Volume 16 Number 5 1988 Nucleic Acids Research

A computer graphics program system for protein structure representation

Andrew M.Ross and Ellis E.Golub

Biochemistry Department, University of Pennsylvania, School of Dental Medicine, Philadelphia, PA 19104, USA

Received August 17, 1987; Revised and Accepted December 31, 1987

ABSTRACT

We have developed a computer graphics program system for the schematic representation of several protein secondary structure analysis algorithms. The programs calculate the probability of occurrence of α-helix, β-sheet and β-turns by the method of Chou and Fasman and assign unique predicted structure to each residue using a novel conflict resolution algorithm based on maximum liklihood. A detailed structure map containing secondary structure, hydrophobicity, sequence identity, sequence numbering and the location of putative N-linked glycosylation sites is then produced. In addition, helical wheel diagrams and hydrophobic moment calculations can be performed to further analyze the properties of selected regions of the sequence. As they require only structure specification as input, the graphics programs can easily be adapted for use with other secondary structure prediction schemes. The use of these programs to analyze protein structure-function relationships is described and evaluated.

INTRODUCTION

Three dimensional structure information is only known for a limited number of proteins, while considerable amino acid sequence data is available. Using statistical and theoretical approaches, algorithms have been created to predict protein secondary structure based on amino acid sequence (1-7). We have developed a group of computer programs that utilize a variety of these algorithms. Most significantly, we have created a graphics package that can readily produce secondary structure maps, helical wheel diagrams (8), hydrophobic moment profiles (9), and hydropathy plots (10-12). These structure maps are a significant improvement compared with the output produced by most other prediction programs including PROTYLZE (13), CHOFAS (14), and DELEAGE (15) which only generate probability tables and x-y graphs. The secondary structure assignments are

produced by a program that uses the method of Chou and Fasman (1-3), who tabulated the probability for each amino acid to contribute to the four major classes of protein secondary structure (α-helix, β-pleated sheet, β-turns and random coil). To produce a unique structure assignment for each residue, we automated the assignment of secondary structure and created routines that resolve conflicts where residues are not uniquely assigned by the basic Chou and Fasman rules.

The programs which we developed have been designed to allow for maximum flexibility in accommodating new predictive algorithms, and utilizing amino acid side chain properties. The flexibility was accomplished by creating a graphics program that is independent of the prediction process, requiring only structural information to create a graphic map. Moreover, our routine can overlay the structure map with a pictorial representaion of amino acid side chain property data (e.g. hydropathy, solvet accessible surface, etc.) which can be used as a rapid means to relate clusters of similar amino acids to secondary structure. Further, output has been made accessible to programs that can import ASCII text files to allow for statistical analysis of numerical data. In this report we describe the foundation behind these programs and evaluate their application to the analysis of structure-function relationships in proteins.

METHODS

We use the basic rules as set forth by Chou and Fasman for calculating probability scores for α-helix, β-sheet, and β-turns. FCHO is a computer program which reads sequence files containing up to 2500 amino acids and calculates all Chou and Fasman probability values. Then using a conflict resolution algorithm based on maximum likelihood, unique secondary structure assignments are made for each residue in the sequence. At the same time, a hydrophobicity score is calculated for each residue using user selected window size and tables, which currently include (but are by no means limited to) the hydrophobicity scales of Kyte and Doolittle (11), Hopp and Woods (10), and Guo et al (12). Putative N-linked glycosylation sites

are identified as any regions conforming to the sequence Asn-X-Thr or Asn-X-Ser, where X is any amino acid residue. Program output includes printed tables, ASCII data files for export to other software and special compressed files for use by the graphics programs.

A unique feature of our system is its ability to draw a variety of schematic diagrams of the predicted structure. This is accomplished by two programs: FCHOX, which produces a graphic map of the predicted secondary structure with various overlays for residue identification, hydrophobicity, and N-linked glycosylation sites, and WHEELS, which can represent any region of the sequence in the form of a helical wheel diagram (8). The WHEELS program also contains a module which can calculate the hydrophobic moment profile (9) for any region of the sequence and write the results to an ASCII file which can be used with software such as LOTUS 1-2-3 for further mathematical or graphical analysis. The combination of these various perspectives (secondary structure maps, helical wheels, and hydrophobic moment profiles) provides powerful heuristic tools for the study of proteins whose amino acid sequences have been determined.

FCHO method of secondary structure assignment

Following computation of the Chou and Fasman probability scores, the structure is predicted from the computed parameters as follows:

1. Random coil assignment. All residues are initially assigned to random coil structure.
2. β-turn detection. When the probability of a β-turn exceeds the pre-set threshold, $Fturn_2$, (1.0×10^{-4}) a β-turn is predicted. If the value is between $Fturn_1$ (7.5×10^{-5}) and $Fturn_2$, a β-turn is predicted if no other secondary structure is more likely. An option to modify the Fturn threshold values is available. The value of these thresholds strongly influences the outcome of the prediction.
3. Helix and sheet prediction. If a nucleus for helix or sheet is detected then the appropriate propagation procedure is invoked. If both helix and sheet are predicted, then

conflict resolution is performed. The probability scores for the putative nucleii (P_α and P_β) are compared, and the more likely structure is assigned.

Hardware and software

All programs are written in C for the IBM Personal Computer and work alikes. The graphics segments require the color graphics (CGA) or enhanced graphics (EGA) adapters. Graphic output can be directed to a printer, monitor, or plotter. A Hewlett Packard 7475A plotter was used to generate the drawings for this publication. An installation program provided with FCHOX allows other plotters using Hewlett Packard Graphics Language (HP-GL) to be used.

Graphic representation of secondary structure

FCHOX is a graphics program which can read the files produced by program FCHO and use the information to draw a picture of the predicted structure (Fig. 1). The symbols used to represent the four basic structure types are the following: α-helix is represented by a sine wave with each residue offset by 100°, β-turns are represented by a chain reversal composed of four amino acids, β-sheets are represented by a cartoon of the bond lengths and angles connecting residues in a β-sheet, and a random coil is represented by gently angled straight lines. In all cases the orientation of the side chain β-carbon is depicted by a dash. For instance, since adjacent residues of a β-sheet are oriented 180° with respect to each other, the dashes representing neighboring residues point in opposite directions. Hydrophobic regions are depicted by filled circles whose radii are proportional to the computed regional hydrophobicity. Similarly, open circles are used to represent hydrophilic residues. The location of putative N-linked glycosylation sites can also be marked, and the amino acid residues can be labeled and numbered, with user control of the label and number positions.

DISPLAY OF HELICAL WHEELS

Helical wheels are drawn by projecting the amino acid side chains onto a plane which is perpendicular to the helical axis (Fig. 2). The side chains are spaced 100° apart since there are

3.6 residues per turn in an α-helix. These figures can also be generated on the computer monitor, printer or a plotter. A list containing all of the predicted helical regions in a sequence is displayed by the program for user selection of a wheel drawing. Alternatively, the user can select any span of residues to view in helical wheel representation. The drawings (Fig. 2) show the perimeter of the polypeptide chain as a circle, with the side chains labelled around the perimeter at 100° intervals. The relative location of each residue in the helix is labelled, its three letter amino acid code is given, and if desired, its hydropathy value is drawn using the same technique as that described above.

Calculation of hydrophobic moment profiles

A function to calculate the hydrophobic moment profile using the method of Eisenberg, Weiss, and Terwilliger (9) was incorporated into the program WHEELS. The moment profile is based on the Fourier equation:

$$\mu=\{[\sum_{n=1}^{N} H \sin(\beta n)]^2 + \sum_{n=1}^{N} H \cos(\beta n)]^2\}^{1/2}$$

This equation is designed to analyze periodicity in protein hydrophobicity. Characteristic curves are generated by structures which reflect the angular advance of the amino acid side chains (β) from the backbone structure (100° for α-helix, 120° for 3_{10} helix, and 160-180° for β-sheet). The program presents the user with a list of all predicted α-helix and β-sheet regions for analysis using the above equation. Alternatively, the user can select any desired span for a hydrophobic moment profile. Presently two hydrophobicity scales are presented for user selection. These are the Eisenberg scale (9) and the statistical PRIFT scale of Cornette et al (16). The values calculated are stored in an ASCII text file which can be imported to a program such as LOTUS 1-2-3 for the production of a hydrophobic moment profile graph (Fig. 3).

RESULTS and DISCUSSION

Evaluation of the predictive and structure assignment algorithms

The accuracy of the FCHO program was assessed by predicting the structure of 62 proteins whose atomic coordinates have been determined by x-ray crystallography and whose secondary structure has been defined by the method of Kabsch & Sander (17). A total of 10747 residues were analyzed using a three state model of secondary structure (α-helix, β-sheet, and coil). A predictive accuracy greater than 33% using this system indicates that the model is more reliable than random assignment. All predictions exceeded 33% with an average accuracy score of 50.7% for the 10747 residues evaluated.

The values found compare favorably with other attempts to create a workable computer model of the Chou and Fasman algorithm. A sample of 20 proteins containing 4244 amino acids was used to compare the FCHO method of prediction against two other computerized implementations of the Chou and Fasman algorithm. The Corrigan and Huang (18) method scored 45.45% with a standard deviation of 7.5 based on the three state model described above. A recently developed Chou and Fasman program written for the Apple IIe by Deleage, Tinland, and Roux (15) scored 51.41% with a standard deviation of 7.6. For these proteins, the FCHO assignment program scored highest with 53.9% accuracy and a standard deviation of 6.7. Structure types were correctly identified as follows: helices, 50.5%; sheets, 35.0%; and coil, 66.6%. The FCHO method was more accurate than the Corrigan and Huang method for 18 of the 20 proteins, and more accurate than the Deleage method for 11 of the 20 proteins (one scored evenly). The conflict resolution routine in FCHO was called 71 times during secondary structure prediction of the 20 protein sample. A total of 570 residues were assigned a secondary structure based on the results of the conflict resolution routine with 242 correctly predicted with an accuracy of 42.5%.

Graphics

While the FCHO implementation of the Chou and Fasman algorithm is at least as good as any available, the importance of these programs is best demonstrated by the graphics output.

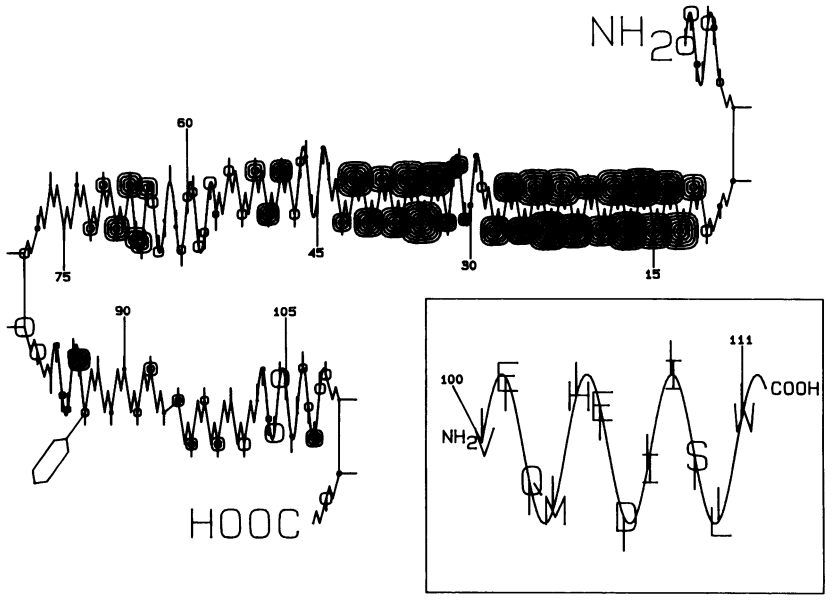

Figure 1. Predicted secondary structure of HIV WMJ2 gp120 residues 1-115. Hydrophobic regions are indicated by filled circles, hydrophilic regions by open circles. A putative N-linked glycosylation site is indicated by the hexagonal balloon. The inset shows the sequence and predicted helical domain between residues 100-111.

Using the assigned structures, program FCHOX displays the predicted secondary structure as a two-dimensional map. This representation (Fig. 1) allows the investigator a holistic view of the prediction, with options to overlay a representation of the hydrophobicity, the sequence numbers, the identity of the residues, as well as predicted N-linked glycosylation sites. When presented in this way, the heuristic value of the structure prediction is significantly enhanced. These color pictures provide useful visual cues which have been particularly successful in suggesting experimental approaches to analysis of protein structure-function relationships (19-25). The latest updates of the graphics software are also capable of drawing multiple structures to the same scale for comparative purposes. In addition, multiple windows can be created on the plotter to allow detailed comparison of selected protein domains (Fig. 1 and inset).

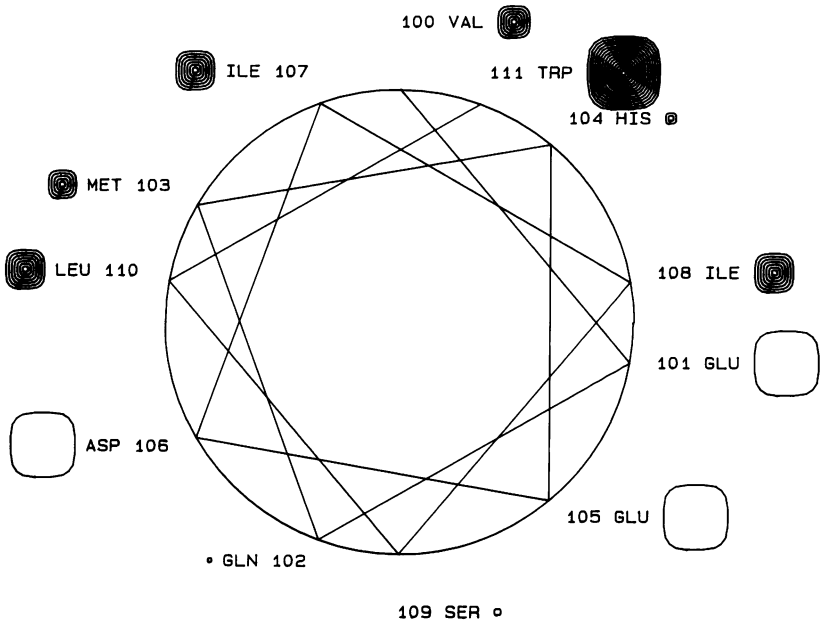

Figure 2. Helical wheel diagram of residues 100-111 of HIV WMJ2 gp120. The identity of the amino acid side chains and their hydrophobicity are depicted according to their angular displacement around the helical axis. Hydrophobicity is indicated as in Figure 1.

From a programming point of view, the suite of programs described here is of particular interest, as it can be readily adapted to any secondary structure prediction process which results in prediction of the three major secondary structure classes and random coil. This is because the graphics program, FCHOX, requires only structure class information in order to produce a picture. The graphics portion of these programs can be harnessed to new prediction algorithms as they become available. New graphics functions can be added as program updates to represent additional structures such as 3_{10} helices and disulfide bridges to meet the needs of algorithms predicting these structures. In a similar fashion, the programs are not limited to the hydrophobicity scales currently used, but could be extended to make use of other tables of physical properties of the amino acid residues. For example, solvent accessible surface parameters could be computed and displayed using the

FCHOX overlay method. The programs described here are really one implementation of a general program design which can easily be adapted to other computer systems. The C language is noted for its portability, although system specific differences in graphics capability would require software modification. However, the modular nature of these programs will facilitate such adaptations.

In addition to the pictorial view of predicted protein secondary structure, the FCHO assignment program produces printed output and ASCII data files which can be used for mathematical analysis of the predictions. We have tailored these files for use with LOTUS 1-2-3, but they can be utilized by any analytic program capable of reading plain text files.

Typical protein structure analysis

To demonstrate the capabilities of the protein structure prediction programs, we have analyzed an N-terminal segment of gp120, one of the envelope glycoproteins of HIV. The sequence depicted includes residues 1-115 of the deduced amino acid sequence of the WMJ2 strain (26). Figure 1 shows the predicted secondary structure of this segment overlayed with symbols representing the hydrophobicity according to the method of Hopp and Woods (10). The signal peptide is clearly depicted by the hydrophobic domain between residues 11 and 29, with the signal peptidase cleavage site located at a local maximum in hydrophobicity. Also shown are α-helical and β-sheet domains, along with three β-turns. A potential N-linked glycosylation site at Asn_{87} is indicated by the hexagonal balloon. The helix between residues 100-111 is of particular interest. The inset to Figure 1 shows this region in detail including the identity of each residue. The ability to display multiple regions in this manner is a particular strength of our programs.

When viewed as a helical wheel diagram (Fig. 2), the amphipathic property of the helix mentioned above (residues 100-111) is evident. The segregation of hydrophobic residues on one edge of this helix while clear in this format is not visible when adjacent residues are averaged as in Figure 1. The wheel perspective demonstrates the spacial juxtaposition of side chains in the helix, where physically adjacent side chains are

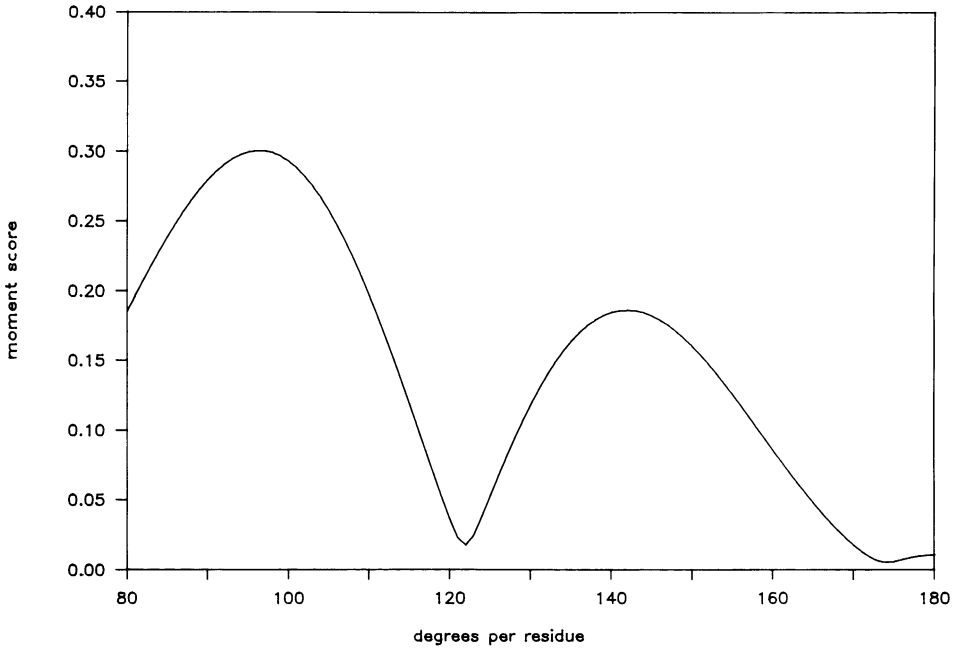

Figure 3. Plot of the hydrophobic moment vs. angular displacement for the region between residues 100-111 of HIV WMJ2 gp120. For detail see text.

four or more sequence positions apart. Thus, Val_{100} is closer to His_{104} and Trp_{111} than to Glu_{101}. Amphipathic helices are believed to be important domains for stabilizing tertiary protein structure or protein-protein interactions (9). Another representation of the amphipathic property of this helical region is shown in Figure 3. This figure was produced by importing the data table output from WHEELS into a LOTUS 1-2-3 worksheet and plotting the hydrophobic moment against the angular displacement. The curve obtained is strikingly similar to the profile shown by Eisenberg (9) for highly amphiphilic α-helix. This plot also reinforces the prediction of this domain as an α-helix, and quantifies the amphipathic appearance seen in the helical wheel diagram.

Evaluation of computer programs

The analysis of a protein sequence by the FCHO program is quite rapid. Execution speed is limited by the time required

for writing of data to disk storage. In a test run of FCHO on the Snowshoe Hare Virus (1441 amino acids), 55 seconds were required to compute all secondary structure probabilities, make unique assignments, calculate hydrophobicity averages, and write this information to disk (27). The test was performed on an IBM PC equipped with an Intel 8087 math co-processor. The program is equipped to handle sequences of up to 2500 amino acids in length, compared with only 1000 for the Deleage program. This is an important limiting factor when studying long amino acid sequences such as the one of Snowshoe Hare virus. The menu driven user interface has been engineered for the general user. The graphics output process allows users to customize the drawings for analytical or presentation purposes. The structure maps produced by these programs are valuable tools in allowing for an overall assessment of the predicted structure and its relation to the assigned hydrophobicity data.

Taken together the methodology presented above constitutes a significant advance in the the computation and presentation of protein secondary structure predictions, and provides the molecular biologist with an important set of tools for the examining the properties of newly determined protein sequences.

ACKNOWLEDGEMENT

This work was supported by a grant from the National Institute for Dental Research, DE-02623, and the IBM Threshold Grant.

REFERENCES

1. Chou, P.Y. and Fasman, G.D. (1974) Biochemistry 13, 211-222.
2. Chou, P.Y. and Fasman, G.D. (1974) Biochemistry 13, 222-245.
3. Chou, P.Y. and Fasman, G.D. (1978) Ann. Rev. Biochem. 47, 251-276.
4. Lewis, P.N., Go, N., Go, M., Kotelchuck, D. and Sheraga, H.A. (1971) Proc. Nat. Acad. Sci. USA 65, 810-815.
5. Robson, B. and Pain, R.H. (1974) Biochem J. 141, 869-882.
6. Kabat, E.A. and Wu, T.T. (1973) Proc. Nat. Acad. Sci. USA 70, 1473-1477.
7. Lim, V.I. (1974) J. Mol. Biol. 88, 857-872.
8. Schiffer, M. and Edmundson, A.B. (1967) Biophys. J. 7, 121-135.
9. Eisenberg, D., Weiss, R.M. and Terwilliger, T.C. (1984) Proc. Nat. Acad. Sci. USA 81, 140-144.
10. Hopp, T.P. and Woods, K.R. (1981) Proc. Nat. Acad. Sci. USA

78, 3824-3828.
11. Kyte, J. and Doolittle, R. F. (1982) J. Mol. Biol. 157, 105-132.
12. Guo, D., Mant, C. T., Taneja, A. K., Parker, J. M. R. and Hodges, R. S. (1986) J. Chromatogr. 359, 499-517.
13. Protylze Protein Structure Predictor (1987) Scientific & Educational Software.
14. CHOFAS (1987) Protein Identification Resource Newsletter 2, 1-8.
15. Deleage, G., Tinland, B. and Roux, B. (1987) Anal. Biochem. 163, 292-297.
16. Cornette, J.L., Kemp, C.B., Margalit, H., Spouge, J.L., Berzofsky, J.A. and DeLisi, C. (1987) J. Mol. Biol. 195, 659-685.
17. Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577-2637.
18. Corrigan, A. J. and Huang, P. C. (1982) Comput. Programs Biomed. 15, 163-168.
19. Wunner, W.H., Dietzschold, B., Macfarlan, R.I., Smith, C.L., Golub, E. and Wiktor, T. (1985) Ann. Inst. Pasteur/Virol. 136 E, 353-362.
20. Wunner, W.H., Dietzschold, B., Smith, C.L., Lafon, M. and Golub, E. (1985) Virology 140, 1-12.
21. Cohen, G.H., Dietzschold, B., Ponce de Leon, M., Long, D., Golub, E., Varrichio, A., Pereira, L. and Eisenberg, R.J. (1984) J. Virol. 49, 102-108.
22. Dietzschold, B., Eisenberg, R.J., Ponce de Leon, M., Golub, E., Hudecz, F., Varrichio, A. and Cohen G.H. (1984) J. Virol. 52, 431-435.
23. Eisenberg, R.J., Long, D., Ponce de Leon, M., Matthews, J.T., Spear, P.G., Gibson, M.G., Lasky, L.A., Berman, P., Golub, E. and Cohen, G.H. (1985) J. Virol. 53, 634-644 (1985).
24. Yoon, K., Davidson, J. M., Boyd, C., May, M., LuValle, P., Orstein-Goldstein, N., Smith, J., Indek, Z. Ross, A., Golub, E. and Rosenbloom, J. (1985) Arch. Biochem. Biophys. 241, 684-691.
25. Wolf, H., Motz, M., Kuhbeck, R., Seibl, R., Jilg, W., Bayliss, G. J., Barrell, B., Golub, E., Zeng, Y. and Gu, S.-Y. (1984) IARC Scientific Publication No. 63 525-539.
26. Modrow, S., Hahn, B. H., Shaw, G. M., Gallo, R. C., Wong-Staal, F. and Wolf, H. (1987) J. Virol. 61, 570-578.
27. Eshita, Y. and Bishop, D. H. (1984) Virology 137, 227-240.