

Estimation of Protein Secondary Structure from Circular Dichroism Spectra: Comparison of CONTIN, SELCON, and CDSSTR Methods with an Expanded Reference Set

Narasimha Sreerama and Robert W. Woody¹

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

Received April 7, 2000

We have expanded the reference set of proteins used in SELCON3 by including 11 additional proteins (selected from the reference sets of Yang and co-workers and Keiderling and co-workers). Depending on the wavelength range and whether or not denatured proteins are included in the reference set, five reference sets were constructed with the number of reference proteins varying from 29 to 48. The performance of three popular methods for estimating protein secondary structure fractions from CD spectra (implemented in software packages CONTIN, SELCON3, and CDSSTR) and a variant of CONTIN, CONTIN/LL, that incorporates the variable selection method in the locally linearized model in CONTIN, were examined using the five reference sets described here, and a 22-protein reference set. Secondary structure assignments from DSSP were used in the analysis. The performances of all three methods were comparable, in spite of the differences in the algorithms used in the three software packages. While CDSSTR performed the best with a smaller reference set and larger wavelength range, and CONTIN/LL performed the best with a larger reference set and smaller wavelength range, the performances for individual secondary structures were mixed. Analyzing protein CD spectra using all three methods should improve the reliability of predicted secondary structural fractions. The three programs are provided in CDPro software package and have been modified for easier use with the different reference sets described in this paper. CDPro software is available at the website: <http://lamar.colostate.edu/~sreeram/CDPro>. © 2000 Academic Press

Key Words: protein secondary structure; CD analysis; protein CD; CDPro software.

The far-UV CD of a protein generally reflects its secondary structure content. One of the most successful applications of CD, the structural characterization of proteins, depends upon the remarkable sensitivity of far-UV CD to the backbone conformation of proteins. Various empirical methods have been developed for analyzing protein CD spectra for quantitative estimation of the secondary structure content (1–20). 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_λ) can be expressed as a linear combination of 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 . The validity and the limitations of this equation have been discussed in the preceding paper (21). Various aspects of secondary structure analysis have been reviewed (22–26). 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 (27, 28).

Different sets of reference proteins, with 15 to 33 proteins and having a good representation of α -rich, β -rich, and mixed- $\alpha\beta$ proteins, have been used in secondary structure analyses. Denatured proteins have also been included in these reference sets (21, 29). Different methods for analyzing protein CD spectra have been developed, and a different reference protein set has generally been used in each of these methods. The use of the different methods in protein CD analysis has been limited to the reference proteins and the secondary structure assignments used in the original publications. It is often difficult to compare the results from different methods because of the different reference sets and different secondary structure assignments used in those methods of analysis. It would also be erroneous to average the results from different

¹ To whom correspondence should be addressed. Fax: (970) 491-0494. E-mail: rww@lamar.colostate.edu.

methods unless secondary structure assignments from the same algorithm [for example, DSSP² (30)] are used in the analyses.

Performing CD analysis using different methods should help improve the reliability of predicted structural features. However, care should be taken to ensure that the protein reference set and the secondary structure assignments used are the same in the methods examined. Most straightforward applications of protein CD analysis use the computer programs implementing a method of analysis with the spectra and structural assignments from the original publications, making the use of results from different methods for improved reliability difficult.

For a good analysis, it is desirable to have a reference protein set with the largest possible representation of CD spectral features and secondary structural combinations. The most widely used reference sets are from Yang and co-workers, Johnson and co-workers, and Keiderling and co-workers. While there are proteins that are common to all three reference sets, some proteins are found in only one of the reference sets. Construction of a larger reference set of proteins by combining proteins from the widely used reference sets should increase the range of spectral data used in the analysis.

In this paper we describe our efforts to achieve these two goals: increasing the range of spectral data used in the analysis and using multiple algorithms for a reliable analysis. We have constructed a reference set of 48 proteins by combining the unique protein CD spectra from four sources, supplemented with secondary structure assignments from Kabsch and Sander's DSSP program (30). Since the reference sets developed from different research groups use different wavelength ranges, the reference set of 48 proteins has the smallest wavelength range of 190–240 nm. We have also constructed four additional reference sets with two different wavelength ranges and including/excluding denatured protein CD spectra. We have also modified three popular programs for CD analysis, CONTIN, SELCON3, and CDSSTR, for use with a reference set of choice. The performances of the three methods are comparable and the use of all three methods is recommended for a reliable analysis.

MATERIALS AND METHODS

CD spectra. We have combined protein CD spectra from different reference sets to construct a reference

set of 48 proteins. Protein CD spectra from five different sources were used. The proteins, the crystal structures used (PDB code in parentheses), and the CD spectral wavelength range are given below.

A larger set of 29 protein CD spectra, in the wavelength range of 178–260 nm, was a gift from W. C. Johnson, Jr. The proteins, and the X-ray structure codes, are: myoglobin (4mbn), hemoglobin (2mhb), hemerythrin (2hmz), T4 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), 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 -bungarotoxin (2abx).

The CD spectra of five proteins, in the wavelength range 185–240 nm, were taken from Pancoska *et al.* (31). The proteins and the X-ray structure codes are: α -chymotrypsinogen (2cga), alcohol dehydrogenase (5adh), carbonic anhydrase (1ca2), glutathione reductase (3grs), and rhodanese (1rhd).

Three protein CD spectra were taken from Sreerama *et al.* (19), also in the wavelength range 185–240 nm. The proteins are: colicin A (1col), green fluorescent protein (1ema), and rat intestinal fatty acid binding protein (1fc).

The CD spectra of six proteins were taken from Yang and co-workers (32) and the spectra are in the wavelength range 190–240 nm. The proteins and the X-ray structure codes are: staphylococcal nuclease (2sns), insulin (4ins), parvalbumin (5cpv), carboxypeptidase A (5cpa), bovine pancreatic trypsin inhibitor (5pti), and adenylate kinase (3adk).

The following five denatured protein CD spectra, in the wavelength range 185–240 nm, were taken from Privalov *et al.* (33): acid-denatured bovine apocytochrome *c* at 5 and 90°C, acid-denatured staphylococcal nuclease at 6 and 70°C, and acid-denatured oxidized bovine ribonuclease at 20°C.

Secondary structure. The secondary structure assignments from DSSP (30) were used to determine the secondary structure fractions of the globular proteins in the reference set, as described by Sreerama *et al.* (19). The α -helix and β -strand structures were split into regular and distorted classes, considering four residues per α -helix and two residues per β -strand distorted. Our grouping of DSSP assignments gave us six secondary structural classes: regular α -helix, α_R ; distorted α -helix, α_D ; regular β -strand, β_R ; distorted

² Abbreviations used: PDB, Protein Data Bank; δ , root mean square deviation; *r*, correlation coefficient; SVD, singular value decomposition; α_R , regular α -helix; α_D , distorted α -helix; β_R , regular β -strand; β_D , distorted β -strand; T, turns; U, unordered; DSSP, a computer program for defining secondary structure of proteins.

β -strand, β_D ; turns, T; and unordered, U. The secondary structure corresponding to the denatured spectra were approximated to be 90% unordered and 2% of each of the remaining five secondary structures, since structures for the denatured proteins are unavailable. The reasons for equating the dynamic denatured structure with the unordered structure of native proteins are discussed in the preceding paper (21).

CD analysis. The CD spectrum of the protein analyzed was removed from the reference set and the secondary structure fractions were determined using the other members of the reference set. Three methods for analyzing protein CD spectra were used, as implemented in computer programs CDSSTR, SELCON3, and CONTIN. Only brief descriptions of the algorithms used in the three methods are given below. Detailed descriptions of these methods are available in the literature.

CDSSTR. This method, developed by Johnson (20), combines many features of previously described methods. One new feature implemented in this method is that only a minimum number of reference proteins (eight in this case) are required for a good analysis. Since one does not know which proteins are essential for analyzing a given CD spectrum, they are selected randomly from the reference set. A large number of combinations of eight proteins can be constructed from a given reference set, giving extreme flexibility to the method, but the solutions may become unstable depending on the eight reference proteins selected. Solutions are obtained in the self-consistent formalism (15) using the singular value decomposition (SVD) algorithm (34) and five SVD components. Solutions from each such combination satisfying the three basic selection rules (the sum of fractions is between 0.95 and 1.05; each fraction is greater than -0.03 ; the RMS deviation between the reconstructed and experimental CD is less than $0.25 \Delta\epsilon$) are considered acceptable. The other new feature in this method is that a certain minimum number of such acceptable solutions (400, in the current implementation) are subjected to another selection rule based on the helical content. The helical content is determined from the helix fraction estimated from the full reference set and the maximum/minimum helix fraction from the acceptable solutions. The final solution is the average of all solutions that satisfy the four selection rules.

SELCON3. This is the latest version of the self-consistent method, SELCON (15). 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 solution replaces the initial guess; and

the process is iterated for convergence. The matrix equation relating the CD spectra to the secondary structure is solved by the singular-value decomposition algorithm (34) and variable selection (9) in the locally linearized model (11); solutions are obtained by varying the reference proteins and/or the SVD coefficients. Acceptable solutions from the different variable selection combinations in the locally linearized model satisfy the three basic selection rules (the sum of fractions is between 0.95 and 1.05; each fraction is greater than -0.025 ; the RMS deviation between the reconstructed and experimental CD is less than $0.25 \Delta\epsilon$). These acceptable solutions are subjected to another selection rule based on the helical content, as defined by Johnson (20). The final solution is the average of all solutions that satisfy the four selection rules.

CONTIN/LL. This is a variant of the CONTIN method developed by Provencher and Glöckner (7). CONTIN uses the ridge regression procedure, which fits the CD spectrum of the test protein (C_λ^{obs}) as a linear combination of the CD spectra of N reference proteins by minimizing the function

$$\sum_{\lambda=1}^n (C_\lambda^{\text{calc}} - C_\lambda^{\text{obs}})^2 + \alpha^2 \sum_{j=1}^N (\nu_j - N^{-1})^2,$$

where C_λ is the spectrum at n wavelengths, α is the regularizer, and ν_j is the coefficient of the CD spectrum for the j th reference protein in the linear combination used to construct the spectrum of the test protein, C_λ^{calc} . The selection rules $f_k \geq 0.0$ and $\sum f_k = 1.0$ are used as constraints. The method gives a range of solutions, depending on the value of α ; smaller values of α give solutions similar to those from the normal least-squares method and larger values tend to give solutions biased toward certain proteins, limiting the number of degrees of freedom. The program CONTIN selects a solution based on several criteria. We found, as have others (23, 35), that this solution is not always the best, and there are solutions closer to the X-ray structure of the test protein that are rejected by the program. Our previous studies (18) have indicated that the solution with the least standard error, from among the set of solutions given by the program CONTIN, gives better results. In CONTIN/LL, the proteins in the reference set are arranged in the order of increasing RMS distance of the CD spectra from that of the protein analyzed, and the more distant proteins are deleted in a systematic manner to construct smaller reference sets. This results in a set of solutions, one for each LL combination, the number of which is determined by the number of reference proteins and the minimum number of proteins used for a solution (six, in our case). The selection rule based on helical content

was used to screen the solutions. The final solution is the average of all solutions that satisfy the four selection rules.

The performance of the analysis was characterized by RMS deviations (δ) and correlation coefficients (r) between the X-ray and CD estimates of secondary structure fractions for different secondary structures. 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 .

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

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

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

RESULTS AND DISCUSSION

We have attempted to improve the reliability of protein CD analysis by using a larger reference set of proteins in the analysis and performing the analysis using more than one method of analysis. A larger reference set gives a better representation of the secondary structural variations in proteins and their influence on CD spectra. Use of multiple methods of analysis, implementing different algorithms to bring out the correspondence between the structural variations and CD spectra, may be required for a good analysis of a given CD spectrum. Results from different methods may also be used as a measure of reliability of the analysis and to determine the error of the analysis.

Reference set. The first step was to construct a larger reference set by combining the proteins from the reference sets currently used in CD analysis. This presented a problem since the wavelength ranges of the CD spectra used in different reference sets were different. The largest wavelength range of 178–260 nm was used in the reference set developed by Johnson and co-workers (29 proteins), and the smallest range of 190–240 nm was used by Yang and co-workers (six proteins). The two other reference sets we considered, from Keiderling and co-workers (five proteins) and

Sreerama *et al.* (three native proteins and five denatured proteins), had the wavelength ranges of 180–250 and 185–240 nm, respectively. For the proteins common in these reference sets, we kept the CD spectrum with the largest wavelength range, which were all from Johnson and co-workers. By combining proteins from these four different sources we constructed the 48-protein reference set, with CD spectra in the wavelength range 190–240 nm. If one considers the CD spectra in the 178- to 260-nm range, only 29 proteins would satisfy the requirement and this forms the 29-protein reference set. We also considered an intermediate range of 185–240 nm, which formed a 42-protein reference set. In the two larger reference sets of 42 and 48 proteins, we have included the five denatured protein CD spectra. By excluding the denatured proteins in the reference set we come up with two additional reference sets of 43 and 37 proteins. As a result, we have five reference sets, depending on the wavelength range and on whether or not the denatured proteins are included in the reference set.

The 42- and 48-protein reference sets include the denatured proteins. The secondary structure of these denatured protein spectra was approximated to be 90% unordered, owing to the spectral similarity of the denatured proteins and the unordered structures, the details of which are presented in the preceding paper (21). The changes in the performance indices for native proteins upon including the denatured proteins are small and not unidirectional.

CD analysis. The next step was to incorporate these reference sets in the computer programs for CD analysis. We considered three programs, CONTIN/LL, SELCON3, and CDSSTR, each using a different algorithm for analyzing a given protein CD spectrum. The three programs were modified so that they all used a single data-file structure.

Comparison of results from these three methods was done by comparing their performance indices. Performance indices, RMS differences (δ) and correlation coefficients (r) between the X-ray and CD estimates of secondary structure fractions, for each method were obtained using a given reference set. These were obtained by removing the CD spectrum of each reference protein from the reference set and analyzing it using other members of the reference set. The performance indices for each of the secondary structures, regular α -helix, distorted α -helix, regular β -strand, distorted β -strand, turns, and unordered, as well as the overall performance indices (calculated by considering all secondary structure fractions) were calculated.

The performance indices for the three methods, obtained for different reference sets, are given in Table 1. The performance indices for the CONTIN method, which corresponds to the solution including all refer-

TABLE 1
Performance of SELCON3, CDSSTR, and CONTIN-LL Programs for Analyzing Protein CD Spectra,
for Different Reference Sets^a

Reference proteins	Method	α_R		α_D		β_R		β_D		T		U		δ	r
		δ_{α_R}	r_{α_R}	δ_{α_D}	r_{α_D}	δ_{β_R}	r_{β_R}	δ_{β_D}	r_{β_D}	δ_T	r_T	δ_U	r_U		
29	SELCON3	0.054	0.946	0.052	0.717	0.087	0.646	0.034	0.742	0.062	0.482	0.101	0.300	0.073	0.795
	CDSSTR	0.050	0.955	0.053	0.805	0.079	0.706	0.029	0.810	0.060	0.536	0.099	0.478	0.066	0.836
	CONTIN	0.046	0.960	0.050	0.727	0.099	0.489	0.031	0.783	0.060	0.476	0.100	0.397	0.070	0.812
	CONTIN-LL	0.050	0.952	0.056	0.695	0.099	0.533	0.034	0.734	0.065	0.448	0.103	0.350	0.072	0.802
37	SELCON3	0.050	0.952	0.043	0.767	0.084	0.705	0.037	0.664	0.056	0.570	0.108	0.154	0.068	0.819
	CDSSTR	0.055	0.946	0.044	0.830	0.096	0.600	0.028	0.811	0.065	0.448	0.101	0.323	0.070	0.809
	CONTIN	0.056	0.940	0.042	0.773	0.101	0.529	0.030	0.787	0.064	0.362	0.087	0.380	0.068	0.814
	CONTIN-LL	0.052	0.948	0.047	0.745	0.098	0.577	0.031	0.763	0.066	0.418	0.094	0.279	0.069	0.811
43	SELCON3	0.053	0.941	0.044	0.776	0.086	0.663	0.031	0.743	0.073	0.367	0.098	0.216	0.068	0.811
	CDSSTR	0.065	0.918	0.045	0.771	0.092	0.611	0.028	0.807	0.068	0.463	0.088	0.369	0.068	0.810
	CONTIN	0.059	0.927	0.046	0.753	0.088	0.631	0.029	0.782	0.078	0.213	0.082	0.397	0.067	0.815
	CONTIN-LL	0.057	0.930	0.043	0.793	0.087	0.649	0.029	0.774	0.077	0.333	0.089	0.253	0.068	0.814
42	SELCON3	0.047	0.956	0.043	0.794	0.082	0.672	0.037	0.690	0.064	0.650	0.140	0.769	0.077	0.873
	CDSSTR	0.052	0.950	0.042	0.847	0.093	0.620	0.029	0.819	0.069	0.585	0.140	0.774	0.080	0.864
	CONTIN	0.054	0.940	0.047	0.755	0.095	0.618	0.032	0.769	0.091	0.187	0.157	0.712	0.090	0.712
	CONTIN-LL	0.049	0.950	0.045	0.781	0.088	0.671	0.030	0.799	0.071	0.575	0.120	0.836	0.074	0.885
48	SELCON3	0.052	0.942	0.044	0.806	0.082	0.694	0.034	0.719	0.076	0.505	0.129	0.775	0.076	0.866
	CDSSTR	0.060	0.930	0.047	0.822	0.087	0.640	0.031	0.770	0.078	0.455	0.135	0.766	0.080	0.852
	CONTIN	0.055	0.934	0.049	0.750	0.091	0.594	0.035	0.685	0.092	0.099	0.154	0.672	0.089	0.814
	CONTIN-LL	0.053	0.941	0.041	0.840	0.081	0.697	0.031	0.765	0.076	0.512	0.114	0.833	0.072	0.884

^a δ , root mean square deviation; r , correlation coefficient; α_R , regular α -helix; α_D , distorted α -helix; β_R , regular β -strand; β_D , distorted β -strand; T, turns; U, unordered.

ence proteins, are given in addition to those for the CONTIN/LL method. The last two columns of Table 1 give the overall performance indices, obtained by considering the secondary structural fractions together, preceded by the performance indices for individual secondary structures.

For the reference sets excluding the denatured proteins, the overall performance indices obtained from all four methods were similar. For the 29-protein reference set, CDSSTR performed the best ($\delta = 0.066$ and $r = 0.836$) and SELCON3 performed the worst ($\delta = 0.073$ and $r = 0.795$). For the 37-protein reference set CDSSTR performed the worst ($\delta = 0.070$ and $r = 0.809$), and SELCON3 and CONTIN performed the best ($\delta = 0.068$ and $r = 0.819$). For the 43-protein reference set all methods gave almost identical performance indices ($\delta = 0.068$ and $r = 0.81$). When denatured proteins were included in the reference set (42- and 48-protein reference sets) CONTIN/LL performed the best ($\delta \approx 0.072$ and $r \approx 0.885$), followed by SELCON3 ($\delta \approx 0.077$ and $r \approx 0.870$) and CDSSTR ($\delta \approx 0.080$ and $r \approx 0.860$).

When we compare the performance indices for the individual secondary structures, we get a slightly different picture. For a given reference set, we find that no method gives the best performance indices for all secondary structures. For example, the superior performance of CDSSTR for the 29-protein reference set

does not translate to a superior performance for all secondary structures; the α -helical fractions were estimated better by the CONTIN method. Similarly for the 37-protein reference set, SELCON3 does better for three of the secondary structures (α_R , β_R , and T), CONTIN for α_D and T, and CDSSTR for β_D . Overall performance indices can be considered as a summary of performance indices of individual secondary structures and these, in general, follow the performances of individual secondary structures. A given method performing better than other methods for more individual secondary structures gives better overall performance indices. However, this does not hold true when we include denatured proteins in the reference set. For the 42-protein reference set CONTIN/LL performs the best only for the unordered fraction, yet gives the best overall performance indices in comparison to other methods, partly because of the larger error in the estimation of unordered fraction by the other methods. SELCON3 performed slightly better than CDSSTR with 42- and 48-protein reference sets that include denatured proteins.

The ridge regression algorithm followed in the CONTIN method gives different weights to different proteins in the reference set in fitting the experimental spectrum. This can be considered as an implicit inclusion of the variable selection principle in CONTIN, which explains the similarity in the results from

TABLE 2

Performance of SELCON3, CDSSTR, and CONTIN-LL Programs for Analyzing Protein CD Spectra, for the 22-Protein Reference Set and King and Johnson's Structural Assignments

Method	α helix		3/10 helix		β sheet		Turns		Poly(Pro)II		Unordered		δ	r
	δ_α	r_α	$\delta_{3/10}$	$r_{3/10}$	δ_β	r_β	δ_T	r_T	δ_{PII}	r_{PII}	δ_U	r_U		
SELCON3	0.052	0.972	0.027	0.572	0.051	0.896	0.036	0.433	0.025	0.779	0.056	0.842	0.043	0.958
CDSSTR	0.038	0.986	0.026	0.635	0.044	0.928	0.041	0.419	0.027	0.759	0.047	0.899	0.039	0.965
CONTIN	0.048	0.976	0.026	0.641	0.060	0.843	0.046	0.351	0.029	0.710	0.044	0.906	0.044	0.956
CONTIN-LL	0.050	0.974	0.027	0.591	0.053	0.880	0.038	0.505	0.027	0.750	0.052	0.869	0.043	0.959

CONTIN and CONTIN/LL, for most cases; CONTIN/LL explicitly includes variable selection in the locally linearized approach. However, the presence of some proteins in the reference set can have an adverse effect on the solution from CONTIN, which is manifested in the differences in the results from CONTIN and CONTIN/LL for reference sets with denatured proteins. In this case CONTIN/LL performs much better than CONTIN for all secondary structures. The solution from CONTIN is one of the solutions included in the averaged-solution from CONTIN/LL.

The reliability of the analysis can be improved by considering solutions from all three methods. This can be done in, at least, three different ways: (1) consider the average of the three solutions from the three methods; (2) choose the median of the three solutions for the final solution; or (3) consider the fractions best determined by a given method based on the performance indices of individual secondary structures. Combination of the best performance indices for each individual secondary structure fraction (Table 1), following the third option, resulted in slight improvement of overall performance indices (δ : 0.065, 0.063, 0.065, 0.070, and 0.072, respectively, for 29-, 37-, 43-, 42-, and 48-protein reference sets.)

Johnson (20) used a reference set of 22 proteins, which is a subset of the 29-protein reference set, in the paper describing the CDSSTR method. He used the secondary structure assignments of King and Johnson (36) in determining six secondary structure fractions (α -helix, 3/10 helix, β -sheet, turns, P2, and unordered) and obtained RMS deviations of less than 0.05. We have performed the analysis by CONTIN/LL and SELCON3 using this 22-protein reference set and secondary structural assignments of King and Johnson (36). The results are compared with those from CDSSTR in Table 2. While the overall performance indices from the CDSSTR method ($\delta = 0.039$ and $r = 0.965$) were better than those from the other methods ($\delta = 0.043$ – 0.044 and $r = 0.956$ – 0.959), the best performance indices for the individual secondary structures were distributed among these methods. We have also included the 22-protein reference set along with

the structural assignments of King and Johnson in the CDPro software package as an additional option. However, these results should not be combined or compared with those from the other five reference sets because of the differences in the secondary structure assignments used.

The performance indices given in Tables 1 and 2 correspond to results for all proteins in the reference set. It is difficult to compare the performances of different reference sets from these values since the number of proteins and/or the wavelength range for each reference set are different. Considering the largest subset of reference proteins (29 proteins) from five reference sets and the smallest wavelength range (190–240 nm), one can perform the analysis using different reference sets and compare their performances. The results of such an analysis are given in Table 3.

The improvements in the analysis resulting from the use of a larger reference set can be discerned from the performance indices given in Table 3. Here the results from the analysis of a set of 29 CD spectra, in the wavelength range 190–240 nm, using five different reference sets and three different methods are summarized as performance indices. Overall performance indices improve as one uses increasingly larger reference set, as evidenced from the results for the 29-, 37-, and 43-protein reference sets, for all three methods (SELCON3: δ , 0.078–0.072 and r , 0.773–0.802; CDSSTR: δ , 0.070–0.065 and r , 0.817–0.833; CONTIN/LL: δ , 0.075–0.069 and r , 0.784–0.817). The results from the 42- and 48-protein reference sets, which include denatured proteins, also show a similar trend.

The performance indices of individual secondary structures also improve, with a few exceptions, when a larger reference set is used. For example, as one goes from 29- to 37- to 43-protein reference sets, the performance indices for α_D , β_R , and β_D structures improve for all three methods. We also find improvements in the performance indices of α_D and T structure with 37-protein reference set compared to 29-protein reference set. The inclusion of denatured proteins in the analysis, in general, improved the performance indices of β structures as evidenced by comparing results from 37-

TABLE 3

Performance of SELCON3, CDSSTR, and CONTIN-LL Programs for Analyzing Protein CD Spectra of 29 Proteins in the Wavelength Range 190–240 nm

Reference proteins	Method	α_R		α_D		β_R		β_D		T		U		δ	r
		$\delta_{\alpha R}$	$r_{\alpha R}$	$\delta_{\alpha D}$	$r_{\alpha D}$	$\delta_{\beta R}$	$r_{\beta R}$	$\delta_{\beta D}$	$r_{\beta D}$	δ_T	r_T	δ_U	r_U		
29	SELCON3	0.052	0.949	0.053	0.689	0.102	0.547	0.036	0.709	0.075	0.302	0.118	0.268	0.078	0.773
	CDSSTR	0.059	0.938	0.052	0.785	0.083	0.655	0.030	0.790	0.074	0.337	0.097	0.491	0.070	0.817
	CONTIN-LL	0.058	0.936	0.055	0.679	0.102	0.486	0.035	0.719	0.074	0.323	0.103	0.317	0.075	0.784
37	SELCON3	0.047	0.960	0.050	0.715	0.094	0.638	0.036	0.704	0.063	0.538	0.116	0.142	0.073	0.795
	CDSSTR	0.059	0.939	0.047	0.811	0.087	0.648	0.030	0.801	0.066	0.452	0.098	0.413	0.069	0.819
	CONTIN-LL	0.054	0.944	0.052	0.706	0.093	0.624	0.033	0.753	0.066	0.447	0.095	0.328	0.069	0.813
43	SELCON3	0.051	0.953	0.048	0.747	0.086	0.659	0.034	0.746	0.073	0.382	0.110	0.181	0.072	0.802
	CDSSTR	0.064	0.929	0.042	0.792	0.081	0.704	0.028	0.843	0.067	0.462	0.089	0.444	0.065	0.833
	CONTIN-LL	0.053	0.942	0.048	0.756	0.084	0.674	0.031	0.781	0.076	0.373	0.096	0.262	0.069	0.817
42	SELCON3	0.048	0.957	0.050	0.724	0.091	0.641	0.037	0.694	0.066	0.527	0.125	0.316	0.076	0.798
	CDSSTR	0.059	0.940	0.047	0.813	0.088	0.616	0.029	0.811	0.073	0.343	0.116	0.495	0.074	0.805
	CONTIN-LL	0.055	0.943	0.052	0.711	0.088	0.650	0.031	0.788	0.069	0.401	0.097	0.485	0.069	0.823
48	SELCON3	0.053	0.949	0.049	0.746	0.081	0.678	0.034	0.748	0.072	0.431	0.119	0.394	0.073	0.810
	CDSSTR	0.062	0.933	0.051	0.819	0.082	0.664	0.028	0.821	0.074	0.341	0.116	0.500	0.074	0.807
	CONTIN-LL	0.053	0.946	0.047	0.785	0.078	0.710	0.031	0.780	0.071	0.441	0.093	0.527	0.065	0.843

and 42-protein reference sets or 43- and 48-protein reference sets. These are offset by the performance of turns and unordered structures, while that of the α structures did not change much.

Another feature that is evidenced by comparing Table 1 and Table 3 is that the results obtained with a larger reference set are equivalent to or slightly better than those obtained with the larger wavelength range. The results for the 29-protein reference set given in Table 1 were obtained with the wavelength range of 178–260 nm, while those given in Table 3 were obtained with 190–240 nm range. It has been suggested that for a reliable analysis one needs CD data at lower wavelengths. Our results indicate that CD data in the wavelength range 190–240 nm can give reliable results when a larger reference set is used. As one referee pointed out, many researchers can only obtain CD data in the limited range of 200–240 nm, owing to experimental difficulties in obtaining spectra at higher energies. Following the referee's suggestion, we have per-

formed the CD analysis in the range 200–240 nm, and the performance indices (Table 4) are comparable to those obtained with CD data from 190–240 nm (Table 3). However, we do not recommend the analysis with data from 200–240 nm, except for extremely difficult samples. The users are encouraged to improve the CD data at lower wavelengths by using smaller pathlength cells and higher concentrations of the sample.

The expansion of the reference set used in protein CD analysis by the addition of new protein CD spectra leads to improvements in the estimation of a majority of secondary structure fractions, as seen in Table 3. These improvements can be attributed to the increased variety in the spectral and structural data used in the analysis resulting from the expanded reference set. The largest improvements were observed for the β_R and T structures (SELCON3: $\delta_{\beta R}$, 0.102 to 0.081; δ_T , 0.075 to 0.063; CONTIN/LL: $\delta_{\beta R}$, 0.102 to 0.078; δ_T , 0.074 to 0.066). The α -helical structures in proteins generally show smaller geometric variations, while the

TABLE 4

Performance Indices for Analyzing Protein CD Spectra of 29 Proteins in the Wavelength Range 200–240 nm

Reference proteins	Method	α_R		α_D		β_R		β_D		T		U		δ	r
		δ	r	δ	r	δ	r	δ	r	δ	r	δ	r		
43	SELCON3	0.056	0.938	0.044	0.809	0.094	0.550	0.037	0.682	0.074	0.273	0.105	0.256	0.073	0.799
	CDSSTR	0.063	0.933	0.043	0.789	0.088	0.609	0.030	0.804	0.076	0.357	0.088	0.452	0.068	0.820
	CONTIN-LL	0.065	0.920	0.047	0.771	0.084	0.658	0.030	0.793	0.081	0.239	0.097	0.311	0.071	0.805
48	SELCON3	0.065	0.920	0.045	0.813	0.099	0.505	0.040	0.638	0.076	0.255	0.132	0.281	0.082	0.768
	CDSSTR	0.056	0.948	0.049	0.825	0.084	0.644	0.033	0.754	0.079	0.322	0.121	0.465	0.076	0.803
	CONTIN-LL	0.066	0.919	0.048	0.769	0.090	0.618	0.033	0.753	0.085	0.139	0.126	0.314	0.080	0.771

variations observed in β -sheets and turns are much larger. The β -sheets in proteins are often bent and/or twisted and show larger variation of the (ϕ, ψ) angles than that in α -helices. Similarly, many types of turns have been identified in proteins, based on the dihedral angles, and several types of turn CD spectra have been observed. The improvements observed in the CD analysis with the expanded reference set can be correlated to the larger structural and spectral variability of β -sheets and turns. The expansion of the reference set enables the inclusion of such variables in the analysis, thus improving the results.

CDPro software. The three computer programs and the data files required to perform the analysis are combined in the software package, CDPro, that is downloadable from the internet. The input file and the data-file structure required by all three individual programs are identical. This should make performing the analysis with any and all of these three programs using different reference sets easier. The option for using a given reference set (e.g., 42-protein reference set) is selected in the input file and the program performs the analysis with the chosen reference set. Detailed descriptions of input/output files for each method and the CDPro software package are available at the CDPro website: <http://lamar.colostate.edu/~sreeram/CDPro>.

We suggest that the largest reference set of proteins available for the wavelength range of the CD spectrum analyzed be chosen. For example, if the CD spectrum is in the range 250–191 nm, then the 43-protein reference set is the suitable reference set. The 48- and 42-protein reference sets are to be used for specific applications, such as analyzing CD spectra taken during the course of protein unfolding.

SUMMARY AND CONCLUSIONS

Reference proteins used in CD analysis software packages from five sources were combined to construct a large reference set of 48 proteins. From this reference set five smaller reference sets, differing in the wavelength range of CD data and inclusion/exclusion of denatured proteins, were constructed. The performances of three popular CD analysis programs, CONTIN, SELCON3, and CDSSTR, were examined using the five reference protein sets and DSSP assignments of secondary structure. The performance indices (RMS differences and correlation coefficients between the CD-predicted and X-ray values) for individual secondary structures were mixed, with each method giving slightly better results for one of the secondary structures. The larger reference set performed better than the smaller reference sets, even with CD data over a smaller wavelength range, and this can be attributed to a larger representation of structural and spectral variability of β -sheets and turns. We recom-

mend that all three methods be used in conjunction for a reliable analysis. The three CD analysis programs described in this paper and the required data files are provided in the CDPro software package to assist such applications.

ACKNOWLEDGMENTS

Thanks are due to Dr. W. C. Johnson, Jr., and Dr. S. Yu. Venyaminov for providing the CD spectra of proteins used in this study. This work was supported by NIH Research Grant GM22994.

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 α -, β - 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.

15. 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.
16. Andrade, M. A., Chacan, P., Merolo, J. J., and Moran, F. (1993) Evaluation of secondary structure of protein from UV circular dichroism spectra using unsupervised learning neural network. *Protein Eng.* **6**, 383–390.
17. Sreerama, N., and Woody, R. W. (1994) Poly(Pro)II helices in globular proteins: Identification and circular dichroic analysis. *Biochemistry* **33**, 10022–10025.
18. 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.
19. Sreerama, N., Venyaminov, S. Y., and Woody, R. W. (1999) Estimation of the number of α -helical and β -strand segments in proteins using circular dichroism spectroscopy *Protein Sci.* **8**, 370–380.
20. Johnson, W. C., Jr. (1999) Analyzing protein circular dichroism spectra for accurate secondary structures. *Proteins: Struct. Funct. Genet.* **35**, 307–312.
21. Sreerama, N., Venyaminov, S. Y., and Woody, R. W. (2000) Estimation of protein secondary structure from CD spectra: Inclusion of denatured proteins with native proteins in the analysis. *Anal. Biochem.* **287**, 243–251.
22. Yang, J. T., Wu, C.-S. C., and Martinez, H. M. (1986) Calculation of protein conformation from circular dichroism. *Methods Enzymol.* **130**, 208–269.
23. Johnson, W. C., Jr. (1988) Secondary structure of proteins through circular dichroism spectroscopy. *Annu. Rev. Biophys. Biophys. Chem.* **17**, 145–166.
24. 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.
25. Greenfield, N. J. (1996) Methods to estimate the conformation of proteins and polypeptides from circular dichroism data. *Anal. Biochem.* **235**, 1–10.
26. 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., Eds.) 2nd ed., pp. 601–620. Wiley, New York.
27. Venyaminov, S. Y., and Vassilenko, K. S. (1994) Determination of protein tertiary structure class from circular dichroism spectra. *Anal. Biochem.* **222**, 176–184.
28. Manavalan, P., and Johnson, W. C., Jr. (1983) Sensitivity of circular dichroism to protein tertiary structure class. *Nature* **305**, 831–832.
29. Venyaminov S. Y., Baikolov 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.
30. Kabsch, W., and Sander, C. (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometric features. *Biopolymers* **22**, 2577–2637.
31. Pancoska, P., Bitto, E., Janota, V., Urbanova, M., Gupta, V. P., and Keiderling, T. A. (1995) Comparison of and limits of accuracy for statistical analyses of vibrational and electronic circular dichroism spectra in terms of correlations to and predictions of protein secondary structure. *Protein Sci.* **4**, 1384–1401.
32. Chang, C. T., Wu, C.-S. C., and Yang, J. T. (1978) Circular dichroism analysis of protein conformation: Inclusion of β -turns. *Anal. Biochem.* **91**, 13–31.
33. Privalov, P. L., Tiktopulo, 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.
34. Forsythe G. E., Malcolm M. A., and Moler C. B. (1977) *Computer Methods for Mathematical Computations*, Prentice Hall, Englewood Cliffs, NJ.
35. Provencher, S. W. (1982) A constrained regularization method for inverting data represented by linear algebraic or integral equations. *Comput. Phys. Commun.* **27**, 213–227.
36. King, S. M., and Johnson, W. C., Jr. (1999) Assigning secondary structure from protein coordinate data. *Proteins: Struct. Funct. Genet.* **35**, 313–320.