An algorithm for secondary structure determination in proteins based on sequence similarity

Jonathan M. Levin, Barry Robson[+] and Jean Garnier*

Laboratoire de Biochimie physique, INRA, Université de Paris Sud, 91405 Orsay Cédex, France

Received 3 June 1986; revised version received 21 July 1986

A secondary structure prediction algorithm is proposed on the hypothesis that short homologous sequences of amino acids have the same secondary structure tendencies. Comparisons are made with the secondary structure assignments of Kabsch and Sander from X-ray data [(1983) Biopolymers 22, 2577–2637] and an empirically determined similarity matrix which assigns a sequence similarity score between any two sequences of 7 residues in length. This similarity matrix differs in many respects from that of the Dayhoff substitution matrix [(1978) in: Atlas of Protein Sequence and Structure, (Dayhoff, M.O. ed). vol. 5. suppl. 3, pp. 353–358, National Biochemical Research Foundation, Washington, DC]. This homologue method had a prediction accuracy of 62.2% over 3 states for 61 proteins and 63.6% for a new set of 7 proteins not in the original data base.

Protein structure *Secondary structure prediction* *Amino acid sequence homology*

1. INTRODUCTION

Methods for the prediction of secondary structure have been in existence for more than 15 years [1–10] with an accuracy of prediction over 3 states (helix, sheet, coil) between 55 and 58% [11]. With the increasing number of amino acid sequences determined by gene sequencing methods the need for improved predictions has never been greater. An aid in the search for sequence-structure relationships has been the appearance of protein data banks containing observed secondary structure based on crystallographic data. The most widely used of which is that of Kabsch and Sander [12], whose assignments are essentially based on hydrogen-bonding patterns.

* To whom correspondence should be addressed
[+] Present address: Epsitron Peptide and Protein Engineering Unit, Department of Biochemistry, University of Manchester, Manchester M13 9PT, England

The algorithm described in this paper works by predicting secondary structure using the information contained in the Kabsch and Sander data base [12]. The algorithm is based on the hypothesis that short homologous sequences of amino acids have the same secondary structure tendencies even if they come from nonhomologous proteins. This was initially explored by P.W. Finn and B. Robson using the Dayhoff substitution matrix [13] (unpublished). There are two major difficulties behind an approach of this sort; one is the definition of homologous sequences and the other is the retrieval of as much information as possible from the data base. The first problem has been resolved using an empirically determined similarity matrix which assigns a sequence similarity score between any two sequences, the second by considering only those sequences with a high degree of homology and then repredicting each residue up to 7 times using a window of 7 residues in length which is shifted along the sequence 1 residue at a time.

Besides its present success in predicting sec-

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/86/$3.50 © 1986 Federation of European Biochemical Societies

ondary structures an interesting feature of this method is its potential for improvement, in a natural way by increasing the data base, by refinement of the similarity matrix, and by combination with the GOR method [9] in a more integrated way.

2. MATERIALS AND METHODS

All calculations were performed on a Sperry 1100/92 using a Fortran 77 source code. The algorithm can be divided into 2 parts; the first is concerned with the search for homologous sequences, the second with the assignment of secondary structure. The similarity matrix which gives a score showing the degree of similarity between any two sequences of amino acids was developed and optimised using the Kabsch and Sander data base. Initially arbitrary assignments for the matrix were made. A value of 2 was given for each element of the principle diagonal containing the identities. A value of 1 was given for pairs of amino acids considered to have properties in common and -1 for dissimilar amino acid pairs. All other pairs were given a value of 0. These original assignments were optimised by making rational changes to the matrix and observing their effect on the prediction accuracy. A more rigorous approach was ruled out due to the number of possible modifications which could be made and the time taken to predict the entire data base. The similarity matrix used is shown in fig.1. The algorithm makes a comparison between every sequence of 7 residues in length in the test protein (there are $n-6$ where n is the number of residues in the protein) and every 7-residue fragment in the data base. This is an iterative procedure, i.e. first residues 1–7 are compared against the data base then residues 2–8, etc. Each time a comparison is made the similarity score between the 2 sequences is calculated. If the calculated score is less than 7 the 2 sequences are considered to be insufficiently similar in terms of their secondary structure tendencies and are thus rejected. For example, if one considers the test sequence STNGIYW then the sequence ATSLVFW which has a score of 6 would be rejected whereas the sequence ATSGVFL which has a score of 7 would be accepted. Every time a sequence is found whose score is greater than or equal to 7 its observed conformation is assigned to the test sequence with its

```
G   2
P   0 3
D   0 0 2
E   0-1 1 2
A   0-1 0 1 2
N   0 0 1 0 0 3
Q   0 0 0 1 0 1 2
S   0 0 0 0 1 0 0 2
T   0 0 0 0 0 0 0 0 2
K   0 0 0 0 0 1 0 0 0 2
R   0 0 0 0 0 0 0 0 0 1 2
H   0 0 0 0 0 0 0 0 0 0 0 2
V  -1-1-1-1 0-1-1-1 0-1-1-1 2
I  -1-1-1-1 0-1-1-1 0-1-1-1 1 2
M  -1-1-1-1 0-1-1-1 0-1-1-1 0 0 2
C   0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2
L  -1-1-1-1 0-1-1-1 0-1-1-1 1 0 2 0 2
F  -1-1-1-1-1-1-1-1-1-1-1-1 0 1 0-1 0 2
Y  -1-1-1-1-1-1-1-1-1-1-1 0 0 0 0-1 0 1 2
W  -1-1-1-1-1-1-1-1-1-1 0-1 0 0 0-1 0 0 0 2

   G P D E A N Q S T K R H V I M C L F Y W
```

Fig.1. The secondary structure similarity matrix. The secondary structure similarity matrix gives a score for the replacement of one amino acid by another. The amino acids Gly, Pro, Thr, His, Cys, Trp, are considered to have unique secondary structure properties, i.e. none of these amino acids have a score greater than 0 when replaced by any other amino acid. The following pairs of amino acids are considered to have secondary structure properties in common: (Asp, Glu) (Asp, Asn) (Glu, Ala) (Ala, Ser) (Gln, Asn) (Gln, Glu) (Asn, Lys) (Lys, Arg) (Val, Ile) (Val, Leu) (Met, Leu) (Ile, Phe) (Phe, Tyr). All amino acids when replaced by themselves have a score of 2 with the exception of Pro and Asn which have a score of 3. Exchanges of different residues on above pairs have a score of 1 apart from the Met, Leu pair which has a score of 2, whilst those amino acids which are very dissimilar have a score of -1. The difference in these values reflects the varying importance of the substitution of one amino acid by another in the formation of secondary structure.

similarity score (fig.2). Once every fragment in the test protein has been compared the secondary structure attributed to each residue is that which has the highest value (see fig.3). The observed conformations are those of the dictionary of Kabsch and Sander [12] and prediction accuracies were calculated as percentage of correctly predicted residues for three states, where helix is 'H', 'G' and 'I', sheet is 'E', turn is 'T' and coil is 'S', 'B'

CONFORMATION

RES	'H'	'E'	'C'	'T'	'S'	'G'	'B'	'I'
1			7+7+8					
2	7		7+8+9+9					
3	7+8+9	7	9					
4	7+8+9	7	9					
5	7+8+9	7			9			
6	7+9	7	8		9			
7			7+8	7+9	9			
8				9	9			

Fig.2. An example of the prediction algorithm. This shows how secondary structure assignments are made. In the above example 3 homologous fragments were found for residues 1–7. The first whose similarity score was 7 had an observed conformation of CHHHHHT, the second had a score of 7 and an observed conformation of CCEEEEC, and the score for the third was 8 with a conformation of CCHHHCC. 2 homologous fragments were found for residues 2–8, each had a score of 9 and a conformation of CHHHHTT and CCCSSSS. In order to avoid over-prediction of helix and under-prediction of aperiodic structure decision constants are used, i.e. the values in the column 'H' are reduced by 20%, the values in the column 'C' are increased by 5%, the values in the column 'T' by 40% and those in the column 'S' by 30%. The prediction is then based on the conformation with the highest score, so for residues 1–8 it is CCHHHHTT (see fig.3).

CONFORMATION

RES	'H'	'E'	'C'	'T'	'S'	'G'	'B'	'I'	PRED
1	0	0	23	0	0	0	0	0	C
2	6	0	35	0	0	0	0	0	C
3	19	7	9	0	0	0	0	0	H
4	19	7	9	0	0	0	0	0	H
5	19	7	0	0	12	0	0	0	H
6	13	7	8	0	12	0	0	0	H
7	0	0	16	22	12	0	0	0	T
8	0	0	0	13	12	0	0	0	T

Fig.3. An example of scoring values for an eight-state prediction. These values are for the example given in fig.2. The predicted conformation for each residue 1 to 8 is listed on column 10 according to the highest score of the eight conformations in columns 2–9.

and 'C' (in the dictionary of Kabsch and Sander the conformation 'C' is ' ').

3. RESULTS

Using 61 of the 62 proteins described by Kabsch and Sander [11], the protein rubredoxin being previously omitted due to insufficient sequence data, the homologue method had a prediction accuracy of 62.2% over 3 states (see table 1).

The results of table 1 are a significant improvement over previous methods of prediction. Some secondary structure prediction methods have been criticised on the grounds that they were better at predicting the proteins in the data base from which they were developed but had less good results when obtained for completely new proteins. We wished to develop a method that would give a good prediction for a test protein when there was no homology between it and any protein in the data base, whilst still retaining sensitivity to a homologous protein should there be one. To test this out the homologue method was used to predict 7 proteins not in the data base (kindly provided by Dr Chris Sander). These proteins were rubredoxin, hydrolase (acid protease), aspartate transcarbamylase, catalase, human hemoglobin, malate dehydrogenase and Bence-Jones protein. The results are shown in tables 2 and 3. As can be seen these results show the homologue method has produced the same level of prediction accuracy with the new proteins. This compares very favourably with one of the most widely used secondary structure prediction methods, the GOR method of Garnier et al. [9], which gave 53% accuracy for the same 7 proteins.

4. DISCUSSION

The ability to design a secondary structure prediction method based on a search for homologous fragments between non-homologous proteins has only recently become feasible with the determination of the 3-dimensional structure of a sufficiently large number of proteins. However, each increase in the data base renders such a method more powerful, as a larger data base means a higher frequency of similar sequences even for very unusual fragment sequences as well as an increased chance of finding a related protein in the data base. A possibility that can be developed with a larger data base is, instead of predicting an unknown protein against the whole

Table 1

Prediction on 61 proteins

Conformation	Numbers of residues observed	Numbers of residues predicted	Numbers of residues correctly predicted	Percentage of correctly predicted residues
Helix	3110	3142	1816	58.4
Sheet	2168	1976	1000	46.12
Aperiodic	5421	5581	3838	70.8
Total	10699	10699	6654	62.2

The proteins are those listed in [11] except rubredoxin. The percentage correct is: (number of residues correctly predicted/number of residues observed) × 100. Each test protein was removed from the protein data base for its own prediction

Table 2

Prediction results for the new proteins

Conformation	Numbers of residues observed	Numbers of residues predicted	Numbers of residues correctly predicted	Percentage of correctly predicted residues
Helix	680	668	422	62.1
Sheet	347	358	161	46.4
Aperiodic	1031	1032	725	70.3
Total	2058	2058	1308	63.6

data base, to predict the unknown protein against a subset of proteins which share some of its properties, for example similar size or similar evolutionary function, etc. These points emphasize an advantage for this kind of prediction method with respect to other methods in that it will continue to improve with an increase in the knowledge base.

An examination of the results shown in tables 1 and 2 shows that the proportions of the 3 conformations in the dictionary are very well predicted even in the case of the 7 new proteins. This is the result of the 'universal' decision constants (not changed for each protein or class) applied which reduce the number of residues predicted as helix and increase the number of residues predicted as aperiodic. Although the proportions of the various conformations are well predicted throughout the data base, the degree of accuracy varies depending on the conformation, with that for the aperiodic being significantly better. What this means is that this method can distinguish between periodic and non-periodic secondary structure whereas the type of periodic conformation, sheet or helix, is more difficult to determine. This might imply that certain sequences within a protein have an equal tendency for sheet or helix, but that the choice between them is determined by the long-range interactions. This conforms with the finding by Kabsch and Sander [14] that identical pentapeptides had different conformations. It also contradicts an assumption often made when considering a homologous fragment based prediction method. The assumption is that a very strict criterion for sequence homology is necessary in order to find 2 sequences with the same conformation. The success of the homologue method lies in looking for a large number of fairly similar

Table 3

Comparison between the GOR and homologue predictions for the new proteins (3 states)

Protein	Homologue prediction (%)	GOR prediction (%)	Number of residues
Acid protease[a] (2APE)	60.4	49.7	318
Aspartate transcarbamylase, chain 1 (4ATC)	59.4	51.0	310
Aspartate transcarbamylase, chain 2 (4ATC)	54.3	43.8	153
Catalase (7CAT)	61.7	60.2	498
Human hemoglobin chain 1 (2HHB)[a]	89.4	54.6	141
Human hemoglobin chain 2 (2HHB)[a]	76.7	52.7	146
Malate dehydrogenase (2MDH)	55.9	47.2	324
Bence-Jones protein[a] (2RHE)	77.2	59.6	114
Rubredoxin (4RXN)	64.8	61.1	54
Total	63.6	53.0	2058

[a] These predictions are very favourable due to the presence of homologous proteins in the data base. If these homologous proteins are removed the accuracy of prediction falls. With the removal of acid protease from the data base the prediction for 2APE falls to 58%, the removal of horse hemoglobin leads to a prediction accuracy of 72 and 57% for human hemoglobin chain 1 and 2, respectively, and the removal of Bence-Jones immunoglobulin and lambda immunoglobulin FAB causes the accuracy of prediction for 2RHE to fall to 61%

sequences and making a consensus prediction rather than looking for a very small number of very similar sequences. The optimal cut-off was found to be at 7. This allows for a large enough sample of peptides to be considered such that no peptide exists in the data base which does not have several homologous counterparts with a score greater than or equal to 7. In fact, it was observed that when the cut-off was increased from 7 to 8 the percentage accuracy fell, this being due to certain portions of the test protein sequence where only 2 or 3 similar peptides could be found. On the other hand, if no cut-off was used the prediction was swamped by the 'noise' within the data base and the entire protein sequence was predicted as aperiodic (the most commonly observed conformation in the data base). The length of the prediction window is also important, with 7 being the optimal length found, although the performance of the method when using a window of 6 was not much lower and windows up to 15 have been tried. Deterioration at higher window values seems due to the reduction in the implied level of data available, i.e. larger data bases could be required to exploit larger windows better.

The success of this method is very much based on the similarity matrix. At the start of the development of this method the Dayhoff similarity

matrix was used [13]. This matrix is based on the frequency of replacement of one amino acid by another within a family of proteins of identical function and is used as a standard when trying to determine sequence homology. However, it soon became apparent that as far as the determination of homologous sequences with similar secondary structure this was not the optimal matrix to use. The similarity matrix published in this paper has been optimised for use in secondary structure prediction and differs in many respects from the Dayhoff matrix. This implies that for applications other than the determination of homologous sequences the Dayhoff matrix might not be the best one to use.

It is hoped that by using the homologue method in conjunction with the GOR method and a pattern recognition process, one will be able to assign correctly the secondary structure motifs for an unknown protein, which can then be used as a starting point for tertiary structure calculations.

ACKNOWLEDGEMENTS

We thank J.F. Gibrat who performed the GOR predictions on the 7 new protein data set and for his help to J.M.L. for starting this work. J.M.L. holds an MRT studentship. Computing was performed at the Paris Sud Informatique Computing Center under a grant from the Ministère de l'Education Nationale.

REFERENCES

[1] Schiffer, M. and Edmundson, A.B. (1967) Biophys. J. 7, 121–135.
[2] Pain, R.H. and Robson, B. (1970) Nature 227, 62–63.
[3] Lewis, P.N., Go, N., Kotelchuk, D. and Scheraga, H.A. (1970) Proc. Natl. Acad. Sci. USA 65, 810–815.
[4] Nagano, K. (1973) J. Mol. Biol. 75, 401–420.
[5] Kabat, E.A. and Wu, T.T. (1973) Proc. Natl. Acad. Sci. USA 70, 1473–1477.
[6] Robson, B. (1974) Biochem. J. 141, 853–867.
[7] Lim, V.I. (1974) J. Mol. Biol. 88, 857–894.
[8] Chou, P.Y. and Fasman, G.D. (1974) Biochemistry 13, 222–244.
[9] Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97–120.
[10] Busetta, B. and Hospital, M. (1982) Biochim. Biophys. Acta 701, 111–118.
[11] Kabsch, W. and Sander, C. (1983) FEBS Lett. 155, 179–182.
[12] Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577–2637.
[13] Schwartz, R.M. and Dayhoff, M.O. (1978) in: Atlas of Protein Sequence and Structure (Dayhoff, M.O. ed.) vol.5, suppl.3, pp.353–358, National Biomedical Research Foundation, Washington, DC.
[14] Kabsch, W. and Sander, C. (1984) Proc. Natl. Acad. Sci. USA 81, 1075–1078.