DNA
Volume 4, Number 4, 1985
Mary Ann Liebert, Inc., Publishers
Pp. 319-326

LABORATORY METHODS

Microcomputer Programs for Prediction and Comparative Evaluation of Protein Secondary Structure From Nucleotide Sequence Data: Application to Ribulose-1,5-Bisphosphate Carboxylase Sequences

SÁNDOR PONGOR,[1] MARY J. GUTTIERI, LISA M. COHEN, and ALADAR A. SZALAY[1]

ABSTRACT

Apple II PASCAL computer programs are described for routine evaluation of protein secondary structures predicted from DNA or amino acid sequence data. The programs predict protein secondary structure using the directional information algorithm of Garnier *et al.* (1978), and calculate hydrophobicity (Kyte and Doolittle, 1982), hydrophilicity (Hopp and Woods, 1981), hydrophobic moment (Eisenberg *et al.*, 1984b), and secondary structure propensity profiles. The novel feature of these programs is the application of numeric and graphic methods, designed to facilitate detection and characterization of structural similarities/divergences using the aforementioned structural parameters. The use of the programs is demonstrated on a set of sequences from the large and small subunits of ribulose-1,5-bisphosphate carboxylase.

INTRODUCTION

Advances in DNA sequencing techniques and oligonucleotide-directed mutagenesis have generated a considerable interest in methods for predicting elements of protein structure from amino acid or DNA sequences. Although current methods of protein secondary structure prediction rarely achieve an accuracy higher than 60–80% (for review, see Schulz and Schirmer, 1979), their simplicity makes them a useful tool for the molecular biologist in the interpretation of gene sequences in terms of protein structure. Empirical methods have been developed to predict membrane-bound domains (Kyte and Doolittle, 1982), antigenic determinants (Hopp and Woods, 1981), and amphiphilic helices (Eisenberg *et al.*, 1984a) from amino acid sequence data. These methods use a common graphic representation to display structural information: a structural parameter P_i, characterizing a structural property of the ith amino acid in a sequence, is plotted against the sequential position i. The resulting scattered graphs are then smoothed by a moving averaging procedure which facilitates the recognition of trends. The procedure of computation does not exceed the capabilities of laboratory microcomputers generally used to process DNA sequence data, yet none of these methods has been incorporated into currently available Apple DNA software.

The Apple PASCAL program package described in this paper provides a collection of empirical computational methods that allows the molecular biologist to interpret amino acid sequences in structural terms. The methods included (Table 1) were selected so as to complement available Apple software. Three Apple DNA software packages are available: the Cornell sequence analysis package (Fristensky *et al.*, 1982; De Banzie *et al.*, 1984), that of Larson and Messing (1983), and that of Malthiery *et al.* (1984). For a discussion of these programs, see Korn and Queen (1985). These programs are all written in PASCAL; none of these contains programs for protein secondary structure analysis. The programs described in this paper are compatible with any of the above Apple PASCAL programs; we did not in-

Boyce Thompson Institute for Plant Research, Cornell University, Ithaca, NY 14853.
[1]Present address: MTA Institute of Enzymology, Budapest pf 7, 1502 Hungary.

TABLE 1. METHODS INCLUDED IN THE PROGRAMS

Method	*Reference*	*Program*
1. Prediction of protein secondary structure	Garnier *et al.* (1978)	PRO1
2. Prediction of membrane-bound regions from hydrophobicity profiles	Kyte and Doolittle (1982)	PRO2
3. Prediction of antigenic determinants from hydrophilicity profiles	Hopp and Woods (1981)	PRO2
4. Secondary structure propensity profiles	Chou and Fasman (1978) and Garnier *et al.* (1978)	PRO2 PRO3
5. Comparison of secondary structures	Pongor and Szalay (1985)	PRO2 PRO3
6. Analysis of membrane and surface-associated sequences by hydrophobic moment plots	Eisenberg *et al.* (1984)	PRO4
7. Hydrophobicity correlation	Sweet and Eisenberg (1983)	PRO2

clude those methods available elsewhere, such as the protein secondary structure prediction method of Chou and Fasman (1978) and the matrix method of homology search (Maizel and Lenk, 1981). Apple software for these methods is described by Corrigan and Huang (1982) and Larson and Messing (1983), respectively.

The novel feature of this program package is the inclusion of numeric and graphic methods for *comparative evaluation* of structural profiles. These methods serve two purposes. First, they help to recognize *similarities* in secondary structure (Pongor and Szalay, 1985) or hydrophobicity profiles (Sweet and Eisenberg, 1983); these features may be useful in estimating the relatedness of proteins that are not obviously related in their primary sequence. Second, they can be used to characterize *differences* of related amino acid sequences in terms of structural parameters (Pongor and Szalay, 1985). This feature makes it possible to predict the effects of amino acid replacements on the secondary structure or hydrophobicity characteristics of a given domain, which can be especially useful in the design of altered proteins. Since comparison of a newly determined amino acid sequence to homologous sequences is a frequent task in the practice of molecular biology, these novel features are described here in a somewhat greater detail. For details of the other methods the reader is referred to the original papers listed in Table 1. The use of the programs is demonstrated on a set of nucleotide and amino acid sequences of the large subunit (LSU) and the small subunit (SSU) of ribulose-1,5-bisphosphate carboxylase (RUBISCO).

THE PROGRAMS

The programs are written in Apple PASCAL and are implemented on an Apple IIe microcomputer (48K) equipped with an IDS 540 type printer (Integral Data Systems) and two disk drives. The limited memory available in the Apple IIe made it necessary to write five independent programs for the different tasks. The present version of the package can handle amino acid sequences up to 510 residues. Both nucleotide and amino acid sequences, written in standard Apple PASCAL files, can be used as input data. This feature makes the package compatible with other Apple PASCAL DNA software. The present version of the program package does not contain a high-resolution printing procedure. With minor modifications, the results can be printed with any printer. The results can also be recorded into Apple PASCAL text or data files for further processing.

The program text and code files are recorded 5¼ in. floppy disks, as are the data files containing the Robson parameters for secondary structure prediction and the structural parameters listed in Table 2. These programs are dependent on the Apple PASCAL operating system and editor. Directions for program operation are available. (For a copy of the program text and code files, send a stamped, self-addressed envelope and a PASCAL-formatted, double-density, soft-sectored 5¼ in. disk to S. Pongor or A.A. Szalay.)

File editing, translation of DNA sequences: PROE

The editor included with the Apple PASCAL system is used to edit DNA and amino acid sequence files for use with the programs. Amino acid sequences represented in the one-letter code (maximum 510 residues in length), can be processed with the programs. Program PROE includes a translation procedure to convert nucleotide sequences (maximum 1530 nucleotides in length) into text files directly usable by the other programs and an option to prepare a (cumulative) codon usage table for the sequence(s) translated. PROE also includes a procedure to print out a maximum of 10 sequence files in a standard format. This feature is included in order to facilitate the alignment of sequences using the Apple PASCAL editor.

Prediction of protein secondary structure: PRO1

Program PRO1 uses the algorithm of Garnier *et al.* (1978) for the prediction of secondary structure using amino acid sequences (represented in one-letter code) as input data. The program includes two optional features: (i) The use of an independently determined secondary struc-

TABLE 2. STRUCTURAL PARAMETERS USED IN PROGRAM PRO2

	Hydrophobicity[a,b]	*Hydrophilicity*[a,c]	*α-Helix*[a,d]	*β-Turn*[a,d]	*β-Structure*[a,d]	*Charge*[e]	*OMH Scale*[f]	*Normalized consensus hydrophobicity scale*[a]
Met	0.60	−0.61	1.63	−1.40	0.05	0	1.02	0.64
Leu	1.36	−0.87	1.00	−0.90	0.73	0	1.22	1.06
Phe	0.74	−1.24	0.15	−0.93	1.38	0	1.92	1.19
Ile	1.32	−0.87	−0.23	−1.10	1.55	0	1.25	1.38
Ala	1.28	−0.19	0.96	−0.43	−0.54	0	0.17	0.62
Cys	0.65	−0.45	0.31	−0.38	0.43	0	−0.40	0.29
Val	1.28	−0.71	−0.46	−1.21	1.82	0	0.91	1.08
Trp	−0.27	−1.72	−0.15	−0.50	1.18	0	0.50	0.81
Thr	−0.08	−0.13	−0.77	0.19	0.43	0	−0.28	−0.05
Ser	−0.10	0.24	−0.77	0.95	−0.75	0	−0.55	−0.18
Pro	−0.09	0.08	−1.92	2.40	−1.30	0	−0.49	0.12
Tyr	−0.32	−1.14	−1.15	0.26	1.19	0	1.67	0.26
Gly	1.37	0.08	−1.77	1.74	−0.75	0	−0.67	0.48
Gln	−0.96	0.19	0.88	0.05	0.18	0	−0.91	−0.85
Arg	−1.28	1.67	−0.27	−0.17	−0.27	+1	−0.59	−2.53
His	−1.13	−0.19	0.69	−0.64	−0.43	0	−0.64	−0.40
Asn	−1.03	0.19	−0.50	0.83	−0.38	0	−0.92	−0.78
Glu	−1.12	1.67	1.54	0.12	−1.79	−1	−1.22	−0.74
Lys	−0.99	1.67	0.73	0.02	−0.79	+1	−1.67	−1.50
Asp	−1.27	1.67	0.04	1.14	−1.33	−1	−1.31	−0.90

[a]Values published in original reference normalized to a mean zero and standard deviation 1. The OMH scale (Sweet and Eisenberg, 1983) and the normalized consensus hydrophobicity scale (Eisenberg *et al.*, 1984b) are already normalized.
[b]Kyte and Doolittle (1982).
[c]Hopp and Woods (1981).
[d]Chou and Fasman (1978).
[e]Novotný and Auffray (1984).
[f]Sweet and Eisenberg (1983).
[g]Eisenberg *et al.* (1984b).

ture content in the prediction; (ii) averaged prediction for a group of (maximum 10) sequences, a procedure suggested to improve the accuracy of the prediction (Garnier *et al.*, 1978; Novotný and Auffray, 1984). (Sequences to be used with option ii have to be aligned for maximum homology.) The output of PRO1 represents the predicted secondary structure using H for α-helix, B for β-structure, T for turn, and C or space for coil. The program also provides predicted cumulative secondary structure content data (% helix, % turn, etc.).

Predicted secondary structures of seven RUBISCO SSU proteins are shown in Fig. 1. Although the primary structure of these proteins is quite variable, the predicted secondary structures show great overall similarity. In higher plants RUBISCO SSU is a membrane-translocated protein which is synthesized in the cytoplasm as a larger precursor. The precursor contains a transit peptide at the amino-terminus which mediates the passage of the molecule through the chloroplast membrane and which is cleaved off to yield the mature protein (Langridge, 1981). In contrast, *Synechococcus* SSU is a cytoplasmic protein synthesized without a transit peptide; its primary structure is homologous to that of the mature SSU in plants. In accordance with this difference in biosynthesis, one of the major differences between the predicted structures is found at the amino-terminal region: the higher plant sequences are predicted to have β-structures here, characteristic of membrane-translocated proteins (Reichelt and Delaney, 1983), while the *Synechococcus* protein is predicted to be helical in this region.

Figure 2 shows how common patterns which are not obvious from the primary sequence can be visualized using secondary structure prediction. The transit peptides from the pea chlorophyll a/b binding polypeptide and those of several RUBISCO SSU precursors are compared in Fig. 2A. The sequences vary in length and there is no extensive homology between them, apart from the conservation of charged residues (Timko and Cashmore, 1984). The predicted secondary structure of chloroplast transit peptides, however, reveals a periodic pattern of β-sheets, which seems to be shared by all of these sequences (Fig. 2B). This pattern is distinctly different from those observed for signal sequences of secreted proteins (for review, see Von Heinje, 1983). The transit sequence of the *Chlamydomonas reinhardii* SSU precursor differs slightly from the general pattern of the higher plant SSU transit peptides. It is interesting to note that the SSU precursor of *C. reinhardii* is not transported into the chloroplasts of higher plants *in vitro* though higher plant SSU precursors are taken up and processed by chloroplasts of other higher plants (Chua and Schmidt, 1978).

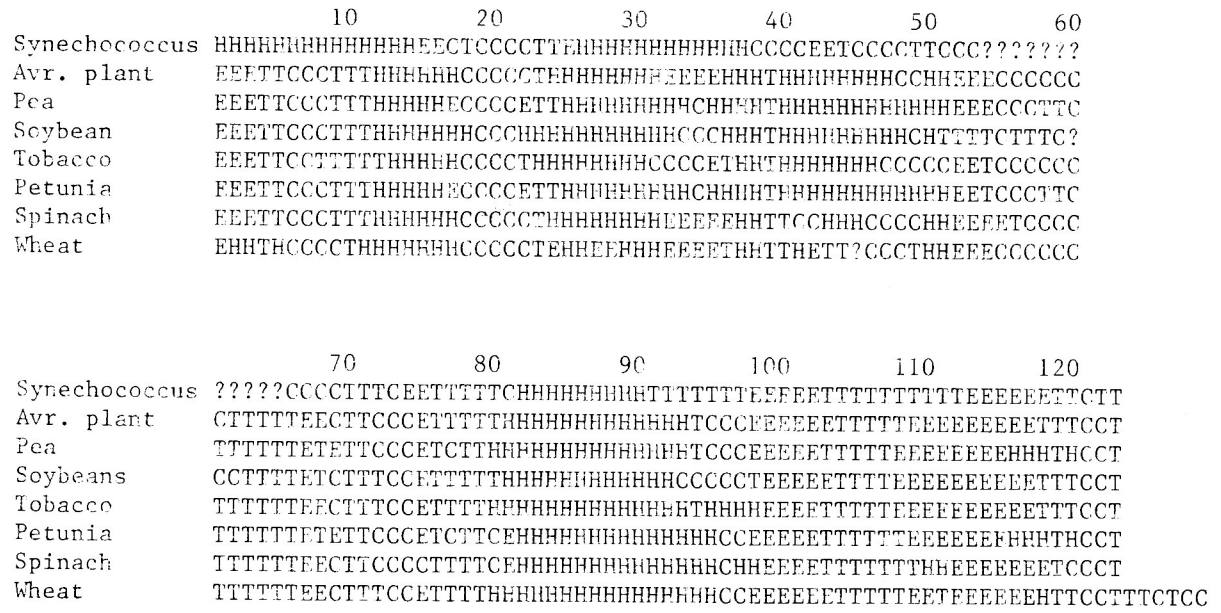

FIG. 1. Secondary structure of RUBISCO SSU proteins as predicted by program PRO1. Method: Garnier *et al.* (1978) (unbiased prediction). H, α-helix; E, extended (β) structure; T, turn; C, coil; ?, alignment gap. Sequences and alignment after Shinozaki and Sugiura (1983). Running times: 2 min, 30 sec for one SSU sequence (123 residues); 15 min for the joint prediction of six SSU sequences.

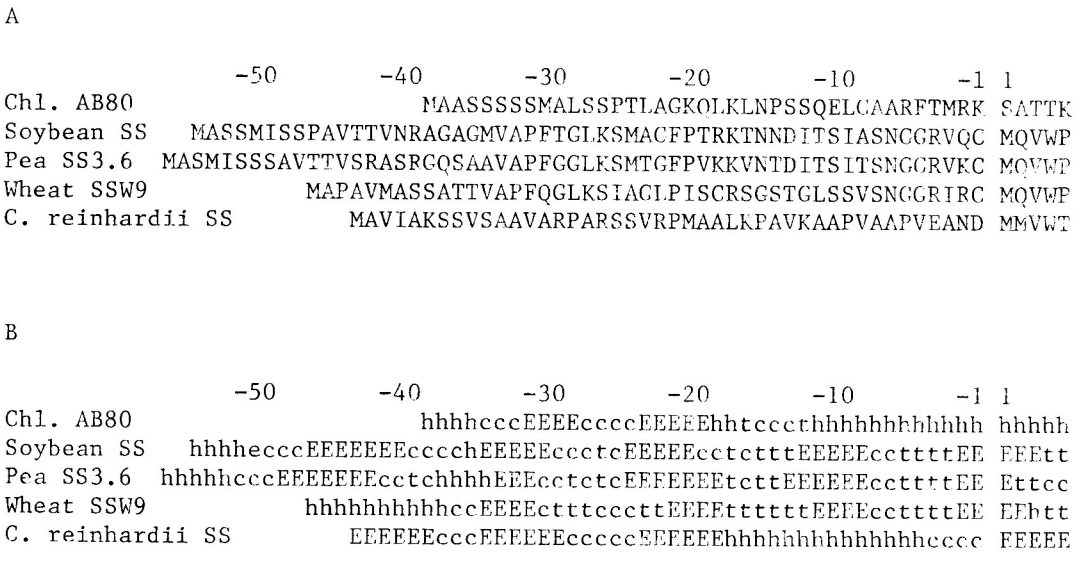

FIG. 2. Amino acid sequence (A) and predicted secondary structure (B) of transit peptides that mediate the post-translational transport of precursor proteins into chloroplast, Chl. AB80, Pea chlorophyll a/b binding polypeptide; SS, RUBISCO small subunit precursor transit peptides from *Chlamydomonas reinhardii*, soybeans, pea, and wheat (Timko and Cashmore, 1984). Prediction by program PRO1. An unbiased prediction was used for the higher plant sequences and DC_H = 158 for the *C. reinhardii* sequence.

Structural profiles: PRO2–PRO4

Structural profiles included in the programs fall into two categories: (i) Single-residue methods (such as methods 1–3 in Table 1 and the Chou-Fasman profiles) use single constants for each amino acid as structural parameters; these methods are included in program PRO2. For better comparison and statistical handling of the data, the structural parameters used by PRO2 are normalized to mean zero and standard deviation one (Table 2). (ii) Multiple-residue methods take neighbor effects into account in the calculation of the structural parameters. Two such methods are included: the directional information method of Garnier *et al.* (1978) in PRO3 and the hydrophobic moment plot (Eisenberg *et al.*, 1984b) in PRO4. The programs use a weighted moving average smoothening procedure with an optional window-width.

Calculation of averaged secondary structure profiles: identification of structurally conserved regions (PRO2, PRO3). PRO3 and PRO4 also include the option to draw average profiles for a group of up to 10 sequences that have been aligned for maximum homology. The standard deviation of the P_i values, σ_i, which is characteristic of the variability at each sequential position, is also displayed in these cases. This feature can provide additional information about the strength of secondary structure prediction: It follows from the probabilistic definition of the P_i values that only those regions for which $P_i > \sigma_i$ (the structure-forming propensity is substantially different from zero) can be considered conserved in the structural sense.

Figure 3 shows averaged α-helix, turn and β-structure propensity profiles calculated for six RUBISCO SSU proteins from different higher plants. It appears that there are only a few regions in which the P_i values are substantially higher than the respective σ_i values; the rest of the protein has no characteristic secondary structure-forming potential. There are two plausible explanations for this finding. First, the structurally determined regions, serving as folding nuclei, may be sufficient for the formation of the native conformation. Second, external factors, such as the large subunit itself, may contribute to the formation of the native conformation of the small subunit, so there may be no rigid selectional constraints on the secondary structure forming propensity of the SSU protein. Region 54–66 displays a pronounced β-turn-forming propensity, which is highly conserved in the SSUs of different higher plants. This region is entirely missing from the cyanobacterial SSU proteins, and is located near an intron–exon junction found in the pea and soybean SSU genes. A comparison between the three-dimensional structures of homologous eukaryotic and prokaryotic proteins has shown that intron–exon junctions in eukaryotic genes frequently coincide with loop structures present on the surface of the eukaryotic proteins (Craik *et al.*, 1983). The variable surface loop structures thus identified were suggested to account for functional differences among the members of a protein family. The strong turn-forming potential of region 54–66 in higher plant SSUs is in fact consistent with a loop structure, which may explain functional differences between higher plant and cyanobacterial SSUs.

Hydrophobicity and hydrophobic moment plots: identification of membrane and surface-associated regions (PRO2,PRO4). The hydrophobicity plot of Kyte and Doolittle (1982) is widely used to identify membrane-associated sequences. This method, included in program PRO2, identifies transmembrane helices by their hydrophobicity which is plotted for each position in the sequence. Eisenberg and co-workers (1984b) recently have expanded this approach by using a hydrophobic moment plot as well as the hydrophobicity values to identify transmembrane helices. The hydrophobic moment measures the amphiphilicity of a helix (Eisenberg *et al.*, 1984a). Its value is high for surface-associated helices (polar and apolar amino acids located on opposite sides of the helix) and low for transmembrane helices which on the other hand have high average hydrophobicities. Program PRO4 contains options to display hydrophobic moment plots as well as to calculate average hydrophobicities and hydrophobic moments required for identification of transmembrane and surface-associated helices by the method of Eisenberg *et al.* (1984b).

COMPARISON OF STRUCTURAL PROFILES

Graphic comparison

If $P_{i,1}$ and $P_{i,2}$ denote structural parameters of the ith amino acid residue in sequences 1 and 2, respectively, then the difference can be written as

$$\Delta P_i = P_{i,2} - P_{i,1} \tag{1}$$

and plotted against the sequential position i (9). Difference plots give a zero baseline for completely homologous regions. Regions that are different in the two sequences will give peaks or valleys depending on the sign of the difference. In all the programs, $P_{i,1}$ can be an average value calculated for a group (maximum 9) of homologous sequences. This feature is useful when a newly derived sequence is compared with a group of known sequences, since a comparison of ΔP_i values and σ_i values indicates if the new sequence is any different, in terms of the given structural feature, from those already known. Examples for difference plots are given in Pongor and Szalay (1985).

Numeric comparison

The concept of using correlation coefficients to compare structural profiles was developed to characterize overall similarities in hydrophobicity (Sweet and Eisenberg, 1983) and secondary structure (Pongor and Szalay, 1985). With the terminology introduced above, the correlation coefficient calculated between two structural profiles can be written as follows:

$$R = \frac{\Sigma(P_{i,1} \times P_{i,2})}{(\Sigma P_{i,1}^2 \times \Sigma P_{i,2}^2)^{1/2}} \tag{2}$$

The correlation indices calculated here range between -1 and $+1$; their interpretation is analogous to that of the linear correlation coefficient (Goulden, 1952). The pro-

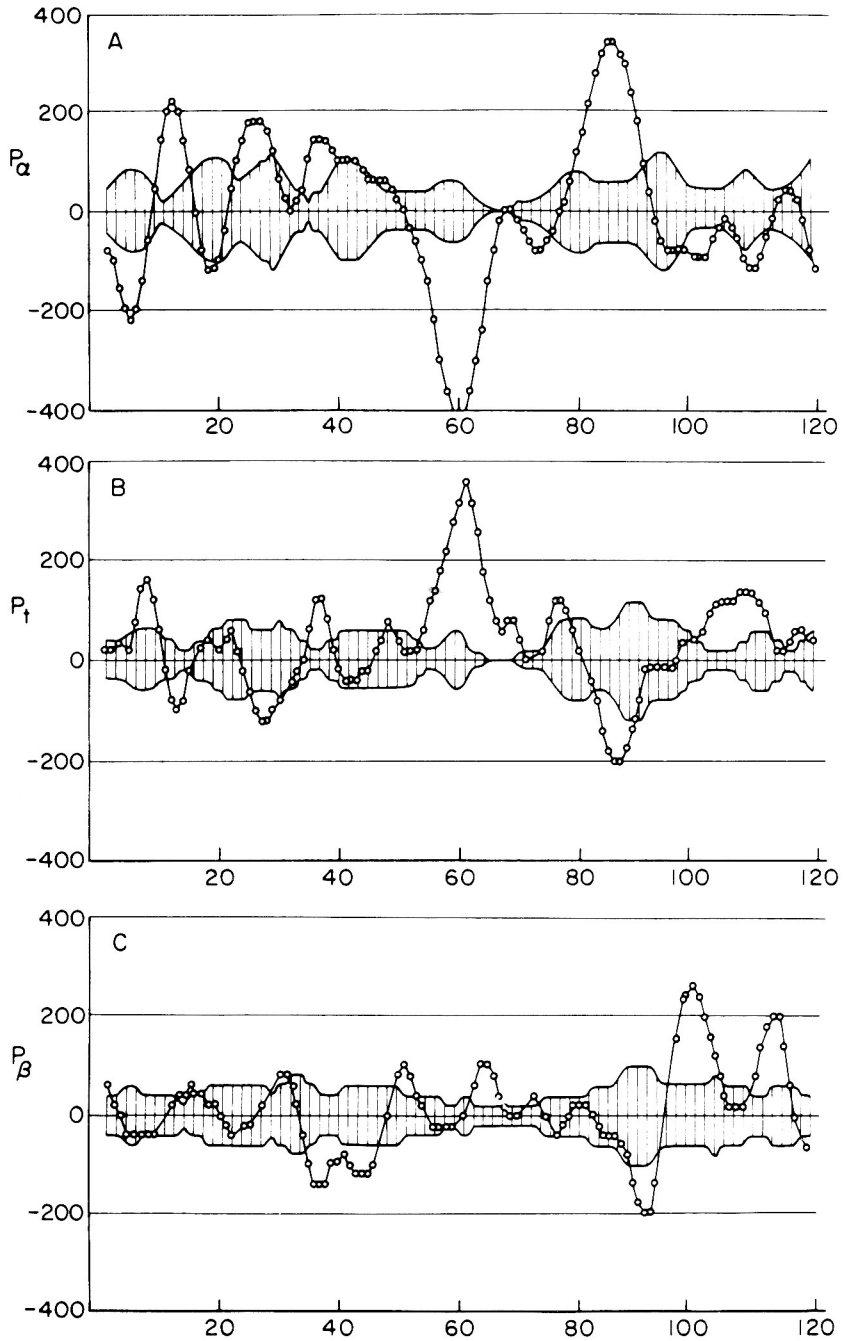

FIG. 3. Averaged α-helix (A), turn (B), and β-structure (C) propensity profiles calculated with program PRO3 from RUBISCO SSU sequences of five species of higher plants (see Fig. 1). Method: Garnier *et al.* (1978). Shaded areas indicate standard deviation ($\pm \sigma_i$) characteristic of the variability of the respective structural parameter. Running times: 10 min for calculation of one averaged structural profile from five SSU sequences.

grams include an option to calculate correlation coefficients for entire sequences or segments thereof using any of the structural profiles mentioned above. Using the averaging option, it is possible to compare a sequence to a group of homologous sequences ($P_{i,1}$ can be an average of maximum 9 values). This correlation index has been used to illustrate evolutionary relatedness of RUBISCO sequences (Pongor and Szalay, 1985). Schaeffer and Sninsky (1984) used a related index to describe similarities of Chou-Fassman profiles for hepatitis B virus open reading frames.

Two new numeric indices are also included programs PRO2–PRO4: (i) $<\Delta P_i>$ is the average difference (the

sum of the ΔP_i values, calculated between two sequences, divided by the number of residues in a segment); and (ii) $< |\Delta P_i| >$ is the average absolute difference (calculated as above but using the absolute values of the difference values). These indices were designed to allow a quick evaluation of amino acid replacements in structural terms (such as increase of decrease in hydrophobicity or secondary structure forming potential).

Table 3 shows the comparative evaluation of the amino-terminal region of RUBISCO LSU using the above criteria. In this example we compare the sequences of three LSU proteins from different higher plants ($P_{i,1}$ is the average of the corresponding parameters of the maize, spinach, and tobacco LSU sequences) to that of the cyanobacterium *Synechococcus R2* ($P_{i,2}$) (see Reichelt and Delaney, 1983, for the sources of sequences). In the chloroplast of higher plants, RUBISCO LSU is synthesized as a larger precursor protein from which a 14-residue amino-terminal extension is removed by proteolysis (Langridge, 1981); the processing has not been experimentally confirmed in cyanobacteria. Data summarized in Table 3 show that the amino-terminal extension of *Synechococcus* LSU (A) bears little resemblance to the homologous higher plant sequences (*R* values low, differences high), in contrast to the putative amino-terminal region (B) which seems to be highly conserved. The pattern of variability is thus similar to that seen among LSUs of different higher plants; this finding supports the suggestion of Reichelt and Delaney (1983) that the *Synechococcus* LSU is also synthesized as a precursor.

CONCLUDING REMARKS

Our work is directed toward a generalized use of structural profiles in assessing the similarities of amino acid sequences. The programs presented here were designed for routine applications of protein secondary structure prediction methods which are typical of the practice of molecular biology. The programs can be used for preliminary characterization of newly determined DNA sequences in terms of predicted protein structure, and also for the *comparative evaluation* of protein sequences, which is a novel feature of these programs. The latter feature can be of use in determining those properties in which the structure of a protein is expected to be different from or similar to its known homologs (proteins encoded by homologous genes), as was illustrated here by the comparative evaluation of ribulose-1,5-bisphosphate carboxylase sequences. Second, the programs allow a quantitative description of amino acid replacements using structural terms which is useful in the design of altered proteins. The secondary structure prediction program (PRO1) is the first Apple implementation of the directional information algorithm (Garnier *et al.*, 1978) that we are aware of.

ACKNOWLEDGMENTS

The expert secretarial assistance of Ms. Debbie Bridwell is gratefully acknowledged. This work was supported by the Boyce Thompson endowment, by grant No. PCM-7820252 from the National Science Foundation, and by a grant from International Minerals and Chemical Corporation to A.A. Szalay.

REFERENCES

ARGOS, P., PEDERSEN, K., MARKS, D.M., and LARKINS, B.A. (1982). A structural model for maize zein proteins. J. Biol. Chem. **257**, 9984–9900.

TABLE 3. COMPARISON OF *Synechococcus* RUBISCO LSU TO HIGHER PLANT LSU SEQUENCES[a]

	Correlation coefficent R		*Average difference* $<\Delta P_i>$		*Average abs. difference* $<|\Delta P_i|>$	
Parameter[b]	*A*[c]	*B*[d]	*A*	*B*	*A*	*B*
1. Hydrophobicity	0.75	1.00	0.11	−0.01	0.32	0.01
2. Hydrophilicity	0.68	1.00	−0.05	0.00	0.40	0.00
3. α-helix	0.78	0.93	−0.13	−0.14	0.45	0.14
4. β-turn	0.82	0.96	0.09	0.10	0.36	0.10
5. β-structure	0.62	0.99	0.02	0.04	0.44	0.04
6. Charges	0.49	1.00	0.21	0.00	0.21	0.00
7. OMH scale	0.94	1.00	−0.01	−0.01	0.23	0.01

[a]Sequences aligned according to Reichelt and Delaney (1983). $P_{i,1}$ in Eqs. [1] and [2] was the average calculated from maize, spinach, and tobacco LSU sequences; $P_{i,2}$ was the corresponding parameter of *Synechococcus* LSU.

[b]Parameters summarized in Table 2.

[c]Putative amino-terminal extension (residues 1–14).

[d]Amino-terminal sequence (residues 15–28).

CHUA, N.-H., and SCHMIDT, G.W. (1978). Post-translational transport into intact chloroplast of a precursor to the small subunit of ribulose-1,5-bisphosphate carboxylase. Proc. Natl. Acad. Sci. USA **75,** 6110–6114.

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

CORRIGAN, A.J., and HUANG, P.C. (1982). A BASIC microcomputer program for plotting the secondary structure of proteins. *Computer Programs in Biomedicine* **15,** 163–168.

CRAIK, C.S., RUTTER, W.J., and FLETTERICK, R. (1983). Splice Junctions: Association with variation in protein structure. Science **220,** 1125–1129.

DE BANZIE, J.S., STEEG, E.W., and LIS, J.T. (1984). Update for users of the Cornell sequence analysis package. Nucleic Acids Res. **12,** 619–625.

EISENBERG, D., WEISS, R.M., and TERWILLIGER, T.C. (1984a). The hydrophobic moment detects periodicity in protein hydrophobicity. Proc. Natl. Acad. Sci. USA **81,** 140–144.

EISENBERG, D., SCHWARZ, E.M., KOMAROMY, M., and WALL, R. (1984b). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J. Mol. Biol. **179,** 125–142.

FRISTENSKY, B., LIS, J., and WU, R. (1982). Portable microcomputer software for nucleotide sequence analysis. Nucleic Acids Res. **10,** 6451–6463.

GARNIER, F., 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.

GOULDEN, C.M. (1952). In *Methods in Statistical Analysis.* (John Wiley & Sons, New York) 2nd Edition, pp. 122–133.

HOPP, T.P., and WOODS, T.P. (1981). Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. USA **78,** 3824–3828.

KORN, L.J., and QUEEN, C. (1984). Analysis of biological sequences on small computers. DNA **3,** 421–436.

KYTE, J., and DOOLITTLE, R.F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157,** 105–132.

LANGRIDGE, P. (1981). Synthesis of the large subunit of the spinach ribulose bisphosphate carboxylase involves a precursor polypeptide. FEBS Lett. **123,** 85–89.

LARSON, R., and MESSING, J. (1983). Apple II computer software for DNA and protein sequence data. DNA **2,** 31–35.

MAIZEL, J.V., and LENK, R.P. (1981). Enhanced graphic matrix analysis of nucleic acid and amino acid sequences. Proc. Natl. Acad. Sci USA **78,** 7665–7669.

MALTHIERY, B., BELLON, B., GIORGI, D., and JACQ, B. (1984). Apple II Pascal computer programs for molecular biologists. Nucleic Acids Res. **12,** 569–579.

NOVOTNÝ, J., and AUFFRAY, C. (1984). A program for prediction of protein secondary structure from nucleotide sequence data: Application to histocompatability antigens. Nucleic Acids Res. **12,** 243–255.

PONGOR, S., and SZALAY, S.S. (1985). Prediction of homology and divergence in the secondary structure of polypeptides. Proc Natl. Acad. Sci. USA **82,** 366–370.

REICHELT, B.Y., and DELANEY, S.F. (1983). The nucleotide sequence for the large subunit of ribulose-1,5-bisphosphate carboxylase from a unicellular cyanobacterium *Synechococcus.* DNA **2,** 121–129.

SCHAEFFER, A., and SNINSKY, J.J. (1984). Predicted secondary structure similarity in the absence of primary amino acid sequence homology: Hepatitis B virus open reading frames. Proc. Natl. Acad. Sci. USA **81,** 2902–2906.

SCHULZ, G.E., and SCHIRMER, R.H. (1979). In *Principles of Protein Structure,* Ch. 6. (Springer Verlag, New York).

SHINOZAKI, K., and SUGIURA, M. (1983). The gene for the small subunit or ribulose-1,5-bisphosphate carboxylase is located close to the gene for the large subunit in the cyanobacterium *Anacystis nidulans.* Nucleic Acids Res. **11,** 6957–6963.

SWEET, R.M., and EISENBERG, D. (1983). Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. J. Mol. Biol. **171,** 479–488.

TIMKO, M.P., and CASHMORE, A.R. (1984). Nuclear genes encoding the constituent polypeptides of the light-harvesting chlorophyll a/b-protein complex from pea. In *Plant Molecular Biology,* R.B. Goldberg, ed. (Alan R. Liss, New York) pp. 403–412.

VON HEINJE, G. (1983). Patterns of amino acids near signal-sequence cleavage sites. Eur. J. Biochem. **133,** 17–21.

Address reprint requests to:
Dr. A.A. Szalay
Boyce Thompson Institute for Plant Research
Cornell University
Ithaca, NY 14853

Received for publication November 28, 1984, and in revised form March 18, 1985.