Protein Science (1999), *8*:370–380. Cambridge University Press. Printed in the USA.

Estimation of the number of α-helical and β-strand segments in proteins using circular dichroism spectroscopy

NARASIMHA SREERAMA,[1] SERGEI YU. VENYAMINOV,[2] AND ROBERT W. WOODY[1]

[1]Department of Biochemistry and Molecular Biology, Colorado State University, Fort Collins, Colorado 80523
[2]Department of Pharmacology, Mayo Foundation, Rochester, Minnesota 55905

(RECEIVED July 15, 1998; ACCEPTED October 28, 1998)

Abstract

A simple approach to estimate the number of α-helical and β-strand segments from protein circular dichroism spectra is described. The α-helix and β-sheet conformations in globular protein structures, assigned by DSSP and STRIDE algorithms, were divided into regular and distorted fractions by considering a certain number of terminal residues in a given α-helix or β-strand segment to be distorted. The resulting secondary structure fractions for 29 reference proteins were used in the analyses of circular dichroism spectra by the SELCON method. From the performance indices of the analyses, we determined that, on an average, four residues per α-helix and two residues per β-strand may be considered distorted in proteins. The number of α-helical and β-strand segments and their average length in a given protein were estimated from the fraction of distorted α-helix and β-strand conformations determined from the analysis of circular dichroism spectra. The statistical test for the reference protein set shows the high reliability of such a classification of protein secondary structure. The method was used to analyze the circular dichroism spectra of four additional proteins and the predicted structural characteristics agree with the crystal structure data.

Keywords: circular dichroism; distorted β-strand; distorted α-helix; helix and strand segments; protein secondary structure

Circular dichroism (CD) spectroscopy is a widely used technique for studying protein and nucleic acid conformations. Over the last three decades, various methods have been developed for the analysis of protein CD spectra based upon the spectral characteristics of protein secondary structures (Greenfield & Fasman, 1969; Chen & Yang, 1971; Bolotina et al., 1980; Brahms & Brahms, 1980; Hennessey & Johnson, 1981; Provencher & Glöckner, 1981; Manavalan & Johnson, 1987; Shubin et al., 1990; van Stokkum et al., 1990; Pancoska et al., 1991; Perczel et al., 1991; Böhm et al., 1992; Sreerama & Woody, 1993, 1994a). In these methods, spectra of either model polypeptides or of a set of reference proteins with known crystal structure are used, and the CD spectrum of a given protein is treated as a linear combination of component secondary structure spectra. For the set of proteins considered, the CD spectra and the secondary structure fractions form either a set of linear equations that is solved by least-squares–based methods or a pattern that is analyzed by pattern recognition methods, and the secondary structure content corresponding to a given CD spectrum is determined. These have been reviewed recently by Venyaminov and Yang (1996) and by Greenfield (1996). Similar methods have also been used in the analyses of infrared (IR) (Kalnin et al., 1990), Raman (Williams, 1983), and vibrational CD (VCD) (Pancoska et al., 1991) spectra of proteins. The information derived from these analyses have been largely limited to the estimation of fractional content of α-helix, β-sheet, β-turns, and (in one case) poly(Pro)II structures in proteins.

The secondary structure fractions for the proteins in the reference set are determined from the corresponding crystal structures. Various algorithms have been developed for assigning the secondary structures in proteins using the coordinates determined from X-ray crystallography, making use of geometric features of the secondary structures (Levitt & Greer, 1977; Kabsch & Sander, 1983; Sklenar et al., 1989; Frishman & Argos, 1995). Among

Reprint requests to: Robert W. Woody, Department of Biochemistry and Molecular Biology, Colorado State University, Fort Collins, Colorado 80523; e-mail: rww@lamar.colostate.edu.

Abbreviations: α, α-helix; α_R, regular α-helix; α_D, distorted α-helix; β, β-strand; β_D, distorted β-strand; β_R, regular β-strand; δ, RMS deviation; CD, circular dichroism; f_x, fractional content of secondary structure X where $X = \alpha, \alpha_R, \alpha_D, \beta, \beta_R, \beta_D, T$, and U; FTIR, Fourier transform infrared spectroscopy; IR, infrared; L_α, average length of α-helical segments; L_β, average length of β-strand segments; *md*, mean deviation; NaP, sodium phosphate; NaAc, sodium acetate; N_α, number of α-helical segments; N_β, number of β-strand segments; N_{res}, number of residues per unique polypeptide chain of a protein; PDB, Protein Data Bank; *r*, correlation coefficient; RMSD, root-mean-square deviation; *T*, turns; *U*, unordered; VCD, vibrational circular dichroism.

these, DSSP (Kabsch & Sander, 1983), which uses hydrogen bonding patterns in the crystal structure to assign the α-helix and β-sheet, is the most widely used. Another algorithm of interest is STRIDE (Frishman & Argos, 1995), which makes use of hydrogen-bonding patterns and backbone-dihedral angles to assign the α-helix and β-sheet. These two methods differ in the algorithm for assigning turns in protein structure. The fractional contents of the secondary structures are determined using the assignments from a given algorithm.

The secondary structures in proteins do not conform to a single geometry. Deviations from ideal conformational angles lead to distortions in the secondary structure such as bends and twists, end fraying, etc. The observed protein CD spectrum is an average of CD signals from all conformations, and reflects the geometric variability in the secondary structure. Another structural variable in proteins is the average length of the α-helices; for example, α-rich proteins have longer α-helices, and β-rich proteins have shorter α-helices. Whether such variations can be recognized from spectroscopic methods is a debatable question. Theoretical studies, however, indicate the influence of geometric variations in the structure on the CD spectra. Attempts to explicitly incorporate the length dependence of the α-helix CD in the analyses of protein CD spectra have had limited success (Chen et al., 1974). Influence of variations in the secondary structure on CD can be overcome in the analyses of protein CD spectra by the variable selection method (Manavalan & Johnson, 1987), where solutions are obtained by randomly selecting/deleting the reference proteins, or the ridge regression method (Provencher & Glöckner, 1981), where CD spectra of reference proteins are weighted differently in the solution. A recent development in protein spectral analysis is the estimation of the number of helix, sheet, and coil segments, which has been done using FTIR and VCD spectra and neural networks (Pancoska et al., 1994, 1996).

In this paper, we extend the analyses of CD spectra of proteins by splitting α-helix and β-sheet fractions into regular and distorted fractions, considering a specific number of terminal residues per segment to be distorted. Our results indicate that on an average four residues per α-helix and two residues per β-strand may be considered distorted in proteins. The fraction of the distorted and regular secondary structures estimated by CD analyses is used to estimate the number of α-helical and β-strand segments and their average length in a given protein. Using the methods developed here, the number of α-helical and β-strand segments were determined from CD spectra of two channel domains for colicin A and colicin E1, green fluorescent protein, and intestinal fatty acid binding protein (rat).

Methods

Reference protein set

The X-ray structures of the following 29 proteins, which formed our reference set for CD analysis, were taken from the Protein Data Bank (PDB) (Bernstein et al., 1977). The proteins and the X-ray structures used (PDB code in parenthesis) are: myoglobin (4mbn), hemoglobin (2mhb), hemerythrin (2hmz), T4-lysozyme (2lzm), triose-phosphate isomerase (3tim), lactate dehydrogenase (6ldh), lysozyme (1lys), thermolysin (8tln), cytochrome *c* (5cyt), phosphoglycerate kinase (3pgk), Eco R1 endonuclease (1eri), flavodoxin (1fx1), subtilisin BPN′ (1sbt), glyceraldehyde 3-phosphate dehydrogenase (3gpd), papain (9pap), subtilisin *novo* (2sbt), ribonuclease A (3rn3), pepsinogen (2psg), β-lactoglobulin (1beb), α-chymotrypsin (5cha), azurin (1azu), elastase (3est), γ-crystallin (4gcr), prealbumin (2pab), concanavalin A (2ctv), Bence-Jones protein (1rei), tumor necrosis factor (1tnf), superoxide dismutase (2sod), and α-bungaratoxin (2abx). The proteins are listed in the order of decreasing fractional content of α-helix.

We used the CD spectra in the range 178–260 nm at 1 nm intervals, which were generously provided by Dr. W.C. Johnson, Jr., in the analyses.

Secondary structure classification

The secondary structure assignments were done using both DSSP (Kabsch & Sander, 1983) and STRIDE (Frishman & Argos, 1995) algorithms. The DSSP method gives assignments of α-helices, 3_{10}-helices, β-sheets, β-bridges, turns, and bends, which were further grouped as follows: the α- and 3_{10}-helix assignments were treated as α-helices; β-sheets as β-strands; turns and bends were treated as turns; a minimum of two adjacent residues were required for such grouping for turns and bends. Single residues assigned to a structure (such as β-bridges, turns, and bends) were grouped under unordered, which contains residues that are not assigned to any defined structural class. Assignments from STRIDE have a similar nomenclature, with the exception of the bend classification, which does not exist in this algorithm. Our grouping of STRIDE assignments was similar to that with DSSP assignments. Such grouping gave us four secondary structure classes: α-helix (α), β-strand (β), turn (T), and unordered (U).

The α-helix structure was split into two classes: regular α-helix (α_R) and distorted α-helix (α_D). Secondary structure class α_D was formed by considering n_α residues for each α-helix segment to be distorted, and was assumed to give a CD signal that is different than the regular α-helix CD. The number of residues per α-helix that are considered under α_D, n_α, was varied from two to six. This corresponds to one to three residues at each end of an α-helical segment, the presumed location of the distorted residues, and the rest of the residues are included in the regular α-helix, the central part of the helical segment. If the number of residues in the α-helical segment is less than n_α, then all residues in that helix are grouped under α_D.

The β-strand structure was also split into two classes, on similar lines as the α-helix, and they are termed β_R and β_D. The number of residues per β-strand that are considered under β_D, n_β, was varied from one to four.

Our grouping of DSSP and STRIDE assignments gave us six secondary structural classes: α_R, α_D, β_R, β_D, T, and U.

Analysis of CD spectra

The analysis of CD spectra for estimating secondary structural fractions was done as follows: the CD spectrum of the protein analyzed for secondary structure was removed from our reference set and the secondary structure fractions were determined using the other members of the reference set, following the self-consistent method (Sreerama & Woody, 1993) version 2 (SELCON2, in prep.). In the self-consistent method, the spectrum of the protein analyzed is included in the matrix of CD spectral data, and an initial guess, the structure of the reference protein having the CD spectrum most similar to that of the protein analyzed, is made for the unknown secondary structure. The matrix equation relating the CD spectra to the secondary structure, $\mathbf{F} = \mathbf{XC}$, is solved by the singular-value

decomposition algorithm (Forsythe et al., 1977) and variable selection in the locally linearized model (van Stokkum et al., 1990). Valid solutions satisfy the conditions that the sum of fractions is between 0.95 and 1.05, each fraction is greater than -0.025 and the RMS deviation (RMSD) between the calculated and experimental CD is less than 0.25 $\Delta\epsilon$. The solution, which is the average of all valid solutions, replaces the initial guess, and the process is repeated until self-consistency is reached. The condition for the sum of fractions was relaxed to 0.90–1.10 for a few proteins. For some proteins the solutions did not converge, and an oscillatory behavior in the RMS difference between solutions from successive iterations was observed. In those cases, the solution obtained before the oscillatory behavior began was selected.

Estimation of number of α-helical and β-strand segments

The number of α-helical and β-strand segments was estimated from the fractions of distorted-helix ($f_{\alpha D}$) and distorted strand ($f_{\beta D}$) determined from the analysis of CD spectra. In a given protein structure, these fractions correspond to n_α residues per α-helix (α_D) and n_β residues per β-strand (β_D). The number of α-helical segments (N_α) was determined using $N_\alpha = (f_{\alpha D} \times N_{\text{res}})/n_\alpha$, and the number of β-strand segments (N_β) using: $N_\beta = (f_{\beta D} \times N_{res})/n_\beta$, where N_{res} is the number of residues in the protein.

The performance of the analysis is characterized by RMSDs (δ) and correlation coefficients (r) between the X-ray and CD estimates of secondary structure fractions for different secondary structure assignments, or estimates of the number of α-helical (N_α) or β-strand (N_β) segments. These are denoted by δ_k and r_k, where k is one of the secondary structural types considered. Overall performance of the analysis for a given set of secondary structure fractions was determined by considering all secondary structure fractions collectively, and these are given by δ and r.

These were calculated using the equations:

$$\delta = \sqrt{\frac{\sum_i (f_i^{\text{CD}} - f_i^X)^2}{N}}$$

and

$$r = \frac{N\sum_i (f_i^{\text{CD}} \times f_i^X) - \sum_{ij} (f_i^{\text{CD}} \times f_j^X)}{\sqrt{\left[N\sum_i (f_i^{\text{CD}})^2 - \left(\sum_i f_i^{\text{CD}}\right)^2\right] \times \left[N\sum_i (f_i^X)^2 - \left(\sum_i f_i^X\right)^2\right]}},$$

where f_i^{CD} and f_i^X are CD and X-ray estimates of secondary structure types of N reference samples.

An RMSD of 0.0 and a correlation coefficient of 1.0 between the X-ray and CD estimates indicates a perfect fit. Improvement in the performance of an analysis is indicated by a reduction of RMSD and an increase in the correlation since these two indices are inversely related. The selection of a better performance is straightforward for the majority of the cases. In a few cases, where the average content or the magnitude of a given secondary structure is small, we need to consider the normalized RMSDs (RMSD/the average content of the secondary structure) to compare the performance indices.

Basis CD spectra

The component CD spectra of the secondary structures (basis spectra) were deconvoluted from CD spectra of the reference proteins. The method followed to obtain secondary structure CD spectra was similar to that used by Compton and Johnson (1986). The matrix containing the CD spectra of reference proteins, **C**, is expressed as a product of three matrices using the singular value decomposition algorithm (Forsythe et al., 1977), $\mathbf{C} = \mathbf{USV^T}$, where **U** and **V** are unitary matrices and **S** is a diagonal matrix. This is incorporated in the matrix equation relating the CD spectra to the secondary structure data matrix, $\mathbf{F} = \mathbf{XC}$. The generalized inverse of **X**, which is $\mathbf{FVS^+U^T}$, gives the spectra corresponding to the secondary structures considered in constructing **F**.

Results

CD analyses of STRIDE and DSSP assignments

The fractions of α-helix, β-strand, turns, and unordered secondary structures and their averages in the reference set, obtained from the DSSP and STRIDE assignments for the 29 reference proteins, are given in Table 1. The fractions of α-helix and β-strand assigned by these two methods are similar, as they follow similar algorithms for assigning α-helix and β-sheet conformations; STRIDE assigns about 1% more residues than DSSP to both α-helix and β-sheet conformations. The average fraction of turns obtained from DSSP for the reference protein set (0.203) is smaller than that obtained from STRIDE (0.255). The turns are assigned in STRIDE using the definitions of Wilmot and Thornton (1990), wherein seven types of turns are defined based on dihedral angles. On the other hand, DSSP uses a stricter definition for turns and bends (which form the turn fraction in our grouping) based on either the hydrogen bonding or at least a 70° bend in the polypeptide chain. The fraction of unordered conformations was determined by subtracting the sum of α-helix, β-sheet, and turn fractions from unity, and the differences between unordered fractions from DSSP and STRIDE reflect their assignments of turns.

The secondary structure fractions were also estimated from the CD spectra, and these assignments, using the SELCON method, and the performance indices are given in Table 2. The performance indices for α-helix and β-strand are similar for these two assignments. The performance of STRIDE assignments of α-helix ($\delta_\alpha = 0.081$, $r_\alpha = 0.937$) is similar to those from DSSP ($\delta_\alpha = 0.080$, $r_\alpha = 0.933$), while that of β-strand are slightly superior (DSSP: $\delta_\beta = 0.113$, $r_\beta = 0.700$; STRIDE: $\delta_\beta = 0.109$, $r_\beta = 0.735$). The performance of turns and unordered fractions are better for DSSP assignments than those for STRIDE. The overall performance indices are slightly better for DSSP assignments (DSSP: $\delta = 0.091$, $r = 0.807$; STRIDE: $\delta = 0.098$, $r = 0.791$).

Regular and distorted fractions

From here on, due to the slightly superior performance of DSSP over STRIDE, only the results from DSSP assignments are presented. CD analyses, similar to those performed with DSSP assignments of secondary structure, were performed with STRIDE assignments and will be discussed later. The α-helical fractions were split into regular and distorted classes by considering two to six (n_α) residues per α-helical segment to be distorted. Only the α-helical segments with more than n_α residues contribute toward the regular α-helical fraction, α_R, while all α-helical segments contribute toward α_D. The β-strand fractions were also split into regular and distorted classes, considering one to four residues per

Table 1. *Secondary structure fractions determined from DSSP and STRIDE assignments of X-ray structures of the reference proteins*

		DSSP				STRIDE			
Protein[a]	N_{res}	f_α	f_β	f_T	f_U	f_α	f_β	f_T	f_U
4mbn	153	0.804	0.000	0.052	0.144	0.804	0.000	0.085	0.111
2mhb	287	0.760	0.000	0.105	0.136	0.777	0.000	0.105	0.118
2hmz	113	0.675	0.000	0.111	0.215	0.719	0.000	0.113	0.168
2lzm	164	0.665	0.085	0.116	0.134	0.665	0.085	0.055	0.195
3tim	250	0.446	0.154	0.124	0.276	0.500	0.156	0.090	0.254
6ldh	329	0.438	0.161	0.155	0.246	0.432	0.192	0.170	0.207
1lys	129	0.419	0.062	0.298	0.221	0.419	0.078	0.357	0.147
8tln	316	0.415	0.165	0.215	0.206	0.427	0.165	0.247	0.161
5cyt	103	0.408	0.000	0.233	0.359	0.437	0.039	0.291	0.233
3pgk	415	0.345	0.110	0.231	0.313	0.354	0.123	0.241	0.282
1eri	276	0.319	0.178	0.210	0.293	0.351	0.181	0.254	0.214
1fx1	148	0.318	0.216	0.264	0.203	0.324	0.243	0.338	0.095
1sbt	275	0.302	0.178	0.225	0.295	0.295	0.178	0.273	0.255
3gpd	334	0.274	0.208	0.217	0.301	0.268	0.219	0.326	0.187
9pap	212	0.259	0.170	0.175	0.396	0.311	0.179	0.156	0.354
2sbt	275	0.215	0.138	0.295	0.353	0.225	0.138	0.415	0.222
3rn3	124	0.210	0.331	0.218	0.242	0.226	0.331	0.226	0.218
2psg	370	0.205	0.386	0.165	0.243	0.208	0.392	0.200	0.200
1beb	162	0.167	0.410	0.216	0.207	0.157	0.420	0.269	0.154
5cha	245	0.114	0.314	0.200	0.371	0.114	0.322	0.345	0.218
1azu	128	0.109	0.250	0.312	0.328	0.094	0.281	0.438	0.188
3est	240	0.108	0.342	0.208	0.342	0.108	0.350	0.363	0.179
4gcr	174	0.092	0.460	0.109	0.339	0.075	0.471	0.132	0.322
2pab	127	0.063	0.449	0.165	0.323	0.063	0.469	0.236	0.232
2ctv	237	0.038	0.464	0.236	0.262	0.051	0.494	0.287	0.169
1rei	107	0.028	0.491	0.229	0.252	0.000	0.500	0.355	0.145
1tnf	157	0.019	0.433	0.219	0.329	0.013	0.469	0.287	0.231
2sod	151	0.018	0.367	0.298	0.316	0.005	0.389	0.374	0.232
2abx	74	0.000	0.108	0.284	0.608	0.000	0.108	0.358	0.534
Average[b]		0.284	0.229	0.203	0.285	0.290	0.240	0.255	0.215

[a]The PDB codes for the protein structures used (see Methods) are given. N_{res} is the total number of residues in the unique polypeptide sequence per protein.

[b]The average content of secondary structures in the set of reference proteins is calculated by dividing the sum of the secondary structure fractions for all proteins by the total number of proteins.

β-strand distorted, on similar lines. The secondary structure fractions corresponding to different n_α and n_β values were determined and used in the analyses of CD spectra, and the performance indices are given in Table 3. When either n_α or n_β is zero, the corresponding distorted fraction is zero and the corresponding entry in columns α_D or β_D is left blank.

The overall performance indices for different secondary structure assignments of regular and distorted fractions are similar. As

Table 2. *Comparison of performance indices[a] from CD analysis for different methods of secondary structure assignments*

	α		β		T		U			
Method[b]	δ_α	r_α	δ_β	r_β	δ_T	r_T	δ_U	r_U	δ	r
DSSP	0.080	0.933	0.113	0.700	0.062	0.482	0.101	0.300	0.091	0.807
STRIDE	0.081	0.937	0.109	0.735	0.090	0.562	0.108	−0.198	0.098	0.791

[a]The performance indices for each of the secondary structures are given as the RMSDs and correlation coefficients between the X-ray and the CD predicted fractions (δ_α, r_α,…). The overall performance indices are calculated as the RMSDs and correlation coefficients (δ and r) for all four secondary structure fractions collectively.

[b]The RMSDs and correlation coefficients, respectively, between the assignments from the two methods were: α-helix: 0.021, 0.997; β-sheet: 0.016, 0.997; turns: 0.073, 0.915; unordered: 0.084, 0.863.

Table 3. *Splitting α-helical and β-sheet fractions: Performance indices from CD analyses for secondary structure fractions determined from DSSP assignments*[a]

				α_R		α_D		β_R		β_D			
n_α	$A_{\alpha D}$	n_β	$A_{\beta D}$	$\delta_{\alpha R}$	$r_{\alpha R}$	$\delta_{\alpha D}$	$r_{\alpha D}$	$\delta_{\beta R}$	$r_{\beta R}$	$\delta_{\beta D}$	$r_{\beta D}$	δ	r
2	0.061	0		0.062	0.947	0.028	0.670	0.116	0.684			0.081	0.831
3	0.092	0		0.056	0.949	0.042	0.671	0.117	0.682			0.081	0.805
4	0.117	0		0.053	0.946	0.051	0.720	0.115	0.687			0.081	0.787
5	0.142	0		0.054	0.933	0.059	0.761	0.114	0.696			0.082	0.774
6	0.157	0		0.056	0.914	0.064	0.793	0.113	0.702			0.082	0.771
0		1	0.044	0.078	0.937			0.099	0.686	0.017	0.741	0.079	0.865
0		2	0.090	0.081	0.933			0.085	0.649	0.035	0.723	0.077	0.856
0		3	0.126	0.079	0.935			0.067	0.677	0.057	0.647	0.075	0.861
0		4	0.158	0.079	0.935			0.049	0.713	0.081	0.571	0.076	0.860
4	0.117	2	0.090	0.054	0.946	0.052	0.717	0.087	0.646	0.034	0.742	0.069	0.817

[a]Number of residues per segment of α-helix, n_α (or β-strand, n_β), determines the fraction α_D (or β_D); if it is zero then the corresponding fraction is not considered. The performance indices for the fraction of turns and unordered are similar to those given in Table 2. The average fractions of α_D and/or β_D in the basis set are given ($A_{\alpha D}$ and $A_{\beta D}$); those of other secondary structures can be obtained with the values given in Table 1. The overall performance indices are given as δ and r (see footnote to Table 2).

n_α varies from two to six, the fraction of α_D increases at the expense of α_R, but δ remains about the same (0.081) and r decreases slightly (0.831 − 0.775). Similarly, the overall performance indices are practically unchanged as n_β is increased from one to four ($\delta = 0.075$, $r = 0.860$). The performance of the distorted fraction, however, improves at the expense of the performance of the regular fraction as either n_α or n_β increases. This is clearly evident with the correlation coefficient for α_D, which increases as n_α increases, at the expense of $r_{\alpha R}$. The corresponding RMS value ($\delta_{\alpha D}$ or $\delta_{\beta D}$) also increases as n_α or n_β increases. This is due to the increase in the average content of α_D or β_D with the increase in the number of distorted residues per segment. A more correct measure of the RMS is the normalized RMS (relative error), obtained by dividing the RMSD by the average content of secondary structure fraction in the reference set, which for α_D decreases with increasing n_α (0.459, 0.456, 0.436, 0.415, 0.407, as n_α varies from two to six), and for β_D increases with increasing n_β (0.380, 0.388, 0.452, 0.513, as n_β varies from one to four). On the other hand, the normalized RMS for α_R or β_R increases with increasing n_α or n_β (α_R: 0.278, 0.292, 0.317, 0.380, 0.441, as n_α varies from two to six; β_R: 0.538, 0.612, 0.650, 0.690, as n_β varies from one to four). The comparison of these performance indices alone was not helpful in determining the average number of distorted residues per α-helical and β-strand segment due to similar overall performance and coupling of performance indices of distorted and regular segments. These were, however, determined from the comparison of the number of segments estimated from CD and determined by DSSP, as described in the following section.

Number of segments

The fraction of residues included in the distorted conformations of α-helix (or β-strand) can be used to estimate the number of segments of α-helices (or β-strands) with the knowledge of n_α (or n_β) and the number of residues in a given protein. For different n_α and n_β values, the secondary structure fractions estimated from the analyses of CD spectra were used to estimate the number of α-helical and β-strand segments. The CD estimates of the number of segments were compared with those from DSSP assignments, and the mean deviation, the RMSD, and the correlation coefficient between them are given in Table 4. The entries that are left blank correspond to values of n_α or n_β equal to zero. The differences between the CD-estimated and the DSSP-determined number of segments were the smallest for $n_\alpha = 4$ and $n_\beta = 1$. However, the average fraction of β_D for the $n_\beta = 1$ assignments was 0.044, which was lower than the average error from the analysis (δ = 0.079, Table 3). The performance indices from $n_\beta = 2$ assignments, with an average fraction of $\beta_D = 0.090$, were similar to those from $n_\beta = 1$ assignments. This led us to select $n_\beta = 2$ assignments over $n_\beta = 1$ assignments for the distorted β-strand fraction.

On an average, **four** residues per α-helix and **two** residues per β-strand can be considered distorted for the purposes of CD analy-

Table 4. *Splitting α-helix and β-sheet fractions: Performance indices from CD analyses for number of α-helical and/or β-strand segments from DSSP*[a]

		α-Helix			β-Strand		
n_α	n_β	$\delta_{N\alpha}$	$r_{N\alpha}$	$md_{N\alpha}$	$\delta_{N\beta}$	$r_{N\beta}$	$md_{N\beta}$
2	0	4.56	0.780	3.55			
3	0	3.42	0.779	2.35			
4	0	3.23	0.790	2.23			
5	0	4.01	0.798	2.88			
6	0	5.11	0.795	3.58			
0	1				2.41	0.934	1.92
0	2				2.63	0.911	2.17
0	3				4.11	0.901	3.36
0	4				7.30	0.888	6.20
4	2	3.24	0.789	2.26	2.50	0.920	2.05

[a]Performance indices are between the CD-determined and DSSP-assigned number of segments. Mean deviation (*md*) was calculated by dividing the sum of absolute differences between the DSSP and CD-determined number of segments by the total number of proteins. The total number of segments in the reference set were: α-helix, 196 (6.8 per protein); β-sheet: 270 (9.3 per protein) (see also the footnote to Table 3).

ses. Using the values of $n_\alpha = 4$ and $n_\beta = 2$ with DSSP assignments, we determined the fractions of regular and distorted fractions of α-helix and β-strand, and performed CD analyses. The summary of the results, in the form of performance indices for α_R, α_D, β_R, and β_D, and those for N_α and N_β, are given in the last rows of Tables 3 and 4. The average fractions of regular and distorted α-helix in our reference set were 0.167 and 0.117, respectively, and that of β-strand were 0.139 and 0.090, respectively. The performance of α_D ($\delta_{\alpha D} = 0.052$, $r_{\alpha D} = 0.717$) was slightly worse than that of α_R ($\delta_{\alpha R} = 0.054$, $r_{\alpha R} = 0.946$), and that of β_D ($\delta_{\beta D} = 0.035$, $r_{\beta D} = 0.723$) was slightly better than that of β_R ($\delta_{\beta R} = 0.085$, $r_{\beta R} = 0.649$). The RMSD and the correlation between the number of α-helical segments in the set of reference proteins determined from CD analyses and DSSP assignments were 3.24 segments and 0.789, respectively, and those for β-strand were 2.50 and 0.920, respectively.

The details of fractional content of the distorted α-helix and β-strand fractions and the number of segments of α-helix and β-strand as determined from CD and DSSP for the reference proteins are given in Table 5. The fractional contents of the regular α-helix and β-strand for these proteins can be obtained in conjunction with Table 1. The number of α-helical and β-strand segments (N_α and N_β) are given per protein molecule. Some X-ray structures have more than one polypeptide chain or more than one molecule in the asymmetric unit. The total number of residues (N_{res}) given in Table 1 corresponds to one unique polypeptide sequence per protein. The fractional values of N_α and N_β in Table 5 are a result of more than one molecule in the asymmetric unit in the X-ray structure.

The number of proteins for which DSSP and CD estimates differ by more than three segments is five for both N_α and N_β. This implies that the CD analyses estimate the number of α-helix segments with a similar degree of accuracy as the β-strand segments. On the contrary, the performance indices given in Table 4 imply that the number of β-strand segments are estimated with a better accuracy. A closer examination of the results clarifies this discrepancy. CD underestimates the number of α-helical segments in pepsinogen by 12 and overestimates the number in phosphoglycerate kinase by seven, and these two dramatically incorrect estimates make the performance indices for N_α worse than those

Table 5. *Comparison of number of α-helical and β-strand segments for the reference proteins determined by CD analyses and DSSP assignments*[a]

			N_α		N_β		L_α		L_β	
Protein	$f_{\alpha D}$	$f_{\beta D}$	DSSP	CD	DSSP	CD	DSSP	CD	DSSP	CD
4mbn	0.222	0.000	9.00	11.00	0.00	0.00	13.67	11.26	—	—
2mhb	0.223	0.000	16.00	19.00	0.00	2.00	13.63	10.76	—	2.25
2hmz	0.197	0.000	6.00	5.75	0.00	2.25	12.71	12.26	—	4.09
2lzm	0.244	0.037	10.00	7.00	3.00	3.00	10.91	13.31	4.70	3.73
3tim	0.210	0.064	13.50	11.00	8.00	5.50	8.27	11.60	4.82	4.45
6ldh	0.161	0.073	14.00	13.00	12.00	10.00	10.30	10.75	4.43	4.55
1lys	0.217	0.047	7.00	4.50	3.00	5.00	7.72	9.42	2.71	4.74
8tln	0.133	0.095	11.00	12.00	15.00	11.00	11.92	10.01	3.48	4.81
5cyt	0.194	0.000	5.00	3.00	0.00	4.00	8.40	10.04	—	4.92
3pgk	0.135	0.067	14.00	21.00	14.00	9.00	10.22	11.30	3.27	4.11
1eri	0.127	0.080	9.00	10.00	11.00	11.00	9.78	9.20	4.47	5.20
1fx1	0.108	0.108	4.00	4.00	8.00	7.00	11.73	8.38	4.00	5.20
1sbt	0.131	0.080	9.00	8.00	11.00	12.00	9.23	9.11	4.45	4.39
3gpd	0.102	0.093	9.00	11.00	15.50	13.50	10.17	8.76	4.48	4.72
9pap	0.123	0.075	7.00	5.00	8.00	7.00	7.87	11.04	4.48	4.65
2sbt	0.102	0.073	7.00	9.00	10.00	12.00	8.45	9.53	3.80	4.68
3rn3	0.097	0.113	3.00	4.00	7.00	4.00	8.68	7.37	5.86	4.97
2psg	0.154	0.151	16.00	4.00	28.00	23.00	4.74	7.39	5.10	5.38
1beb	0.111	0.123	5.00	3.50	10.00	8.50	5.41	8.15	6.64	5.13
5cha	0.045	0.106	3.00	5.00	13.00	13.50	9.31	6.39	5.92	5.03
1azu	0.062	0.109	2.00	2.00	7.00	9.00	6.98	10.05	4.57	5.48
3est	0.087	0.117	6.00	2.00	14.00	11.00	5.18	6.74	5.86	4.60
4gcr	0.086	0.161	4.00	0.00	14.00	12.00	4.00	5.72	—	5.78
2pab	0.031	0.142	1.00	2.00	9.00	8.50	7.87	6.32	6.34	5.01
2ctv	0.038	0.135	3.00	4.00	16.00	16.00	2.97	6.98	6.87	5.21
1rei	0.028	0.196	1.00	0.00	10.50	6.00	3.00	—	4.99	4.70
1tnf	0.019	0.140	1.00	1.67	11.00	13.67	2.98	4.36	6.18	7.17
2sod	0.018	0.119	0.75	4.00	9.00	10.00	3.67	6.08	6.17	5.96
2abx	0.000	0.095	0.00	2.00	3.50	6.50	—	4.05	2.30	5.07

[a]The fractions of α_D and β_D for $n_\alpha = 4$ and $n_\beta = 2$, which are used in the CD analyses are given. The fractions of other secondary structures can be obtained with the help of Table 1. The number of segments of α-helix (N_α) and β-strand (N_β) are given per single molecule of the protein. The fractional values of N_α and N_β are a result of more than one molecule in the asymmetric unit. L_α and L_β are the average lengths (number of residues) of α-helical and β-strand segments. The average length of α-helical segments in the reference proteins is 9.24 residues, and that of a β-strand segments is 5.02 residues.

for N_β. Removal of these two proteins from the analysis improves the estimates of N_α substantially ($\delta_{N\alpha} = 2.04$; $r_{N\alpha} = 0.894$) and that of N_β marginally ($\delta_{N\beta} = 2.21$; $r_{N\beta} = 0.895$). It should also be noted that while we consider four residues per α-helix segment to be distorted in determining the number of segments from the CD-determined fraction of the distorted α-helix, a few three-residue helical segments (assigned by DSSP as 3_{10}-helices) were considered in determining the DSSP-assigned fraction of the distorted α-helix. Our method may be considered to slightly underestimate the number of α-helical segments.

We have also given the CD-estimated and DSSP-determined average lengths of the α-helical and β-strand segments for the reference proteins in Table 5. These were determined by dividing the total number of residues in α-helix or β-strand conformations (obtained by multiplying the fraction of α or β by the total number of residues) by the number of segments. The average lengths of α-helical or β-strand segments for the set of reference proteins, estimated by CD, are in good agreement with those determined from the X-ray structure. The proteins with higher α-helix content have longer α-helices (about 9–13 residues), and those with higher β-strand content have shorter α-helices (about 5–8 residues), and CD estimates generally reflect these observations. The average length of β-strands does not show much variation between α-rich and β-rich proteins, and is about 4–6 residues. For the reference set of proteins, the average length of an α-helical segment is 9.2 residues, and that of a β-strand segment is 5.0 residues.

Basis CD spectra

The basis CD spectra of regular and distorted α-helix and β-strand conformations, corresponding to $n_\alpha = 4$ and $n_\beta = 2$, were deconvoluted from the CD spectra of reference proteins. Two sets of basis CD spectra of α_R and α_D are given in Figure 1, and these were obtained either using CD spectra of all reference proteins (Fig. 1A), or those of proteins with α-helix fraction greater than 0.3 (Fig. 1B). Similarly, either all reference proteins or only those with β-strand fraction greater than 0.3 were used for obtaining the CD spectra corresponding to β_R and β_D. These are given in Figure 2.

The CD spectrum of regular α-helix, from both Figures 1A and 1B, resembles the typical α-helix CD spectrum with a positive band around 194 nm and negative bands around 210 and 220 nm, and a crossover point at 203 nm. The amplitudes are very similar to those for high molecular weight polypeptide models, and the shoulder on the high energy side of the positive band is discernible. The positive band is slightly red-shifted from these models (194 nm vs. 192 nm). The distorted α-helix CD spectrum, as extracted from two sets of reference proteins, has similar features but with reduced amplitudes and shifted wavelengths. It has a positive band around 186–193 nm, negative bands around 199–209 and 227–228 nm, and crossover point between 193–203 nm. As the average content of α_D increases in the reference protein set, the characteristics of the α_D CD spectrum become more pronounced.

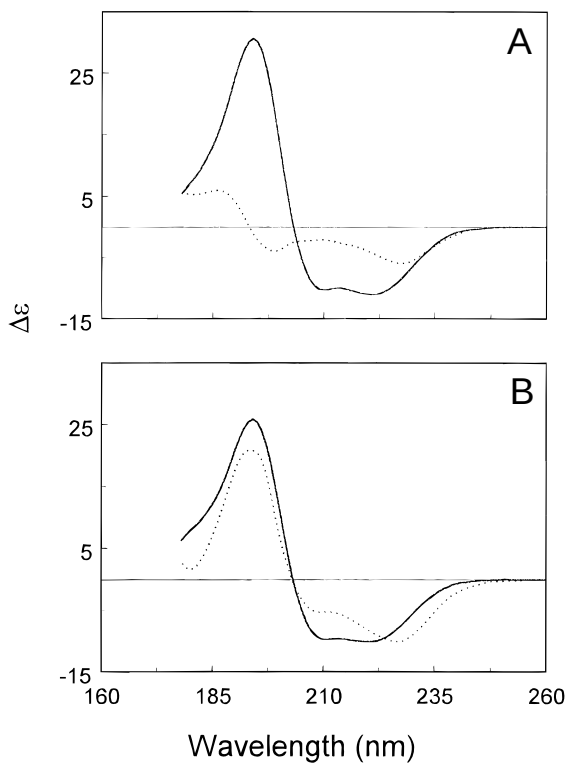

Fig. 1. The CD spectra associated with the regular (solid line) and distorted α-helix (dotted line) structure deconvoluted from the reference proteins. **A**: The CD spectra were calculated with all 29 reference proteins. **B**: The CD spectra were calculated with reference proteins with the total α-helix fraction greater than 0.3.

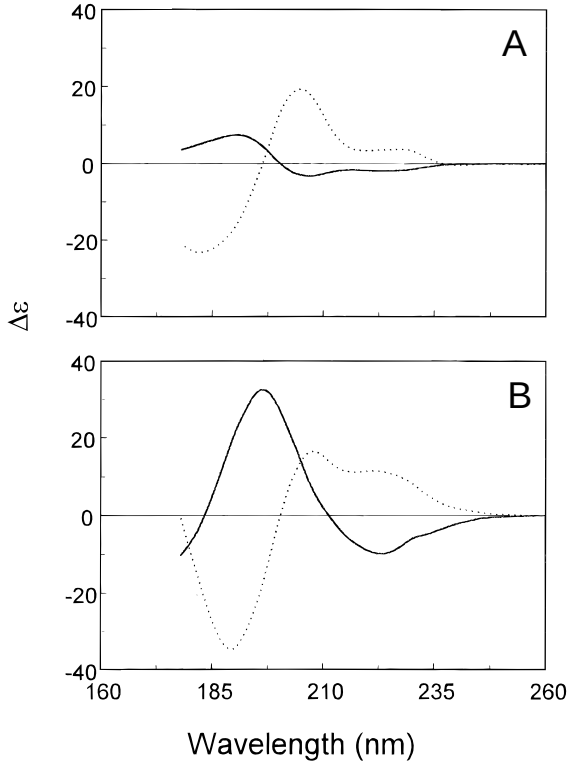

Fig. 2. The CD spectra associated with the regular (solid line) and distorted β-strand (dotted line) structure deconvoluted from the reference proteins. **A**: The CD spectra were calculated with all 29 reference proteins. **B**: The CD spectra were calculated with reference proteins with the total β-strand fraction greater than 0.3.

The deconvoluted CD spectrum of the regular β-strand is dependent on the reference protein set used, unlike the CD spectrum of regular α-helix. The CD spectrum obtained with proteins having β-strand fraction greater than 0.3 has a positive band around 196 nm and a negative band around 223 nm, which resembles model β-sheet CD spectra, although the negative band in the models is generally near 217 nm. The spectrum obtained with all reference proteins has these bands shifted to higher energy and their amplitudes reduced. The CD spectrum of distorted β-strand has a strong negative band between 182–189 nm, followed by two positive bands around 204–208 and 223–228 nm.

Estimates from the STRIDE assignments

We performed a complete series of analyses with assignments from the STRIDE algorithm. The STRIDE assignments also led to the conclusion that, on an average, four residues per α-helix and two residues per β-strand can be considered distorted. Using these criteria we determined the secondary structure fractions of proteins in the reference set and performed CD analyses. The results from STRIDE assignments for α_R, α_D, β_R, β_D, T, and U are compared with those from DSSP assignments in Table 6.

In general, the results from STRIDE and DSSP assignments are similar for α-helix and β-strand conformations. The performance of STRIDE assignments is slightly better for β_D and α_D, and slightly worse for T and U, than that of DSSP. The overall performance of STRIDE assignments is slightly poorer than that of DSSP assignments. The estimation of the number of α-helical segments from STRIDE assignments is similar in quality to that from DSSP, but that of β-strand segments is slightly worse.

Applications to other proteins

We applied the method of analysis developed in this paper to four proteins not included in the reference protein set: the channel domains of colicin A and colicin E1, green fluorescent protein, and intestinal fatty acid binding protein (rat). Of these four proteins, the coordinates of the X-ray structure were available for three proteins, and the PDB codes are 1col (colicin A), 1ema (green fluorescent protein), and 1fc (intestinal fatty acid binding protein).

Table 6. *Comparison of performance indices for different secondary structures obtained from alternate reference sets*[a]

	DSSP		STRIDE
	178–260 nm	185–240 nm	178–260 nm
$\delta_{\alpha R}$; $r_{\alpha R}$	0.054; 0.946	0.054; 0.945	0.055; 0.947
$\delta_{\alpha D}$; $r_{\alpha D}$	0.052; 0.717	0.051; 0.702	0.046; 0.776
$\delta_{\beta R}$; $r_{\beta R}$	0.087; 0.646	0.087; 0.613	0.077; 0.750
$\delta_{\beta D}$; $r_{\beta D}$	0.034; 0.742	0.033; 0.753	0.035; 0.710
δ_T; r_T	0.061; 0.500	0.067; 0.379	0.091; 0.562
δ_U; r_U	0.103; 0.249	0.100; 0.335	0.115; −0.289
δ; r	0.069; 0.817	0.069; 0.815	0.075; 0.789
$\delta_{N\alpha}$; $r_{N\alpha}$	3.24; 0.789	3.13; 0.789	3.18; 0.786
$\delta_{N\beta}$; $r_{N\beta}$	2.50; 0.920	2.42; 0.918	2.91; 0.880

[a]The alternate reference sets are either from DSSP and STRIDE assignments or with different wavelength ranges.

For the channel domain of colicin E1, the structure has been published (Elkins et al., 1997), but the PDB structure was unavailable. The CD spectra of these proteins were measured at the Mayo Foundation on a J-710 spectropolarimeter in quartz cells of 0.01–0.02 cm, using the following parameters: 2 s response; 20 nm/min scan speed; 0.1 nm data acquisition interval; five accumulations; 2 nm bandwidth. The CD spectra were smoothed using the noise reduction routines provided with the J-710 spectropolarimeter. The solvent and conditions of spectral acquisition were as follows: green fluorescent protein, 20 mM NaP, pH 7.0, $T = 10\,^{\circ}C$; intestinal fatty acid binding protein, 10 mM NaP, pH 7.0, $T = 25\,^{\circ}C$; colicin channel domains A and E1, 20 mM NaAc, 100 mM NaP, pH 4.0, $T = 25\,^{\circ}C$. The CD spectra are given in Figure 3.

The CD spectra of these proteins are in the 185–240 nm wavelength range. To examine the performance of the truncated wavelength range, we performed the CD analyses for the number of α-helical and β-strand segments with the truncated CD spectra of reference proteins, and the results are summarized in Table 6. The performance indices obtained with the 185–240 nm wavelength range were comparable to those obtained with the 178–260 nm wavelength range.

The CD spectra of the channel domains of colicin A and colicin E1, green fluorescent protein, and intestinal fatty acid binding protein were analyzed using the truncated CD spectra of reference proteins. The results are compared with the secondary structure characteristics of X-ray structure obtained from DSSP in Table 7. The number of α-helical segments estimated by CD agree very well with the X-ray data, while average lengths of α-helical segments agree reasonably well. The agreement for the number and average length of β-strands is slightly poorer.

Discussion

The majority of the investigations of protein spectral analysis have been directed toward estimating the fractional contents of the major secondary structures in proteins, viz., α-helix, β-sheet, and β-turns. The rest of the structures form a fraction that has been

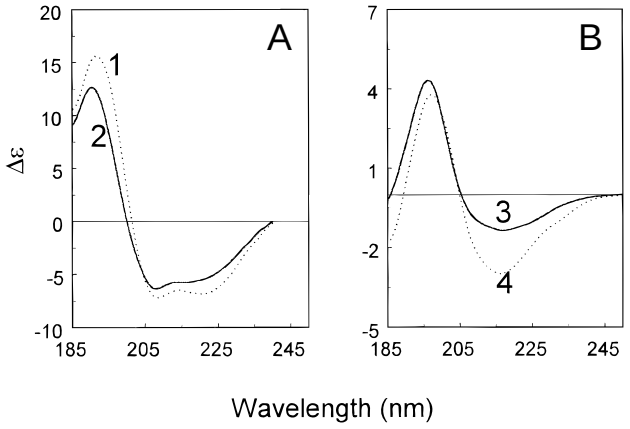

Fig. 3. The experimental CD spectra of channel domains of colicin A (curve 1), colicin E1 (curve 2), green fluorescent protein (curve 3), and intestinal fatty acid binding protein (curve 4), proteins not included in the reference set of proteins. The CD spectra in (**A**), curves 1 and 2, are similar to those of α-rich proteins. The CD spectra in (**B**), curves 3 and 4, are similar to those of β-rich proteins.

Table 7. *Comparison of secondary structure estimates predicted by CD with DSSP assignments of X-ray structure for the four test proteins not included in the reference proteins*[a]

			Secondary structure fractions						α-Helical segments		β-Strand segments	
Protein	N_{res}	Method	$f_{\alpha R}$	$f_{\alpha D}$	$f_{\beta R}$	$f_{\beta D}$	f_T	f_U	N_α	L_α	N_β	L_β
Colicin A	204	DSSP	0.529	0.225	0.000	0.000	0.044	0.202	12	12.8	0	0.0
		CD	0.494	0.213	0.009	0.011	0.088	0.189	11	13.3	1	4.0
Colicin E1	190	DSSP	—	—	—	—	—	—	—	—	—	—
		CD	0.337	0.180	0.051	0.053	0.147	0.284	9	11.5	5	4.0
Green fluorescent protein	236	DSSP	0.004	0.064	0.347	0.093	0.191	0.301	4	4.0	11	9.5
		CD	0.039	0.051	0.216	0.120	0.242	0.280	3	7.1	14	5.6
Intestinal fatty acid binding protein	131	DSSP	0.053	0.061	0.432	0.152	0.152	0.152	2	7.5	10	7.7
		CD	0.068	0.070	0.221	0.121	0.245	0.296	2	9.1	8	5.6

[a]The fractional contents of six secondary structure conformations are given. DSSP corresponds to the X-ray structure assignments and CD to predictions from the analysis of CD spectra. The number (N) and average length (L) of α-helix (α) and β-strand (β) are given. N_{res} is the total number of residues in the protein. Columns are left blank where data are unavailable.

referred to as irregular regions, other structures, remainder, random coil, or unordered conformation. We use the term unordered (Sreerama & Woody, 1994b). The left-handed poly(Pro)II structure was shown to be a major fraction of the unordered conformation (Sreerama & Woody, 1994a). β-Turn structures were split into turns and bends, based on DSSP assignments (Kabsch & Sander, 1983), by Pancoska et al. (1991). The α-helix fractions were first subdivided into ordered and disordered classes by Williams (1983) for the analysis of Raman spectra. This was later adopted by Kalnin et al. (1990), studying infrared spectra, who considered two residues at each end of an α-helix as disordered. Although not included in their paper, Kalnin et al. (1990) also examined the inclusion of one and three residues from each end of an α-helix in the disordered fraction, but the RMS and correlation between the X-ray and infrared estimates, for a set of 19 reference proteins, were poorer than their results from considering two residues from each end of an α-helix disordered. Kalnin et al. (1990) subdivided the β-sheet fractions also by considering residues lacking a classical hydrogen bond in a β-sheet as disordered. There have also been investigations considering 3_{10}-helix, parallel, and antiparallel β-sheets, and a few other secondary structures that form a subset of the four major secondary structure types listed above (for a summary, see review by Venyaminov & Yang, 1996). Most spectral analyses, however, determine the four major secondary structural fractions: α, β, turns, and unordered.

We have performed a systematic investigation of whether a subdivision of the α-helix and β-sheet fractions is useful in the analysis of protein CD spectra. As indicated previously, we varied the number of distorted residues per segment of an α-helix from two to six, and that of a β-strand from one to four. From the CD analysis, the number of residues in the distorted structures and the number of α-helical and β-strand segments were determined. The relative performance of the CD analyses indicated that on an average four residues per α-helix and two residues per β-strand are distorted in proteins.

The classical α-helix (Pauling et al., 1951) is characterized by ϕ, ψ angles of $-57°$ and $-47°$, respectively, and a hydrogen bond between the peptide C=O group of the n^{th} residue and the N–H group of the $(n + 4)^{\text{th}}$ residue. The geometries of the α-helical conformations in proteins, however, are flexible, and show variations from the characteristics of a classical α-helix. The first three and the last three peptide groups in an α-helix have one hydrogen bond with another peptide group of the α-helix (between the C=O bond of the n^{th} residue and the N–H bond of the $(n + 4)^{\text{th}}$ residue, at one end, and between the N–H bond of the k^{th} residue and the C=O bond of the $(k - 4)^{\text{th}}$ residue, at the other end), while the peptide groups in the interior of the α-helix have two hydrogen bonds between the corresponding C=O and N–H bonds. In the case of the 3_{10}-helix, two peptide groups at each end have one hydrogen bond with another peptide group of the 3_{10}-helix. The central residues of the α-helix, which has two hydrogen bonds per each residue, are more geometrically constrained than the end residues. The protein far-UV CD spectrum is largely due to the exciton interaction between the transitions on the peptide chromophores, and the CD spectrum due to end residues in an α-helix is expected to be different than that due to residues in the interior of the α-helix.

Our results indicate that one can consider two residues at each end of an α-helix to be distorted, and this is exactly the number used by Williams (1983) and Kalnin et al. (1990) in defining the disordered helix from Levitt and Greer (1977) assignments. Comparison of our performance indices with those of Williams (1983) or of Kalnin et al. (1990) is not possible because of the differences in the protein set, the secondary structure assignments, and the spectroscopic method. The CD spectrum of the regular α-helix, corresponding to the central part of the α-helix and calculated from the reference set of proteins (Fig. 1), shows the CD characteristics of a typical α-helix. On the other hand, the calculated CD spectrum of the distorted α-helix, due to the residues at the ends of α-helices, has corresponding bands broadened and their amplitudes diminished.

The β-sheets in proteins exhibit a greater geometric variability than the α-helices and they can be bent and twisted. The two types of β-sheets, antiparallel and parallel, are combined in our β-strand fraction, and these two have different hydrogen-bonded structures. Depending on how many β-strands form the β-sheet and the location of a given residue in the β-sheet, a peptide group in a β-strand can have either one or two hydrogen bonds. Our results indicate that we can consider two residues per β-strand as distorted, which corresponds to one residue at each end of a β-strand. The argument for the end residues of a β-strand having a different

CD signal than that due to the central residues is slightly weaker than that for the case of α-helix, owing to the greater geometric variability of the β-sheet structure. This is reflected in the CD spectra we calculate for the distorted β-strand from the reference protein set. In this case, even the CD spectrum of the regular β-strand is dependent on the set of reference proteins used, and only that obtained with proteins having the total β-strand fraction greater than 0.3 resembles model β-sheet CD spectra. The CD spectrum of the distorted β-strand is unlike any model CD spectrum. We are unable to explain this CD spectrum, which has a negative band between 182–189 nm and two positive bands around 204–208 and 223–228 nm. The CD contributions from the aromatic and/or cystinyl side chains, which are not explicitly considered in our analysis, may be influencing this spectrum.

The main result of this study is the estimation of the number of the α-helical and β-strand segments in a protein from the analysis of its CD spectrum. For the proteins in the reference set, we obtain reasonable estimates of the number of segments; CD and X-ray estimates of α-helical segments differ by four or more segments for four proteins (3pgk, 2psg, 3est, 4gcr), and that for β-strand segments for four proteins (8tln, 5cyt, 3pgk, 2psg). The two proteins, 3pgk and 2psg, are common between these two sets. Removal of these two proteins from the analysis improves the statistics: $\delta_{N\alpha} = 2.04$; $r_{N\alpha} = 0.894$; $\delta_{N\beta} = 2.21$; $r_{N\beta} = 0.895$. We have also examined the performance of a truncated wavelength range (185–240 nm). The performance of the truncated wavelength range was comparable to that of the complete wavelength range (178–260 nm), which is in accord with the results of Venyaminov et al. (1991). It is generally believed that the reduced spectral information due to truncation of the short-wavelength CD data results in poorer performance indices. However, our results and those obtained by Venyaminov et al. (1991) indicate little or no effect of the truncated wavelength range on the analysis. Contributions of aromatic side chains and the twisting of β-sheets to the CD spectra below 185 nm may be poorly correlated with secondary structure content. These effects could nullify the presumed advantage of including data from this region of the spectrum. This aspect of the CD analysis needs further investigation.

Information about the number of the α-helical and β-strand segments has also been obtained by Pancoska et al. (1994) from the analysis of VCD spectra of proteins using neural networks. They summarized the number of segments of helix, sheet, and coil, and the connectivities between them in the form of a matrix, which they called the matrix descriptor of the protein super-secondary structure. The VCD spectra and the corresponding matrix descriptors for a set of 23 reference proteins formed the input and output layer of a neural network, respectively. Each protein VCD spectrum in the reference set was analyzed using the rest of the proteins. They obtained a standard deviation of 3.52 and 3.00 for the number of α-helical and β-strand segments for the proteins in their reference set. A direct comparison of our results with their results is not possible due to the differences in the reference proteins and the methodology. There were 16 proteins in common between their reference set and ours. Among these, the number of segments of either α-helix or β-strand estimated by our method differed from the X-ray determined values by at least three segments for six proteins, while Pancoska et al. (1994) obtained such differences for seven proteins.

We have used the method developed in this study to estimate the number of secondary structural segments in four additional proteins, not included in the reference set, and the CD spectra of these proteins corresponded to the truncated wavelength range of 185–240 nm. The results of the analyses compared well with the X-ray structure data available for three of the proteins analyzed.

Although the method developed in this study estimates the number of α-helical and β-strand segments in proteins, it is based on the determination of fractional contents of secondary structures from the CD spectra and suffers from the same limitations as the methods that estimate secondary structural fractions from protein spectra. The following assumptions are involved in such an empirical analysis: (1) the protein CD spectrum is represented as a linear combination of secondary structure component spectra; (2) the ensemble-averaged solution structure and the time-averaged solid-state structure are equivalent; (3) the CD contributions from nonpeptide chromophores do not influence the analysis; (4) the effect of the tertiary structure on CD is negligible; (5) effects of the geometric variability of the secondary structures are not explicitly considered. These have been reviewed (Manning, 1989; Venyaminov & Yang, 1996). By and large, these assumptions are valid. Inadequacies of these assumptions are generally overcome by the variable selection principle (Manavalan & Johnson, 1987) and the ridge regression method (Provencher & Glöckner, 1981), with some sort of averaging of the contributions from the reference proteins. However, one can still obtain a poor analysis if the CD spectrum has significant nonpeptide contributions. The analysis is also dependent on the choice of reference proteins, even with the variable selection method, because the CD of reference proteins represent the variations in the protein structures only to a limited extent. One needs a sufficiently large reference set to make the analysis completely independent of the reference protein set.

The results obtained in this study and those of Pancoska et al. (1994) have proven that the structural information that can be obtained from spectroscopic methods goes beyond the fractional contents of various protein secondary structures. One area where this additional information about the number of secondary structural segments is useful is in sequence-based secondary structure prediction methods. Carrara et al. (1992) obtained a slight improvement in the sequence-based secondary structure prediction by the Garnier et al. (1978) method by optimizing the assignments to the CD-derived fractions of α-helix and β-sheet. Such an approach, wherein one incorporates the experimentally derived structural information in the theoretical sequence-based structural prediction method, is much awaited and might lead to more reliable assignments of secondary structures at the residue level.

Another useful application of the methods developed here is in the protein folding and/or protein unfolding studies. High-quality time-resolved CD spectral data can be obtained on the ms time scale using synchrotron radiation (Jones, 1998). The present method should be useful in analyzing such data in protein folding experiments.

Conclusions

We have extended the structural information that can be obtained from the analysis of protein CD spectra, which had been limited to the fractional content of secondary structures, to the number of α-helical and β-strand segments. This was achieved by splitting the α-helix and β-strand conformations in a protein into regular and distorted α-helix and β-strand structures. A certain number of terminal residues per segment of α-helix or β-strand was considered distorted, and the resulting secondary structure fractions were utilized in the analysis of CD spectra by the SELCON method.

From the performance of the CD analyses, it was concluded that on an average four residues per α-helix and two residues per β-strand can be considered distorted in proteins. The fractional content of the regular and distorted α-helical and β-strand structures determined from CD analysis was further utilized in estimating the number of α-helical and β-strand segments in a given protein.

We have shown that the number of secondary structural segments can be obtained from the analysis of protein CD spectra to a reasonable degree of accuracy. The approach developed here can also be used in the analyses of IR, Raman, and VCD spectra of proteins. The additional information about protein structure obtained from experimental studies is important in bridging the gap between the secondary structure analyses of protein spectra and the sequence-based protein secondary structural predictions.

The latest version of the SELCON program for analyzing protein CD spectra and the related data files (spectral and secondary structure data) are available via anonymous *ftp* at the internet address: bciris.bmb.colostate.edu (129.82.125.151), login_name: anonymous, password: your_name; directory: pub/SELCON3. Please use binary mode of transfer to *get* (or copy) SELCON3.EXE, which can be run on a PC. The source code, SELCON3.FOR, can be compiled with a FORTRAN 77 compiler on any platform. The details of setting up the program are given in the README file. The authors may also be contacted via e-mail at: sreeram@lamar.colostate.edu, venyaminov.sergei@mayo.edu, rww@lamar.colostate.edu. SELCON3 is also available at the web page http://lamar.colostate.edu/~sreeram/SELCON3.

Acknowledgments

Thanks are due to Dr. W.C. Johnson, Jr., for providing the CD spectra of proteins used in this study. S.Y.V. thanks Prof. F.G. Prendergast for his interest and support during the preparation of this work. This work was supported by NIH Research Grants GM22994 (R.W.W.) and GM34847 (S.Y.V.).

References

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rogers JR, Kennard O, Shimonouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Böhm G, Muhr R, Jaenicke R. 1992. Quantitative analysis of protein far UV circular dichroism spectra by neural networks. *Protein Eng 5*:191–195.

Bolotina IA, Chekhov VO, Lugauskas VY, Finkel'shtein AV, Ptitsyn OB. 1980. Determination of the secondary structure of proteins from the circular dichroism spectra. 1. Protein reference spectra for α-, β- and irregular structures. *Mol Biol (Eng Transl Mol Biol) 14*:701–709.

Brahms S, Brahms J. 1980. Determination of protein secondary structure in solution by vacuum ultraviolet circular dichroism. *J Mol Biol 138*:149–178.

Carrara EA, Gavotti C, Casati P, Nozza F. 1992. Improvement of protein secondary structure prediction by combination of statistical algorithms and circular dichroism. *Arch Biochem Biophys 294*:107–114.

Chen YH, Yang JT. 1971. A new approach to the calculation of secondary structures of globular proteins by optical rotatory dispersion and circular dichroism. *Biochem Biophys Res Commun 44*:1285–1291.

Chen YH, Yang JT, Chau KH. 1974. Determination of the helix and β-form of proteins in aqueous solution by circular dichroism. *Biochemistry 13*:3350–3359.

Compton LA, Johnson WC Jr. 1986. Analysis of protein circular dichroism spectra for secondary structure using a simple matrix multiplication. *Anal Biochem 155*:155–167.

Elkins P, Bunker A, Cramer WA, Stauffacher CV. 1997. A mechanism for toxin insertion into membranes is suggested by the crystal structure if the channel-forming domain of colicin E1. *Structure 5*:443–458.

Forsythe GE, Malcolm MA, Moler CB. 1977. *Computer methods for mathematical computations*. Englewood Cliffs, NJ: Prentice-Hall.

Frishman D, Argos P. 1995. Knowledge-based protein secondary structure assignment. *Proteins Struct Funct Genet 23*:566–579.

Garnier J, Osguthorpe DJ, 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.

Greenfield NJ. 1996. Methods to estimate the conformation of proteins and polypeptides from circular dichroism data. *Anal Biochem 235*:1–10.

Greenfield NJ, Fasman GD. 1969. Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry 8*:4108–4116.

Hennessey JP Jr, Johnson WC Jr. 1981. Information content in the circular dichroism of proteins. *Biochemistry 20*:1085–1094.

Jones G. 1998. Current state-of-the art and future possibilities for synchrotron radiation circular dichroism (SRCD). *Proc 1st mini-conference on CD with synchrotron radiation*. Warrington, UK: Daresbury Laboratory. p 5.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometric features. *Biopolymers 22*:2577–2637.

Kalnin NN, Baikalov IA, Venyaminov SY. 1990. Quantitative IR spectrophotometry of peptide compounds in water solution III. Estimation of the protein secondary structure. *Biopolymers 30*:1273–1280.

Levitt M, Greer J. 1977. Automatic identification of secondary structures in globular proteins. *J Mol Biol 114*:181–293.

Manavalan P, Johnson WC Jr. 1987. Variable selection method improves the prediction of protein secondary structure from circular dichroism. *Anal Biochem 167*:76–85.

Manning MC. 1989. Underlying assumptions in the estimation of secondary structure of proteins by circular dichroism spectroscopy—A critical review. *J Pharm Biomed Anal 7*:1103–1119.

Pancoska P, Bitto E, Janota V, Keiderling TA. 1994. Quantitative analysis of vibrational circular dichroism spectra of proteins. Problems and perspectives. *Faraday Discuss 99*:287–310.

Pancoska P, Fabian H, Yoder G, Baumruk V, Keiderling TA. 1996. Protein structural segments and their interconnections derived from optical spectra. Thermal unfolding of ribonuclease T1 as an example. *Biochemistry 35*:13094–13106.

Pancoska P, Yasui SC, Keiderling TA. 1991. Statistical analysis of the vibrational circular dichroism of selected proteins and relationship to secondary structure. *Biochemistry 30*:5089–5103.

Pauling L, Corey RB, Branson HR. 1951. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci USA 37*:205–211.

Perczel A, Hollosi M, Tusnady G, Fasman GD. 1991. Convex constraint analysis: A natural deconvolution of circular dichroism curves of proteins. *Protein Eng 4*:69–79.

Provencher SW, Glöckner J. 1981. Estimation of protein secondary structure from circular dichroism. *Biochemistry 20*:33–37.

Shubin VV, Khazin ML, Efimovskaya TB. 1990. Prediction of protein secondary structure of globular proteins using circular dichroism spectra. *Mol Biol (Eng Transl Mol Biol) 24*:165–176.

Sklenar H, Etchebest C, Lavery R. 1989. Describing protein structure: A general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins Struct Funct Genet 6*:46–60.

Sreerama N, Woody RW. 1993. A self-consistent method for the analysis of protein secondary structure from circular dichroism. *Anal Biochem 209*:32–44.

Sreerama N, Woody RW. 1994a. Poly(Pro)II helices in globular proteins: Identification and circular dichroic analysis. *Biochemistry 33*:10022–10025.

Sreerama N, Woody RW. 1994b. Protein secondary structure from circular dichroism spectroscopy. Combining variable selection principle and cluster analysis with neural network, ridge regression and self-consistent methods. *J Mol Biol 242*:497–507.

van Stokkum IHM, Spoelder HJW, Bloemendal M, van Grondelle R, Groen FCA. 1990. Estimation of protein secondary structure and error analysis from circular dichroism spectra. *Anal Biochem 191*:110–118.

Venyaminov SY, Baikalov IA, Wu CSC, Yang JT. 1991. Some problems of CD analysis of protein conformation. *Anal Biochem 198*:250–255.

Venyaminov SY, Yang JT. 1996. Determination of protein secondary structure. In: Fasman GD, ed. *Circular dichroism and the conformational analysis of biomolecules*. New York: Plenum. pp 69–108.

Williams RW. 1983. Estimation of protein secondary structure from the laser Raman amide I spectrum. *J Mol Biol 166*:581–603.

Wilmot CM, Thornton JM. 1990. β-Turns and their distortions: A proposed new nomenclature. *Protein Eng 3*:479–493.