Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 366–370, January 1985
Biochemistry

Prediction of homology and divergence in the secondary structure of polypeptides

(protein secondary structure/protein homology/secondary structure prediction/ribulose 1,5-bisphosphate carboxylase)

SÁNDOR PONGOR AND ALADAR A. SZALAY*

Boyce Thompson Institute for Plant Research, Cornell University, Ithaca, NY 14853

Communicated by Gordon H. Sato, September 27, 1984

ABSTRACT **A quantitative procedure is described for the comparison of secondary structure of homologous proteins. Standard predictive methods are used to generate probability profiles from pairs of homologous amino acid sequences; correlation coefficients (R) are then computed between each pair of amino acids for α-helix (R_α), extended structure (R_β), turn (R_t), and coil (R_c). R values are >0.2 for correctly aligned homologous sequences. Unrelated or incorrectly aligned sequences give R values near zero. Lack of correlation for a segment of otherwise well-correlated sequences is used to identify structural divergence, which is then evaluated graphically by using difference profiles. A combination of these techniques correctly predicts secondary structural differences between melittin or β-endorphin and their respective synthetic analogs. The method is potentially useful to describe evolutionary changes in protein secondary structure as well as in the design of peptide analogs.**

Homology of proteins has been studied at various levels of structural organization. Analysis of amino acid sequences played a key role in establishing phylogenetic relationships (1, 2); comparison of three-dimensional structures made it possible to identify similarities between proteins that are not demonstrably homologous in primary structure (3, 4). Whereas these procedures provide quantitative measures of homology in amino acid sequence or in three-dimensional structure, comparison of secondary structures is usually carried out on a qualitative basis, such as visual comparison of predicted secondary structures (5). Hydropathy profiles (6), "helical wheels" (7), and "helical nets" (8) proved to be very useful in recognizing common structural patterns in homologous sequences such as membrane-bound segments (9) or amphiphilic helices (10), but they do not provide quantitative information concerning secondary structure. Furthermore, they provide little help in characterizing structural differences that may exist between homologous proteins, a problem crucial in the design of peptides or protein analogs. Recent progress in this area has shown that biological activity of peptides can be improved or modulated by modifying their secondary-structure-forming potential (10, 11). Similar changes may have taken place in the functional evolution of proteins, but current understanding of these processes is limited by the lack of quantitative methods.

In this paper we describe a quantitative procedure for the comparison of predicted secondary structures. We show that correlation coefficients of structural profiles, generated by standard predictive methods (12, 13), can be used to describe similarity in secondary structure. Differences in secondary structure are indicated by a low correlation value and can be graphically evaluated using difference profiles. Here we describe the method and its application to theoretical model sequences, synthetic peptides of known structural differences, and evolutionarily related, functionally identical proteins.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

METHOD OF CALCULATION

Calculation of Secondary Structure Correlation. We define the secondary structure correlation index R as a measure of similarity between two structural profiles. If $P_{j,1}$ and $P_{j,2}$ denote a structural propensity parameter of the jth amino acid in amino acid sequences 1 and 2, respectively, then R can be written as follows:

$$R = \frac{\Sigma P_{j,1} \times P_{j,2}}{\sqrt{\Sigma P_{j,1}^2 \times \Sigma P_{j,2}^2}} \quad [1]$$

If one of the P values is missing because of a deletion in one of the sequences, the other P value is not taken into account in the calculation. We have chosen Eq. **1** because it is analogous to the linear correlation coefficient calculated between $P_{j,1}$ and $P_{j,2}$ values. In the simplest case, P_j values are the Chou–Fasman parameters (12) giving thus separate R values for α-helix (R_α), β-sheet (R_β), and turn (R_t). Alternatively, P_js can be calculated by the directional information method of Garnier *et al.* (13) as shown here for α-helix:

$$P_{j,\alpha} = \sum_{i=j-8}^{i=j+8} I(A_i) + D_\alpha, \quad [2]$$

where $I(A_i)$ is the directional information measure of an amino acid A_i, and D_α is the decision constant (taken as 0 in these calculations) (13).

In addition to secondary structure parameters, we also used hydrophilicity values (14), charge numbers [−1 for aspartate and glutamate, +1 for lysine and arginine, as used in plots by Novotný and Auffray (15)], and hydrophobicity values (6, 16) to calculate correlation coefficients according to Eq. **1**. In the last case, the expression of R becomes analogous to that recently suggested by Sweet and Eisenberg (16).

$R = 1$ for two identical sequences, $R \approx 0$ for two unrelated sequences, and R approaches −1 if the structural profiles are anticorrelated. The z-test of Fisher (11) can be used as a test of significance. To meet the criteria of the z-test, the Chou–Fasman parameters as well as the hydrophobicity (14) and hydrophilicity (16) values were rescaled to mean zero and standard deviation 1 for the calculation of Eq. **1**. However, any two sequences that could be rationally aligned were found to result in R values with very high statistical significance. (Some examples are given in the legends to Tables 1–

Abbreviations: R, secondary structure correlation index; P, structural propensity parameter; Rb*P*Case, ribulose 1,5-bisphosphate carboxylase. The single-letter amino acid code used is A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp.

*To whom correspondence and reprint requests should be addressed.

3). Consequently, we prefer to use the correlation values as relative, ranking indicators of structural similarity. Comparison to known structures could be used to propose criteria of evaluation.

Difference Profiles. R values give a general measure of similarity of secondary structure for two homologous sequences; we used difference profiles to establish the sequential location and direction of the differences underlying a particular R value. With the terminology introduced above, these can be expressed as ΔP_j vs. j plots, where

$$\Delta P_j = (P_{j,2} - P_{j,1}). \quad [3]$$

These plots give a zero baseline for completely homologous regions and give peaks or valleys wherever differences exist between the two structural profiles. Quantitative evaluation of ΔP values is meaningful only on a comparative basis. Such an evaluation is possible if a sequence is compared to a group of selected sequences. $P_{j,1}$ is then an average value calculated from several related sequences with a σ_j standard deviation. σ_j indicates the variability of P_j within a given group of sequences and ΔP_j should substantially exceed σ_j for structural differences to be predicted in a given region. (See Fig. 5 for an example.)

Computer Programs. Calculations were carried out by programs written in Apple Pascal. The programs are run on an Apple II plus microcomputer (64K memory) and calculate structural correlation and difference profiles by using the structural parameters of Garnier *et al.* (13) and of Chou and Fasman (12) as well as hydrophobicity values (6, 16), hydrophilicity values (14), and charge numbers. The programs are available upon request.

RESULTS AND DISCUSSION

Alignment of Sequences Is Critical. The value of the structural correlation index R depends on the alignment of the sequences to be compared. This is illustrated in Fig. 1, where several R values are plotted against an alignment shift, using the tobacco ribulose 1,5-bisphosphate carboxylase (EC 4.1.1.39; Rb*P*Case) small subunit sequence (123 amino acids) as an example. R values calculated by the directional information methods (13) give smooth maxima; all the single-residue information methods show a sharp drop on a misalignment by one residue, with smaller maxima on further shift, probably arising from internal periodicities of the

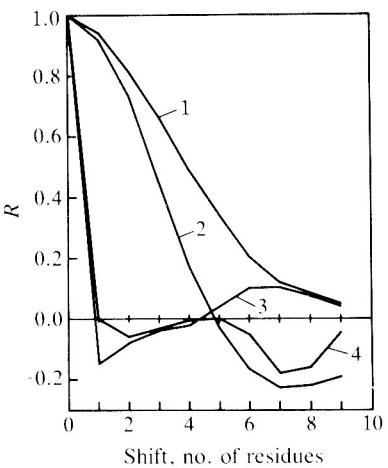

FIG. 1. Secondary structural correlation values plotted against a misalignment shift. Curve 1, R_α [Garnier *et al.* (13)]; curve 2, R_β (13); curve 3, R_H [hydrophobicity correlation, Sweet and Eisenberg (16)]; curve 4, R_α [Chou and Fasman (12)]. Sequence used was that of Rb*P*Case small subunit from tobacco (17).

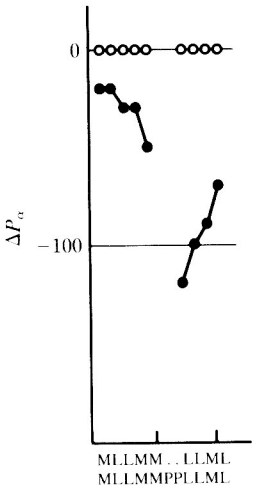

FIG. 2. Secondary structural difference profiles for hypothetical sequences [4] as determined by directional information method of Garnier *et al.* (13) (●) and Chou–Fasman method (12) (○).

sequence such as the alternation of polar and nonpolar residues in α-helices (16).

We suggest that only sequences for which a single rational alignment is possible be compared by the present method. This is generally straightforward with synthetic peptide analogs; in the case of protein sequences, the respective DNA sequence may be used as a template for alignment.

Single-Residue vs. Directional Information Methods. In principle both the Chou–Fasman method (12) and the directional information method of Garnier *et al.* (13) can be used to calculate structural correlation values. There is an important difference between these methods, however, in the handling of deletions and insertions. For example, the two hypothetical sequences

M L L M M · · L L M L

and M L L M M P P L L M L [4]

give $R = 1$ if the Chou–Fasman structural parameters are used in the calculation. On the other hand, the directional information method of Garnier *et al.* (13) takes neighbor interactions into account and, accordingly, the structural cor-

Table 1. Structural correlation of melittin and synthetic analog

Structure (method)	R
α-Helix [Garnier *et al.* (13)]	0.11
Extended structure [Garnier *et al.* (13)]	0.89
β-Turn [Garnier *et al.* (13)]	0.81
Coil [Garnier *et al.* (13)]	0.54
α-Helix [Chou–Fasman (12)]	0.18
β-Sheet [Chou–Fasman (12)]	0.79
β-Turn [Chou–Fasman (12)]	0.53
Hydrophobicity [Kyte–Doolittle (6)]	0.94
Hydrophobicity [Sweet–Eisenberg (16)]	0.83
Hydrophilicity [Hopp–Woods (14)]	0.90
Charge distribution	0.89

Sequences (21):

Mellitin 10 20

G I G A V L K V L T T G L P A L I S W I K R K R Q Q

Synthetic analog

L L Q S L L S L L Q S L L S L L Q W L K R K R Q Q

Values given were calculated by using Eq. 1 (P_1, melittin; P_2, synthetic analog). R values >0.40 are significant at the 99.99% level (11).

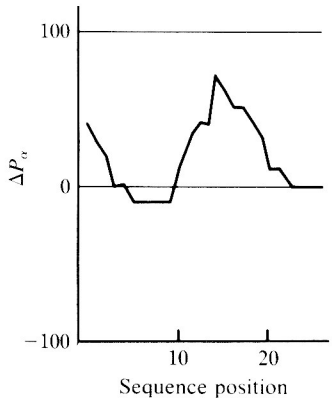

FIG. 3. Helicity difference profiles of melittin (P_1 in Eq. **3**) and synthetic analog (P_2 in Eq. **3**), as determined from α-helix probability values of Garnier *et al.* (13). See sequences in legend of Table 1.

relation coefficients are <1 (R_α = 0.79 and R_β = −0.52). Similarly, the expected difference is graphically detected by the difference profiles calculated from the directional information parameters but not by those calculated from the single-residue information methods (Fig. 2). Based on this finding, it can be supposed that difference profiles calculated from the directional information parameters may be able to detect deletions and insertions that are not readily apparent from the primary sequences.

Another difference is apparent from the comparison of difference profiles (see Fig. 4). While the single-residue information methods (Fig. 4, profiles A) indicate the variability, they do not show the direction of the trend as clearly as does the directional information method (profiles B). In Fig. 4, the fluctuations of the difference profiles A mask the overall pattern evident in the smoother profiles B.

Tests with Synthetic Peptides. The procedures of comparison described above were tested first with synthetic peptide analogs that have known differences in secondary structure as compared to their natural counterparts. The synthetic peptides were designed to have the following general characteristics relative to the corresponding natural peptides: (*i*) equal or increased helix content; (*ii*) conserved charged residues and hydrophobic/hydrophilic balance; (*iii*) low (20–50%) primary sequence homology (10, 19).

Melittin is a 26-residue hemolytic peptide isolated from bee venom, which has the potential to form an amphiphilic helix (20). When the sequence of the amphiphilic segment was modified according to the above specifications, a biologically active analog was obtained that was higher in α-helix content (35%) than the native melittin (18%), as determined by circular dichroism (21). Structural correlation values listed in Table 1 reflect the difference in α-helix formation: The value of R_α is the lowest among all values compared. A positive peak in the difference plot clearly shows the increase in helix-forming potential (Fig. 3).

Structural correlation values calculated for β-endorphin and its three synthetic analogs are summarized in Table 2. The analogs were designed in a manner to retain or increase the propensity of β-endorphin to form an amphiphilic helix while leaving the hydrophobic/hydrophilic balance intact (22). The structural correlation values show indeed a general similarity in secondary structure, whereas the low correlation of charges reflects the fact that the charged amino acids were not all paired in the analogs. The α-helix content of the peptides follows the order 2 > 1 > 3 > β-endorphin (10, 22). According to the difference profiles shown in Fig. 4*B*, analog 2 is indeed predicted to have the highest α-helix content; analogs 1 and 3 show relatively small ΔP_α values in the helical region (residues 14–31). Interestingly, the value of hydrophobicity correlation indicates some dissimilarity even though the hydrophobic/hydrophilic balance is retained in the analogs; this calculated dissimilarity is probably due to the sensitivity of this method to sequential rearrangements (see Fig. 1, curve 3).

Tests with Homologous Proteins. We have calculated secondary structural correlation coefficients between Rb*P*Case sequences of plant and bacterial origin (for a review, see ref. 17). Since secondary structure prediction indicates a high α-helix content for both the small and the large subunit of this enzyme (ref. 23 and unpublished observations), here we present only the R_α helical propensity correlation coefficients. The values summarized in Table 3 are statistically significant at the 99.99% level for all possible comparisons. This is expected since Rb*P*Case fulfills the same enzymatic functions in all photosynthetic organisms. Low R_α values are found between the bacterial (*R. rubrum, Synechococcus*) sequences on one hand and the higher plant sequences on the other, which is in accordance with the large evolutionary dis-

Table 2. Structural correlation of β-endorphin and synthetic analogs

	R		
Structure (method ref.)	Analog 1	Analog 2	Analog 3
α-Helix (13)	0.97	0.76	0.64
Extended structure (13)	0.79	0.70	0.79
β-Turn (13)	0.70	0.81	0.66
Coil (13)	0.80	0.79	0.79
α-Helix (12)	0.78	0.62	0.58
β-Sheet (12)	0.90	0.85	0.83
β-Turn (12)	0.86	0.85	0.78
Hydrophobicity (6)	0.35	0.79	0.23
Hydrophobicity (16)	0.49	0.85	0.39
Hydrophilicity (14)	0.45	0.67	0.23
Charge distribution	0.30	0.46	−0.18

Sequences (22) (% sequence homologies with β-endorphin are given in parentheses):

	10								20								30								
β-Endorphin	Y	G	G	F	M	T	S	E	K	S	Q	T	P	L	V	T	L	F	K	N	A	I	I	K	L	A	Y	K	K	G	E
Analog 1 (61%)	Y	G	G	F	M	T	S	E	K	S	Q	T	P	L	V	T	L	F	K	Q	L	L	K	Q	L	Q	K	L	L	Q	K
Analog 2 (51%)	Y	G	G	F	M	T	S	E	K	S	Q	T	P	L	L	K	L	L	Q	K	L	L	L	Q	L	L	F	K	Q	K	Q
Analog 3 (29%)	Y	G	G	F	M	S	G	S	G	S	G	S	P	L	L	Q	L	W	Q	K	L	L	K	Q	L	Q	K	L	L	Q	K

R values were calculated as in Eq. **1** (P_1, β-endorphin, P_2, analog 1, 2, or 3, respectively). *R* values >0.37 are significant at the 99.9% level (11).

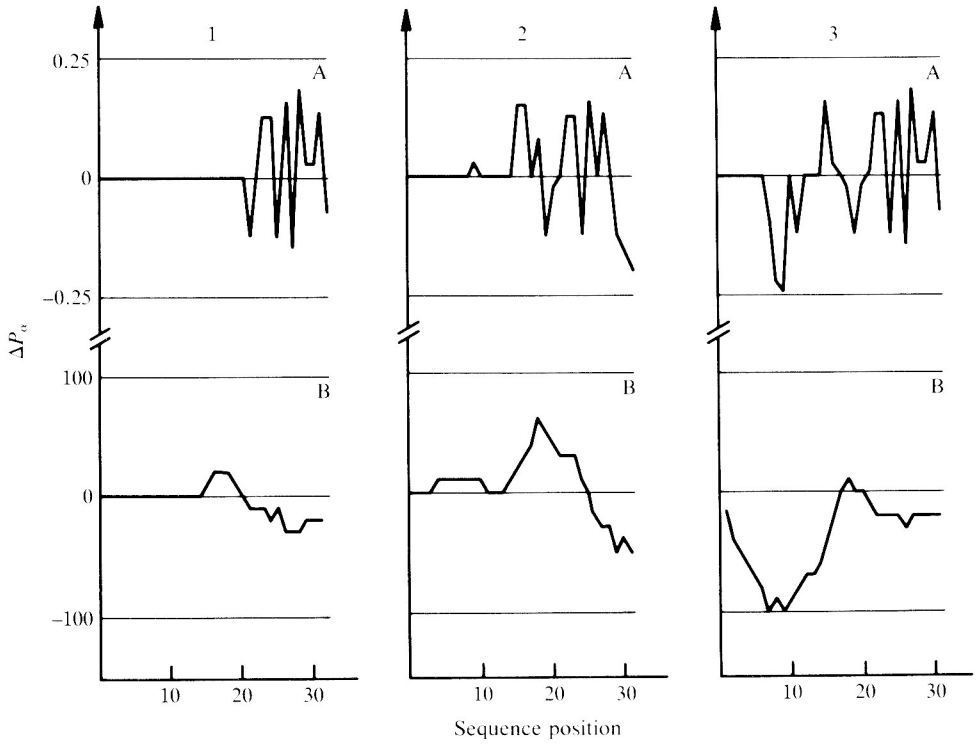

FIG. 4. Helicity difference profiles comparing β-endorphin (P_1 in Eq. **3**) to its synthetic analogs 1, 2, and 3 (P_2 in Eq. **3**, respectively). Profiles A, Chou–Fasman method (12); profiles B, method of Garnier *et al.* (13). Number above each pair of profiles indicates the analog involved in the comparison. See sequences in legend of Table 2.

tance between these organisms. A comparison of the R_α values obtained between various higher plants shows that monocotyledonous and dicotyledonous plants also can be distinguished on the same basis, although the R_α values are substantially higher than in the case of the plant/bacterium comparisons.

To test the ability of the method to detect differences (as opposed to similarities) in secondary structure, we carried out a detailed comparison between the *Synechococcus* Rb*P*Case small subunit sequence and the corresponding higher plant sequences. The Rb*P*Case small subunit is a membrane-translocated protein in higher plants, which is synthesized in the cytoplasm in the form of a larger precursor. The precursor contains a transit peptide at the NH_2 terminus, which mediates the passage of the molecule through the chloroplast membranes and which is cleaved off to yield the mature small subunit protein (24). In contrast, *Synechococcus* Rb*P*Case small subunit is a cytoplasmic protein homologous to the mature small subunit in plants and is synthesized without transit peptide. To determine whether this difference is manifested in the predicted secondary structure, we compared *Synechococcus* Rb*P*Case small subunit to an average of the higher plant sequences as follows. Average structural profiles were calculated from six higher plant sequences according to Eq. **2** and substituted into Eq. **1** for $P_{j,1}$; the $P_{j,2}$ values were those of the *Synechococcus* Rb*P*Case small subunit sequence. This comparison gave R_α and R_β values of 0.36 and 0.26, respectively. A comparison of individual segments of the sequences revealed that the NH_2-terminal regions (residues 1–40) are the least similar: $R_\alpha = 0.02$ and $R_\beta = 0.09$. Difference plots of the NH_2-terminal region show, that the higher plant Rb*P*Case small subunit proteins have a greater tendency for β-structure (Fig. 5*b*) and a lower α-helix-forming potential (Fig. 5*a*) than the *Synechococcus* protein. This finding is in accordance with the known propensity of membrane-translocated proteins to form β-structure at their NH_2 termini (25).

Table 3. R_α calculated between Rb*P*Case sequences from plant and bacterial sources

	Small subunit							Large subunit			
	To	Sp	Pe	So	Pt	Wh		*SC*	To	Sp	Ma
SC	0.25	0.35	0.42	0.42	0.42	0.44	*RR*	0.41	0.43	0.45	0.30
To		0.88	0.88	0.85	0.88	0.87	*SC*		0.87	0.86	0.74
Sp			0.84	0.88	0.88	0.87	To			0.95	0.85
Pe				0.88	0.95	0.78	Sp				0.85
So					0.82	0.78					
Pt						0.77					

R_α was calculated as in Eq. **1**, using the P_α values of Garnier *et al.* (13). R_α values >0.18 (small subunit) or >0.09 (large subunit) are significant at the 99.99% level. *SC*, *Synechococcus* R2; To, tobacco; Sp, spinach; Pe, pea; So, soybean; Pt, petunia; Wh, wheat; Ma, maize; *RR*, *Rhodospiryllum rubrum*. Alignment according to refs. 17 and 18.

Emr and Silhavy (26) proposed that α-helix formation of the signal peptide plays a key role in the export of the LamB protein to the outer membrane of *Escherichia coli*. Recently, Briggs and Gierasch (27) have shown by chemical synthesis and circular dichroism measurements that the exported signal peptides of the wild type (Wt) and the two revertants (R1 and R2) have higher α-helix contents in hydrophobic environments than the nonexported protein from the mutant (M). Calculations made in our laboratory on the same signal sequences show that the α-helix-forming potential of these signal peptides follow the order R2,R1,Wt > M; these findings are in agreement with the circular dichroism data and provide further support for the suggestion made by Emr and Silhavy (26).

Scope and Limitations of the Method. This paper describes a procedure for the comparison of secondary structures predicted from homologous amino acid sequences. The underlying principle of the method is that a comparative use of predictive techniques should be more reliable than prediction itself, since the long-range interactions—a large source of error unaccounted for in these methods—are often largely

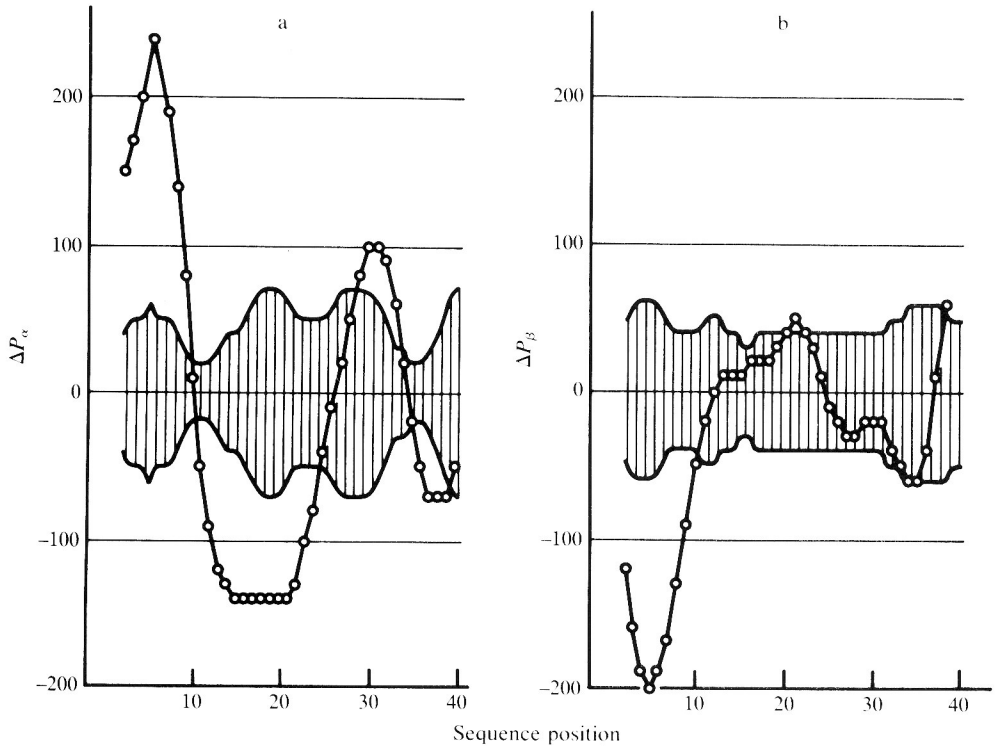

FIG. 5. Helicity (*A*) and β-structure (*B*) difference profiles comparing the NH_2-terminal region of the *Synechococcus* Rb*P*Case small subunit (P_2 in Eq. **3**) to the mean (P_1 in Eq. **3**) of membrane-translocated Rb*P*Case small subunit sequences from tobacco, spinach, petunia, maize, soybean, and pea enzymes (17). The hatched area designates the standard deviation of the P_1 profile (i.e., the range of variability that exists among the plant proteins). The curves were smoothed by a weighted moving average procedure (15).

equivalent in homologous proteins. We compare structural profiles rather than assigned conformations, because conformational assignment is necessarily somewhat arbitrary in secondary structure prediction. Of the two standard methods used to generate structural profiles in this paper, the one of Garnier *et al.* (13) proved to be generally applicable.

Results presented here suggest that secondary structure, as predicted by standard methods, is strongly correlated between homologous proteins. This is consistent with the expectation that secondary structure may be conserved in cases where the primary sequence has changed. The parallel between structural correlation and evolutionary trends suggests that secondary-structure-forming potential is, in fact, conserved in amino acid substitutions, although it is apparent that this phenomenon is not isolated from the sequential context. As an alternative method of assessing protein similarity, structural correlation is probably more useful in detecting distant similarities than methods based on amino acid identities, as was illustrated with our results for synthetic peptide analogs.

The results also indicate problems associated with the method. The procedure can only be used on precisely aligned amino acid sequences. Another difficulty is the lack of an absolute scale to measure similarities or differences; therefore, the results can only be evaluated on a relative basis. These difficulties are greatly reduced if closely related sequences are compared.

To date, structural profiles have been compared mainly by visual inspection. The method presented in this paper allows quantitative expression of structural similarities, which is potentially useful in understanding functional evolution of proteins as well as in optimizing secondary structure characteristics in the design of peptide and protein analogs.

We thank Ms. M. J. Guttieri for valuable discussions, Dr. G. Némethy for helpful remarks on the manuscript, Ms. L. M. Cohen for help in the calculations, and Mrs. J. Ruocco for secretarial assistance. This work was supported by the Boyce Thompson Institute Endowment and by Grant PCM-7820252 from the National Science Foundation to A.A.S.

1. Fitch, W. W. (1966) *J. Mol. Biol.* **16,** 9–16.
2. McLachlan, A. D. (1971) *J. Mol. Biol.* **61,** 409–424.
3. Rossman, M. G. & Argos, P. (1977) *J. Mol. Biol.* **109,** 99–129.
4. Remington, S. J. & Matthews, B. W. (1980) *J. Mol. Biol.* **140,** 77–99.
5. Hung, M.-C. & Wensink, P. (1983) *J. Mol. Biol.* **164,** 481–492.
6. Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157,** 105–132.
7. Schiffer, M. & Edmundson, A. B. (1967) *Biophys. J.* **7,** 121–135.
8. Dunnill, P. (1968) *Biophys. J.* **8,** 865–875.
9. Widger, W. R., Cramer, W. A., Herrmann, R. G. & Trebst, A. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 674–678.
10. Kaiser, E. T. & Kézdy, F. J. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 1137–1143.
11. Goulden, C. M. (1952) *Methods in Statistical Analysis* (Wiley, New York), 2nd Ed., pp. 122–133.
12. Chou, P. Y. & Fasman, G. D. (1978) *Adv. Enzymol.* **47,** 45–148.
13. Garnier, F., Osguthorpe, D. J. & Robson, B. (1978) *J. Mol. Biol.* **120,** 97–120.
14. Hopp, T. P. & Woods, T. P. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 3824–3828.
15. Novotný, J. & Auffray, C. (1984) *Nucleic Acids Res.* **12,** 243–255.
16. Sweet, R. M. & Eisenberg, D. (1983) *J. Mol. Biol.* **171,** 479–488.
17. Miziorko, H. M. & Lorimer, G. M. (1983) *Annu. Rev. Biochem.* **52,** 507–535.
18. Shinozaki, K. & Sugiura, M. (1982) *Gene* **20,** 91–102.
19. Kaiser, E. T. & Kézdy, F. J. (1984) *Science* **223,** 249–255.
20. Terwilliger, T. C., Weissmann, L. & Eisenberg, D. (1982) *Biophys. J.* **37,** 353–361.
21. DeGrado, W. F., Musso, G. F., Lieber, M., Kaiser, E. T. & Kézdy, F. J. (1982) *Biophys. J.* **37,** 329–338.
22. Taylor, J. W., Miller, R. J. & Kaiser, E. T. (1982) *J. Mol. Pharmacol.* **22,** 657–666.
23. Müller, K.-D., Salnikow, F. & Vater, F. (1983) *Biochim. Biophys. Acta* **742,** 78–83.
24. Chua, N.-M. & Schmidt, G. W. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 6110–6114.
25. Garnier, F., Gaye, P., Mercier, F.-C. & Robson, B. (1980) *Biochimie* **62,** 231–239.
26. Emr, S. D. & Silhavy, T. J. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 4599–4603.
27. Briggs, M. S. & Gierasch, L. M. (1984) *Biochemistry* **23,** 3111–3113.