CABIOS

Vol. 11 no. 6 1995
Pages 681–684

SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments

C.Geourjon and G.Deléage[1]

Abstract

Recently a new method called the self-optimized prediction method (SOPM) has bèen described to improve the success rate in the prediction of the secondary structure of proteins. In this paper we report improvements brought about by predicting all the sequences of a set of aligned proteins belonging to the same family. This improved SOPM method (SOPMA) correctly predicts 69.5% of amino acids for a three-state description of the secondary structure (α-helix, β-sheet and coil) in a whole database containing 126 chains of non-homologous (less than 25% identity) proteins. Joint prediction with SOPMA and a neural networks method (PHD) correctly predicts 82.2% of residues for 74% of co-predicted amino acids. Predictions are available by Email to deleage@ibcp.fr or on a Web page (http://www.ibcp.fr/predict.html).

Introduction

Numerous methods have been developed to predict the secondary structure of proteins from their amino acid sequences (for a recent review see Eisenhaber *et al.*, 1995). Currently available methods have success rates ranging from 56 to 72% for a three-state (α-helix, β-sheet and aperiodic states) description of secondary structure. The number of proteins with known structure has increased at an average rate of more than 150 new structures elucidated per year (Lattman, 1994). Even though the size of the database of secondary structures has not grown with the same rate, since all the proteins should not present too much identity to be incorporated in it, this increase does not lead to a concomittant increase in prediction accuracy. At the same time the number of proteins that belong to a given family has grown with the increasing size of the protein sequence database and this classification into known families has been used with success (Boscott *et al.*, 1993). Moreover, several groups have already incorporated multiple alignment information to increase the success rate in secondary structure prediction (Levin *et al.*, 1986; Rost and Sander, 1994a; Di Francesco *et al.*, 1995). Recently, we have described a new method called SOPM (self-optimized prediction method) to predict the secondary structure of a given protein (Geourjon and Deléage, 1994). Briefly, this method: (i) builds a limited database of protein sequences with their known secondary structures; (ii) predicts the secondary structure of all the proteins of the database using a similarity algorithm; (iii) determines the prediction parameters that maximize the accuracy of the prediction; (iv) applies the prediction parameters to the given protein.

In this paper we have investigated the possibility of increasing the prediction accuracy of SOPM by taking into account the information brought about by multiple alignment of related protein sequences. The results are a 4% additional gain in predictive power of the self-optimized prediction method (a single sequence-based predictive scheme) when measured on the same database. Thus the global method, called SOPMA (self-optimized prediction from multiple alignment) now reaches 73.2%. On a more restrictive database (25% identity threshold) towards homology (Rost and Sander, 1993) the success rate is 69.5%, and 82% if joint prediction with the PHD neural method (Rost and Sander, 1993) is taken into account.

Systems

All calculations were carried out on an IBM rs6000 560 workstation. All the programs were written in a Fortran F77 compatible language (IBM VS-FORTRAN 2.3 compiler) using AIX 3.2 (Unix) as the operating system. Thus portability is warranted for most machines working under Unix (Silicon Graphic, SUN and Hewlett Packard).

Methods

The general flow chart is given in Figure 1 for each sequence of the reference database. The first step consists of searching for homologous proteins in the SWISSPROT sequence database (Bairoch and Boeckman, 1994) using the FASTA (Pearson and Lipman, 1988) program. A sequence is retained if its OPT score is greater than 80. The most homologous (limited to a set of 25) sequences are extracted from the SWISSPROT database. The second

Institut de Biologie et de Chimie des Protéines, UPR 412-CNRS, 7 passage du Vercors, F-69367 Lyon cedex 07, France

[1] *To whom correspondence should be addressed. Email deleage@ibcp.fr*

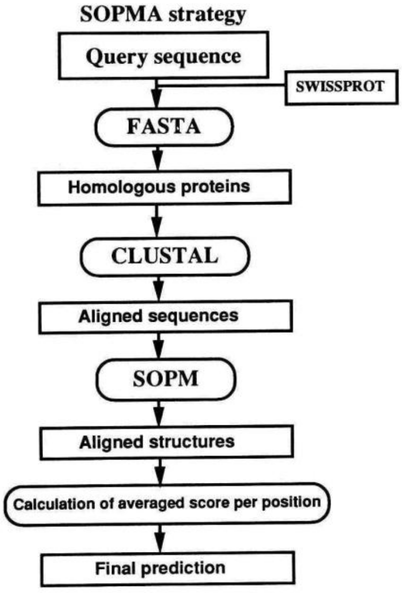

Fig. 1. Logical strategy of the SOPMA method.

step consists of alignment of the sequences constituting the set of homologous proteins using the CLUSTAL program (Higgins and Sharp, 1988) with default parameters. Then, the third step is to apply the SOPM method to each of the aligned sequences. For each amino acid position in the multiple alignment the conformational score of each state is averaged over all the sequences at the given position. In the multiple alignment a nil value is attributed to each gap. Finally, the conformational state yielding the highest score is attributed to the given amino acid with the averaged conformational score.

Prediction accuracy

Different ways of calculating the prediction accuracy have been used.

The most commonly used way of expressing prediction accuracy is the percentage of correctly predicted residues, $Q_3\%$ or $Q_4\%$ for 3 and 4 states respectively. The second way of expressing the accuracy is the correlation coefficient C described by Matthews (1975):

$$C(i) = \frac{(p_i.n_i) - (u_i.o_i)}{\sqrt{(n_i + u_i)(n_i + o_i)(p_i + u_i)(p_i + o_i)}}$$

where i designates one state among k possibilities, p_i is the number of residues correctly predicted and observed to belong to state i, n_i is the number of residues correctly predicted and not observed to belong to state i, u_i is the number of residues not predicted but observed to belong to state i and o_i is the number of residues predicted but not observed to belong to state i. A value of 1 indicates a fully correlated prediction, a zero value a non-correlated prediction and a negative value a negatively correlated prediction.

Another interesting parameter is the root mean square deviation (r.m.s.) σ of the estimation of secondary structure content from a prediction method. The third parameter is the segment overlap (SOV) (Rost and Sander, 1994b). This parameter measures the segmental accuracy rather than a residue per residue accuracy. It is calculated as follows:

$$\mathrm{SOV}^{\delta} = \frac{1}{N} \times \sum_{s} \frac{\min[end(s1); end(s2)] - \max[beg(s1); beg(s2)] + 1 + \delta}{\max[end(s1); end(s2)] - \min[beg(s1); beg(s2)] + 1} * len(s1)$$

where $\min[a; b]$ is the minimum of the values a and b and $\max[a; b]$ is the maximum: *len* is the length of the given segment. δ is a parameter that takes into account the uncertainty of secondary structure boundaries definition from three-dimensional data.

$$\delta \leqslant \min \left\{ [\max ov(s1; s2) - \min ov(s1; s2)]; \min ov(s1; s2); \frac{len(s1)}{2} \right\}$$

where $\min ov$ and $\max ov$ are the nominator and the denominator of the SOV equation respectively.

Results and discussion

The new version of the SOPM method has been applied using a jack-knife procedure to two recently described

Table I. Accuracy of the SOPMA method on the 239 protein chains of the DATABASE.DSSP (50% identity cut-off)

State	Observed	Predicted	Correct	Q_3	Sigma	C	SOV	Q_4
Helix	14 002	13 641	10 101	72.1	8.7	0.62	0.74	71.9
Sheet	10 396	9960	6794	65.4	9	0.57	0.78	64.2
Coil	21 825	22 622	16 918	77.5	9.4	0.54	0.67	68.9
Turn								29.5
Total	46 223	46 223	33 813	73.2			0.66	64.2

Table II. Accuracy of the SOPMA method on the Rost and Sander database (25% identity cut-off)

State	Observed	Predicted	Correct	Q_3	Sigma	C	SOV	Q_4
Helix	7390	7023	5024	70.4	12.0	0.56	0.74	70.0
Sheet	4958	4786	2991	60.3	10.8	0.51	0.72	60.2
Coil	10 742	11 282	8040	74.8	12.0	0.48	0.63	66.5
Turn								25.0
Total	23 091	23 091	16 055	69.5			0.68	61.5

databases; DATABASE.DSSP (Geourjon and Deléage, 1994) and the Rost and Sander (1993) databases. The main difference between the two databases lies in the identity cut-off for removing proteins chains. The DATABASE.DSSP reference database has been built using a 50% identity cut-off level and contains 239 chains that represent 46 223 amino acids. The Rost and Sander database has been established using a 25% identity cut-off level and contains 126 chains that yield 23 091 amino acids. It has to be mentioned that β-turn information has been taken into account, since users are most often interested in a four-state prediction, thus including turn information.

The results obtained by applying the SOPMA method to DATABASE.DSSP are given in Table I for 3 and 4 states. The global Q_3 value (see Methods) is as high as 73.2% using a jack-knife procedure (leaving one protein out and making predictions on that protein and repeating the procedure for all proteins in the data set). Considering the turn information drops Q_4% down to 64.2%. The correlation coefficients C reported in Table 1 are 0.62 for the α-helix state, 0.57 for the β-sheet state and 0.54 for the coil state. The r.m.s. deviation is as low as 8.7% for the α-helix state, 9% for the β-sheet state and 9.4% for the coil state, showing a rather good estimation of the secondary structure content of a protein by this method. The SOV parameters (see Methods) are 0.74, 0.78 and 0.67 for the helix, sheet and coil states respectively. However, one has to check the secondary structure predictive methods on proteins sequences that are clearly not related (when homology modelling is not confidently applicable). In other words, the method has to be checked on a database containing protein sequences that share less than 25% identity. A reference database that contains 126 non-related protein sequences corresponds to this criteria and has been given by Rost and Sander (1993).

The results obtained by applying the SOPMA method to the 126 proteins are given in Table II. The global Q_3 value (see Methods) is 69.5%. When the turn state is also considered Q_4 is 61.5%. Thus the decrease in accuracy is lower on the Rost and Sander database (7%) than on our database (9%; see above). The correlation coefficients C reported in Table II are 0.56 for the α-helix state, 0.51 for the β-sheet state and 0.48 for the coil state. The r.m.s. deviation is as low as 12% for the α-helix state, 10.8% for the β-sheet state and 12% for the coil state. Obviously, the success rate is dependent upon the similarity level between the protein sequences contained in the database. This fact is very useful for a user who wants to derive the secondary structure of related sequences (to look for mutations that potentially affect regular secondary structures). These results have to be compared with those obtained by other methods. To date the most powerful methods to predict the secondary structure of proteins are based on neural networks. Indeed, the PHD method has a success rate of 72% for a three-state description of protein secondary structure. When checked on the same database as the PHD method our SOPMA method yields 69.5% of correctly predicted residues (three states, see Table II). Thus our SOPMA method constitutes a valuable alternative to the neural networks-based methods, which can be biased by a dependence between the learning and the training sets of proteins. Anyway, having two methods with high accuracy based on different principles available for the experimentalist is a decisive advantage that can (should) be used in joint prediction. Although not new, the joint prediction approach allows a cross-validation that improves individual methods and is particularly useful for the user.

Table III. Predictive success of joint prediction between SOPMA and neural net method PHD checked on the Rost and Sander database (25% identity cut-off)

State	Observed	Predicted	Percent joint	Correct	Q_3	C	SOV
Helix	7390	5191	70.2	4475	86.2	0.63	0.73
Sheet	4958	3313	66.8	2583	78.0	0.56	0.73
Coil	10 742	8552	79.6	6957	81.3	0.54	0.64
Total	23 091	17 056	73.9	14 015	82.2		0.68

Joint prediction with neural network

Generally a user is more interested in a knowledge of the secondary structure of given segments, rather than a global prediction of the complete sequence. In this context, joint prediction is a powerful way to address this problem, provided the joined methods are based on different principles. Since the neural network-based methods have proved to be very efficient, we have made a joint prediction with the PHD method (Rost and Sander, 1993, 1994) and our SOPMA method (this work). The results of this joint prediction obtained on the Rost and Sander database are given in Table III. The global Q_3 value (see Methods) is as high as 82% for 74% of jointly predicted residues (residues predicted as in the same conformational state by both methods). That means that the user has access to a cross-validation of different methods. The most significant improvement brought out by co-prediction is an increase of 10% in the success rate on jointly predicted segments. This clearly shows that both methods are not redundant and that they take benefit from the joint prediction, one from the other. This is valuable information for all users of secondary structure prediction methods, to know that two methods based on different principles yield identical results on some easily identifiable parts of the sequence. Moreover, a better agreement between predicted and observed length in the helical structures (see SOV parameters) is obtained, indicating that one would expect a good predictive power from both methods. This means that methods of predicting protein secondary structures are now able to locate most of the stretches with regular structures and that recognition of folding patterns can be investigated in the best way. However, efforts have still to be made to improve turn prediction.

Availability and mailservers

The SOPM and SOPMA methods are available by anonymous ftp to ibcp.fr or can be obtained through a mailserver (HELP to deleage@ibcp.fr to get information). Alternatively, the SOPM/SOPMA methods are reachable on our Web page (http://www.ibcp.fr/predict.html). Firstly, our mailserver makes a prediction of the secondary structure using SOPMA and forwards the request to predictprotein@EMBL-heidelberg.de for PHD prediction (Rost *et al.*, 1993). A consensus prediction is generated on the fly. The user will also receive the FASTA search result file, the multiple alignment (CLUSTAL) of the related sequences and the potential sites and signatures detected from the PROSITE library (Bairoch, 1994) using our pattern algorithm (Geourjon and Deléage, 1993). The SOPMA program with the corresponding database can be obtained for non-commercial use by anonymous ftp (ibcp.fr) and can be invoked within the ANTHEPROT suite of programs for protein sequence analysis (Geourjon and Deléage, 1995).

Acknowledgements

C.Geourjon is a recipient of a fellowship from the Région Rhône-Alpes.

References

Bairoch,A. (1994) PROSITE: recent developments. *Nucleic Acids Res.*, **22**, 3578–3580.

Bairoch,A. and Bockmann,B. (1994) The SWISS-PROT protein sequence data bank: current status. *Nucleic Acids Res.*, **22**, 3578–3580.

Boscott,P.E., Barton,G.J. and Richards,W.G. (1993) Secondary structure prediction for modelling by homology. *Protein Engng*, **6**, 261–266.

Di Francesco,V., Munson,P.J. and Garnier,J. (1995) Use of multiple alignments in protein secondary structure prediction. *Proceedings of the 28th Annual Hawaii International Conference on System Sciences*, Vol. 5, pp. 285–291.

Eisenhaber,F., Persson,B. and Argos,P. (1995) Protein structure prediction: recognition of primary, secondary and tertiary structural features from amino acid sequence. *Crit. Rev. Biochem. Mol. Biol.*, **30**, 1–94.

Geourjon,C. and Deléage,G. (1993) Interactive and graphic coupling between multiple alignments, secondary structures prediction and motif/pattern scanning into proteins sequences. *Comput. Applic. Biosci.*, **9**, 87–91.

Geourjon,C. and Deléage,G. (1994) SOPM: a self-optimised method for protein secondary structure prediction. *Protein Engng*, **7**, 157–164.

Geourjon,C. and Deléage,G. (1995) ANTHEPROT 2.0: a three dimensional module fully coupled with protein sequence analysis methods. *J. Mol. Graphics*, **13**, 209–212.

Higgins,D.G. and Sharp,P.M. (1988) CLUSTAL: a package for performing multiple sequence alignments on a microcomputer. *Gene*, **73**, 237–244.

Lattman,E.E. (1994) Protein crystallography for all. *Proteins*, **18**, 103–106.

Levin,J.M. and Garnier,J. (1988) Improvements in a secondary prediction method based on a search for local sequence homologies and its use as a model building tool. *Biochim. Biophys. Acta*, **955**, 283–295.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.

Rost,B., Sander,C. and Schneider,R. (1993) PHD—an automatic mail server for protein secondary structure prediction. *Comput. Applic. Biosci.*, **10**, 53–60.

Rost,B. and Sander,C. (1994a) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, **19**, 55–72.

Rost,B. and Sander,C. (1994b) Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.*, **235**, 13–26.

Received on June 30, 1995; revised on September 4, 1995; accepted on September 7, 1995