Protein Engineering vol.11 no.4 pp.249–251, 1998

SHORT COMMUNICATION

Protein secondary structure prediction in different structural classes

M.Michael Gromiha[1,2] and S.Selvaraj[3]

[1]Tsukuba Life Science Center, Institute of Physical and Chemical Research (RIKEN), 3–1–1 Koyadai, Tsukuba, Ibaraki 305, Japan and [3]Department of Physics, Bharathidasan University, Tiruchirapalli 620 024, Tamil Nadu, India

[2]To whom correspondence should be addressed. E-mail: gromiha@rtc.riken.go.jp

Information about the secondary structure of a protein can be helpful in understanding its native folded state. In previous work, it was shown that the medium-range interactions predominate in all-α class and the long-range interactions predominate in all-β class proteins. Based on this, in this work the performance of several structure prediction methods in different structural classes of globular proteins was analyzed. It was found that all the methods predict the secondary structures of all-α proteins more accurately than other classes.

Keywords: globular proteins/long-range interactions/medium-range interactions/secondary structure prediction/structural class

Introduction

The prediction of the native conformation of proteins from their amino acid sequences is one of the most challenging problems in molecular biology. Secondary structure prediction is an important intermediate step in this process. The known protein structures have been classified into four structural classes, namely all-α, all-β, α/β and α + β (Levit and Chothia, 1976; Richardson and Richardson, 1989). The information about the structural class may help to improve the accuracy levels of secondary structure prediction schemes in globular proteins.

Several methods have been proposed for predicting the structural classes (Chou and Zhang, 1993; Metfessel *et al.*, 1993; Boberg *et al.* 1995; Chou, 1995; Zhang *et al.*, 1995) and secondary structures of globular proteins. They are based on different algorithms, such as statistical analysis (Chou and Fasman, 1974), information theory (Garnier *et al.*, 1978), neural networks (Qian and Sejnowski, 1988; Kneller *et al.*, 1990; Chandonia and Karplus, 1996), nearest neighbour methods (Yi and Lander, 1993), multiple alignment (Russell and Barton, 1993; King and Sternberg, 1996), combination of multiple alignment and neural networks (Rost and Sandor, 1994), hydrophobicity profiles (Cid *et al.*, 1992; Gromiha and Ponnuswamy, 1995), pairwise alignment (Frishman and Argos, 1997) and 3D–1D compatibility (Ito *et al.*, 1997). The accuracy levels in these methods are in the range 65–80%.

In previous work (Gromiha and Selvaraj, 1997), it was observed that for all-α proteins the average contact in the medium range is comparably higher and that in the long range is comparably lower. Further, a markedly opposite trend was observed for all-β proteins. This implied that secondary structure formation is more influenced by medium-range interactions in all-α proteins whereas it is influenced by long-range interactions in all-β proteins. As most of the secondary structure prediction algorithms take into account only the effect of neighbouring residues along the sequence which include short- and medium-range interactions, it is hypothesized that secondary structure prediction of all-α proteins would be better than that of the other classes of proteins.

In this work, we categorize the proteins into three different structural classes (Gromiha and Ponnuswamy, 1995; Michie *et al.*, 1996; Orengo *et al.*, 1997) all-α, all-β and mixed (combination of α/β and α + β) and compute the accuracy of secondary structure prediction for each structural class. The predicted accuracies are compared among them and the hypothesis 'better prediction of all-α proteins' is tested.

Materials and methods

Assignment of structural class

The proteins are classified into three structural classes based on the criterion of Chou (1995) and Kneller *et al.* (1990). All-α proteins have at least 40% α-helical content and <5% β-strand content. All-β proteins have at least 40% β-strand content and <5% α-helical content. Mixed class is the combination of α + β and α/β class which contains more than 15% α-helical and 15% β-strand contents.

Computation of secondary structure prediction accuracy for the structural classes

Several different prediction reports available in the literature were collected (Chou and Fasman, 1978; Garnier *et al.*, 1978; Qian and Sejnowski, 1988; Kneller *et al.*, 1990; Rost and Sandor, 1994; Gromiha and Ponnuswamy, 1995; Chandonia and Karplus, 1996; Ito *et al.*, 1997). The predicted accuracies for each protein were grouped according to their structural class (we used the above method unless it is specified explicitly).The average accuracy (weighted average) was calculated using the relation

$$\text{Accuracy } (\%) = \Sigma n_i Q_i / \Sigma n_i \qquad (1)$$

where n_i is the number of residues in each protein and Q_i is the predicted accuracy of the protein.

Medium- and long-range interactions

The average medium and long-range contacts are computed as detailed in our previous paper (Gromiha and Selvaraj, 1997). For a given residue, the composition of surrounding residues within a sphere of 8 Å (Manavalan and Ponnuswamy, 1977, 1978) was computed. The residues that are within a distance of two residues (in the sequence) from the central residue are considered to contribute to short-range interactions (Ponnuswamy *et al.*, 1980), those within a distance of four residues to medium-range interactions (Ponnuswamy *et al.*, 1973) and those more than four residues away to long-range interactions.

Table I. Accuracy of prediction in different structural classes with eight methods

Method	All-α	All-β	Mixed
Chou and Fasman (1974) (statistical analysis)	62.2	49.5	55.3
Garnier *et al.* (1978) (information theory)	61.9	58.2	58.1
Qian and Sejnowski (1988) (neural network)	67.0	64.0	64.0
Kneller *et al.* (1990) (enhanced neural network)	79.0	70.0	64.0
	94.0 (3ICB)	79.0 (2KAI)	73.0 (3PGM)
Rost and Sandor (1994) (neural network, PHD)	80.8	68.8	72.4
Gromiha and Ponnuswamy (1995) (hydrophobicity profile method)	84.1	83.6	79.8
	97.2 (1PPT)	88.5 (1APR)	86.2 (1BP2)
Chandonia and Karplus (1996) (neural network)	71.1	66.3	66.9
Ito *et al.* (1997) (3D–1D compatibility)	74.5	70.6	67.3
	86.0 (2SPC)	75.2 (2PCD)	78.6 (2NAC)

The maximum predictive accuracy of a protein reported by three methods is given in the second row with the protein name: 3ICB, calcium binding protein; 2KAI, kallikrein A; 3PGM, phosphoglycerate mutase; 1PPT, avian pancreatic polypeptide; 1APR, acid protease; 1BP2, phospholipase A2; 2SPC, spectrin; 2PCD, protocatechuate 3,4-dioxygenase; 2NAC, NAD-dependent formate dehydrogenase.

Table II. Average short-, medium- and long-range contacts in different structural classes of globular proteins

Class	N_p	N_r	Average residue contacts		
			Short	Medium	Long
All-α	14	1728	3.951	2.802	2.361
All-β	16	2353	3.950	0.921	5.201
Mixed	33	6810	3.965	1.902	3.982

N_p and N_r are the number of proteins and number of residues in each structural class, respectively.

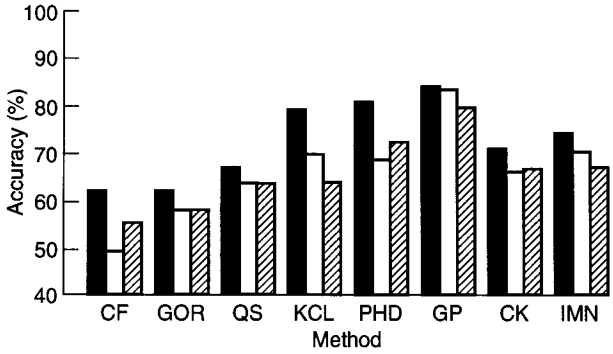

Fig. 1. Average predictive accuracy levels of the three structural classes of globular proteins by eight different methods. Black, all-α; white, all-β; hatched, mixed. CF, Chou and Fasman (1978); GOR, Garnier *et al.* (1978); QS, Qian and Sejnowski (1988); KCL, Kneller *et al.* (1990); PHD, Rost and Sandor (1994);GP, Gromiha and Ponnuswamy (1995); CK, Chandonia and Karplus (1996); IMN, Ito *et al.* (1997).

Table III. Predicted accuracy of helices and strands by different methods

Method	Helices	Strands
Chou and Fasman (1974)	64.45	60.01
Deleage *et al.* (1988)	67.62	64.1
Garnier *et al.* (1978)	64.47	62.97
Ptitsyn and Finkelstein (1983)	74.19	57.69

Data from Fiser *et al.* (1997).

Results and discussion

The average predicted accuracy of all the three structural classes with different methods are given in Table I. It can be seen that the 3D–1D compatibility method predicts a maximum of 86% (2SPC) in the set of all-α proteins, better than the other two structural classes. Also, the algorithm with the enhanced neural network and the hydrophobicity profile method predict the α-helical segments in all-α proteins with a highest accuracy of 94% (3ICB) and 97% (1PPT), respectively. The highest levels of accuracy for all-β proteins are 75.2, 79.0 and 88.5%, respectively, less than 10% of that of all-α proteins.

The medium- and long-range interactions computed for all three structural classes are given in Table II. It can be seen that average residue contacts in the medium-range are greater in the all-α proteins than the all-β proteins, whereas average long-range contacts are greater for the all-β type proteins.

The average predicted accuracy levels of all the structural classes by different methods are displayed in Figure 1. From this figure (and also from Table I) we observe that all the methods predict the secondary structures of the all-α class better than those of the other classes. The enhanced neural network method (Kneller *et al.*, 1990) and the PHD algorithm (Rost and Sandor, 1994) predict with a difference of 9%. Also other methods predict about 5% greater accuracy in the all-α class.

Recently, Fiser *et al.* (1997) stressed the importance of long-range interactions for protein structure prediction. They compared the predictive accuracies of secondary structures by different methods. The results for helices and strands are given in Table III. It can be seen that the helical segments are predicted better than β-strands. The ALB method (Ptitsyn and Finkelstein, 1983) identifies the helices with 74% accuracy whereas the accuracy for β-strands is 58%, with a difference of 16%.

Conclusions

This evaluation of different secondary structure prediction methods revealed that in spite of the differences in their approach, they are able to predict uniformly the secondary structures of proteins belonging to the all-α class better than those of other classes of proteins. A plausible reason for this tendency is the predominant role of short- and medium-range interactions in all-α proteins. Similarly, uniformly lower accuracy in the prediction of other classes of proteins implies the dominance of long-range interactions. Hence developing secondary structure prediction techniques that are specific for

each structural class incorporating the influence of short-, medium- and long-range interactions may pave the way for improving the secondary structure prediction of proteins.

References

Boberg,J., Salakoski,T. and Vihinen,M. (1995) *Protein Engng*, **8**, 505–512.
Chandonia,J.-M. and Karplus,M. (1996) *Protein Sci.*, **5**, 768–774.
Chou,K.C. (1995) *Proteins*, **21**, 319–344.
Chou,K.C. and Zhang,C.T. (1993) *Eur. J. Biochem*., **207**, 429–433.
Chou,P.Y. and Fasman,G.D. (1974) *Biochemistry*, **13**, 222–245.
Chou,P.Y. and Fasman,G.D. (1978) *Adv. Enzymol*., **47**, 45–148.
Cid,H., Bunster,M., Canales,M. and Gazitura,F. (1992) *Protein Engng*, **5**, 373–375.
Deleage,G., Clerc,F.F., Roux,B. and Gautheron,D.C. (1988) *CABIOS*, **4**, 351–356.
Fiser,A., Dosztanyi,Z. and Simon,I. (1997) *CABIOS*, **13**, 297–301.
Frishman,D. and Argos,P. (1997) *Proteins*, **27**, 329–335.
Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) *J. Mol. Biol*., **120**, 97–120.
Gromiha,M.M. and Ponnuswamy,P.K. (1995) *Int. J. Pept. Protein Res*., **45**, 225–240.
Gromiha,M.M. and Selvaraj,S. (1997) *J. Biol. Phys*., **23**, 151–162.
Ito,M., Matsuo,Y. and Nishikawa,K. (1997) *CABIOS*, **13**, 415–423.
King,R.D. and Sternberg,M.J.E. (1996) *Protein Sci*., **5**, 2298–2310.
Kneller,D.G., Cohen,F.E. and Langridge,R. (1990) *J. Mol. Biol*., **214**, 171–182.
Levitt,M. and Chothia,C. (1976) *Nature*, **261**, 552–557.
Manavalan,P. and Ponnuswamy,P.K. (1977) *Arch. Biochem.Biophys*., **184**, 476–587.
Manavalan,P. and Ponnuswamy,P.K. (1978) *Nature*, **275**, 673–674.
Metfessel,B.A., Saurugger,P.N., Connelly,D.P. and Rich,S.S. (1993) *Protein Sci*., **2**, 1171–1182.
Michie,A.D., Orengo,C.A. and Thornton,J.M. (1996) *J. Mol. Biol*., **262**, 168–185.
Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structure*, **5**, 1093–1108.
Ponnuswamy,P.K., Warme,P.K. and Scheraga,H.A. (1973) *Proc. Natl Acad.Sci. USA*, **70**, 830–833.
Ponnuswamy,P.K., Prabakaran,M. and Manavalan,P. (1980) *Biochim. Biophys. Acta*, **623**, 301–316.
Ptitsyn,O.B. and Finkelstein,A.V. (1983) *Biopolymers*, **22**, 15–25.
Qian,N. and Sejnowski,T.J. (1988) *J. Mol. Biol*., **202**, 865–884.
Richardson,J.S. and Richardson,D.C. (1989) In Fasman,G.D. (ed.), *Prediction of Protein Structure and Principles of Protein Conformation*. Plenum Press, New York, pp. 1–98.
Rost,B. and Sandor,C. (1994) *Proteins*, **19**, 55–72.
Russell,R.B. and Barton,G.J. (1993) *J. Mol. Biol*., **234**, 951–957.
Yi,T.-M. and Lander,E.S. (1993) *J. Mol. Biol*., **232**, 1117–1129.
Zhang,C.T., Chou,K.C. and Maggiora,G.M. (1995) *Protein Engng*, **8**, 425–435.

Received November 14, 1997; revised January 8, 1998; accepted January 12, 1998