CABIOS COMMUNICATIONS

Vol.4. no.4. 1988
Pages 473–477

A protein secondary structure prediction scheme for the IBM PC and compatibles

Stavros J.Hamodrakas

Abstract

A prediction scheme has been developed for the IBM PC and compatibles containing computer programs which make use of the protein secondary structure prediction algorithms of Nagano (1977a,b), Garnier et al. (1978), Burgess et al. (1974), Chou and Fasman (1974a,b), Lim (1974) and Dufton and Hider (1977). The results of the individual prediction methods are combined as described by Hamodrakas et al. (1982) by the program PLOTPROG to produce joint prediction histograms for a protein, for three types of secondary structure: α-helix, β-sheet and β-turns. The scheme requires uniform input for the prediction programs, produced by any word processor, spreadsheet, editor or database program and produces uniform output on a printer, a graphics screen or a file. The scheme is independent of any additional software and runs under DOS 2.0 or later releases.

It is well known that the biological activity of proteins depends primarily on their spatial conformation.

The three-dimensional structure of several proteins, at atomic or near-atomic resolution, has been elucidated mainly by single crystal X-ray crystallographic methods (Bernstein *et al.*, 1977). These methods are laborious, time-consuming and expensive and require the use of suitable single crystals which are not easily produced for several proteins. Therefore, the need has arisen to obtain information about protein folding from other methods.

In the last two decades the amino acid sequences of >3000 proteins have been determined (Barker *et al.*, 1986). Since experimental findings indicate that all the necessary information for a protein to fold into its native structure is coded into its amino acid sequence (Anfinsen, 1973), several attempts have been made to predict three-dimensional structure from this sequence (Taylor, 1987 and references therein). In these attempts the correct prediction of elements of secondary structure frequently plays a key role, since these elements may represent the initial nuclei in the process of protein folding.

For various other reasons, not mentioned here but described extensively in recent reviews (Argos and Mohana Rao, 1986; Hodgman, 1986; Taylor, 1987; Argos and McCaldon, 1988), it is also important to predict correctly the secondary structure of proteins from their sequence alone.

Several prediction algorithms have been published and can be classified mainly into two categories—statistical or stereochemical—but their success has been rather limited (Kabsch and Sander, 1983a; Argos and McCaldon, 1988). It has previously been claimed that combined prediction schemes provide a higher degree of accuracy than individual prediction methods (Schulz *et al.*, 1974; Argos *et al.*, 1976). Since the microcomputer has become a powerful and inexpensive laboratory tool, it would be convenient to have compact, in-house programs for predicting protein secondary structure.

This paper describes a joint prediction scheme developed for the IBM PC/XT/AT and compatibles, which employs some popular prediction methods. The package can be modified to run on other personal computers. The prediction scheme contains computer programs, making use of the secondary structure prediction methods of Nagano (1977a,b), Garnier *et al.* (1978), Burgess *et al.* (1974), Chou and Fasman (1974a,b), Lim (1974a,b) and Dufton and Hider (1977). These programs were written in FORTRAN 77 and they run on the IBM PC/XT/AT and compatibles under DOS 2.0 or later releases.

The results of the individual prediction methods are combined as described by Hamodrakas *et al.* (1982) and Hamodrakas and Kafatos (1984) using a BASIC computer program, to produce joint prediction histograms (JPH) for three types of secondary structure, α-helix (H), β-sheet (B) and β-turns (T), which are presented separately for each type of secondary structure.

The programs can handle up to 50 amino acid sequences simultaneously. Each sequence may contain up to 500 amino acid residues. The programs require at least 128 kbyte RAM memory, one floppy disk drive, a standard graphics adapter (CGA) and a monitor capable of displaying graphics. They run without the need of any additional features, such as FORTRAN or BASIC compilers, since they are in the form of directly executable modules.

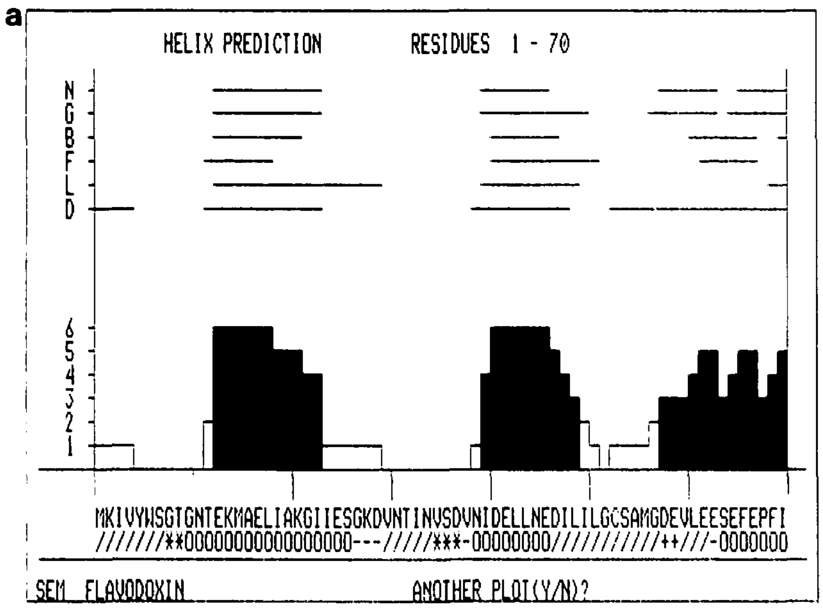
The six prediction programs require uniform input and produce uniform output.

Input to the prediction programs is given in the form of an input ASCII file which can be produced by any word processor, spreadsheet, database or editor program capable of producing ASCII type files. It should contain (i) a title line for the run; (ii) the number of amino acid sequences to be examined; (iii) the name of the first protein; (iv) the number of amino acid residues

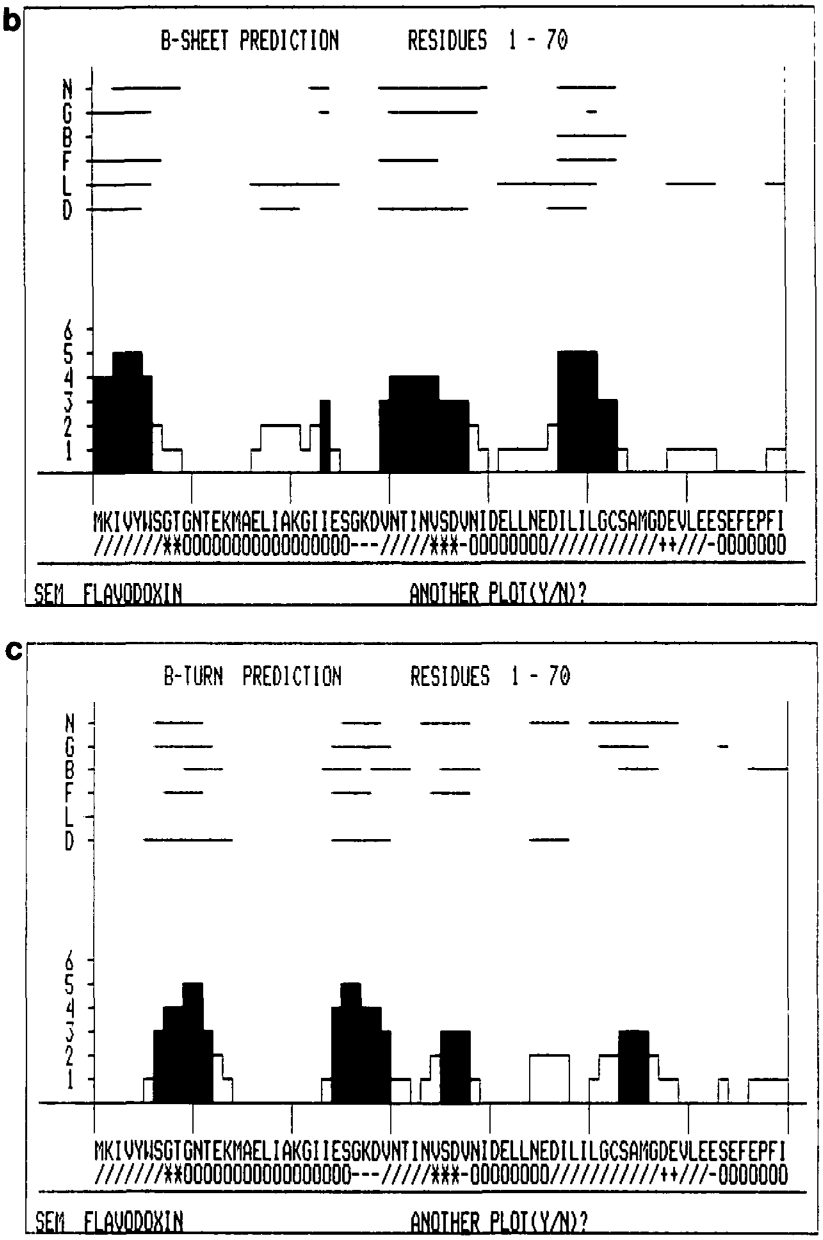
Department of Biochemistry, Cell and Molecular Biology and Genetics, University of Athens, Panepistimiopolis, Kouponia, Athens 157.01, Greece

```
  TEST RUN  FOR 4 PROTEINS
  4
  TRYPSIN  INHIBITOR  BOVINE
 58
RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA
 GGGGS     S    EEEEEEETTTTEEEEEEE SSS  SS BSSHHHHHHHHS


  RIBONUCLEASE  S
 124
KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHESLADVQAVCSQKNVACKNGQTNCYQSYSTMS
ITDCRETGSSKYPNCAYKTTQANKHIIVACEGNPYVPVHFDASV
  HHHHHHHHHB   SS       HHHHHHHHTTSSSSS  SEEEEE S HHHHHGGGGSEEEEEETTTTEEEEEEE SS EE
EEEEEE TT BTTB  EEEEEEEEEEEEEE SSS  EEEEEEE


  A. POLYPHEMUS  PC18  A FAMILY
 99
YGCGCGCGLGGYGGLGYGGLGYGGLGYEGTGACLGEYGGTGIGNVAVAGELPVAGKTAVGGQVPIIGAVGFGGTAGAAGC
VSIAGRCGGCGCGCGRGIY



  FLAVODOXIN
 138
MKIVYWSGTGNTEKMAELIAKGIIESGKDVNTINVSDVNIDELLNEDILILGCSAMGDEVLEESEFEPFIEEISTKISGK
KVALFGSYGWGDGKWMRDFEERMNGYGCVVVETPLIVQNEPDEAEQDCIEFGKKIANI
///////**000000000000000000---/////***-00000000///////////++///-00000000000***///
////////++*-00000000000000/////*+/////+*-**0000000000000*-
```

Fig. 1. An example of an input data file to the secondary structure prediction programs. This data file contains four protein sequences. The three-dimensional structures of three proteins, bovine pancreatic trypsin inhibitor, ribonuclease S and flavodoxin, are known (Bernstein *et al.*, 1977). Secondary structure assignment (shown below each amino acid sequence, presented in the one-letter code) for bovine pancreatic trypsin inhibitor and ribonuclease S, is after Kabsch and Sander (1983b), that for flavodoxin after Levitt and Greer (1977). The fourth protein, silk moth chorion protein PC18 (Hamodrakas *et al.*, 1982), has unknown three-dimensional structure and this is indicated by the two empty lines below its sequence. Two empty lines have been inserted below the data for each protein to improve readability.



```
MKIVYWSGTGNTEKMAELIAKGIIESGKDVNTINVSDVNIDELLNEDILILGCSAMGDEVLEESEFEPFI
///////**000000000000000000---/////***-00000000///////////++///-0000000

SEM  FLAVODOXIN                        ANOTHER PLOT(Y/N)?
```


b
B-SHEET PREDICTION RESIDUES 1 - 70
N
G
B
F
L
D
6
5
4
3
2
1
c
B-TURN PREDICTION RESIDUES 1 - 70
N
G
B
F
L
D
6
5
4
3
2
1

```
MKIVYWSGTGNTEKMAELIAKGIIESGKDVNTINVSDVNIDELLNEDILILGCSAMGDEVLEESEFEPFI
///////**00000000000000000---/////***-0000000////////////++///-0000000
SEM FLAVODOXIN                 ANOTHER PLOT(Y/N)?
```

```
MKIVYWSGTGNTEKMAELIAKGIIESGKDVNTINVSDVNIDELLNEDILILGCSAMGDEVLEESEFEPFI
///////**00000000000000000---/////***-0000000////////////++///-0000000
SEM FLAVODOXIN                 ANOTHER PLOT(Y/N)?
```

Fig. 2. Prediction plot for (a) α-helix, (b) β-sheet and (c) β-turn for residues 1–70 of flavodoxin (see also Figure 1). Individual predictions, as derived according to Nagano (N, 1977), Garnier *et al.* (G, 1978), Burgess *et al.* (B, 1974), Chou and Fasman (F, 1978), Lim (L, 1974) and Dufton and Hider (D, 1977) are shown by horizontal lines. Joint prediction histograms, constructed by tallying the individual predictions, are also shown. The most probable structures, predicted by three or more methods, are shaded. Observed secondary structure after Levitt and Greer (1977) is also shown below the amino acid sequence. The plot is a 'hard copy', on a dot-matrix printer, of a monitor screen.

of the first sequence; (v) the amino acid sequence, given in the one-letter code; (vi) the observed secondary structure (if any) of the first protein given in the form of one-letter codes, one for each residue. Data (iii)–(vi) are repeated as many times as the number of proteins to be examined.

Two types of output are produced from each prediction program: an output file, uniform for all methods, destined to become input to the joint prediction plot producing program PLOTPROG; the other, also uniform for all prediction algorithms, consists of the results of the prediction, which can be displayed on the monitor, printed on a printer or stored in a file.

Three types of secondary structure are predicted: α-helix, β-sheet and β-turn. Each residue is assigned one of the three conformational states, but the results of prediction are presented separately for helix, sheet and turns, as described by Hamodrakas *et al.* (1982).

The joint prediction program PLOTPROG produces joint prediction plots, displayed on the monitor, for each type of secondary structure, for the six prediction methods, in a way analogous to that described by Hamodrakas *et al.* (1982). The most probable structures predicted by three or more methods are shaded. An additional program, which may be very useful in the implementation of the PLOTPROG, is the DOS utility (or command) 'GRAPHICS', a screen-dump program which can be used to 'hard copy' the contents of a graphics screen to a dot-matrix printer. We have not incorporated into the program any routines for plotters, since these may vary from one system to another. However, the format of the output files of the prediction programs renders them very convenient as input files to simple plot programs for plotters.

The prediction algorithms are described in detail in the relevant papers. However, some modifications have been made to the algorithms and some of them are listed below:

(i) There are no changes to the method of Nagano (1977a,b). However, the loop-forming residues are presented as turn-forming residues (T).

(ii) The directional method of Garnier *et al.* (1978) is used by choosing run constants of 1 and decision constants of 0. This method, unambiguously assigns one of four conformational states to each residue in an amino acid sequence. However, in this work, the residues predicted to be coil (C) are presented as turn-forming residues (T).

(iii) The method of Burgess *et al.* (1974) simultaneously predicts helix, extended structure and bends according to their nonamer model. The assignments for helix and extended structure (β-sheet) are retained and residues 4–7 of the nonamer bend model are presented as turn-forming residues (T).

(iv) For the method of Chou and Fasman the latest set of parameters (Chou and Fasman, 1978, p. 45) was used. Turn prediction is made according to the algorithm, presented in Chou and Fasman (1979, p. 367).

(v) In the method of Lim (1974a,b) both helix and β-sheet prediction are presented independently, even if they overlap. In Lim's algorithm helix prediction prevails over β-sheet prediction.

(vi) For the method of Dufton and Hider (1977), tetrapeptides are used instead of hexa- and pentapeptides, for helix and sheet prediction respectively: experience shows that the use of hexa- and pentapeptides leads to overprediction. Tetrapeptides are predicted as probable β-turns if their $P(t)$ values are ≥ 2.0.

A typical example of an input data file to each prediction program is shown in Figure 1. This data file was created and printed utilizing a word processor program.

Figure 2a–c presents joint prediction plots, together with individual predictions, for three types of secondary structure, α-helix, β-sheet and β-turns, for a segment of the protein flavodoxin. These plots were produced on a monitor screen utilizing the program PLOTPROG, which combines individual prediction methods, and were hard copied on a dot-matrix printer.

The three-dimensional structure of flavodoxin is known and secondary structure assignments have been made by several investigators (see, for example Bernstein *et al.*, 1977; Levitt and Greer, 1977 and references therein; Cohen *et al.*, 1983; Kabsch and Sander, 1983b). In Figure 2a–c, the secondary structure assignment of Levitt and Greer (1977) is given below the amino acid sequence (one-letter code). Any other secondary structure assignment (e.g. the one introduced by Kabsch and Sander, 1983b) might have been used instead.

Several methods have been proposed to measure the agreement between prediction and observation to assess the power of the predictive schemes. An auxiliary program, INDEX, is included in the package to calculate the most commonly used accuracy indices Q (Chou and Fasman, 1974b) and C (Mathews, 1975; Argos *et al.*, 1976). Our tests clearly showed that the joint prediction scheme is always comparable to the best individual prediction method, for all proteins studied and for each type of secondary structure. Since there are several problems concerning secondary structure assignment and the prediction methods themselves, discussed extensively by Argos and Mohana Rao (1986) no accuracy estimates are reported here. In most cases the joint prediction plots successfully indicate the main structural elements of a protein.

The prediction programs are fast. Secondary structure prediction for a protein with 200 amino acid residues takes <20 s for each method, with the exception of the method of Nagano which requires <3 min. PLOTPROG is also very fast; it produces each plot in <2 s.

The programs are distributed on request, together with a user's guide and sample data input and output, at a very low cost, to cover diskette cost, postage and handling.

References

Anfinsen,C.B. (1973) Principles that govern the folding of protein chains. *Science*,**181**, 223–230.

Argos,P. and Mohana Rao,J.K. (1986) Prediction of protein structure. *Methods Enzymol.*, **130**, 185–207.

Argos,P. and McCaldon,P. (1988) Theoretical and computer analysis of protein primary sequences: structure comparison and prediction. *Genetic Engineering, Principles and Methods.* Vol. 10, in press.

Argos,P., Schwarz,J. and Schwarz,J. (1976) An assessment of protein secondary structure prediction methods based on amino acid sequence. *Biochim. Biophys. Acta*, **439**, 261–273.

Barker,W.C., Hunt,L.T., George,D.G., Yeh,L.S., Chen,H.R., Blomquist,M.C., Seibel-Rose,E.I., Elzanowski,A., Hong,M.K., Ferrick,D.A., Bair,J.K., Chen,S.L. and Ledley,R.S. (1986) *Protein Sequence Database.* National Biomedical Research Foundation, Georgetown University, Washington, DC.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasoumi,M. (1977) The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.

Burgess,A.W., Ponnuswamy,P.K. and Scheraga,H.A. (1974) Analysis of conformations of amino acid residues and prediction of backbone topography in proteins. *Isr. J. Chem.*, **12**, 239–286.

Chou,P. and Fasman,G.D. (1974a) Conformational parameters for amino acids in helical, β-sheet and random coil regions calculated from proteins. *Biochemistry*, **13**, 211–221.

Chou,P. and Fasman,G.D. (1974b) Prediction of protein conformation. *Biochemistry*, **13**, 222–245.

Chou,P. and Fasman,G.D. (1978) Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.*, **47**, 45–148.

Chou,P. and Fasman,G.D. (1979) Prediction of β-turns. *Biophys. J.*, **26**, 367–384.

Cohen,F.E., Abarbanel,R.M., Kuntz,I.D. and Fletterick,R.J. (1983) Secondary structure assignment for α/β proteins by a combinatorial approach. *Biochemistry*, **22**, 4894–4904.

Dufton,M.J. and Hider,R.C. (1977) Snake toxin secondary structure predictions: structure activity relationships. *J. Mol. Biol.*, **115**, 117–193.

Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, **120**, 97–120.

Hamodrakas,S.J. and Kafatos,F.C. (1984) Structural implications of primary sequences from a family of Balbiani ring encoded proteins in chironomus. *J. Mol. Evol.*, **20**, 206–303.

Hamodrakas,S.J., Jones,C.W. and Kafatos,F.C. (1982) Secondary structure predictions for silkmoth chorion proteins. *Biochim. Biophys. Acta*, **700**, 42–51.

Hodgman,T.C. (1986) The elucidation of protein function from its amino acid sequence. *Comput. Applic. Biosci.*, **2**, 181–187.

Kabsch,W. and Sander,C. (1983a) How good are predictions of protein secondary structure? *FEBS Lett.*, **155**, 179–182.

Kabsch,W. and Sander,C. (1983b) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Levitt,M. and Greer,J. (1977) Automatic identification of secondary structure in globular proteins. *J. Mol. Biol.*, **114**, 181–239.

Lim,V.I. (1974a) Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. *J. Mol. Biol.*, **88**, 857–872.

Lim,V.I. (1974b) Algorithms for prediction of α-helical and β-structural regions in globular proteins. *J. Mol. Biol.*, **88**, 873–894.

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

Nagano,K. (1977a) Logical analysis of the mechanism of protein folding. IV Supersecondary structures. *J. Mol. Biol.*, **109**, 235–250.

Nagano,K. (1977b) Triplet information in helix prediction applied to the analysis of supersecondary structures. *J. Mol. Biol.*, **109**, 251–274.

Schulz,G.E., Barry,C.D., Friedman,J., Chou,P., Fasman,G.D., Finkelstein,A.V., Lim,V.I., Ptitsyn,O.B., Kabat,E.A., Wu,T.T., Levitt,M., Robson,B. and Nagano,K. (1974) Comparison of predicted and experimentally determined secondary structure of adenyl kinase. *Nature*, **250**, 140–142.

Taylor,W.R. (1987) Protein structure prediction. In Bishop,M.J. and Rawlings,C.J. (eds), *Nucleic Acid and Protein Sequence Analysis: A Practical Approach*. IRL Press, Oxford, pp. 285–322.

Received on January 26, 1988; accepted on May 24, 1988