

## Estimation of Protein Secondary Structure from Circular Dichroism Spectra: Inclusion of Denatured Proteins with Native Proteins in the Analysis

Narasimha Sreerama,\* Sergei Yu. Venyaminov,† and Robert W. Woody\*.<sup>1</sup>

\*Department of Biochemistry and Molecular Biology, Colorado State University, Fort Collins, Colorado 80523; and

†Department of Pharmacology, Mayo Foundation, Rochester, Minnesota 55905

Received April 7, 2000

**We have expanded our reference set of proteins used in the estimation of protein secondary structure by CD spectroscopy from 29 to 37 proteins by including 3 additional globular proteins with known X-ray structure and 5 denatured proteins. We have also modified the self-consistent method for analyzing protein CD spectra, SELCON3, by including a new selection criterion developed by W. C. Johnson, Jr. (*Proteins Struct. Funct. Genet.* 35, 307–312, 1999). The secondary structure corresponding to the denatured proteins was approximated to be 90% unordered, owing to the spectral similarity of the denatured proteins and unordered structures. We examined the thermal denaturation of ribonuclease T1 by CD using both the original and expanded sets of reference proteins and obtained more consistent results with the expanded set. The expanded set of reference proteins will be helpful for the determination of protein secondary structure from protein CD spectra with higher reliability, especially of proteins with significant unordered structure content and/or in the course of denaturation.** © 2000

Academic Press

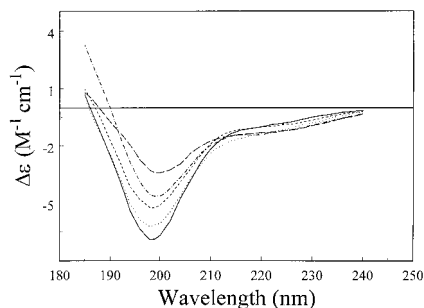
**Key Words:** protein secondary structure; SELCON; CD analysis; protein CD; unordered structure.

One of the most successful applications of CD, the structural characterization of proteins, depends upon the remarkable sensitivity of the far-UV CD to the backbone conformation of proteins; the far-UV CD of a protein generally reflects the secondary structure content of the protein. Various empirical methods have been developed for analyzing protein CD spectra for

quantitative estimation of the secondary structure content (1–20), and various aspects of secondary structure analysis have been reviewed (21–25). Methods have also been developed to estimate the number of secondary structural segments in proteins (19), and to assign the protein tertiary structure class from the analysis of far-UV CD spectra (26, 27).

The basic principle involved in the analysis of protein CD spectra, and used in the estimation of secondary structure fractions, is that the protein CD spectrum ( $C_\lambda$ ) can be expressed as a linear combination of CD spectra of individual secondary structure components ( $k$ ),  $B_{k\lambda}$ ,  $C_\lambda = \sum f_k B_{k\lambda}$ , where  $f_k$  is the fraction of the secondary structure  $k$ . Spectra of model polypeptides or of a set of reference proteins with known structures are used to determine the component secondary structure spectra,  $B_{k\lambda}$ . The most successful methods use the CD spectra of a set of reference proteins with known structures, and the secondary structure fractions for the reference proteins are determined from the corresponding crystal structure. The following assumptions are involved in these methods: (1) the contributions from individual secondary structures are additive; (2) the ensemble-averaged solution structure and the time-averaged solid-state X-ray structure are equivalent; (3) the effect of the tertiary structure on CD is negligible; (4) the CD contributions from nonpeptide chromophores do not influence the analysis; (5) effects of the geometric variability of the secondary structures need not be explicitly considered. Although assumptions 1–3 are generally valid, assumptions 4 and 5 are likely to be incorrect in many cases (28, 29). This is the reason that flexible basis methods (variable weighting of the reference proteins (7), variable selection (9, 11, 15, 17–20) in which a large number of subsets of the reference set are used with the results subjected to tests to identify good solutions, and neural networks

<sup>1</sup> To whom correspondence should be addressed. Fax: (970) 491-0494. E-mail: [rww@lamar.colostate.edu](mailto:rww@lamar.colostate.edu).



**FIG. 1.** CD spectra of five denatured proteins included in the Exp37 reference protein set. The proteins are: apocytochrome *c*, 5°C (short dashes); apocytochrome *c*, 90°C (long dashes); staphylococcal nuclease, 6°C (solid); staphylococcal nuclease, 70°C (dot-dash); and oxidized ribonuclease, 20°C (dotted).

(14, 16, 18)) have largely displaced fixed basis methods (1–6, 8, 10, 12, 13). Although the flexible basis methods do not explicitly determine the CD spectra of individual secondary structural components ( $B_{kl}$ ), they implicitly assume such spectra, containing the effects of distortions, side chains, and end effects, adapted to the protein being analyzed. By contrast, a fixed reference set cannot take these variations into account.

A reference protein set with a good representation of CD spectral features and secondary structural combinations is desirable for a good analysis. Different sets of reference proteins, with the number of proteins varying from 15 to 33 and having a good representation of  $\alpha$ -rich,  $\beta$ -rich, and mixed- $\alpha\beta$  proteins, have been used in secondary structure analyses. In one case, four denatured protein CD spectra were included in a set of 16 reference proteins (30). We have extended the set of 29 reference proteins, used in our SELCON3 method, by adding 3 native proteins and 5 denatured proteins. We have also modified the SELCON3 method, introducing a new criterion for selecting valid solutions based on helical content developed by Johnson (20). The method and the reference protein set described in this paper were used to analyze changes of secondary structure in the course of thermal unfolding of ribonuclease T1.

## MATERIALS AND METHODS

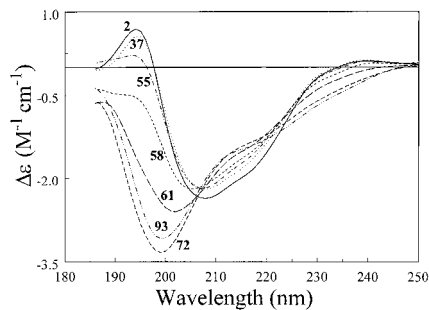
**Reference proteins.** Our original reference set consisted of 29 proteins. The proteins and the X-ray structures used (PDB<sup>2</sup> code in parenthesis) are: myoglobin (4mbn), hemoglobin (2mhb), hemerythrin (2hmz), T4

<sup>2</sup> Abbreviations used: PDB, Protein Data Bank;  $\delta$ , root mean square deviation;  $r$ , correlation coefficient; Org29, original 29 reference proteins; Exp32, set of 32 globular reference proteins; Exp37, set of 37 reference proteins; SVD, singular value decomposition;  $\alpha_R$ , regular  $\alpha$  helix;  $\alpha_D$ , distorted  $\alpha$  helix;  $\beta_R$ , regular  $\beta$  strand;  $\beta_D$ , distorted  $\beta$  strand; T, turns; U, unordered; DSSP, a computer program for defining secondary structure of proteins.

lysozyme (2lzm), triosephosphate isomerase (3tim), lactate dehydrogenase (6ldh), lysozyme (1lys), thermolysin (8tln), cytochrome *c* (5cyt), phosphoglycerate kinase (3pgk), *EcoRI* endonuclease (1eri), flavodoxin (1fx1), subtilisin BPN' (1sbt), glyceraldehyde 3-phosphate dehydrogenase (3gpd), papain (9pap), subtilisin *novo* (2sbt), ribonuclease A (3rn3), pepsinogen (2psg),  $\beta$ -lactoglobulin (1beb),  $\alpha$ -chymotrypsin (5cha), azurin (1azu), elastase (3est),  $\gamma$ -crystallin (4gcr), prealbumin (2pab), concanavalin A (2ctv), Bence-Jones protein (1rei), tumor necrosis factor (1tnf), superoxide dismutase (2sod), and  $\alpha$ -bungarotoxin (2abx). The original set is referred to as Org29. Three globular proteins, colicin A (1col), green fluorescent protein (1ema), and rat intestinal fatty acid-binding protein (1ifc), were added to the Org29 reference proteins, bringing the total number of globular proteins to 32; this set is referred to as Exp32.

**CD spectra.** CD spectra of the original 29 proteins were kindly provided by Dr. W. C. Johnson, Jr. The CD spectra of the three globular proteins added to the set of reference proteins were taken from Sreerama *et al.* (19). CD spectra of five denatured proteins were added to Exp32, bringing the number of CD spectra to 37, and this set is referred to as Exp37. The CD spectra of acid-denatured bovine apocytochrome *c* at 5 and 90°C, acid-denatured staphylococcal nuclease at 6 and 70°C, and acid-denatured oxidized ribonuclease at 20°C were taken from Privalov *et al.* (31). These five CD spectra are shown in Fig. 1.

The CD spectra of ribonuclease T1 (RNase T1) at pH 5.0 and at different temperatures were measured at the Mayo Foundation on a J-710 spectropolarimeter in quartz cells of 0.01- to 0.02-cm pathlengths, using the following parameters: 2-s response; 20 nm/min scan speed; 0.1-nm data acquisition interval; 5 accumulations; 2-nm bandwidth. The CD spectra were smoothed using the noise reduction routines provided with the J-710 spectropolarimeter. CD spectra of RNaseT1 at different temperatures are given in Fig. 2.



**FIG. 2.** CD spectra of ribonuclease T1 during the course of thermal denaturation in 20 mM NaAc buffer at pH 5.0. The temperatures (°C) are shown along with the spectra.

**Secondary structure.** The secondary structure assignments from DSSP (32) were used to determine the secondary structure fractions of the globular proteins in the reference set as described by Sreerama *et al.* (19). The  $\alpha$ -helical and  $\beta$ -sheet structures were split into regular and distorted classes by considering four residues per  $\alpha$  helix and two residues per  $\beta$  strand distorted. Our grouping of DSSP assignments gave us six secondary structural classes: regular  $\alpha$  helix,  $\alpha_R$ ; distorted  $\alpha$  helix,  $\alpha_D$ ; regular  $\beta$  strand,  $\beta_R$ ; distorted  $\beta$  strand,  $\beta_D$ ; turns, T; and unordered, U. The secondary structure of the denatured proteins was approximated to be 90% unordered and 2% of each of the remaining secondary structures, since structures for the denatured proteins are unavailable.

**CD analysis.** The protein to be analyzed for secondary structure fractions was removed from the reference protein set and the secondary structure fractions were determined from its CD spectrum using the other members of the reference set, following the self-consistent method, SELCON3. In the self-consistent method (15) 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 is solved by the singular value decomposition (SVD) algorithm (33), with variable selection (9) in the locally linearized model (11). Solutions are obtained by varying the reference proteins and/or the number of SVD components retained.

The performance of the analysis was characterized by RMS deviations ( $\delta$ ) and correlation coefficients ( $r$ ) between the X-ray (DSSP, 32) and CD estimates of secondary structure fractions for different secondary structure types. These are denoted by  $\delta_k$  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  $\delta$  and  $r$ .

The RMS deviations and correlation coefficients were calculated using the equations

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

and

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

where  $f_i^{\text{CD}}$  and  $f_i^{\text{X}}$  are CD and X-ray estimates for secondary structure types of  $N$  reference proteins.

## RESULTS AND DISCUSSION

### SELCON3 Program

The variable selection procedure as implemented in the self-consistent method yields a set of solutions as the number of reference proteins and the number of SVD components are varied. Four selection criteria are used to find valid solutions among the solutions given by the self-consistent method. They are: Sum rule,  $|\sum f_k - 1.0| \leq 0.05$ , where  $f_k$  is the fraction of secondary structure  $k$ ; Fraction rule,  $f_k \geq -0.025$ ; Spectral rule,  $\delta_{\text{CD}}$  (RMS difference between the experimental spectrum and the calculated spectrum)  $\leq 0.25\Delta\epsilon$ ; and Helix rule (20), fraction of helix ( $f_\alpha$ ) is in a range determined from the helix fraction obtained with all reference proteins in the analysis ( $f_{\alpha\text{HJ}}$ ). The first two criteria have been derived by relaxing the physically meaningful constraints requiring that each fraction should be positive and the sum of all fractions should be unity.

These selection criteria are incorporated in a step-wise manner in SELCON3 as follows: (a) Considering all proteins and five SVD components, the helix fraction ( $f_{\alpha\text{HJ}}$ ) is determined by the SELCON method. The output corresponds to that of the original method of Hennessey and Johnson (8). (b) The self-consistent method, with the incorporation of variable selection in the locally linearized formalism (11), is used to estimate all secondary structure fractions. Valid solutions satisfy the Sum rule and the Fraction rule. If at least one solution is not obtained, the rules are relaxed. The process is iterated for self-consistency. The results correspond approximately to those from the original SELCON (15) and SELCON1 (18) programs. (c) Solutions obtained from the SELCON method, at the end of step b, are screened using the Spectral rule. If at least one solution is not obtained, this rule is relaxed. The results are similar to those from SELCON2 program (Sreerama and Woody, unpublished results). (d) The Helix rule (20) is now applied to further screen solutions from step c. From  $H_{\text{min}}$ ,  $H_{\text{max}}$ , and  $H_{\text{ave}}$ , determined from solutions at the end of step c, and the helix content from step a,  $f_{\alpha\text{HJ}}$ , helix limits for valid solutions are determined. If  $f_{\alpha\text{HJ}} > 0.65$ , then valid solutions have  $f_\alpha > 0.65$ . If  $0.65 > f_{\alpha\text{HJ}} > 0.25$ , then valid solutions have  $f_\alpha = (f_{\alpha\text{HJ}} + H_{\text{max}})/2 \pm 0.03$ . If  $0.25 > f_{\alpha\text{HJ}} > 0.15$ , then valid solutions have  $f_\alpha = (f_{\alpha\text{HJ}} + H_{\text{ave}})/2 \pm 0.03$ . If  $f_{\alpha\text{HJ}} < 0.15$ , then valid solutions have  $f_\alpha = (f_{\alpha\text{HJ}} + H_{\text{min}})/2 \pm 0.03$ .

Performance indices, RMS differences, and correlation coefficients between the X-ray and CD estimates of secondary structure fractions from the SELCON3 and SELCON2 methods, obtained using CD spectra in the

TABLE 1

Comparison of Performance Indices of SELCON2 and SELCON3 Methods for Org29 Set of Reference Proteins<sup>a</sup>

Method	$\alpha_R$		$\alpha_D$		$\beta_R$		$\beta_D$		T		U		$\delta$	$r$
	$\delta_{\alpha R}$	$r_{\alpha R}$	$\delta_{\alpha D}$	$r_{\alpha D}$	$\delta_{\beta R}$	$r_{\beta R}$	$\delta_{\beta D}$	$r_{\beta D}$	$\delta_T$	$r_T$	$\delta_U$	$r_U$		
SELCON2	0.054	0.946	0.052	0.717	0.087	0.646	0.034	0.742	0.061	0.500	0.103	0.249	0.069	0.817
SELCON3	0.051	0.950	0.049	0.737	0.084	0.655	0.033	0.744	0.062	0.446	0.101	0.294	0.067	0.824

<sup>a</sup>  $\delta$ , root mean square deviation;  $r$ , correlation coefficient;  $\alpha_R$ , regular  $\alpha$ -helix;  $\alpha_D$ , distorted  $\alpha$ -helix;  $\beta_R$ , regular  $\beta$ -strand;  $\beta_D$ , distorted  $\beta$ -strand; T, turns; U, unordered.

range 260–178 nm and the Org29 reference set, are compared in Table 1. These were obtained by removing each reference protein from the set and analyzing it using other members of the reference set. The performance indices for each of the secondary structures (regular  $\alpha$  helix, distorted  $\alpha$  helix, regular  $\beta$  strand, distorted  $\beta$  strand, turns, and unordered) as well as the overall performance indices (calculated by considering all secondary structure fractions) are given. SELCON3 performs slightly better than SELCON2 for all structures except turns. The main difference between SELCON2 and SELCON3 methods is that SELCON3 has an additional selection rule, the Helix rule, which screens solutions based on the total  $\alpha$ -helix content, leading to selection of solutions that have a smaller range of  $\alpha$ -helical content. The approximate inverse relation (34) between the total  $\alpha$ -helical and total  $\beta$ -sheet content in proteins results in selection of solutions that have similar total  $\beta$ -sheet content. Selection of solutions that have a smaller range of total  $\alpha$ -helical and total  $\beta$ -sheet contents by the Helix rule in SELCON3 is the reason for its better performance.

#### Performance of Different Reference Protein Sets

Performance indices for the three reference protein sets (Org29, Exp32, and Exp37), obtained using SELCON3, are compared in Table 2. Exp32 and Exp37 were constructed by adding three and eight additional proteins, respectively, to Org29. The CD spectra of

these additional proteins were in the 240- to 185-nm range, and the analyses reported in Table 2 were performed with CD spectra in this reduced wavelength range. As expected, the performance indices obtained using CD spectra of Org29 proteins in the 240- to 185-nm range are inferior to those obtained with the 260- to 178-nm range (Table 1). Reduction in the wavelength range, particularly in the short-wavelength region, results in a reduction in the information content of the CD spectrum and, in general, the analysis gives poorer results.

With the addition of reference proteins to Org29, the performance indices for the  $\alpha$ -helical fractions (for  $\alpha_R$  and  $\alpha_D$ :  $\delta$ , ~0.058 to ~0.047) and turns ( $\delta_T$ , ~0.080 to ~0.060) improved, and those for  $\beta$ -strand fractions either remained the same (Exp37:  $\delta_{\beta R}$ , 0.081;  $\delta_{\beta D}$ , 0.036) or became worse (Exp32:  $\delta_{\beta R}$ , 0.095;  $\delta_{\beta D}$ , 0.040), while those for the unordered fraction became worse ( $\delta_U$ , ~0.120). The overall performance indices for Exp37 ( $\delta$ , 0.072;  $r$ , 0.899) were slightly better than those for Org29 ( $\delta$ , 0.073;  $r$ , 0.795) and Exp32 ( $\delta$ , 0.073;  $r$ , 0.805). With the addition of denatured proteins, the average content of each secondary structure in the reference proteins was reduced by ~10%, while that of the unordered fraction increased by ~25%. The changes in the average content of secondary structures can partially explain the observed changes in the performance indices of individual secondary structures. For example, an increase of ~25% and a decrease of

TABLE 2

RMS Differences and Correlation Coefficients between the CD-Estimated and X-Ray-Determined Secondary Structure Fractions for the Three Reference Protein Sets<sup>a</sup>

Reference protein set <sup>b</sup>	$\alpha_R$		$\alpha_D$		$\beta_R$		$\beta_D$		T		U		$\delta$	$r$
	$\delta_{\alpha R}$	$r_{\alpha R}$	$\delta_{\alpha D}$	$r_{\alpha D}$	$\delta_{\beta R}$	$r_{\beta R}$	$\delta_{\beta D}$	$r_{\beta D}$	$\delta_T$	$r_T$	$\delta_U$	$r_U$		
Org29	0.057	0.939	0.058	0.661	0.084	0.653	0.034	0.747	0.080	0.226	0.105	0.309	0.073	0.795
Exp32	0.049	0.958	0.049	0.751	0.095	0.646	0.040	0.657	0.059	0.558	0.115	0.167	0.073	0.805
Exp37	0.047	0.960	0.047	0.783	0.081	0.762	0.036	0.735	0.064	0.687	0.122	0.848	0.072	0.899

<sup>a</sup> Wavelength range used (240–185 nm) is smaller than that in Table 1.

<sup>b</sup> Statistics for Exp37 includes results for the five denatured proteins in the reference set.

TABLE 3  
RMS Differences and Correlation Coefficients between the CD-Estimated and X-Ray-Determined Secondary Structure Fractions for the 29 Proteins of Org29 Set<sup>a</sup>

Reference protein set	$\alpha_R$		$\alpha_D$		$\beta_R$		$\beta_D$		T		U		$\delta$	$r$
	$\delta_{\alpha_R}$	$r_{\alpha_R}$	$\delta_{\alpha_D}$	$r_{\alpha_D}$	$\delta_{\beta_R}$	$r_{\beta_R}$	$\delta_{\beta_D}$	$r_{\beta_D}$	$\delta_T$	$r_T$	$\delta_U$	$r_U$		
Org29	0.057	0.939	0.058	0.661	0.084	0.653	0.034	0.747	0.080	0.226	0.105	0.309	0.073	0.795
Exp32	0.051	0.950	0.051	0.722	0.091	0.592	0.041	0.620	0.059	0.518	0.119	0.143	0.074	0.793
Exp37	0.052	0.948	0.053	0.708	0.080	0.688	0.038	0.659	0.067	0.417	0.123	0.360	0.074	0.805

<sup>a</sup> CD data in the 240- to 185-nm wavelength range was used.

~10% in the average contents of unordered structure and helical structures, respectively, can be correlated to the increase of ~18% in  $\delta_U$  and the decrease of ~20% in  $\delta_{\alpha_R}$  and  $\delta_{\beta_D}$ .

Another way of examining the performance of these reference protein sets is by analyzing the CD spectra of a given set of proteins using all three reference sets and comparing the results. Such a comparison is presented in Table 3, where the Org29 proteins were analyzed with all three reference protein sets and the performance indices obtained from each of the three sets is reported. An improvement in the estimation of  $\alpha$ -helical and turns fractions was obtained with Exp32 (for  $\alpha_R$  and  $\alpha_D$ :  $\delta$ , ~0.051;  $\delta_T$ , ~0.059), in comparison with Org29 (for  $\alpha_R$  and  $\alpha_D$ :  $\delta$ , ~0.058;  $\delta_T$ , ~0.080), which comes at the expense of the estimates of  $\beta$ -strand fractions ( $\delta_{\beta_R}$ , 0.091;  $\delta_{\beta_D}$ , 0.041). The performance indices for the regular  $\beta$ -strand fraction obtained with Exp37 ( $\delta_{\beta_R}$ , 0.080) was the best among the three reference protein sets. The estimates obtained with Exp37 were comparable to the best estimates with Org29 and Exp32, except for the unordered fraction.

In developing the Exp37 reference protein set, we have assigned the denatured proteins 90% unordered structure; the 2% assigned to each of the other five structures is less than the residual error normally observed in these types of analyses. In doing so, we are not assuming that the CD spectrum of the static unordered component of globular proteins is identical to the CD spectrum of the dynamically unordered denatured proteins. We are assuming, however, that there is sufficient similarity within and between these two types of unordered structures to permit a successful CD analysis using a method based upon flexible basis spectra. The CD spectrum for a given secondary structural class averaged over a set of reference proteins can be calculated (35). Although the CD spectra obtained for the unordered component of the globular proteins show substantial variation (11, 35, and present work, data not shown) with the set of reference proteins, they have in common a strong negative band in the 195- to 200-nm region. This feature is also the most prominent characteristic of the CD spectra of the denatured pro-

teins (Fig. 1). It is generally attributed to the presence of a significant fraction of residues in the poly(Pro)II conformation in the unordered regions of globular proteins and in denatured proteins and dynamically disordered synthetic polypeptides (36). The range of CD spectra for the unordered regions of globular proteins plus denatured proteins is unlikely to exceed that for  $\beta$ -turns and perhaps even for  $\beta$ -sheets.

It is important to distinguish between the unordered structural component in a globular protein and the unordered conformation in a denatured protein. The former is a relatively static structure, undergoing the fluctuations typical of the polypeptide chain in a globular protein. It is defined as the part of the chain that is left over after the well-defined secondary structural elements ( $\alpha$  helix,  $\beta$  sheet, and  $\beta$  turn) are excluded. The  $\phi, \psi$  angles for residues in this unordered component are distributed over the Ramachandran map. The unordered conformation in a denatured protein, by contrast, is a highly dynamic structure, undergoing large-scale fluctuations in conformation. The  $\phi, \psi$  angles for residues in a denatured protein are also distributed over the Ramachandran map, but in addition these angles vary rapidly with time, undergoing frequent transitions between various regions of the map. A denatured protein is sometimes referred to as a random coil, but in a random coil there is no correlation between nearest neighbor monomer units, whereas denatured proteins do exhibit some correlation (37, 38)

The validity of incorporating the denatured proteins in the reference set is supported by the results obtained for globular proteins using the Exp37 reference set, which are slightly better than those obtained with the reference sets lacking these proteins. A stronger argument for the expanded reference set is provided by the analysis of denatured proteins. Table 4 shows the secondary structure content for the five denatured proteins obtained using the three sets of reference proteins. Using reference sets consisting only of globular proteins, the five denatured proteins are predicted to average about 10%  $\alpha$  helix, 20%  $\beta$  sheet, 25%  $\beta$  turns, and 45% unordered. The predominance of ordered structures and the minority status of the unordered

TABLE 4  
Analysis of Denatured Protein CD Spectra with Different Reference Protein Sets

	Reference protein set	$\alpha_R$	$\alpha_D$	$\beta_R$	$\beta_D$	T	U
Apocytochrome <i>c</i> (5°C)	Org29	0.026	0.026	0.097	0.072	0.237	0.542
	Exp32	0.030	0.031	0.095	0.073	0.230	0.532
	Exp37	0.017	0.021	0.043	0.036	0.066	0.819
Apocytochrome <i>c</i> (90°C)	Org29	0.051	0.054	0.139	0.095	0.212	0.461
	Exp32	0.051	0.055	0.133	0.094	0.213	0.459
	Exp37	0.038	0.033	0.072	0.053	0.067	0.720
Oxidized ribonuclease (20°C)	Org29	0.026	0.090	0.087	0.061	0.235	0.481
	Exp32	0.035	0.070	0.056	0.075	0.250	0.509
	Exp37	0.016	0.025	0.042	0.041	0.077	0.800
Staphylococcal nuclease (6°C)	Org29	0.026	0.056	0.110	0.081	0.211	0.486
	Exp32	0.030	0.044	0.023	0.067	0.241	0.571
	Exp37	0.002	0.019	0.008	0.024	0.051	0.899
Staphylococcal nuclease (70°C)	Org29	0.056	0.058	0.177	0.092	0.215	0.409
	Exp32	0.056	0.058	0.176	0.092	0.215	0.409
	Exp37	0.043	0.053	0.079	0.048	0.079	0.698

conformation is difficult to reconcile with the usual picture of denatured proteins. Using Exp37, the five denatured proteins average about 6%  $\alpha$ -helix, 11%  $\beta$ -sheet, 8%  $\beta$ -turns, and 75% unordered conformation.

Comparison of the analyses of high- and low-temperature CD data for apocytochrome *c* and staphylococcal nuclease using Exp37 reference set does reveal a remaining defect in the analyses. The high-temperature results give larger fractions of the defined structures and smaller fraction of the unordered structure, relative to the low-temperature results. This counterintuitive result is due to a redistribution of residues in the unordered category over the Ramachandran map as the temperature increases. Residues in the  $P_{II}$  conformation, dominant at low temperatures, move into the  $\alpha$ -helix and  $\beta$ -sheet regions of  $\phi, \psi$  space at higher temperatures. This problem could probably be avoided by distinguishing  $P_{II}$  from a truly unordered conformation (17, 20). However, this would require a reliable estimate of the  $P_{II}$  content of the denatured proteins in the reference set, which is not available. Despite this undesirable temperature dependence, we believe that the

present method represents a distinct advance in the analysis of the unordered conformations of native proteins and proteins in various stages of unfolding.

#### Applications

With the inclusion of five denatured CD spectra, Exp37 is particularly useful in the analysis of protein CD spectra that are suspected to have a significant unordered component. We have examined the usefulness of Exp37 in analyzing the CD spectra of RNaseT1 in various stages of thermal unfolding by analyzing CD data at different temperatures using both Org29 and Exp37 reference protein sets.

The CD spectra of RNaseT1 at different temperatures (2 to 93°C) are shown in Fig. 2. The results of analysis of these CD spectra with Org29 and Exp37 are given in Tables 5 and 6, respectively. There are systematic changes in the secondary structure fractions with temperature. As the temperature increases, the unordered fraction increases and the helical and strand fractions decrease. The extents of

TABLE 5  
CD Analysis of RNaseT1 CD Spectra at Different Temperatures from Org29 Reference Protein Set

Temperature (°C)	$\alpha_R$	$\alpha_D$	$\beta_R$	$\beta_D$	T	U	$\Sigma f$	Number of segments	
								$\alpha$ -helix	$\beta$ -strand
2	0.046	0.063	0.188	0.120	0.231	0.356	1.003	1.63	6.22
37	0.040	0.058	0.181	0.116	0.226	0.348	0.969	1.50	6.01
55	0.038	0.055	0.165	0.106	0.212	0.336	0.913	1.44	5.52
58	0.030	0.048	0.166	0.108	0.229	0.377	0.958	1.24	5.61
61	0.026	0.040	0.156	0.104	0.243	0.444	1.013	1.04	5.43
72	0.030	0.042	0.133	0.095	0.238	0.476	1.013	1.08	4.93
93	0.038	0.044	0.122	0.093	0.231	0.478	1.006	1.13	4.84

TABLE 6

CD Analysis of RNaseT1 CD Spectra at Different Temperatures from Exp37 Reference Protein Set

Temperature (°C)	$\alpha_R$	$\alpha_D$	$\beta_R$	$\beta_D$	T	U	$\Sigma f$	Number of segments	
								$\alpha$ -helix	$\beta$ -strand
2	0.045	0.057	0.181	0.106	0.192	0.410	0.991	1.49	5.49
37	0.036	0.056	0.186	0.115	0.213	0.392	0.999	1.47	6.00
55	0.035	0.052	0.164	0.105	0.196	0.408	0.959	1.35	5.46
58	0.023	0.044	0.150	0.102	0.206	0.460	0.985	1.15	5.28
61	0.017	0.033	0.102	0.076	0.163	0.588	0.979	0.85	3.96
72	0.022	0.028	0.073	0.060	0.103	0.701	0.987	0.72	3.10
93	0.028	0.031	0.073	0.059	0.100	0.688	0.978	0.80	3.07

decrease of the helical and strand fractions predicted from the two analyses are, however, different. While Exp37 predicts a decrease of more than 50% of both  $\alpha$  helix (0.102 to 0.050) and  $\beta$  strand (0.301 to 0.132), Org29 predicts only a 30–40% reduction ( $\alpha$  helix: 0.109 to 0.066;  $\beta$  strand: 0.308 to 0.215). The predicted changes in the number of segments also indicate similar differences. Exp37 predicts a larger reduction in the number of segments ( $\alpha$  helix: 1.5 to 0.8;  $\beta$  strand: 5.5 to 3.1) than Org29 ( $\alpha$  helix: 1.6 to 1.1;  $\beta$  strand: 6.2 to 4.8). The turns fraction decreases with temperature according to the analysis with Exp37 (0.213 to 0.101) while it remains about the same in the analysis with Org29 ( $\sim$ 0.230). In general, Exp37 predicts a reduction in both the number of segments and the total content of  $\alpha$ -helix and  $\beta$ -strands, as well as turns, with an increase of unordered structure.

The CD analysis of thermal unfolding of RNaseT1 can be interpreted as melting of all structures with increasing temperature, accompanied by an increase of unordered fraction. Our results disagree with the analysis of thermal unfolding of RNaseT1 by VCD and FTIR spectra carried out by Pancoska *et al.* (39), who concluded that thermal melting of RNaseT1 results in the loss of  $\alpha$ -helical component, an increase in the number of  $\beta$ -strand segments, and minimal changes in turns and unordered structures. The FTIR analysis of Fabian *et al.* (40) indicated a substantial loss of  $\alpha$ -helical and  $\beta$ -strand components, similar to our results. The hydrogen–deuterium exchange kinetics study by Mullins *et al.* (41) suggested a global unfolding pathway, with an initial melting of all  $\alpha$  helix and some  $\beta$  sheet.

The inclusion of denatured proteins in the set of reference proteins provides a better representation of the unordered structure and should be particularly useful in the analysis of CD spectra of proteins with a larger unordered component, including partially unfolded proteins.

## SUMMARY AND CONCLUSIONS

Secondary structure estimation from optical spectroscopic studies (such as CD, IR, vibrational CD, Raman) is empirical, and largely depends on the proper selection of reference proteins. A “good” reference protein set should be representative of the variables in the data. Current reference protein sets provide good representations of the range of major secondary structure contents in proteins, such as  $\alpha$  helix ( $f_\alpha$ : 0.0 to 0.80) and  $\beta$  sheet ( $f_\beta$ : 0.0 to 0.49), but that of unordered structure is less varied ( $f_U$ : 0.13 to 0.39). In one case,  $\alpha$ -bungarotoxin with  $f_U = 0.61$  was included (W. C. Johnson, Jr., personal communication). Our present study is an attempt to better represent the unordered structure in the CD analysis by including denatured proteins in the set of reference proteins, and approximating the structure as 90% unordered.

It has been shown that inclusion of four denatured proteins in the set of 16 native reference proteins improves the performance of protein secondary structure determination by CD spectroscopy (30). Analogous results have been obtained by vibrational spectroscopy in the analysis of protein secondary structure by infrared absorption (S. Yu. Venyaminov, unpublished results; 42). Secondary structure estimation for native globular proteins by SELCON3, including denatured proteins in the reference protein set, gives a slight improvement in some respects and, at worst, a slight deterioration in others. Taken together, these results suggest that although the unordered regions of globular proteins are likely to make CD contributions that are not identical to those of dynamically disordered proteins and polypeptides, there is sufficient similarity to warrant their treatment as a single category in methods that utilize a flexible basis set (9, 11, 14–20). The reference protein set that includes both native and denatured proteins should be particularly useful in the analysis of CD spectra with larger unordered component, such as proteins in the course of denaturation.

The SELCON3 program, different reference sets, and related data files are available at the websites:

<http://lamar.colostate.edu/~sreeram/SELCON3> and  
<http://lamar.colostate.edu/~sreeram/CDPro>,

## 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 Professor 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

- Greenfield, N., and Fasman, G. D. (1969) Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry* **8**, 4108–4116.
- Saxena, V. P., and Wetlaufer, D. B. (1971) A new basis for interpreting the circular dichroism spectra of proteins. *Proc. Natl. Acad. Sci. USA* **68**, 969–972.
- Chen, Y. H., and Yang, J. T. (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, Y. H., Yang, J. T., and Martinez, H. M. (1972) Determination of the secondary structures of proteins by circular dichroism and optical rotatory dispersion. *Biochemistry* **11**, 4120–4131.
- Bolotina, I. A., Chekhov, V. O., Lugauskas, V. Y., Finkel'shtein, A. V., and Ptitsyn, O. B. (1980) Determination of the secondary structure of proteins from the circular dichroism spectra. 1. Protein reference spectra for  $\alpha$ -,  $\beta$ - and irregular structures. *Mol. Biol.* **14**, 701–709. [English translation]
- Brahms, S., and Brahms, J. (1980) Determination of protein secondary structure in solution by vacuum ultraviolet circular dichroism. *J. Mol. Biol.* **138**, 149–178.
- Provencher, S. W., and Glöckner, J. (1981) Estimation of protein secondary structure from circular dichroism. *Biochemistry* **20**, 33–37.
- Hennessey, J. P., Jr., and Johnson, W. C., Jr. (1981) Information content in the circular dichroism of proteins. *Biochemistry* **20**, 1085–1094.
- Manavalan, P., and Johnson, W. C., Jr. (1987) Variable selection method improves the prediction of protein secondary structure from circular dichroism. *Anal. Biochem.* **167**, 76–85.
- Shubin, V. V., Khazin, M. L., and Efimovskaya, T. B. (1990) Prediction of protein secondary structure of globular proteins using circular dichroism spectra. *Mol. Biol.* **24**, 165–176. [English translation]
- van Stokkum, I. H. M., Spoelder, H. J. W., Bloemendal, M., van Grondelle, R., and Groen, F. C. A. (1990) Estimation of protein secondary structure and error analysis from circular dichroism spectra. *Anal. Biochem.* **191**, 110–118.
- Perczel, A., Hollosi, M., Tusnady, G., and Fasman, G. D. (1991) Convex constraint analysis: A natural deconvolution of circular dichroism curves of proteins. *Protein Eng.* **4**, 669–679.
- Pancoska, P., and Keiderling, T. A. (1991) Systematic comparison of statistical analysis of electronic and vibrational circular dichroism for secondary structure prediction of selected proteins. *Biochemistry* **30**, 6885–6895.
- Böhm, G., Muhr, R., and Jaenicke, R. (1992) Quantitative analysis of protein far UV circular dichroism spectra by neural networks. *Protein Eng.* **5**, 191–195.
- Sreerama, N., and Woody, R. W. (1993) A self-consistent method for the analysis of protein secondary structure from circular dichroism. *Anal. Biochem.* **209**, 32–44.
- Andrade, M. A., Chacán, P., Merolo, J. J., and Morán, F. (1993) Evaluation of secondary structure of protein from UV circular dichroism spectra using unsupervised learning neural network. *Protein Eng.* **6**, 383–390.
- Sreerama N., and Woody, R. W. (1994) Poly(Pro)II helices in globular proteins: Identification and circular dichroic analysis. *Biochemistry* **33**, 10022–10025.
- Sreerama, N., and Woody, R. W. (1994) 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.
- Sreerama, N., Venyaminov, S. Y., and Woody, R. W. (1999) Estimation of the number of  $\alpha$ -helical and  $\beta$ -strand segments in proteins using circular dichroism spectroscopy *Protein Sci.* **8**, 370–380.
- Johnson, W. C., Jr. (1999) Analyzing protein circular dichroism spectra for accurate secondary structures. *Proteins: Struct. Funct. Genet.* **35**, 307–312.
- Yang, J. T., Wu, C.-S. C., and Martinez, H. M. (1986) Calculation of protein conformation from circular dichroism. *Methods Enzymol.* **130**, 208–269.
- Johnson, W. C., Jr. (1988) Secondary structure of proteins through circular dichroism spectroscopy. *Annu. Rev. Biophys. Biophys. Chem.* **17**, 145–166.
- Venyaminov, S. Y., and Yang, J. T. (1996) Determination of protein secondary structure. In *Circular Dichroism and the Conformational Analysis of Biomolecules* (Fasman, G. D., Ed.), pp. 69–107. Plenum, New York.
- Greenfield, N. J. (1996) Methods to estimate the conformation of proteins and polypeptides from circular dichroism data. *Anal. Biochem.* **235**, 1–10.
- Sreerama, N., and Woody, R. W. (2000) Circular dichroism of peptides and proteins. In *Circular Dichroism: Principles and Applications* (Berova, N., Nakanishi, K., and Woody, R. W., Ed.), 2nd ed., pp. 601–620. Wiley, New York.
- Venyaminov, S. Y., and Vassilenko, K. S. (1994) Determination of protein tertiary structure class from circular dichroism spectra. *Anal. Biochem.* **222**, 176–184.
- Manavalan, P., and Johnson, W. C., Jr. (1983) Sensitivity of circular dichroism to protein tertiary structure class. *Nature* **305**, 831–832.
- Woody, R. W., and Dunker, A. K. (1996) Aromatic and cystine side-chain circular dichroism in proteins. In *Circular Dichroism and the Conformational Analysis of Biomolecules* (Fasman, G. D., Ed.), pp. 109–157. Plenum, New York.
- Manning, M. C., Illangasekare, M., and Woody, R. W. (1988) Circular dichroism studies of distorted  $\alpha$ -helices, twisted  $\beta$ -sheets, and  $\beta$ -turns. *Biophys. Chem.* **31**, 77–86.
- Venyaminov, S. Y., Baikalov, I. A., Shen, Z. M., Wu, C. S. C., and Yang, J. T. (1993) Circular dichroic analysis of denatured proteins: Inclusion of denatured protein in the reference set. *Anal. Biochem.* **214**, 17–24.
- Privalov, P. L., Tiktópulo, E. I., Venyaminov, S. Y., Griko, Y. V., Makhatadze, G. I., and Khechinashvili, N. N. (1989) Heat capacity and conformation of proteins in the denatured state. *J. Mol. Biol.* **204**, 737–750.
- Kabsch, W., and Sander, C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometric features. *Biopolymers* **22**, 2577–2637.

33. Forsythe, G. E., Malcolm, M. A., and Moler, C. B. (1977) *Computer Methods for Mathematical Computations*, Prentice Hall, Englewood Cliffs, NJ.
34. Pancoska, P., Blazek, M., and Keiderling, T. A. (1992) Relationships between secondary structure fractions for globular proteins. Neural network analysis of crystallographic data. *Biochemistry*, **31**, 10250–10257.
35. Compton, L. A., and Johnson, W. C., Jr. (1986) Analysis of protein circular dichroism spectra for secondary structure using a simple matrix multiplication. *Anal. Biochem.* **155**, 155–167.
36. Woody, R. W. (1992) Circular dichroism and conformation of unordered peptides. *Adv. Biophys. Chem.* **2**, 37–79.
37. Wüthrich, K. (1994) NMR assignments as a basis for structural characterization of denatured states of globular proteins. *Curr. Opin. Struct. Biol.* **4**, 93–99.
38. Barbar, E. (1999) NMR characterization of partially folded and unfolded conformational ensembles of proteins. *Biopolymers* **51**, 191–207.
39. Pancoska, P., Fabian, H., Yoder, G., Baumruk, V., and Keiderling, T. A. (1996) Protein structural segments and their interconnections derived from optical spectra. Thermal unfolding of ribonuclease T1 as an example. *Biochemistry* **35**, 13094–13106.
40. Fabian, H., Schultz, C., Naumann, D., Landt, O., Hahn, U., and Saenger, W. (1993) Secondary structure and temperature induced unfolding and refolding of ribonuclease T1 in aqueous solution—A Fourier transform spectroscopic study. *J. Mol. Biol.* **232**, 967–981.
41. Mullins, L. S., Pace, C. N., and Raushel, F. M. Conformational stability of ribonuclease T1 determined by hydrogen–deuterium exchange. *Protein Sci.* **6**, 1387–1395.
42. Kalnin, N. N., Baikalov, I. A., and Venyaminov, S. Yu. (1990) Quantitative IR spectrophotometry of peptide compounds in water solution. III. Estimation of protein secondary structure. *Biopolymers* **30**, 1273–1280.