[24] The Use of Structural Profiles and Parametric Sequence Comparison in the Rational Design of Polypeptides

By SÁNDOR PONGOR

'Introduction

"Protein engineering," i.e., the design and construction of novel or mutated proteins, may involve different levels of structural information. In the case of well-studied proteins the design can rely on a wealth of X-ray, NMR, and chemical modification data. On the other hand, mutations often have to be designed in cases when the amino acid sequence is the only structural information available. For example, some proteins are not readily available in quantities sufficient for structural analysis, and some short peptides have no characteristic structures in solution. Mutated proteins, on the other hand, can now be obtained more and more readily due to advances in recombinant DNA techniques, which make it possible to test structural hypotheses in practice. If the goal is the exchange of a single amino acid residue then the problem can be attacked by replacing the residue in question with all the 19 remaining amino acids in parallel mutagenesis experiments. On the other hand, this approach would be clearly too expensive and time consuming if several residues of a region have to be changed. The number of possibilities can be narrowed down if a potentially important structural pattern, such as an amphiphilic α-helix, can be recognized in the given region. In this case the mutations can be rationally designed based on the hypothesis that those amino acid replacements leaving the original structural pattern intact are likely to be accepted. There are two generally used methods available for the recognition of local structures in amino acid sequences: (1) the use of symbolic diagrams such as the Schiffer–Edmundson[1] "helical wheels" or the "helical nets,"[2] and (2) the use of structural profiles.

This chapter briefly describes how the structural profiles are constructed and used. A particular goal of this chapter is to review the numeric and graphic methods developed for the comparative evaluation of structural profiles.[3,4] The advantage of comparing structural profiles rather than primary sequences originates from the simple fact that pro-

[1] M. Schiffer and A. B. Edmundson, *Biophys. J.* 7, 121 (1967).

[2] P. Dunnill, *Biophys. J.* 8, 865 (1968).

[3] S. Pongor and A. A. Szalay, *Proc. Natl. Acad. Sci. U.S.A.* 82, 366 (1985).

[4] S. Pongor, M. J. Guttieri, L. M. Cohen, and A. A. Szalay, *DNA* 4, 319 (1985).

files, unlike primary sequences, can be subjected to arithmetic operations (averaging, subtraction, etc.) and their similarities can be characterized in quantitative terms such as correlation coefficients and standard deviation. This is essentially a *parametric approach of sequence comparison* which makes it possible, e.g., to compare groups of sequences, to carry out a semiquantitative comparison (ranking) of sequences in structural terms, etc. These operations which are not feasible with primary sequences can be easily programed on laboratory microcomputers and allow quantitative characterization of the mutations planned. The molecular biologist can use these techniques in two closely related areas: (1) comparison of mutated or "designed" sequences to their wild-type counterparts; (2) comparison of newly determined amino acid sequences to their known homologs.

Principle and Scope

Structural profiles make it possible to compare wild-type and designed sequences in quantitative terms. Average profiles and standard deviation profiles calculated from homologous sequences allow identification of conserved structural features that can help to formulate a structural hypothesis for the design of engineered analogs. The present chapter contains a compilation of the most important parameter sets used to construct structural profiies from primary sequence data. Numeric and graphic methods designed for the comparative evaluation of structural profiles are described. The methods can be used for the simple quantitative evaluation (ranking) of sequences containing multiple amino acid replacements.

Definitions

The *structural profile* is one of the simplest representations of a protein structure. It is constructed by plotting a structural parameter P_i against the sequential position i. The P_i structural parameters are obtained either statistically, such as the Chou-Fasman parameters,[5] or by physical measurement, such as the various hydrophobicity parameters. (Profile representations of three-dimensional structures are not discussed here).

In the computational sense it is useful to divide the structural profiles into two groups, single residue and multiple residue methods. In the *single residue methods* each of the 20 amino acids is characterized by a constant structural parameter which is independent from the sequence. In this case the profile is constructed simply by plotting these constant val-

[5] P. Y. Chou and G. M. Fasman, *Adv. Enzymol.* 47, 45 (1978).

ues against the sequential position. In the *multiple residue methods* the P_i parameters of the ith amino acid depends both on the ith amino acid and on the sequential neighbors, and is calculated separately for every position in the sequence. For example, the algorithm of Garnier *et al.*[6] uses contributions of the eight preceding and the eight subsequent sequential neighbors for the calculation of the secondary structure propensity parameter of the ith amino acid [Eq. (3)].

For the presentation of the data, the plotted P_i parameters have to be processed by *curve smoothing* that visualizes trends within the profiles by removing dispersion. For our purposes basically any smoothing technique is acceptable, and commercially available softwares usually contain such options. In these procedures each P_i parameter is replaced by a new P'_i value which is an average calculated within a *window*, i.e., a sequential environment of the ith residue. Usually, the *window width* is an optional odd number $(2k + 1)$ and the resulting average value is assigned to the central position of the window. In the simplest case, P'_i can be an arithmetic average [Eq. (1a)] or a weighted arithmetic average [e.g., Eqs. (1b) and (c)]:

$$P'_i = [1/(2k + 1)] \sum_{j=i-k}^{j=i+k} P_j \tag{1a}$$

$$P'_i = [17P_i + 12(P_{i+1} + P_{i-1}) - 3(P_{i+2} + P_{i-2})]/35 \tag{1b}$$

$$P'_i = \{7P_i + 3[2(P_{i-1} + P_{i+1}) + P_{i-2} + P_{i+2}] - 2(P_{i-3} + P_{i+3})\}/21 \tag{1c}$$

More advanced methods can produce continuously differentiable curves.[7-9] While the latter procedures more efficiently remove dispersion than arithmetic averaging [Eq. (1a)], they also distort the values numerically and may give rise to chain-end artifacts at the first and last residues. Repeated use of the same algorithm also improves smoothing efficiency. We note that (1) if the profiles are to be used for prediction, then the optimized procedures recommended by the original papers have to be employed (see below). (2) The unprocessed P_i parameters should be used for numeric and graphic comparison of profiles as well as for the calculation of average and standard deviation profiles.

The different structural parameters (Table I) are usually scaled so as to meet some statistical or other requirement. For better comparison and

[6] F. Garnier, D. J. Osguthorpe, and B. Robson, *J. Mol. Biol.* **120**, 97 (1978).

[7] P. R. Bevington, "Data Reduction and Error Analysis for the Physical Sciences." McGraw-Hill, New York, 1969.

[8] A. Savitzky and M. Golay, *Anal. Chem.* 36, 1627 (1974).

[9] G. D. Rose, *Nature (London)* **272**, 586 (1978).

TABLE I

STRUCTURAL PARAMETERS (SINGLE RESIDUE METHODS)[a]

Amino acid residue	a	b	c	d	e	f	g	h	i	j	k	l	m
Gly	0.57	0.75	1.56	−0.40	0.48	0.62	1.00	0.00	0.00	0.49	3.40	0.00	0
Ala	1.42	0.83	0.74	1.80	0.62	1.56	2.20	−0.50	−4.20	1.81	11.50	0.00	0
Val	1.06	1.70	0.59	4.20	1.08	1.14	2.90	−1.50	−8.40	1.08	21.57	0.13	0
Leu	1.21	1.30	0.50	3.80	1.06	2.93	5.00	−1.80	−10.10	3.23	21.40	0.13	0
Ile	1.08	1.60	0.47	4.50	1.38	1.67	3.30	−1.80	−10.50	1.45	21.40	0.13	0
Ser	0.77	0.75	1.43	−0.80	−0.18	0.81	−0.87	0.30	6.30	0.97	−9.47	1.67	0
Thr	0.83	1.19	0.98	−0.70	−0.05	0.91	0.04	−0.40	3.80	0.84	15.77	1.66	0
Asp	1.01	0.52	1.52	−3.50	−0.90	0.14	−2.10	3.00	31.00	0.05	11.68	49.70	−1
Glu	1.51	0.37	0.95	−3.50	0.74	0.23	−2.30	3.00	24.70	0.11	13.57	49.90	−1
Asn	0.67	0.89	1.46	−3.50	0.78	0.27	−2.40	0.20	12.20	0.23	12.82	3.38	0
Gln	1.11	1.10	0.95	−3.50	−0.85	0.51	−2.40	0.20	10.10	0.72	14.45	3.53	0
Lys	1.16	0.74	1.19	−3.90	−1.50	0.15	−2.40	3.00	17.60	0.06	15.71	49.50	+1
His	1.00	0.87	0.97	−3.20	−0.40	0.29	−2.80	−0.50	14.30	0.31	13.69	51.60	0
Arg	0.98	0.93	1.01	−4.50	−2.53	0.45	−2.40	3.00	47.30	0.20	14.28	52.00	+1
Phe	1.13	1.38	0.66	2.80	1.19	2.03	1.00	−2.50	−14.20	1.96	19.80	0.35	0
Tyr	0.69	1.47	1.14	−1.30	0.26	0.68	−0.79	−2.30	4.70	0.39	18.03	1.61	0
Trp	1.08	1.37	0.60	−0.90	0.81	1.08	−1.10	−3.40	−8.40	0.77	21.67	2.10	0
Cys	0.70	1.19	0.96	2.50	0.29	1.23	−2.00	−1.00	−6.30	1.89	13.46	1.48	0
Met	1.45	1.05	0.60	1.90	0.64	2.96	−0.21	−1.30	−11.30	2.67	16.25	1.43	0
Pro	0.57	0.55	1.56	−1.60	0.12	0.76	−1.60	0.00	13.90	0.76	17.43	1.58	0

[a] Key to column heads: (a) Chou–Fasman helical propensity[5]; (b) Chou–Fasman β-sheet propensity[5]; (c) Chou–Fasman turn propensity[5]; (d) hydrophobicity according to Kyte and Doolittle[16]; (e) consensus hydrophobicity scale according to Eisenberg *et al.*[22]; (f) membrane-buried helical potential according to Argos and Palau[18]; (g) membrane propensity scale according to Kuhn and Leigh[19]; (h) hydrophilicity scale according to Hopp and Woods[22]; (i) signal sequence helical potential scale according to Von Heinje[24] (kJ/mol); (j) signal sequence propensity scale according to Argos and Palau[18]; (k) side-chain bulkiness; (l) polarity; (m) charge.

uniformity, we have normalized the single residue amino acid parameter sets to a mean 0 and a standard deviation 1 (Table II).

Description of Structural Profiles

Secondary Structure Propensity Profiles. The most widely used parameter set, that of Chou and Fasman,[5] is statistically derived from protein structures determined by X-ray crystallography. The parameters represent the propensity of each amino acid to form an α-helix, β-sheet, or turn (Table I). For example, the α-helix propensity parameter of an amino acid a, $P_{a,\alpha}$ is calculated by dividing the frequency of the amino acid a in α-helix, $f_{a,\alpha}$ by the average frequency of residues in the helix conformation:

$$P_{a,\alpha} = f_{a,\alpha}/\langle f_\alpha \rangle \tag{2}$$

In their original paper Chou and Fasman used structural data on 15 proteins for the calculation of the conformational parameters. The Chou–Fasman parameters in Table I were calculated from a database containing 29 proteins. In the original method the profiles are smoothed with a four-residue window, and the average value is assigned to the first position within the window. Levitt[10] has published a Chou–Fasman type parameter set derived from a database of 60 protein structures.

The method of Chou and Fasman is a single residue method, in which neighbor interactions are not included. Robson and co-workers developed the so-called directional information algorithm,[6] a statistically based multiple residue method in which contributions from eight preceding and eight subsequent sequential neighbors are included. For example, the propensity of the ith amino acid to be in an α-helix, $P_{i,\alpha}$ is calculated by the following equation:

$$P_{i,\alpha} = \sum_{i-8}^{i+8} I_{j,\alpha} - D_\alpha \tag{3}$$

where I_j values are the contributions of the neighbors as well as of the ith residue to the propensity value, taken from Table III for α-helix. There are 20 × 17 P values for each of the four conformational states (α-helix, β-sheet, β-turn, and coil, Tables III to VI). D_α is the so-called decision constant which is selected from Table VII if the secondary structure content of the protein is known from independent measurement. For comparative purposes, unbiased profiles ($D_\alpha = 0$, etc.) can be used. Strongly predicted segments have P_i values of several hundred centinats.

[10] M. Levitt, *Biochemistry* **18**, 4277 (1978).

TABLE II

NORMALIZED[a] STRUCTURAL PARAMETERS (SINGLE RESIDUE METHODS)[b]

Amino acid residue	a	b	c	d	e	f	g	h	i	j	k	l
Gly	−1.53	−0.73	1.55	1.37	0.48	−0.48	0.61	0.08	−0.35	−0.56	−2.58	−0.6
Ala	1.49	−0.54	−0.68	1.28	0.62	0.64	1.14	−0.19	−0.61	0.91	−0.84	−0.6
Val	0.21	1.82	−1.09	1.28	1.08	0.14	1.45	−0.71	−0.87	0.09	1.34	−0.6
Leu	0.75	0.74	−1.13	1.36	1.06	2.27	2.38	−0.87	−0.97	2.48	1.30	−0.6
Ile	−0.28	1.55	−1.55	1.32	1.38	0.77	1.63	−0.87	−1.00	0.51	1.30	−0.6
Ser	−0.82	−0.76	1.19	−0.10	−0.18	−0.25	−0.21	0.24	0.04	−0.03	−1.27	−0.5
Thr	−0.60	−0.44	−0.03	−0.08	−0.05	−0.13	0.19	−0.13	−0.11	−0.17	0.09	−0.5
Asp	0.04	−1.33	1.44	−1.27	−0.90	−1.05	−0.75	1.67	1.60	−1.05	−0.80	1.6
Glu	1.81	−1.79	−0.11	−1.12	−0.74	−0.94	−0.84	1.67	1.18	−0.98	−0.39	1.6
Asn	1.17	−0.38	1.28	−1.03	−0.77	−0.89	−0.89	0.19	0.41	−0.85	−0.55	−0.4
Gln	0.39	0.19	−0.09	−0.96	−0.85	−0.61	−0.89	0.19	0.28	−0.31	−0.20	−0.4
Lys	0.57	−0.78	0.54	−0.99	−1.50	−1.04	−0.89	1.67	0.74	−1.04	0.07	1.6
His	0.00	−0.43	−0.11	−1.13	−0.40	−0.87	−1.06	−0.19	0.54	−0.76	−0.36	1.7
Arg	−0.07	−0.27	0.05	−1.28	−2.53	−0.68	−0.89	1.67	2.58	−0.88	−0.24	1.7
Phe	0.46	0.95	−0.90	0.74	1.19	1.20	0.61	−1.24	−1.23	1.07	0.96	−0.6
Tyr	−1.10	1.20	0.40	−0.32	0.26	−0.41	−0.18	−1.14	−0.06	−0.67	0.57	−0.5
Trp	0.28	0.93	−1.07	−0.27	0.81	0.07	−0.31	−1.72	−0.87	−0.25	1.36	−0.5
Cys	−1.07	0.44	−0.09	0.65	0.29	0.25	−0.71	−0.45	−0.74	0.99	−0.41	−0.5
Met	1.60	0.06	−1.07	0.06	0.64	2.31	0.08	−0.61	−0.05	1.86	0.19	−0.5
Pro	−1.53	−1.30	1.55	−0.09	0.12	−0.31	−0.53	0.08	0.51	−0.26	0.44	−0.5

[a] Structural parameters normalized to a mean 0 and a standard deviation 1.

[b] Key to column heads: (a) Chou–Fasman helical propensity[5]; (b) Chou–Fasman β-sheet propensity[5]; (c) Chou–Fasman turn propensity[5]; (d) hydrophobicity according to Kyte and Doolittle[16] (e) consensus hydrophobicity scale according to Eisenberg *et al.*[22]; (f) membrane-buried helica potential according to Argos and Palau[18]; (g) membrane propensity scale according to Kuhn anc Leigh[19]; (h) hydrophilicity scale according to Hopp and Woods[22]; (i) signal sequence helica potential scale according to Von Heinje[24]; (j) signal sequence propensity scale according to Argos and Palau[18]; (k) side-chain bulkiness; (l) polarity.

Both the Chou–Fasman and the Robson profiles were originally designed for secondary structure prediction; for the details of prediction methodology the reader is referred to the original papers. In addition, both methods can be used for the predictive *recognition* of supersecondary structures as suggested by Taylor and Thornton.[11,12] Super–secondary structures are sequentially linked pieces of secondary structures that are in contact in three dimensions. The simple supersecondary structures include the βαβ unit (found in the β/α protein family), the β-hairpin

[11] W. L. Taylor and J. M. Thornton, *Nature (London)* 301, 540 (1983).

[12] W. L. Taylor and J. M. Thornton, *J. Mol. Biol.* 173, 487 (1984).

TABLE III

STRUCTURAL PARAMETERS (DIRECTIONAL INFORMATION MEASURES) FOR THE α-HELICAL CONFORMATION[a]

Amino acid	Residue position																
residue	$j - 8$		$j - 6$		$j - 4$		$j - 2$		j		$j + 2$		$j + 4$		$j + 6$		$j + 8$
Gly	−5	−10	−15	−20	−30	−40	−50	−60	−86	−60	−50	−40	−30	−20	−15	−10	−5
Ala	5	10	15	20	30	40	50	60	65	60	50	40	30	20	15	10	5
Val	0	0	0	0	0	0	5	10	14	10	5	0	0	0	0	0	0
Leu	0	5	10	15	20	25	28	30	32	30	28	25	20	15	10	5	0
Ile	5	10	15	20	25	20	15	10	6	0	−10	−15	−20	−25	−20	−10	−5
Ser	0	−5	−10	−15	−20	−25	−30	−35	−39	−35	−30	−25	−20	−15	−10	−5	0
Thr	0	0	0	−5	−10	−15	−20	−25	−26	−25	−20	−15	−10	−5	0	0	0
Asp	0	−5	−10	−15	−20	−15	−10	0	5	10	15	20	20	20	15	10	5
Glu	0	0	0	0	10	20	60	70	78	78	78	78	78	70	60	40	20
Asn	0	0	0	0	−10	−20	−30	−40	−51	−40	−30	−20	−10	0	0	0	0
Gln	0	0	0	0	5	10	20	20	10	−10	−20	−20	−10	−5	0	0	0
Lys	20	40	50	55	60	60	50	30	23	10	5	0	0	0	0	0	0
His	10	20	30	40	50	50	50	30	12	−20	−10	0	0	0	0	0	0
Arg	0	0	0	0	0	0	0	0	−9	−15	−20	−30	−40	−50	−50	−30	−10
Phe	0	0	0	0	0	5	10	15	16	15	10	5	0	0	0	0	0
Tyr	−5	−10	−15	−20	−25	−30	−35	−40	−45	−40	−35	−30	−25	−20	−15	−10	−5
Trp	−10	−20	−40	−50	−50	−10	0	10	12	10	0	−10	−50	−50	−40	−20	−10
Cys	0	0	0	0	0	0	−5	−10	−13	−10	−5	0	0	0	0	0	0
Met	10	20	25	30	35	40	45	50	53	50	45	40	35	30	25	20	10
Pro	−10	−20	−40	−60	−80	−100	−120	−140	−77	−60	−30	−20	−10	0	0	0	0

[a] According to Garnier *et al.*[6]

TABLE IV

STRUCTURAL PARAMETERS (DIRECTIONAL INFORMATION MEASURES) FOR THE EXTENDED CONFORMATION[a]

Amino acid	Residue position																
residue	$j-8$		$j-6$		$j-4$		$j-2$		j		$j+2$		$j+4$		$j+6$		$j+8$
Gly	10	20	30	40	40	20	0	−20	−42	−20	0	20	40	40	30	20	−10
Ala	0	0	0	0	−5	−10	−15	−20	−23	−20	−15	−10	−5	0	0	0	0
Val	0	0	−10	−20	0	20	40	60	68	60	40	20	0	−20	−10	0	0
Leu	0	0	0	0	0	5	10	20	23	20	10	5	0	0	0	0	0
Ile	0	−10	−20	−10	0	20	40	60	67	60	40	20	0	−10	−20	−10	0
Ser	0	10	20	10	0	−5	−10	−15	−17	−15	−10	−5	0	10	20	10	0
Thr	5	10	15	20	15	15	10	10	13	10	10	15	15	20	15	10	5
Asp	0	5	10	15	20	0	−20	−30	−44	−30	−20	0	0	0	0	0	0
Glu	−10	−15	−20	−25	−30	−35	−40	−45	−50	−55	−60	−60	−50	−40	−30	−20	−10
Asn	10	30	50	30	20	0	−15	−30	−41	−30	−15	0	20	30	50	30	10
Gln	0	0	0	0	0	−5	−10	0	12	20	30	40	50	50	40	30	15
Lys	−5	−10	−15	−20	−30	−40	−50	−40	−33	−20	−10	0	10	10	0	0	0
His	−10	−20	−40	−20	−10	0	−10	−20	−25	−35	−30	−25	−20	−15	−10	−5	0
Arg	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0
Phe	0	0	0	0	0	5	10	20	26	10	−10	−30	−60	−65	−60	−40	−20
Tyr	0	5	10	15	20	25	30	35	40	35	30	25	20	15	10	5	0
Trp	0	0	0	0	0	−10	−10	−10	−10	−10	−10	−15	−20	−25	−30	−20	−10
Cys	0	0	0	0	0	10	20	30	44	30	20	10	0	0	0	0	0
Met	−10	−20	−30	−40	−40	−30	0	10	23	10	0	−30	−40	−40	−30	−20	−10
Pro	10	20	30	30	20	10	0	−10	−18	−20	−10	10	30	40	30	20	10

[a] According to Garnier *et al.*[6]

TABLE V
STRUCTURAL PARAMETERS (DIRECTIONAL INFORMATION MEASURES) FOR THE TURN CONFORMATION[a]

Amino acid	Residue position																
residue	*j* − 8		*j* − 6		*j* − 4		*j* − 2		*j*		*j* + 2		*j* + 4		*j* + 6		*j* + 8
Gly	0	0	0	0	10	30	55	55	57	40	0	0	0	0	0	0	0
Ala	0	0	0	−10	−20	−30	−40	−50	−50	−40	−30	−20	−10	0	0	0	0
Val	0	0	0	0	−10	−20	−30	−40	−60	−40	−30	−20	−10	0	0	0	0
Leu	0	0	0	−10	−20	−30	−40	−50	−56	−20	−10	0	0	0	0	0	0
Ile	0	0	0	0	0	−10	−20	−30	−46	−40	−10	0	0	20	30	20	10
Ser	0	−10	−20	−20	10	15	20	25	26	25	20	15	10	0	0	0	0
Thr	0	10	20	20	20	15	18	5	3	5	10	15	20	20	20	10	0
Asp	0	0	0	0	0	0	5	10	31	10	5	0	0	0	0	0	0
Glu	0	−5	−10	−15	−20	−30	−40	−45	−47	−20	0	10	5	0	0	0	0
Asn	0	0	0	10	20	30	35	40	42	40	35	30	20	10	5	0	0
Gln	10	20	30	25	20	15	10	5	4	20	30	40	50	60	50	40	20
Lys	−10	−20	−30	−40	−25	−10	0	10	10	10	0	−20	−30	−20	−10	5	0
His	0	0	0	0	0	0	0	0	−3	0	10	20	30	20	10	0	0
Arg	0	0	0	0	0	0	0	10	21	30	40	30	20	10	0	0	0
Phe	0	0	0	0	0	−5	−10	−15	−18	−15	0	15	30	25	20	10	0
Tyr	0	0	0	5	15	15	20	25	29	25	20	15	15	5	0	0	0
Trp	0	0	0	10	20	30	40	80	36	−30	30	40	50	60	70	40	20
Cys	20	40	50	60	60	55	50	45	44	40	35	30	25	20	15	10	5
Met	−5	−15	−20	−25	−30	−35	−40	−45	−48	−45	−40	−35	−30	−25	−20	−15	−5
Pro	10	20	30	40	50	70	10	−90	36	90	10	0	0	0	0	0	0

[a] According to Garnier *et al.*[6]

TABLE VI

STRUCTURAL PARAMETERS (DIRECTIONAL INFORMATION MEASURES) FOR THE COIL CONFORMATION[a]

Amino acid residue	Residue position																
	$j-8$		$j-6$		$j-4$		$j-2$		j		$j+2$		$j+4$		$j+6$		$j+8$
Gly	0	0	0	0	10	30	40	45	49	45	40	30	10	0	0	0	0
Ala	0	0	0	0	−5	−10	−20	−25	−25	−25	−20	−15	−10	−5	0	0	0
Val	0	0	0	0	−10	−20	−25	−30	−35	−30	−25	−20	−10	0	0	0	0
Leu	0	0	0	−10	−20	−30	−40	−30	−20	−20	−10	0	0	0	0	0	0
Ile	0	0	0	0	0	−10	−20	−30	−33	−30	−10	0	10	20	30	20	0
Ser	0	−10	−20	−20	10	15	20	25	50	25	20	15	10	0	0	0	0
Thr	0	10	20	30	20	15	10	15	17	15	10	15	20	30	20	10	0
Asp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Glu	0	0	10	20	40	20	0	−10	−44	−40	−20	−10	0	0	0	0	0
Asn	0	0	0	10	20	30	35	40	46	40	35	30	20	10	0	0	0
Gln	10	20	30	25	20	15	10	0	−5	20	30	40	50	60	50	40	20
Lys	−10	−20	−30	−40	−25	−20	−10	−8	−8	0	0	−20	−30	−20	−10	−5	0
His	0	0	0	0	0	0	0	10	16	15	10	10	10	10	5	0	0
Arg	0	0	0	0	0	0	0	0	−12	0	20	30	20	10	0	0	0
Phe	0	0	0	0	0	−5	−10	−20	−41	−20	0	15	30	25	20	10	0
Tyr	0	0	0	0	0	0	0	0	−6	0	0	0	0	0	0	0	0
Trp	0	0	0	10	20	30	40	20	12	20	30	40	50	60	70	40	20
Cys	0	0	0	0	0	0	−10	−30	−47	−30	−10	0	0	0	0	0	0
Met	0	−5	−10	−15	−20	−25	−30	−40	−41	−40	−30	−25	−20	−15	−10	−5	0
Pro	0	0	10	20	30	40	50	55	58	50	10	0	0	0	0	0	0

[a] According to Garnier *et al.*[6]

TABLE VII
DECISION CONSTANTS RELATED TO THE SECONDARY STRUCTURE CONTENT OF THE PROTEIN[a]

% Secondary structure (α-helix or extended)	Decision constants[b]	
	α-Helix, D_α	Extended, D_β
I. Less than 20%	158	50
II. Between 20 and 50%	−75	−87.5
III. Over 50%	−100	−87.5

[a] According to Garnier *et al.*[6]

[b] These values have to be substituted into Eq. (3). Decision constants for the turn and coil conformations are equal to 0.

(β-turn-β or $\beta \times \beta$, frequently found in the all-β family), and the $\alpha \times \alpha$ hairpin. An examination of 31 proteins showed that two-thirds of the identified pieces of secondary structures were parts of such units.[13] The principle is illustrated in Fig. 1. For example, the $\beta\alpha\beta$ structure can be recognized from a peak in the α-helix propensity profile flanked on both sides by peaks in the β-sheet profile. Naturally, not all supersecondary structures may be clearly detectable from structural profiles; their importance lies in the fact that they represent interpretation of predictive data (structural profiles) at a higher level of complexity (supersecondary structures).

Membrane Preference Profiles. Transmembrane segments in proteins are characterized by a large number of nonpolar residues. Based on the experimental data of Nazaki and Tanford[14] and Wolfenden *et al.*,[15] Kyte and Doolittle[16] developed a hydropathy scale that characterizes the hydrophobicity of the amino acid side chains. Nonpolar side chains have large positive hydropathy parameters (Ile = 4.50), and polar ones are assigned negative values (Arg = −4.50). Hydropathy profiles constructed with these parameters and smoothed with a 19-residue window [Eq. (1), $k = 9$], will display large positive peaks at membrane-bound segments. Proteins located within the membranes are frequently characterized by a series of transmembrane regions that appear in the graphs as positive

[13] M. Levitt and C. Chotia, *Nature* (*London*) **261**, 552 (1976).

[14] T. Nozaki and C. Tanford, *J. Biol. Chem.* **246**, 2211 (1971).

[15] R. Wolfenden, L. Anderson, P. M. Cullis, and C. C. B. Southgate, *Biochemistry* **20**, 849.

[16] J. Kyte and R. F. Doolittle, *J. Mol. Biol.* **157**, 105 (1982).

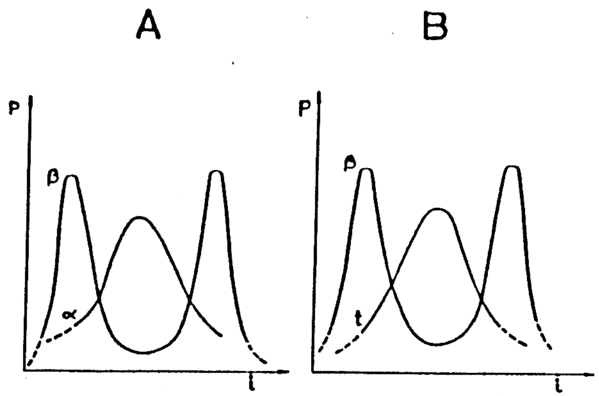


FIG. 1. Schematic structural profiles representing two simple supersecondary structures, the $\beta\alpha\beta$ (A) and the β-turn-β units (B). A conserved β-turn-β unit can be tentatively identified in Fig. 2 at the C-terminus of the ribulose 1,5-bisphosphate carboxylase small subunit polypeptide.

peaks interspersed by negative valleys corresponding to regions located outside the membrane.

Several other hydrophobicity scales were developed. Tables I and II include that of Eisenberg *et al.*,[17] the "consensus hydrophobicity scale," which incorporates features of several other scales. We note that in our applications differences between the various scales are negligible and that the scale of Kyte and Doolittle[16] seems to have the widest acceptance by molecular biologists.

Transmembrane segments are frequently helical. Argos and Palau developed a membrane-buried helix propensity scale, using statistical data on 1125 residues.[18] Kuhn and Leigh developed another statistical membrane propensity scale, using structural data on 24 known transmembrane segments.[19] The parameters range from −2.8 (His) to 5.0 (Leu) and are scaled so that a protein of average composition will have an average membrane propensity of 0. The profiles are smoothed with a running average algorithm [Eq. (1a)], using a window width of 7. Eisenberg *et al.* used a combination of hydrophobic moment profiles and hydrophobicity profiles for the identification of transmembrane helices[20–23] (see below).

[17] D. Eisenberg, R. M. Weiss, T. C. Terwilliger, and W. Wilcox, *Faraday Symp. Chem. Soc.* **17**, 109.

[18] P. Argos and J. Palau, *Int. J. Pept. Protein Res.* **19**, 380 (1982).

[19] L. Kuhn and J. S. Leigh, *Biochim. Biophys. Acta* **828**, 351 (1985).

[20] D. Eisenberg, E. Schwarz, M. Komaromy, and R. Wall, *J. Mol. Biol.* **179**, 125 (1984).

[21] D. Eisenberg, *Annu. Rev. Biochem.* **53**, 595 (1984).

[22] D. Eisenberg, R. M. Weiss, and T. C. Terwilliger, *Nature (London)* **299**, 371 (1982).

[23] D. Eisenberg, R. M. Weiss, and T. C. Terwilliger, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 140 (1984).

Signal Sequence Helical Potential Profiles. Signal sequences are characterized by a stretch of nonpolar residues of high helix-forming propensity, and their amino acid composition significantly differs from the average protein composition. Von Heinje developed a helix hydrophobicity scale based on the free energy required to transfer a membrane-bound helix into solution.[24] The parameters range from −14.42 kJ/mol for Phe to 47.3 kJ/mol for Arg. In the profiles the helical part of the signal sequence displays a negative trough. Argos *et al.* developed a signal sequence helical propensity parameter set from a 705-residue database compiled from signal sequences using Eq. (2).[25]

The Hydrophilicity Profile: Prediction of Antigenic Determinants. Antigenic determinants of proteins are located on the surface of the molecule and are frequently found in regions highly exposed to the solvent. This, together with the fact that charged, hydrophilic amino acids are common features of antigenic determinants led Hopp and Woods to use a hydrophilicity profile to locate antigenic determinants in a number of proteins.[26] Their hydrophilicity parameters are derived from the solvent parameters of Levitt.[27] These parameters are based on the same experimental data as the hydrophobicity scales and indeed contain the same information. In this scale hydrophilic amino acids have large positive parameters (Arg = 3.0) whereas hydrophobic residues have negative ones (Trp = −3.4) (Table III). The hydrophilicity profiles are smoothed by simple arithmetic averaging using a window width of 6 residues, and the tentative antigenic regions are identified as those displaying the highest positive peaks. As peptide chain turns are often hydrophilic,[8] this technique is likely to select turns as potential epitopes.

The Hydrophobic Moment Profile. Amphiphilic structures in proteins and peptides are those in which hydrophobic and hydrophilic side chains are separated on opposite sides. Amphiphilic structures are "surface seeking"; i.e., they bind to membrane–solution surface boundaries. Amphiphilic segments of soluble proteins are frequently located on the protein surface with the hydrophilic side chains pointing toward the solution. Amphiphilic structures in peptides can be easily recognized by diagramatic representations such as the Schiffer–Edmundson wheel diagrams[1] (see Chap. [25], this volume). Graphic analysis of long sequences would be cumbersome, however, so Eisenberg and co-workers developed a numeric method.[20-23] The method is based on the concept of *hydrophobic*

[24] G. Von Heinje, *Eur. J. Biochem.* **116**, 419 (1981).

[25] P. Argos, J. K. Mohana Rao, and P. A. Hargrave, *Eur. J. Biochem.* **128**, 565 (1982).

[26] T. Hopp and K. R. Woods, *Proc. Natl. Acad. Sci. U.S.A.* **78**, 3824.

[27] M. Levitt, *J. Mol. Biol.* **104**, 59 (1976).

moment, a numeric index of amphiphilicity defined by the following equation:

$$\mu_H = \left(\left[\sum_{j=1}^{n} H_j \sin(\delta j)\right]^2 + \left[\sum_{j=1}^{n} H_j \cos(\delta j)\right]^2\right)^{1/2} \tag{4}$$

H_j is the hydrophobicity (Table I, column e) of the jth amino acid, δ the periodicity parameter (100° for α-helix and 180° for β-sheet, expressed in radians), and n is the total number of residues. μ_H is a Fourier transform type expression that has large values if residues of adverse hydrophobicity are clustered on opposite sides of the helix. Membrane-seeking amphiphilic helices can be located in protein sequences according to their hydrophobic moment μ_H and average hydrophobicity, $\langle H\rangle$. μ_H of amphiphilic helices is greater than 0.47 and $\langle H\rangle$ is in the range of -0.22 to 0.34. Transmembrane regions on the other hand have greater hydrophobicity values ($\langle H\rangle \geq 0.68$ for a 21-residue segment or ≥ 1.10 for two adjacent segments) and lower hydrophobic moments ($\mu_H < 0.3$).[21]

For the construction of the hydrophobic moment profile, the value of μ_H is calculated for a window of 21 residues, and its value is plotted at the central position of the window:

$$\mu_{H,i} = \left(\left[\sum_{j=i-10}^{i+10} H_j \sin(\delta j)\right]^2 + \left[\sum_{j=i-10}^{i+10} H_j \cos(\delta j)\right]^2\right)^{1/2} \tag{5}$$

The profile calculated with $\delta = 100°$ gives maxima at amphiphilic helices; amphiphilic β-sheets give peaks at $\delta = 180°$. A simple analysis of amphiphilicity can be carried out by plotting hydrophobicity (smoothed by arithmetic averaging over a 21-residue window) and the hydrophobic moments for α-helix ($\delta = 100°$) and β-sheet ($\delta = 180°$) in the same coordinate system. Finer-Moore and Stroud introduced a two-dimensional hydrophobic moment plot in which the sequential position i and δ are the two dimensions.[28] In this representation the amphiphilic segments appear as "hills" at their respective δ values.

Physical Parameters. Basically any physical parameter of the amino acids could be used to construct a structural profile. The available softwares usually contain three of these, *electric charge*, *side-chain bulkiness*, and the *polarity index* (Tables I and II). The bulkiness parameter is defined as the ratio of side-chain volume to length (in Å^2).[29] The polarity index is proportional to the electric force (originating from charge and dipole effects) due to an amino acid side chain acting on its immediate

[28] J. Finer-Moore and R. M. Stroud, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 155 (1984).

[29] J. M. Zimmerman, N. Eliezer, and R. Simha, *J. Theor. Biol.* **21**, 170 (1968).

surroundings.[29] These parameter sets can be useful if a newly derived sequence is compared to its known homologs.

Average Profiles and Standard Deviation Profiles: Identification of Structurally Conserved and Variable Regions[4]

Average profiles can be used to identify structurally conserved regions in homologous sequences such as members of a gene family. This feature is especially useful when the degree of primary sequence homology is low between the proteins in question. For the calculation of an average profile, the homologous sequences have to be aligned for maximum homology. The average profile of n aligned sequences is defined as

$$\langle P_i \rangle = \left(\sum_{j=1}^{n} P_{i,j} \right) \Big/ n \tag{6}$$

The structural variability that exists among the sequences at a given sequential position can be characterized by standard deviation of the mean $\langle P_i \rangle$:

$$\sigma_i = \left[\sum P_{i,j}{}^2 - \left(\sum P_{i,j} \right)^2 \Big/ n \right]^{1/2} / (n - 1) \tag{7}$$

The number n is not constant throughout the sequence; it includes only those sequences in which the P_i value is not missing because of an alignment gap. The $\langle P_i \rangle$ and the σ_i profiles, defined by Eqs. (6) and (7), respectively, can be shown in the same coordinate system by one of the following methods: (1) The σ_i profile is drawn symmetric to 0 as a shaded area ($\pm\sigma_{i\,\text{profile}}$). This makes it easy to identify regions where the $\langle P_i \rangle$ profile is significantly different from 0 (i.e., it is outside the shaded area, see Fig. 2). (2) The σ_i profile can be added symmetrically to the $\langle P_i \rangle$ profile as a shaded area ($\langle P_i \rangle \pm \sigma_{i\,\text{profile}}$). This makes it possible to compare an additional sequence to those included in the average. For example, a structural profile of a newly determined sequence can be visually compared to a group of homologous sequences (presented as an average ± standard deviation profile) and segments identified where the new sequence is expected to be different from those already known (i.e., the profile is outside the shaded area). Since the secondary structures are stabilized by short-range interactions, regions displaying no conserved structural propensity ($P_i < \sigma_i$) may be those stabilized by long-range interactions. These regions are likely to "accept" modifications without affecting biological activity of the protein.

The use of average profiles is illustrated on the ribulose 1,5-bisphosphate carboxylase small subunit (Rubisco SSU) gene family.[4] Figure 2 shows averaged α-helix, β-sheet, and turn propensity profiles calculated

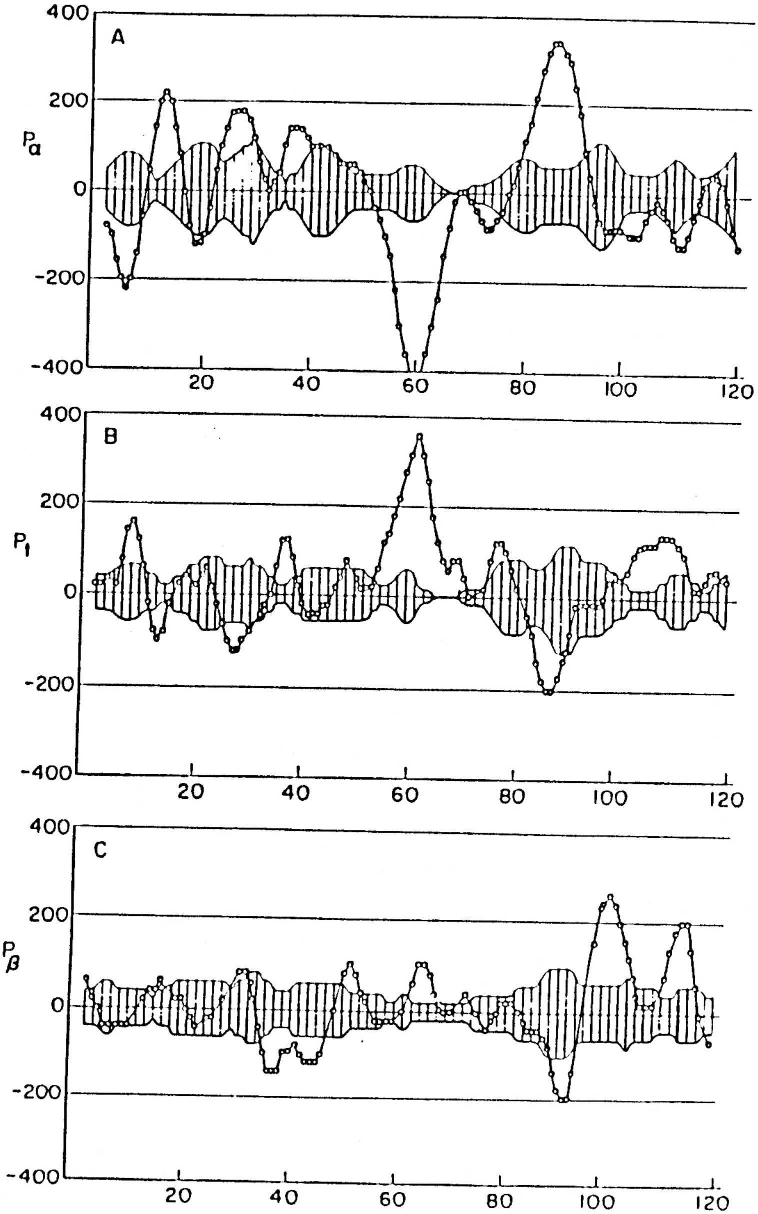


FIG. 2. Averaged α-helix (A), turn (B), and β-structure (C) propensity profiles calculated from ribulose 1,5-bisphosphate carboxylase small subunit sequences of five higher plants (tobacco, spinach, petunia, soybean, and pea).[4] Shaded areas indicate the standard deviation ($\pm\sigma_i$) of the respective structural parameter.

for six ribulose 1,5-bisphosphate carboxylase small subunit proteins from different higher plants. It follows from the probabilistic definition of the P_i values that only those regions for which $P_i > \sigma_i$ (i.e., the structure forming propensity is substantially different from 0) can be considered conserved in the structural sense. It appears that there are only a few regions in which this condition is fulfilled; the rest of the protein has no characteristic 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. For example, a conserved ($P_i > \sigma_i$) β-turn-β supersecondary structure can be tentatively identified in the C-terminal region. Second, external factors, such as the large subunit of the enzyme, or long-range interactions may contribute to the formation of the native conformation, so there may be no rigid selectional constraints on the secondary structure forming propensity of the small subunit. Region 54–66 displays a pronounced β-turn forming propensity which is conserved in all the six sequences. However, this region is entirely missing in two homologous proteins isolated from prokaryotic organisms (the cyanobacteria *Synechococcus* R2 and *Anabaena variabilis*). In the higher plant genes this region is located near an intron–exon junction. A comparison of 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 these proteins.[10] These "variable surface loop" structures were implied to account for functional differences among prokaryotic and eukaryotic members of a gene family. The strong turn forming potential of region 54–66 is in fact consistent with a surface loop structure, which may explain functional differences between higher plant and cyanobacterial Rubisco SSUs.

Comparison of Structural Profiles[3,4]

Homology of proteins has been studied at various levels of structural organization. Comparison of gene sequences at the DNA or amino acid level provides quantitative measures of homology which is essential for establishing phylogenetic relationships. Comparison of three-dimensional structures makes it possible to reveal similarities between proteins that are not demonstrably homologous in their primary structure. Comparison of structural profiles is essentially a parametric approach to the analysis of sequence homology that has several advantages over the comparison of

[10] C. S. Craik, W. J. Rutter, and R. Fletterick, *Science* **220**, 1125 (1983).

[11] R. M. Sweet and D. Eisenberg, *J. Mol. Biol.* **171**, 479 (1983).

primary sequences. First, it makes it possible to describe sequence differences in terms of protein structure. For example one can rank homologous sequences according to a structural parameter (average helicity, hydrophobicity, hydrophobic moment, etc.) and then correlate biological activity with the structural parameter in question. Second, the use of average and standard deviation profiles makes it possible to compare groups of sequences, which is not possible through primary sequence comparison. For example, one can answer questions whether a newly determined protein sequence is within the range of the known homologous sequences in terms of secondary structure, hydrophobicity, etc. Also, the expected structural differences can be located and characterized. Third, the numeric indices described below allow optimization of sequence alignment in equivocal cases.

Before structural profiles are compared, the two sequences have to be aligned for maximum homology by introducing gaps at appropriate positions. Similarly, gaps should be introduced into the calculated structural profiles at the same positions so as to maintain homology. As a rule, positions where one of the residues is missing due to an alignment gap are omitted from the further calculations.

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

$$\Delta P_i = P_{i,2} - P_{i,1} \tag{8}$$

and plotted against the sequential position i. 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. The difference profiles are usually scattered graphs that need to be smoothed for data presentation. Quantitative evaluation of ΔP_i is meaningful only on a comparative basis. Such an evaluation is possible if a profile ($P_{i,2}$) is compared to an average profile ($P_{i,1}$) calculated from known homologs with a standard deviation σ_i. In this case σ_i can be considered as a measure of "acceptable" variability and ΔP_i should substantially exceed σ_i for structural differences to be predicted in a given region.

The difference plots shown in Fig. 3 were calculated between α-helix and β-sheet propensity profiles of Rubisco SSU from the prokaryote *Synechococcus* R2 ($P_{i,1}$), with the average of six higher plant Rubisco SSU proteins ($P_{i,2}$).[3] The figure shows the N-terminal part of the ΔP_i profile and the $\pm\sigma_i$ profile which characterizes the variability of the plant sequences. In this region, the *Synechococcus* R2 protein seems to have a higher α-helix forming potential ($\Delta P_{i,\alpha} > 0$) whereas the higher plant proteins

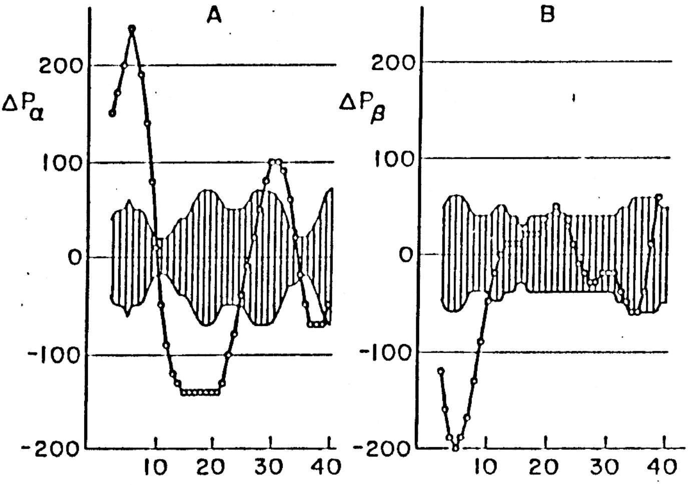


FIG. 3. α-Helix and β-structure propensity difference profiles comparing the N-terminal region of the *Synechococcus* R2 ribulose 1,5-bisphosphate carboxylase small subunit [$P_{i,2}$ in Eq. (8)] to the average profile [$P_{i,1}$ in Eq. (8)] of five membrane-translocated ribulose 1,5-bisphosphate carboxylase small subunit sequences (from tobacco, spinach, petunia, soybean, and pea).

seem to have a greater tendency to form β-structures at the N-terminus ($\Delta P_{i,\beta} < 0$). Rubisco SSU is a membrane-translocated protein in higher plants that is synthesized with a transit peptide whereas its prokaryotic counterparts are cytoplasmic proteins. The finding is thus in accordance with the known propensity of membrane-translocated proteins to form β-structures at their N-termini.

Numeric Comparison. The concept of using *correlation coefficients* to compare structural profiles was developed overall similarities in hydrophobicity[30] and secondary structure.[3,4] With the terminology introduced at Eq. (8) the correlation coefficient calculated between two structural profiles can be written as follows:

$$R = \frac{\sum P_{i,1}P_{i,2}}{\left(\sum P_{i,1}{}^2 P_{i,2}{}^2\right)^{1/2}} \tag{9}$$

R is equal to 1 for two identical sequences, R approximates for two unrelated sequences, and R approaches -1 if the structural profiles are anticorrelated. The z-test of Fisher can be used as a test of significance,[32] i.e., to establish if R is significantly different from 0. To meet the criteria of the z-test, the structural parameters were normalized to mean 0 and

[32] C. M. Goulden, "Methods in Statistical Analysis," 2nd Ed., p. 122. Wiley, New York, 1952.

standard deviation 1. The rescaled structural parameters are summarized in Table I. If the original (not normalized) structural parameters are used to calculate correlation coefficients, the results can be erroneous. For example, the average of the Chou–Fasman parameters is 1 (i.e., >0), so sequences of similar composition are likely to give high correlation coefficients in the absence of any structural similarity if these parameters are used for the calculation of Eq. (9) without normalization. Several authors use the following formula for calculating correlation coefficients from nonnormalized parameters:

$$R = \frac{\Sigma\,([P_{i,1} - \langle P_{i,1}\rangle][P_{i,2} - \langle P_{i,2}\rangle])}{\left(\Sigma\,[P_{i,1} - \langle P_{i,1}\rangle]^2[P_{i,2} - \langle P_{i,2}\rangle]^2\right)^{1/2}} \tag{10}$$

where $\langle P_{i,1}\rangle$ and $\langle P_{i,2}\rangle$ are mean values calculated in the sequential region included in the summation.

The correlation coefficient can be used as an indicator of structural similarity. The procedure is illustrated on synthetic peptide analogs that have known differences in secondary structure as compared to their natural counterparts.[33] The synthetic peptides were designed so as to have the following characteristics relative to the corresponding natural peptides: (1) equal or increased helix content; (2) conserved charged residues and hydrophobic/hydrophilic balance; (3) low (20–50%) primary sequence homology.

Melittin is a 26-residue hemolytic peptide with a segment that has the potential to form an amphiphilic helix. When the sequence of the amphiphilic segment was rearranged according to the above criteria, a biologically active analog was obtained that was higher in α-helix content (35%) than the native melittin (18%) as determined by circular dichroism.[34] Structural correlation values listed in Table VIII reflect the differences in α-helix formation: The value of the α-helix correlation R_α is the lowest among all values compared. The correlation coefficient can also be used to rank sequences according to their expected structural similarity. For example, analogs of β-endorphin representing a range of α-helix content were correctly ranked using R_α alone.[35] The correlation coefficient is an index of variability and gives no information about the direction of the difference (this can be established from a difference plot). Finally we mention that the structural correlation coefficient can also be used to describe evolutionary changes in protein secondary structure.[3]

There are two other numeric indices that allow quick evaluation of

[33] E. T. Kaiser and F. J. Kézdy, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 1137.

[34] W. F. DeGrado, G. F. Musso, M. Lieber, and E. T. Kaiser, *Biophys. J.* **37**, 329.

[35] J. W. Taylor, R. J. Miller, and E. T. Kaiser, *J. Mol. Pharmacol.* **22**, 657.

TABLE VIII
STRUCTURAL CORRELATION CALCULATED BETWEEN MELITTIN AND ITS SYNTHETIC ANALOG[a]

Structural profile	R^b
α-Helix propensity (Garnier *et al.*[6])	0.11
Extended structure (Garnier *et al.*[6])	0.89
β-Turn (Garnier *et al.*[6])	0.81
Coil (Garnier *et al.*[6])	0.54
α-Helix (Chou–Fasman[5]	0.18
β-Sheet (Chou–Fasman[5])	0.79
β-Turn (Chou–Fasman[5])	0.53
Hydrophobicity (Kyte–Doolittle[16])	0.94
Hydrophilicity (Hopp–Woods[22])	0.90
Charges	0.89

[a] Sequences[34]: Melittin, 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 L Q W L K R K R Q Q.

[b] R values given were calculated by using Eq. (9) (P_1, melittin; P_2, synthetic analog). R values above 0.40 are significant at the 99.99% level.

multiple amino acid replacements in shorter segments: (1) $\langle \Delta P_i \rangle$ is the average difference (sum of the ΔP_i values calculated between the two sequences divided by the number of residues in the segment and (2) $\langle |\Delta P_i| \rangle$, the average absolute difference $\langle \Delta P_i \rangle$ is the average value of the difference profile, i.e., it contains the same information. $\langle |\Delta P_i| \rangle$ is an indicator of variability. The use of these indices is illustrated in Ref. 4.

Computer Programs

Sequence management programs used in most molecular biology laboratories (for reviews, see Ref. 36 and F. Lewitter and W. Rindone, Vol. 155 [36]) are generally capable of drawing structural profiles from primary sequence data, even though they may not contain all parameter sets summarized in this chapter. Averaging of profiles is feasible in some computer programs (e.g., Ref. 37), but the features crucial for engineering applications (standard deviation profiles and the numeric indices of structural

[16] L. J. Korn and C. Queen, *DNA* 3, 421 (1984).

[17] J. Novotny and C. Auffray, *Nucleic Acids Res.* **12**, 243 (1984).

comparison) are not standard options in the generally used molecular biology softwares. These capabilities can be either incorporated into separate programs (such as the protein sequence evaluation package developed by the author[4]), or integrated into available sequence management programs. The design of the programs should make it possible to carry out any combination of the numeric and graphic procedures outlined above. It is also recommended that segments of sequences could be separately analyzed. Curve smoothing and graphic output can be performed by commercially available microcomputer software.

Comments: Evaluation of the Results

Since engineering of a protein segment can have a variety of goals (e.g., reinforcement or disruption of structural elements, removal of sites for enzymatic or chemical attack), there are no generally applicable rules for the evaluation of the results. In most cases the problem itself defines the solution, and the experimenter can simply select the most promising replacement alternatives, using the numeric and graphic methods outlined here. General guidelines can be given, however, for two experimental situations: (1) The sequence to be redesigned has no known homologs. In this case it is recommended to carry out secondary structure prediction and to draw helical wheel and helical net diagrams for the segment of interest. On this basis one might be able to recognize a structural motif (e.g., Fig. 1) which makes it possible to formulate a structural hypothesis for amino acid replacements. Then the sequences containing the proposed alterations can be ranked in computer experiments using the methods presented here. (2) If the sequence to be redesigned has several known homologs (with differences in their primary sequence) then average profiles and standard deviation profiles can be constructed with the parameters thought to be important in protein folding, i.e., secondary structure propensities, hydrophobicity, polarity, and side-chain bulkiness. Through visual comparison of the profiles it can then be established which of these properties is the most conserved within the group. (For the comparison between the parameters, the normalized parameter sets summarized in Table II have to be used.)

Additional information can be obtained from the predicted secondary structure of the individual sequences as well as from averaged predictions.[6,36] As an example, Fig. 4A shows the primary sequence of transit peptides that mediate passage into the chloroplasts of proteins synthesized in the cytoplasm. Although chloroplasts of higher plants reciprocally recognize each other's transit peptides, the primary sequences show little homology. On the other hand, the predicted secondary structures

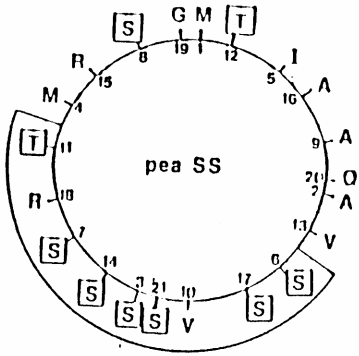
A

```
                    -50        -40        -30        -20        -10        -1  1
Chl. AB80                              MAASSSSSMALSSPTLAGKOLKLNPSSQELCAARFTMRK  SATTK
Soybean SS    MASSMISSPAVTTVNRAGAGMVAPFTGLKSMAGFPIRKTNNDITSIASNGGRVQC  MQVWP
Pea SS3.6   MASMISSSAVTTVSRASRGQSAAVAPFGGLKSMTGFPVKKVNTDITSITSNGCRVKC  MQVWP
Wheat SSW9                   MAPAVMASSATTVAPFQGLKSIACLPISCRSGSTGLSSVSNGGRIRC  MQVWP
```

B

```
                    -50        -40        -30        -20        -10        -1  1
Chl. AB80                              hhhhcccEEEEccccEEEEEhhtcccthhhhhhhhhhhh  hhhhh
Soybean SS    hhhhecccEEEEEEEEcccchEEEEEccctcEEEEEcctctttEEEEEcctttttEE  EEEtt
Pea SS3.6   hhhhhcccEEEEEEEEcctchhhhEEEcctctcEEEEEEEEtcttEEEEEEccttttEE  Ettcc
Wheat SSW9                   hhhhhhhhhhccEEEEctttcccttEEEEttttttEEEEEccttttEE  EEhtt
```

FIG. 4. Amino acid sequence (A) and predicted secondary structure (B) of transit peptides that mediate the posttranslational transport of precursor proteins into the chloroplast.[4] Chl AB80, chlorophyll *a/b* binding polypeptide; SS, ribulose 1,5-bisphosphate carboxylase small subunit transit peptide. Method was by Garnier *et al.* (unbiased prediction).[6]



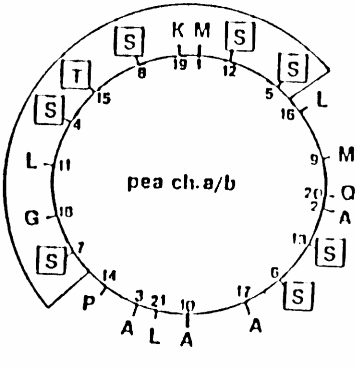


FIG. 5. Helical wheel representation of the N-termini of two chloroplast transit peptides. The hydroxyamino acids are boxed (abbreviations are as in Fig. 4).

show a distinct similarity: they all contain an α-helix at the N-terminus followed by a repetitive pattern of β-strands (Fig. 4B). In addition, the helical wheel diagrams shown in Fig. 5 show that the hydroxyamino acids, serine and threonine, seem to form asymmetric clusters on the N-terminal helices. Naturally, these and similar findings are only hypothetical and need experimental confirmation. On the other hand, they can provide a rational framework for experimental design and thereby reduce the number of directed mutagenesis experiments necessary for obtaining biologically active protein analogs.

[25] Structure–Function Analysis of Proteins through the Design, Synthesis, and Study of Peptide Models

By JOHN W. TAYLOR and E. T. KAISER

In recent years, a large number of intermediate-sized biologically active peptides have been identified and their amino acid sequences determined. Unless they contain multiple disulfide linkages, these peptides usually do not form very stable secondary or tertiary structures in aqueous solution. However, they often have the potential to form segments of amphiphilic secondary structures that might be induced on binding to a matching amphiphilic environment provided by the biological interfaces at which their activities are expressed.[1,2] Understanding structure–function relationships in such peptides requires the identification of these potential structures and rigorous testing of their importance.

A variety of peptides, including serum apolipoproteins,[3,4] toxins,[5-7] chemotactic factors,[8] and peptide hormones,[9-12] contain segments of their

[1] E. T. Kaiser and F. J. Kezdy, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 1137 (1983).

[2] E. T. Kaiser and F. J. Kezdy, *Science* **223**, 249 (1984).

[3] W. M. Fitch, *Genetics* **86**, 623 (1977).

[4] A. D. McLachlan, *Nature (London)* **267**, 465 (1977).

[5] W. F. DeGrado, F. J. Kezdy, and E. T. Kaiser, *J. Am. Chem. Soc.* **103**, 679 (1981).

[6] A. Argiolas and J. J. Pisano, *J. Biol. Chem.* **260**, 1437 (1985).

[7] D. Andreu, R. B. Merrifield, H. Steiner, and H. G. Boman, *Biochemistry* **24**, 1683 (1985).

[8] D. G. Osterman, G. F. Griffin, R. M. Senior, E. T. Kaiser, and T. F. Deuel, *Biochem. Biophys. Res. Commun.* **107**, 130 (1982).

[9] J. W. Taylor, D. G. Osterman, R. J. Miller, and E. T. Kaiser, *J. Am. Chem. Soc.* **103**, 6965 (1981).

[10] G. R. Moe, R. J. Miller, and E. T. Kaiser, *J. Am. Chem. Soc.* **105**, 4100 (1983).

[11] S. H. Lau, J. Rivier, W. Vale, E. T. Kaiser, and R. J. Kezdy, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 7070 (1983).