising, are feasible only for small-size proteins; there have been no major breakthroughs in the molecular modeling techniques and threading techniques need further development.

During this contest, protein secondary structure prediction was reevaluated and recognized as a useful tool for establishing starting points for tertiary structure calculation determination of protein structures. Early approaches to

protein secondary structure prediction from the primary sequence had prediction accuracy, which is the percentage of correctly predicted residues in the three states: α-helix, β-strand, and coil, of about 57% (Chou & Fasman, 1978; Garnier et al., 1978). Various later attempts to improve the accuracy (Gibrat et al., 1987; Biou et al., 1988; Levin & Garnier, 1988; Holley & Karplus, 1989; Qian & Sejnowski, 1989; King & Sternberg, 1990; Salzberg & Cost, 1992; Stolorz et al., 1992; Zhang et al., 1992; Munson et al., 1994) with innovative artificial intelligence techniques, such as neural networks, machine learning, nearest neighbors, and combined approaches, have not achieved prediction accuracies greater than 66%. The inclusion of evolutionarily related sequences into the prediction scheme has given a significant boost in prediction accuracy, up to values of about 68-72% (Zvelebil et al., 1987; Levin et al., 1993; Rost & Sander, 1993, 1994; Rost et al., 1994a; Di Francesco et al., 1995; Salamov & Solovyev, 1995). In general, the suggested explanation for these improvements in prediction accuracy is that sequence alignments of homologous proteins should emulate as closely as possible the structural alignment. Thus, aligned residues, in particular those in the core proteins, should belong to the same secondary structure elements. Sequence alignments may be utilized to obtain a consensus from the predictions based on each homologous sequence, or they may be used to build sequence profiles at each aligned position. In addition to the identity of the aligned residues, which is a feature exploited by all the predictive schemes, other information is available from sequence alignments, such as the location of gaps or the patterns of residue mutation in the aligned protein families. Some authors have used such information to refine their prediction models (Zvelebil et al., 1987; Rost & Sander, 1993; Rost et al., 1994a; Salamov &Solovyev, 1995). However, the reasons why the inclusion of this additional information improves the quality of the prediction have not been understood.

In his extensive review Rost (Rost, 2001) asks the following question: 88% is a limit, but shall we ever reach close to there?

A database of 200 random proteins with known secondary structure formations is prepared. To find the secondary structure of a protein, test substrings of consecutive residues of length 13 of this protein are formed. Then in proteins in the database, substrings of length 13 with high enough similarities to the test string are collected in a pool. The most common secondary structure formation corresponding to the central amino acid of substrings in the pool is attached to the central amino acid of the test substring as secondary structure formation.

2. FORMULATION OF THE PROBLEM

To estimate the conformation of the protein at a given residue, we consider 6 right and 6 left neighbors of this residue. Our hypothesis is that the conformation at the central residue is determined by these neighbors and by itself.

Primary structure: DETTALVCDNGSG
Secondary structure: CCCCCCSSSSSS

Figure 1 Primary and secondary structures of a protein of length 13 residues.

(a) Database

Primary structures of 200 proteins are obtained from the PDB website. Secondary structures of these proteins are obtained in the form of the xray analyses in three conformations helix "h", sheet "s" , and others ".". Others are interpreted as coils "c".

Figure 2α-helices, β-sheets, and coils on the same picture (PDB code for the protein: *1OC0*).

(b) Symbols for Amino Acids

Proteins are chains in the three dimensional space built from smaller chemical molecules called amino acids. There are 20 different amino acids. Each of them is denoted by a different letter in the Latin alphabet as shown below.

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

Table 1 Names and symbols of 20 amino acids

Based on the protein chain it is easy to create its relevant sequence of amino acids replacing an amino acid in chain by its code in Latin alphabet. As a result a word on the amino acids' alphabet is received. This word can be called a protein primary structure on the condition that letters in this word are in the same order as amino acids in the protein chain are.

A secondary structure of a protein is a subsequence of amino acids coming from the relevant protein. These sub chains form in the three dimensional space regular structures which are the same in shape for different proteins. In the analysis, a similar representation for the secondary structures as for the primary ones has been used. A secondary structure is represented by a word on the relevant alphabet of secondary structures – each kind of a secondary structure has its own unique letter α-helix, H; β-sheet S, and coil C. An alphabet of secondary structures consisting of three different secondary structures has been considered in the analysis.

(c) Coding the Data

In this paper, data corresponding to an amino acid consists of 6 right, and 6 left neighboring amino acids of this amino acid in the primary chain of the protein as in Table 2. In the second row, secondary structure conformations of these neighboring amino acids are given.

A	E	E	K	E	A	**V**	L	G	L	W	G	K
H	H	H	H	H	E	**E**	E	E	C	C	C	E

Table 2 Six right, and six left neighboring amino acids of the central amino acid V.

Secondary structure letters H, E, and C are coded as in the table below;

H	E	C
1	0	0
0	1	0
0	0	1

Table 3 Codes for secondary structure letters H, E, and C.

(d) Similarity Measures

To find the secondary structure of a protein, test substrings of consecutive residues of length 13 of this protein are cut. Then in proteins in the database, substrings of consecutive residues of length 13 are cut as well. To infer the conformation of the central amino acid of test substring, we search for similar substrings of the same length of 13 from the proteins in the database. For this purpose, two similarity measures are used.

(1) Modified Hamming Similarity

Hamming distance of two substrings of the same length is the number of the mismatches as seen in Table 4.

G	R	L	P	A	C	**V**	V	D	C	G	T	A
M	L	S	P	A	D	**K**	V	N	V	K	A	A
0	0	0	1	1	0	**0**	1	0	0	0	0	1

Table 4 Hamming distance of two substrings of the same length is the number of the mismatches

Now a similarity measure of the given two strings can be defined as

$$Hsim = 13 - Hdis. \qquad (1)$$

There is a consensus about the affect of amino acid composition of the primary sequence on the secondary structure of a protein. But clearly this affect is local. That is amino acids far away of the central amino acid have less affect on the conformation at the central amino acid, compared to the nearer ones. It means that the match "VV" at the 8th position is more important than the match "AA" at 13th position.

To weight matches we propose a Gaussian curve

$$F(x) = e^{-(x-7)^2/s^2}, \qquad (2)$$

where s is a measure for the spread of the curve.

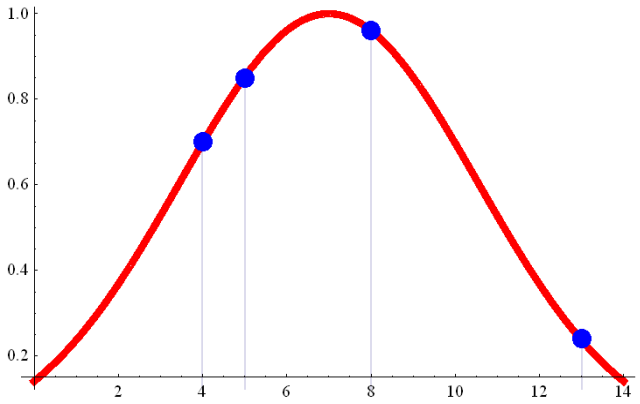

Figure 3 Weighted matches for s=5.

For the substrings in Table 4, hamming similarity is $Hsim = 4$, modified similarity is $MHsim = 2.75$.

(2) Alignment Score

The alignment of two strings

GRLPACVVDCGTG
MLSPADKTNVKAA (3)

is obtained as

G	–	R	L	–	P	A	C	V	V	D	–	C	G	T	_	G
_	M	_	L	S	P	A	_	_	_	D	K	_	_	T	N	_

_	_	_	_
V	K	A	A

(4)

"–"in the first reference string is called a deletion, while a "–" in the second query string is called an insertion. Same amino acids in a column are called "matches", different amino acids in a column are called "mismatches". If we reward matches by $ms = 1$, penalize mismatches by $mm = 0$, deletions $d = -0.5$, insertions by $i = -.25$, the alignment score of the above alignment is

$$Score = 4 * 1 + 9 * 0 + 8 * (-0.5) + 8 * (-0.25) = -2 \quad (5)$$

3. IMPLEMENTATION

To obtain secondary structure at an amino acid in a protein, taking six right, and six left neighbors of this amino acid, we compose an ordered 13 tuple of amino acids as a test string. Then from protein database at hand we take proteins whose secondary structures already known, in an orderly way, and choose a substring of consecutive amino acids of length 13 as a target string. Then we compute similarities of this pair of test-target substrings according to one of the similarity measures given in the above. If similarity is higher than the prescribed threshold, we put the conformation of the central amino acid of the target string in a basket. We repeat this procedure for all 13tuples of consecutive amino acids, of the proteins in the database. Eventually the commonest of conformations in the basket is attached as the conformation of the central amino acid in the test string.

Database of Proteins

200 proteins of known structures, with a total 169 026 amino acid residues collected from PDB almost randomly. To test the accuracy of the method each time one of the proteins is chosen as the testing protein, and other 199 proteins are taken as target proteins.

4. RESULTS AND DISCUSSION

In a biological context, the term homology, defines similarity of structure, physiology based upon a genetic factor. The protein homology most recognized by similarities in their amino acid sequence. There is a widely accepted hypothesis that: "the greater the sequence similarity; the more closely related are the scaffold structure". Based on this approach, proteins primary sequence similarity was investigated with searching for similar substrings of the same length of 13 from the proteins in the database. Each of the similar 13 tuples in the database is found and collected into the basket with modified hamming similarity threshold 5 and 3 separately. The proteins that have high similarity and high accuracy with some certain proteins in database has been detected and separated for further analysis. Further analyzes is going to cover some fundamental questions such as; what is the structure of the similar regions in highly similar proteins? In what bases correct structural classification of proteins can be performed? We believe that answering these questions will enable us to classify proteins, existed protein classification approaches are going to be analyzed and methodology is going to be strengthening. The following question; "what is the advantage of the structural classification of proteins over randomly chosen proteins" will be addressed. In the other hand proteins that have no similar or low similar sequences in the database also detected. This protein`s structure, physiological characters and their physicochemical properties are going to be analyzed in order to reveal information about the influence of this parameter. The essence of particular parameter aimed to be found which make this protein structure unique. We believe that it will provide us a new attribute in order to increase the prediction capacity of our algorithms.

For each test substring of length 13, around 16000 comparisons are made with 13tuples of amino acid residues of target proteins. In an average desktop computer this operation is performed in around five seconds. Therefore it is not feasible to increase the number proteins in the database. For this reason, for high thresholds for the similarity, some of the test 13tuples may not have similar enough 13tuples in target proteins. In such a case, the conformation of the central amino acid of the test 13tuple remains undefined. On the other hand, high similarity brings high accuracy in the secondary structure estimation. In Table 4, for certain values of the modified hamming similarity threshold, the percentage of the indecisive residues, and accuracy in the three conformations α-helices, β-sheets, and c-coils are given.

MHsim	Indecisives %	Accuracy %
≥ 5.0	76.59	56.93
≥ 4.5	54.15	54.15
≥ 3.0	4.74	53.41

Table 5 Similarity thresholds vs. accuracy in MHsim

For the half of the proteins in database whose correct secondary structure estimations are better, the correct estimation rates are as in Table 6.

MHsim	Indecisives %	Accuracy %
≥ 5.0	58.87	79.52
≥ 4.5	40.74	67.06
≥ 3.0	5.02	65.56

Table 6 Similarity thresholds vs. accuracy in MHsim for better half.

For the similarity measure computed by the use of the alignment score, the built in function Needleman Wunsch Similarity in MATHEMATICA is used. As scoring matrix, PAM70 is chosen. Test and target substrings are of length 17 to give more stable and reliable alignment scores.

In Table 7, for the values 1, and −5 of the alignment score thresholds, the percentages of the indecisive residues, and

average accuracies in the three conformations α-helices, β-sheets, and c-coils are given.

A. Score	Indecisives %	Accuracy %
≥ 1	69.69	63.65
≥ −5	6.31	54.01

Table 7 Alignment score thresholds vs. accuracy in ASsim

For the half of the proteins in database whose correct secondary structure estimations are better, the correct estimation rates are as in Table 8.

A. Score	Indecisives %	Accuracy %
≥ 1.	51.78	93.24
≥ -5	7.31	62.16

Table 8 Alignment score thresholds vs. accuracy in ASsim for better half.

These results show that the analysis which relies on a database of 200 proteins has a estimation power that is comparable with the famous online estimation tools. Table 6, and Table 8 display the correct estimation rates of the half of the proteins in database whose correct secondary structure estimations are better.

REFERENCES

Biou V, Gibrat JF, Levin JM, Robson B, Garnier J. 1988. Secondary structure prediction: Combination of three different methods. Protein Eng 2:185-191.

Di Francesco V, Munson PJ, Garnier J. 1995. Use of multiple sequence alignments in protein secondary structure prediction. 28th Hawaii International Conference on System Sciences. IEEE Computer Society Press.

King RD, Sternberg MJ. 1990. Machine learning approach for the prediction of protein secondary structure. J Mol Biol 216: 441-457.

Levin JM, Garnier J. 1988. Improvements in secondary structure prediction method based on search for local sequence homologies and its use as a model building tool. Biochim Biophysics Acta 955: 283-295.

Levin JM, Pascarella S, Argos P, Garnier J. 1993. Quantification of secondary structure prediction improvement using distantly related proteins. Protein Eng. 6 (8): 849-854.

Munson PJ, Di Francesco V, Porrelli R. 1994. Protein secondary structure prediction using periodic-quadratic-logistic models: Statistical and technical issues. 27th Hawaii International Conference on System Sciences. IEEE Computer Society Press.

Qian N, Sejnowski TJ. 1989. Predicting the secondary structure of globular proteins using neural network models. J Mol Biol. 202 (4): 865-884.

Rost B, Sander C. 1993. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol. 232:584-599.

Rost B, Sander C. 1994. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins Struct Funct Genet 19: 55-72.

Rost B, Sander C, Schneider R. 1994. Evolution and neural networks. Protein secondary structure prediction above 71% accuracy. 27th Hawaii International Conference on System Sciences. IEEE Computer Society Press.

Salamov AA, Solovyev VV. 1995. Prediction of protein secondary structure by combining nearest-neighbor algorithms, multiple sequence alignments. J Mol. Biol. 247:11-15.

Salzberg S, Cost S. 1992. Predicting protein secondary structure with a nearest-neighbor algorithm. J Mol Biol 227: 371-374.

Stolorz P, Lapedes A, Xia Y. 1992. Predicting protein secondary structure using neural net and statistical methods. J Mol Biol 225: 363-377.

Zhang X, Mesirov JP, Waltz DL. 1992. Hybrid system for protein secondary structure prediction. J Mol Biol 225: 1049-1063.

Zvelebil MJ, Barton GJ, Taylor WR, Stenberg MJE.1987. Prediction of protein secondary structure and active sites using alignment of homologous sequences. J Mol Biol 195: 957-961