

A Self-Consistent Method for the Analysis of Protein Secondary Structure from Circular Dichroism

Narasimha Sreerama and Robert W. Woody

Department of Biochemistry, Colorado State University, Fort Collins, Colorado 80523

Received September 21, 1992

A self-consistent procedure for estimating the secondary structure content from circular dichroism spectra of proteins is presented. In this method the spectrum of the protein to be analyzed is included in the basis set and an initial guess is made for the unknown structure as a first approximation. The resulting matrix equation is solved using the singular value decomposition algorithm and the initial guess is replaced by the solution. The process is repeated until self-consistency is attained. The best features of the variable selection and the locally linearized methods are incorporated in this procedure. We have applied this method to examine the inconsistencies in the CD data, to compare the predictions with different ranges and resolutions of the CD data, and to compare different assignments of secondary structures from X-ray structure analyses in the context of secondary structure predictions. The results are compared using the root mean square differences and correlation coefficients. The results obtained are as good as or better than the previous analyses. For most of the proteins considered the self-consistent solutions obtained with different initial guesses were similar. We find the Kabsch and Sander protein crystal structure analysis to be most suitable for our prediction method. © 1993 Academic Press, Inc.

The circular dichroism (CD)¹ spectrum of a globular protein, $c(\lambda)$, can be expressed as a linear equation, $c(\lambda)$

¹ Abbreviations used: CD, circular dichroism; ir, infrared; SC, self-consistent; PG, Provencher and Glöckner method for estimating secondary structure; SVD, singular value decomposition; VS, variable selection; LL, locally linearized model; RMS, root mean square; CCA, convex constraint analysis; LG, Levitt and Greer method for X-ray structure analysis; KS, Kabsch and Sander method for X-ray structure analysis; KSH, Kabsch and Sander method for X-ray structure analysis with α and β fractions equal to the percentage of hydrogen bonds present in those conformations; HJ, Hennessey and Johnson method for X-ray structure analysis; α , α -helix; β , β -sheet; T, turns; U, unordered structure; δ_r , root mean square differences between

$= \sum_k f_k b_k(\lambda)$, where f_k are the fractions and $b_k(\lambda)$ are the CD spectra of secondary structure classes (1-4). Over the past two decades methods for correlating the CD spectra and the secondary structure of proteins have been developed (5-22). The earlier methods made use of the polypeptide spectra in specific secondary structure classes as the basis spectra $b_k(\lambda)$ to fit the spectrum $c(\lambda)$ (5-7). Most of the recent methods make use of the CD spectra of proteins with known secondary structures and calculate the secondary structure spectra by a least squares procedure (8-19,21,22) or by neural network methods (20). Of these methods the most widely used are the constrained statistical regularization (also known as ridge regression) procedure of Provencher and Glöckner (15) (PG), and the singular value decomposition (16,23) algorithm (SVD) in conjunction with the variable selection method (18) (VS) of Johnson and co-workers. Perczel *et al.* (21,22) have developed a method called "convex constraint analysis" (CCA) which does not require the input of structural information from X-ray diffraction.

A modification of the VS method, called the locally linearized model (LL), has been proposed by van Stokkum *et al.* (19). The only essential difference between the LL and VS methods is that, in addition to the number of basis proteins, the number of significant singular values retained is allowed to vary in the LL method, whereas the latter parameter is fixed in the VS method. The principal advantage of the LL method is the large reduction in computer time required for the analysis. This is primarily due to the strategy of arranging the basis set proteins in order of increasing root-mean-

spectra; δ_r , root mean square differences between secondary structures; r , correlation coefficient; δ_λ , ratio of root mean square differences between secondary structures to the fraction present in the basis set; IG1, initial guess equal to the X-ray structure; IG2, initial guess equal to the structure with smallest δ_r ; IG3, initial guess equal to 100% α ; IG4, initial guess equal to 25% each of the four secondary structures.

square (RMS) differences (δ_c) of the CD spectra of the proteins with that of the protein to be analyzed, and systematically deleting the more distant proteins from the basis set, thus making sure that the proteins which are important are always included in the analysis. The comparison of results made by van Stokkum *et al.* (19) seemed to indicate that the LL method gave a significant improvement over the VS method. However, W. C. Johnson, Jr. (personal communication) has pointed out that in their test of the VS method, van Stokkum *et al.* deleted no more than three proteins, in contrast to the VS method described by Manavalan and Johnson (18). When properly applied, i.e., when all possible deletions of proteins have been considered, the results of the VS and LL methods are essentially equivalent, as is also the case for the PG method (15).

These methods have generally been evaluated by deleting the protein to be analyzed from the basis set and comparing the predicted structure with the known structure. It is well known (4,15,16,24) that the prediction improves when the protein that is analyzed is included in the basis set. This is because the solution will be biased toward the input structure for the protein being analyzed if this protein is included in the basis set. One of the underlying assumptions of these prediction methods is that the CD spectra of two proteins with similar secondary structures are similar. The method we describe here is based upon the VS and LL methods, with an additional feature. We include the CD spectrum of the protein analyzed in the basis set and solve the set of linear equations by a self-consistent procedure.

One of the major questions in these prediction methods is how to identify a failed solution in the absence of known structure. The PG method uses the conditions $\sum f_k = 1.0$ and $f_k \geq 0.0$ as constraints and gives a set of solutions to any spectrum with different regularization parameters. We follow Johnson and co-workers (VS) and van Stokkum *et al.* (LL) and use these conditions as criteria for selecting the best solution. While there is no guarantee that solutions meeting these criteria are correct, it is clear that solutions which do not meet them are incorrect.

Another major concern is the selection of a proper basis set of proteins. A majority of the publications on the subject have used different sets of proteins as the basis. Recently Venyaminov *et al.* (25), based on the five most significant CD spectra of two sets of reference proteins calculated using the SVD method, concluded that the selection of the reference set is not a very serious problem. Toumadje *et al.* (26) have shown that the RMS differences are decreased by extending the wavelength to 168 nm. The relative amounts of the secondary structures present in the basis set are important in analyzing the solutions and might influence the RMS differences and/or the correlation coefficients between the X-ray and the predicted structure, which are used to

assess the accuracy of the prediction. This is very much evident when one considers the fractions of parallel and antiparallel β -sheets separately in the analysis. The average amount of the parallel β -sheet (β_p) present in the basis set is quite small compared to the other fractions, and the predictions obtained for β_p show the least correlation with the X-ray structures (17), yet the RMS differences are small. One should, in principle, consider the averages of the secondary structures present in the basis set leading to the observed RMS difference and also the correlation between the predicted and X-ray structures to compare different basis sets in assessing the performance of an analysis. None of the analyses take into account the possible experimental errors in the CD spectra. In the manual for the analysis program CONTIN (27), Provencher suggests that the spectrum be multiplied by factors of 0.9 and 1.1 to account for as much as a 10% experimental error in determining the protein concentration. We will examine these factors with our self-consistent (SC) method.

Different assignments of the secondary structure of proteins have been used in the previous analyses of prediction of secondary structures from spectroscopic data (9-19,28-30). Three of the widely used methods for assigning secondary structures from the X-ray data and used in the prediction from CD are: the Levitt and Greer (31) method (LG), which makes use of C_α coordinates; the Kabsch and Sander (32) method (KS), based on the hydrogen bond patterns; and the Hennessey and Johnson (16) method (HJ) based on visual inspection of the structure. The first two of these methods are automated. The secondary structure assignments from these methods differ and it is still an open question as to which of these assignments is the most suited for the prediction methods. We will compare the secondary structure assignments from these methods and compare their performance for the secondary structure prediction from CD spectra.

METHODS

The set of linear equations relating the CD spectra with the secondary structure can be represented as the matrix equation

$$\mathbf{F} = \mathbf{X} \mathbf{C}, \quad [1]$$

where \mathbf{C} is the $m \times n$ matrix of CD data; \mathbf{F} is the $l \times n$ secondary structure data matrix; \mathbf{X} is an $l \times m$ matrix relating the spectra to the structure. Here, n is the number of proteins, l is the number of secondary structure elements, and m is the number of wavelengths used to represent the CD spectra.

We included the spectrum of the protein analyzed, \mathbf{c} , in the matrix \mathbf{C} to obtain the matrix \mathbf{C}_1 , an $m \times (n + 1)$ matrix. The proteins in the matrix \mathbf{C}_1 were then

arranged in increasing order of the RMS distance (δ_c) from \mathbf{c} , the protein to be analyzed, which forms the first row in the newly ordered matrix \mathbf{C}_2 . The \mathbf{F} matrix was also arranged in the same order to construct matrix \mathbf{F}_2 . The elements of the first row in \mathbf{F}_2 , corresponding to the protein to be analyzed, are unknown and must be determined. We make an initial guess for this structure, \mathbf{f}_0 , and complete the matrix \mathbf{F}_2 . Equation 1 was thus modified to

$$\mathbf{F}_2 = \mathbf{X} \mathbf{C}_2 \quad [2]$$

and was solved for \mathbf{X} , which was used to calculate the structure \mathbf{f}

$$\mathbf{f} = \mathbf{X} \mathbf{c} \quad [3]$$

The matrix \mathbf{C}_2 was decomposed using the SVD algorithm (23) as the product of three matrices,

$$\mathbf{C}_2 = \mathbf{U} \mathbf{S} \mathbf{V}^T, \quad [4]$$

where \mathbf{U} and \mathbf{V} are unitary matrices, \mathbf{S} is a diagonal matrix, and the superscript T indicates the transpose of the matrix. Equations 2 and 4 give

$$\mathbf{X} = \mathbf{F}_2 \mathbf{V} \mathbf{S}^+ \mathbf{U}^T, \quad [5]$$

where \mathbf{S}^+ is the inverse of \mathbf{S} .

Only the first few diagonal elements of \mathbf{S} , the significant singular values, are required to reconstruct the matrix \mathbf{C}_2 to within experimental error (16) and not all proteins contribute significantly toward the solution (18,19). We incorporated these ideas in our method and, in fact, a set of solutions \mathbf{f} were obtained by varying the number of proteins considered (N_p from 3 to $N + 1$) for analysis, which is similar to the VS method, and the number of singular values (N_s from 1 to $N_p - 2$ or 7, whichever is the smaller) used to construct \mathbf{X} . Only those solutions which satisfy the conditions

$$|\Sigma f_k - 1.0| \leq 0.05 \quad \text{and} \quad f_k \geq -0.025$$

were considered. The limits 0.025 and 0.05, which are smaller than those Manavalan and Johnson (18) used, were considered since the protein to be analyzed was included in the matrices \mathbf{C} and \mathbf{F} . For each value of N_p , the number of proteins retained in the basis set, a set of solutions meeting the above criteria was obtained corresponding to various values of N_s . For each value of N_p , the single solution with Σf_k closest to 1.0 was retained. These solutions were then averaged to obtain the final solution, \mathbf{f}_1 . The solution \mathbf{f}_1 now replaces \mathbf{f}_0 in the matrix \mathbf{F}_2 as the second approximation to the unknown structure. Equation 2 was then solved for a new solution

\mathbf{f}_2 . The process was repeated until the RMS difference between the successive solutions was less than 0.0025. Lower values of the convergence limit led to oscillatory behavior of solutions in some cases and extremely slow convergence in proteins having high α -helix content.

The RMS deviations (δ) and correlation coefficients (r) were calculated using

$$\delta = \sqrt{\frac{1}{N} \Sigma (x_i - y_i)^2}$$

and

$$r = \frac{N \Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{(N \Sigma x_i^2 - (\Sigma x_i)^2)(N \Sigma y_i^2 - (\Sigma y_i)^2)}}$$

The basis set consisted of an all- α polypeptide, poly(L-glutamic acid), and the following 15 proteins: Bence-Jones protein, prealbumin, rubredoxin, chymotrypsin, elastase, papain, thermolysin, lysozyme (egg white), subtilisin BPN', glyceraldehyde-3-phosphate dehydrogenase, flavodoxin, lactate dehydrogenase, triose-phosphate isomerase, cytochrome c, hemoglobin, and myoglobin. Of these, the first 5 are $\beta\beta$ proteins, the next 3 are $\alpha + \beta$ proteins, the next 5 are α/β proteins, and the last 3 are $\alpha\alpha$ proteins. The CD spectra of these proteins were kindly provided by Dr. W. C. Johnson, Jr. The secondary structure assignments of α -helix (α), β -sheet (β), turns (T), and unordered (U) were according to the Kabsch and Sander method (32) and the CD data used were in the range 178–260 nm at a resolution of 1 nm, unless otherwise noted. Four different initial guesses were considered in the analysis: (a) The X-ray structure (IG1), (b) the structure of the protein having the smallest δ_c (IG2), (c) 100% α -helix (IG3), and (d) 25% each of α -helix, β -sheet, turns, and unordered structures (IG4).

RESULTS AND DISCUSSION

We have used the RMS differences (δ_r) and correlation coefficients (r) between the X-ray and predicted secondary structures, which are henceforth called prediction or performance indices, to compare the results obtained with different methods and different sets of data. The results obtained with different initial guesses are compared with those obtained with other methods in Table 1. It is evident that the predictions for α are always the best, with the smallest δ_r (0.071–0.087) and largest r (0.951–0.964), which is consistent with the fact that the helix, if present to a considerable extent, dominates the CD spectrum of a protein, and CD is always at its best in predicting the percentage of helix. For the other three secondary structure elements prediction indices are similar (β : 0.073–0.086 and 0.840–0.864; T:

TABLE 1
Performance of the Self-Consistent Method: Correlation Coefficients (r) and RMS Differences (δ_r)
between the Predicted and X-Ray Structures

Method ^a	Secondary structure							
	α		β		T		U	
	δ_r	r	δ_r	r	δ_r	r	δ_r	r
SC (IG1)	0.080	0.951	0.073	0.864	0.043	0.855	0.060	0.728
SC (IG2)	0.087	0.953	0.077	0.842	0.054	0.769	0.062	0.716
SC (IG3)	0.071	0.964	0.075	0.854	0.045	0.840	0.057	0.757
SC (IG4)	0.080	0.963	0.086	0.840	0.048	0.824	0.059	0.763
LL model ^b	0.073	0.960	0.074	0.850	0.049	0.803	0.054	0.788
LL model I ^c	0.090		0.12; 0.08 ^d		0.07		0.09	
VS (168-260) ^e	0.043	0.986	0.070	0.876	0.050	0.561	0.089	0.755
VS (178-260) ^e	0.060	0.977	0.068	0.889	0.059	0.321	0.085	0.789
VS (178-260) ^f	0.06	0.97	0.10	0.76	0.07	0.49	0.07	0.86
VS (190-260) ^f	0.07	0.95	0.13	0.45	0.05	0.54	0.09	0.69
PG model II ^g	0.09		0.14; 0.06 ^d			0.07		0.08
HJ (178-260) ^f	0.06	0.98	0.14	0.54	0.07	0.30	0.12	0.61
CCA ^h	0.112	0.93	0.095	0.71	0.204	0.73	0.094; 0.174	0.35; 0.48
Neural network ⁱ		1.0		0.91;0.63		0.64		0.96

^a SC (IG1)–SC (IG4) are the results from the self-consistent method with different initial guesses and 178 to 260-nm spectral range. See text for the details of other abbreviations and symbols.

^b The solutions are obtained with test protein deleted from the basis set. The relaxed selection criteria, $|\Sigma f_k - 1.0| \leq 0.15$ and $\Sigma f_k \geq -0.05$, are used since the strict selection criteria, $|\Sigma f_k - 1.0| \leq 0.05$ and $\Sigma f_k \geq -0.025$, did not give solutions for a few proteins. The final solution was obtained as the average of the best solution under each set of proteins (see Methods).

^c From van Stokkum *et al.* (19), corresponding to the final solution with Σf_k closest to 1.0.

^d The two numbers correspond to the fractions of antiparallel and parallel β -sheets respectively. The values are from van Stokkum *et al.* (19) from an analysis with five secondary structural elements.

^e Toumadje *et al.* (26).

^f Manavalan and Johnson (18).

^g van Stokkum *et al.* (19).

^h Perczel *et al.* (21). The two sets of values under U are due to a slightly different classification of secondary structures.

ⁱ Böhm *et al.* (20) from an analysis of 13 proteins with five secondary structural elements.

0.043–0.054 and 0.769–0.855; U: 0.057–0.062 and 0.716–0.763).

The solutions obtained with different initial guesses should have been similar since the self-consistent solution for a given set of **C** and **F** matrices should be the same, irrespective of the initial guess, although they may differ within the convergence limit. This was true for most of the proteins in the basis set, for which the solutions obtained with different initial guesses differed very little. The number of iterations needed to reach the final solution, however, was different. An illustrative example was myoglobin, which has 78% α and no β . The guesses IG1, IG2, and IG3 were similar, having high α content, since myoglobin has a high α content (IG2: the protein with smallest δ_c was hemoglobin with 75% α), and self-consistency was achieved in less than four iterations. However, for IG4 (25% α) about eight iterations were needed to reach the final solution.

The solutions obtained with different initial guesses are compared in Table 2. The solutions for Bence-Jones protein did not satisfy our selection criteria after two iterations (successive solutions differed by $\delta = 0.0093$;

convergence limit 0.0025) with all four initial guesses. Relaxed selection criteria, however, gave a self-consistent solution but $|\Sigma f_k - 1.0|$ was greater than 0.05 and we did not include this solution in calculating the performance indices. The differences between IG1 and IG3 are minimal, and so are those between IG2 and IG4. The performance indices for these initial guesses also fall into two groups with IG1 and IG3 in one and IG2 and IG4 in the other. Comparing the predictions for the proteins individually, we found that the differences in the performance indices for different initial guesses are caused mainly by the differences in predictions for hemoglobin and poly(Glu). Both are high- α structures and the results from initial guesses IG1 and IG3 were similar in α to the X-ray structure, while those from IG2 and IG4 were lower in α . Surprisingly, the protein having the smallest δ_c with hemoglobin was not myoglobin, in spite of similar secondary structure assignments (75 and 78% α , and no β), but triose-phosphate isomerase (47% α and 18% β). The closest CD spectrum to that for poly(Glu) was that of myoglobin. In these two cases the choice of initial guess was important and the final solutions

TABLE 2
Comparison of Solutions Obtained from Different Initial Guesses^a

Initial guess	Secondary structure							
	α		β		T		U	
	δ_f	r	δ_f	r	δ_f	r	δ_f	r
IG1-IG2	0.033	0.994	0.024	0.981	0.016	0.984	0.017	0.969
IG1-IG3	0.015	0.998	0.009	0.997	0.009	0.994	0.018	0.975
IG1-IG4	0.036	0.992	0.010	0.997	0.013	0.990	0.024	0.965
IG2-IG4	0.013	0.998	0.025	0.983	0.011	0.999	0.010	0.984

^a See text for explanation of abbreviations and symbols.

reached from opposite directions of the X-ray structure differed by approximately 10%. The better performance indices for the initial guess IG3 arises due to better predictions for the high α proteins. In the basis set, most of the proteins have substantial amounts of all four secondary structures (20–40%). The initial guess IG4, which is 25% of each, gave slightly better results than IG2 for these proteins, which resulted in slightly better performance indices for IG4 over IG2. However, for a test protein about which one has no structural information, IG2, the structure of the protein with CD spectrum most similar to that of the test protein, seems to be the most appropriate choice. In general, different initial guesses give essentially similar results, thus indicating that the choice of initial guess is not a factor affecting the final solution. We have used IG2 as the initial guess in the remainder of the paper, unless otherwise mentioned.

We have compared these results with related and available published results in Table 1. Some of the values from the literature were obtained with different basis sets of proteins or used different analyses of the X-ray diffraction structure. Therefore the results cannot be compared rigorously. Nevertheless, distinct differences in performance criteria are significant. It is interesting to note that the results obtained by the LL method (19) with our procedure for selecting the final solution are comparable to those of the self-consistent method. Both the LL and our method gave slightly better results than the VS method (row 9). However, VS with an extended wavelength range (168 nm) (26) did better in predicting α and β as compared to our method but not as well in T and U. It should be noted that the latter VS results were obtained with a different basis set and are taken from Toumadje *et al.* (26), while the LL results were obtained in our laboratory. Our method gives the best allowed solution for the spectrum which is consistent with the basis set of proteins considered and performs as well as or better than the previous analyses.

Though the solutions obtained with the X-ray structure as the initial guess (IG1) and 100% α (IG3) gave the best results overall, they are still different from the X-

ray structure. This led us to consider the possibility of a secondary structure data matrix consistent with the CD data matrix. To construct such a matrix we started with initial guess IG1, termed iteration 0, and obtained the self-consistent solutions for the proteins in the basis set, which is iteration 1. This set, iteration 1, was now used as the secondary structure matrix F_2 and a new set of solutions, iteration 2, was obtained. The process was repeated until convergence was achieved and the results are compared in Table 3. As expected, the differences between the X-ray structure (iteration 0) and the results of successive iterations increase but reach a limit, corresponding to self-consistency. When this set is used as the F matrix the solutions obtained for the basis set proteins were identical to the secondary structure in the F matrix; this F matrix is therefore consistent with the C matrix. It can be seen that the performance indices between iterations 4 and 5 have reached convergence, while those between iterations 0 and 5 are slightly worse for β and slightly better for other three secondary structures than those between iterations 0 and 1. The secondary structure set thus obtained, the self-consistent structure, is compared with the X-ray structure in Table 4. The proteins for which one or more components differ by more than 10% between these two sets are cytochrome c, hemoglobin, flavodoxin, prealbumin, and lysozyme. It is interesting to note that the results obtained by Toumadje *et al.* (26) after extending the wavelength range to 168 nm show similar differences between the predicted and the X-ray structures for some proteins common in the two sets. Are these differences due to differences between the solution and X-ray structure? Are the aromatic side-chains and/or disulfide linkages and distortions in the secondary structures responsible for these differences? Or are they due to possible errors in determining the concentration? It is not possible to definitively answer the first two questions at present. Work is in progress in our laboratory to include the aromatic contributions in prediction methods.

We examined the possibility of error in determining

TABLE 3
Convergence and Performance Indices of the Self-Consistent Structure^a

Iterations	Secondary structure							
	α		β		T		U	
	δ_r	r	δ_r	r	δ_r	r	δ_r	r
0 and 1	0.080	0.951	0.073	0.864	0.043	0.855	0.060	0.728
1 and 2	0.037	0.986	0.030	0.965	0.022	0.960	0.044	0.780
1 and 5	0.020	0.996	0.044	0.921	0.015	0.979	0.037	0.871
4 and 5	0.014	0.998	0.006	0.998	0.004	0.999	0.015	0.958
0 and 5	0.077	0.956	0.093	0.773	0.041	0.873	0.056	0.807

^a See text for explanation.

concentration by multiplying the CD spectra by factors of 0.9, 0.95, 1.0, and 1.05, which will account for 5–10% experimental error, and comparing the solutions systematically with the X-ray structure. For most proteins little or no improvement was obtained, but substantial improvements were obtained in some proteins, and the optimal multiplication factor was generally 1.05 (0.95 in a few cases). The greatest improvement was obtained with the Bence-Jones protein and hemoglobin. It may be recalled that no acceptable solutions were obtained for Bence-Jones protein with strict selection criteria, and hemoglobin had the smallest δ_c with triose-phosphate isomerase, rather than the closely homologous myoglobin. With a multiplication factor of 1.05, both give better results, with Bence-Jones protein giving acceptable solutions and hemoglobin having the smallest δ_c with myoglobin. We systematically multiplied the spectrum of these proteins in the CD data matrix and compared the performance indices with the previous sets. The results are presented in Table 5. Upon multiplying the CD spectrum of Bence-Jones protein in the matrix **C** by 1.05 we obtained acceptable predictions for it, and the performance indices also improved. Multiplying both Bence-Jones protein and hemoglobin spectra in the matrix **C** by 1.05 led to further improvement in the performance indices, which are now comparable with those obtained with wavelength extended to 168 nm (26), albeit with a slightly different basis set. A similar computer experiment with the spectra of cytochrome c and flavodoxin did not lead to any further improvement for α and β , but gave lower correlation coefficients for T and U. We are unable to explain the small inconsistencies which still persist between the predicted and the X-ray structure of some proteins.

To examine the effect of short-wavelength data, the importance of which was questioned by Venyaminov *et al.* (25), we have tested the performance of our method on truncated CD spectra (182–260, 185–260, and 190–260 nm). The results are shown in Table 6. The performance indices for α are high even in the range 190–260

nm and those for T are comparable to those with 178–260 nm. The results with 185–260 nm are similar to those with 178–260 nm, with β being slightly worse. The T and U performance indices are lower for 185- to 260- and 190- to 260-nm sets, but still better in comparison to the results obtained by Venyaminov *et al.* (25) with the PG method. Table 6 also shows the effect of CD data at different levels of resolution (0.5, 1.0, and 2.0 intervals). The performance indices obtained with different resolutions are comparable, with those at 0.5 nm resolution being the best.

In Table 7 we have compared the predictions from the self-consistent method at different ranges and resolutions of the CD data with those obtained by the VS method (26) for three representative proteins: elastase ($\beta\beta$ class), papain ($\alpha + \beta$ class), and myoglobin ($\alpha\alpha$ class). The X-ray data used in these two analyses are also given, since they were different. The results obtained at different resolutions of the data were similar and those with different wavelength ranges did not differ greatly from the X-ray values. The differences between the predicted and the X-ray structures from our method are smaller than those from Toumadje *et al.* (26). Our method can be used with a higher level of confidence for the CD data recorded with a longer wavelength cutoff.

As indicated earlier, the results from our laboratory reported in Tables 1–7 were obtained using the secondary structures from the Kabsch and Sander (KS) analysis (32) of the X-ray structures and those with the VS and LL methods used the Hennessey and Johnson (HJ) analysis (16). We used the KS analysis because of its better performance, in a previous study (33), as compared to other analyses of the X-ray data. We carried out a similar comparison with the self-consistent method using initial guess IG2, and the results are presented in Table 8. In addition to the three methods of X-ray structure analysis (KS, HJ, and LG), we also considered a variant of the Kabsch and Sander method (KSH), in which the number of hydrogen bonds in α and

TABLE 4
Comparison of the Self-Consistent Structure (SS)
with the X-Ray Structure^a

Protein ^b	Secondary structure				Total
	α	β	T	U	
Prealbumin					
X Ray	0.070	0.490	0.230	0.210	
SS	0.223	0.286	0.276	0.211	0.996
Rubredoxin					
X Ray	0.170	0.190	0.320	0.320	
SS	0.154	0.299	0.292	0.268	1.007
Chymotrypsin					
X Ray	0.100	0.340	0.260	0.300	
SS	0.176	0.294	0.279	0.243	0.992
Elastase					
X Ray	0.100	0.370	0.250	0.280	
SS	0.111	0.327	0.298	0.258	0.994
Papain					
X Ray	0.250	0.190	0.260	0.300	
SS	0.271	0.244	0.254	0.235	1.003
Thermolysin					
X Ray	0.340	0.180	0.300	0.180	
SS	0.391	0.185	0.215	0.198	0.989
Lysozyme					
X Ray	0.390	0.110	0.320	0.180	
SS	0.334	0.215	0.240	0.218	1.007
Subtilisin BPN'					
X Ray	0.310	0.200	0.250	0.240	
SS	0.289	0.237	0.250	0.217	0.993
Glyceraldehyde-3-phosphate dehydrogenase					
X Ray	0.260	0.250	0.260	0.230	
SS	0.297	0.233	0.248	0.218	0.996
Flavodoxin					
X Ray	0.380	0.240	0.240	0.140	
SS	0.287	0.246	0.255	0.214	1.002
Lactate dehydrogenase					
X Ray	0.370	0.140	0.250	0.240	
SS	0.417	0.178	0.212	0.196	1.003
Triose-phosphate isomerase					
X Ray	0.470	0.180	0.150	0.200	
SS	0.470	0.155	0.195	0.186	1.006
Cytochrome c					
X Ray	0.430	0.020	0.220	0.330	
SS	0.322	0.228	0.237	0.199	0.986
Hemoglobin					
X Ray	0.760	0.000	0.110	0.130	
SS	0.598	0.107	0.152	0.158	1.015
Myoglobin					
X Ray	0.780	0.000	0.100	0.120	
SS	0.723	0.054	0.101	0.124	1.002
Poly(L-glutamic acid)					
X Ray	1.00	0.000	0.000	0.000	
SS	1.00	-0.015	0.003	0.058	1.046

^a The X-ray structure is from the Kabsch and Sander analysis of the X-ray crystal structure. See text for the explanation of abbreviations and symbols.

^b Bence-Jones protein is not in the list of proteins since the solutions did not meet our selection criteria.

β structures per 100 residues from the Kabsch and Sander analysis (32) was taken for the fractions of α and β , respectively. The difference between the fractions of α and β in KS and KSH was added to the U. The results from the CD analysis with the KSH F matrix are also included in the Table 8. Among these, the LG and HJ methods gave slightly poorer results and KSH gave better results for α , while those for KS was intermediate. For β , LG showed the best correlation and RMS difference. Among the other structural types, LG gave the better analysis for T and HJ for U. Performance indices for T and U were mixed, with HJ showing the worst correlation for T and LG for U. Those for KS and KSH were comparable, with KS having slightly poorer correlation than KSH for T. Overall, the results obtained from KSH were better and those from HJ were poorer than those from KS and LG. In comparison with KS, LG gave better results for β and poorer results for U, while the results for α and T were comparable. The KS method was preferred due to consistently higher correlations for all four secondary structures ($r > 0.7$) and because it agreed better with the other two methods than the KSH method. A similar comparison was carried out considering only three secondary structures, α , β , and unordered (T and U combined), in the analysis. Restricting the analysis to three secondary structural classes did not alter the above conclusions.

Venyaminov *et al.* (25) have also carried out a similar analysis, in which they have compared the performance of secondary structure classification by the original crystallographic investigators (12), LG, and KS, using the PG method for CD analysis. They obtained dramatically different results, finding that KS performs very poorly and the other two methods give comparable performance indices. The correlation coefficients for T were less than 0.5, but they did not report results for U. A direct comparison is not possible for two reasons: their CD analysis method is different than ours, though, in principle, the results from PG method should be comparable to our results; the reference set of proteins is different from our reference set.

Table 9 shows the comparison of the methods for analyzing the secondary structures from X-ray data. We have presented the RMS differences and correlation coefficients between the F matrices from the different analyses. The α and β fractions are highly correlated (r , 0.93–0.99) while T and U are much less so (r , 0.74–0.88). (The correlation coefficient of 1.0 and the RMS deviation of zero for T in KS and KSH results from the identical definitions for the conformation in the two methods.) The RMS differences vary from 0.034 to 0.141 for α , 0.032 to 0.182 for β , 0.051 to 0.08, for T and 0.081 to 0.214 for U, with the largest differences between KSH and LG and the smallest between HJ and KS, or HJ and KSH. Table 10 gives the average values of different secondary structures from these analyses, calculated by

TABLE 5
Performance Indices of Modified CD Data Matrices^a

CD(m) ^b	Secondary structure							
	α		β		T		U	
	δ_f	r	δ_f	r	δ_f	r	δ_f	r
Unmodified ^c	0.087	0.953	0.077	0.842	0.054	0.769	0.062	0.716
Bence-Jones protein	0.086	0.960	0.070	0.894	0.051	0.781	0.062	0.676
Hemoglobin	0.072	0.971	0.074	0.886	0.049	0.799	0.060	0.711
Cytochrome c	0.071	0.973	0.068	0.903	0.048	0.802	0.057	0.737

^a See text for explanation of abbreviations and symbols.

^b CD matrix with specified protein (and those mentioned in the previous rows) spectra multiplied by a factor 1.05.

^c Row 2, Table 1.

adding the fractions under each secondary structure element and dividing by the total number of proteins. From the average values of the secondary structure elements, it is evident that LG and KSH form two extremes for α , β , and U, while HJ and KS (KSH) form extremes for T. The LG method classifies most of the residues in one of the ordered conformations and only about 12% in U, while HJ classifies about 16% residues each in β and T and about 30% residues in U. The KS method classifies a higher percentage of residues as T (22%) than others and the percentages of the other three elements are intermediate. These numbers have a direct bearing on the values given in Table 8. The high correlation of α and β indicates a systematic increase or decrease of the repetitive structure among the different methods of analyses. However, the correlation among T and U was not so good and is variable, which is consistent with the differences in the averages. The low RMS difference between HJ and KSH for U is explained by similar percentages of U structure in these two sets and so is the low RMS difference between HJ and LG for α .

The analyses of X-ray structures depend on the parameters chosen for defining the secondary structures (C_α - C_α distances, hydrogen bonding pattern, C_α torsion angle, etc.) and the criteria used to determine acceptable structures (34). The KSH analysis uses a rigid criterion based on the hydrogen bonds present in the secondary structures and the LG analysis uses a flexible criterion based on the C_α - C_α distances, thus forming the extremes in the secondary structure assignment. These assignments are correct within their definitions. The comparison of these assignments discussed above indicate a very good correlation for α and β . Similar results were obtained by Pancoska *et al.* (35) between the KS and LG assignments for a larger set of proteins. The KS and LG assignments give similar performance indices for all but the unordered conformation (Table 8), implying that either of these can be used in the prediction methods.

The RMS differences are generally considered as the indicator of the accuracy of the analysis (25,26), but to compare the performance of different sets of data, one

TABLE 6
Performance of Prediction Method for Different Ranges and Resolutions of CD Data^a

λ (nm)	λ_1 (nm)	Secondary structure							
		α		β		T		U	
		δ_f	r	δ_f	r	δ_f	r	δ_f	r
190	1.0	0.088	0.942	0.093	0.733	0.048	0.844	0.067	0.655
185	1.0	0.081	0.946	0.086	0.782	0.046	0.835	0.069	0.600
182	1.0	0.088	0.939	0.084	0.806	0.051	0.792	0.058	0.748
178	1.0	0.087	0.953	0.077	0.842	0.054	0.769	0.062	0.716
178	0.5	0.079	0.959	0.077	0.845	0.048	0.826	0.059	0.736
178	2.0	0.081	0.957	0.081	0.834	0.047	0.832	0.052	0.749

^a Columns under λ and λ_1 show the lower value of the wavelength range (λ -260 nm) and the resolution of CD data in the data matrix. See text for explanation of other symbols.

TABLE 7

Predictions for Selected Proteins at Different Ranges and Resolutions of CD Data^a Compared with Published Results

Protein	λ (nm)	λ_1 (nm)	α	β	T	U	Total
Elastase	X ray (KS)		0.10	0.37	0.25	0.28	
	190	1.0	0.070	0.400	0.290	0.240	1.000
	185	1.0	0.077	0.379	0.288	0.256	1.000
	182	1.0	0.053	0.421	0.281	0.252	1.007
	178	1.0	0.046	0.428	0.283	0.248	1.005
	178	0.5	0.078	0.383	0.280	0.256	0.997
	178	2.0	0.079	0.381	0.282	0.256	0.998
	X ray (HJ) ^b		0.10	0.37	0.22	0.31	
	178 ^b	0.5	0.06	0.35	0.18	0.40	0.99
	168 ^b	0.5	0.06	0.35	0.19	0.40	1.00
Myoglobin	X ray (KS)		0.78	0.00	0.10	0.12	
	190	1.0	0.720	0.051	0.140	0.080	0.991
	185	1.0	0.747	0.048	0.111	0.108	1.014
	182	1.0	0.741	0.056	0.097	0.104	0.998
	178	1.0	0.720	0.067	0.084	0.127	0.998
	178	0.5	0.726	0.065	0.081	0.128	1.00
	178	2.0	0.728	0.068	0.074	0.124	0.994
	X ray (HJ) ^b		0.78	0.00	0.12	0.10	
	178 ^b	0.5	0.83	0.11	0.03	0.11	1.00
	168 ^b	0.5	0.79	-0.02	0.04	0.16	1.02
Papain	X ray (KS)		0.25	0.19	0.26	0.30	
	190	1.0	0.280	0.220	0.230	0.260	0.99
	185	1.0	0.318	0.166	0.217	0.307	1.008
	182	1.0	0.297	0.169	0.231	0.298	0.995
	178	1.0	0.296	0.170	0.244	0.292	1.002
	178	0.5	0.303	0.176	0.247	0.271	0.997
	178	2.0	0.303	0.178	0.247	0.271	0.999
	X ray (HJ) ^b		0.28	0.09	0.14	0.49	
	178 ^b	0.5	0.29	0.15	0.19	0.38	1.01
	168 ^b	0.5	0.30	0.09	0.19	0.40	0.98

^a Columns under λ and λ_1 show the lower value of the wavelength range (λ -260 nm) and the resolution of CD data in the data matrix. See text for explanation of other symbols.

^b The results are from Toumadje *et al.* (26) using VS method and HJ classification (16) of secondary structures.

should consider the total secondary structure present in the data set. A better indicator would be the RMS difference for a given type of secondary structure divided by the average amount of that secondary structure present in the basis set, termed δ_A . In Table 8 we have given the values of δ_A in comparing the performances of dif-

ferent X-ray analyses of secondary structures. The comparison of δ_A values does not, however, change the conclusions drawn earlier. The KS and KSH perform better than the other methods, overall.

Another interesting observation concerns the number of acceptable solutions for a given protein and the num-

TABLE 8
Performance Indices of Different Analyses of the X-Ray Structure^a

X ray set	Secondary structure											
	α			β			T			U		
	δ_f	r	δ_A	δ_f	r	δ_A	δ_f	r	δ_A	δ_f	r	δ_A
KSH	0.077	0.967	0.252	0.063	0.854	0.400	0.050	0.809	0.229	0.070	0.717	0.219
KS	0.087	0.953	0.227	0.077	0.842	0.422	0.054	0.769	0.245	0.062	0.716	0.288
HJ	0.090	0.940	0.231	0.094	0.748	0.506	0.083	0.224	0.506	0.071	0.852	0.244
LG	0.094	0.945	0.226	0.083	0.903	0.298	0.047	0.837	0.261	0.054	0.525	0.451

^a See text for the explanation of abbreviations and symbols.

TABLE 9
Comparison of Methods for Secondary Structure Analyses of X-Ray Structures^a

Methods compared	α		β		T		U	
	δ_f	r	δ_f	r	δ_f	r	δ_f	r
KSH-KS	0.092	0.98	0.032	0.99	0.000	1.00	0.120	0.88
KSH-HJ	0.092	0.98	0.040	0.97	0.080	0.74	0.081	0.81
KSH-LG	0.141	0.96	0.182	0.95	0.069	0.81	0.214	0.76
KS-HJ	0.034	0.99	0.039	0.97	0.080	0.74	0.105	0.87
KS-LG	0.068	0.98	0.145	0.93	0.069	0.81	0.107	0.78
HJ-LG	0.059	0.99	0.157	0.93	0.051	0.81	0.188	0.80

^a See text for explanation of symbols.

ber of proteins with similar spectra in the basis set. Most of the proteins in the basis set gave a large number of acceptable solutions, with the majority of the solutions having Σf_k close to 1.0. Multiplying the CD spectra of these proteins to correct for possible errors in concentration did not have much effect on the final solution. The CD spectrum of these proteins had several proteins in the basis set with similar CD spectra (δ_c less than 1.0 $\Delta\epsilon$). The proteins hemoglobin, myoglobin, Bence-Jones protein, papain, and prealbumin were exceptions and, indeed, the number of acceptable solutions was smaller and, in the case of Bence-Jones protein, zero. It is important to have a good representation of the spectrum analyzed in the basis set in order to get a good prediction.

Recently, Bobba *et al.* (36) have reported the overprediction of β in comparison to α and the inability to distinguish between the defined and the remainder structures from CD analysis using the PG method. This is probably caused by an underrepresentation of the U structure in the basis set. A similar situation would arise with our method if there were no proteins in the basis set with spectra similar to the test spectrum. As a result, the method would not give any solution, resulting in a failed analysis. One way to overcome this problem would be to increase the number of basis set proteins with different compositions of secondary structures so that all secondary structures are well represented.

TABLE 10
Average Values of Secondary Structures in Different Assignments of the X-Ray Structures^a

X ray set	α	β	T	U
KSH	30.6	15.7	22.0	31.8
KS	38.8	18.6	22.0	20.6
HJ	39.4	15.3	16.3	29
LG	42.9	27.4	17.9	11.8

^a See text for the explanation of abbreviations and symbols.

Perczel *et al.* (21,22), in their recently developed "convex constraint analysis" method, predict the secondary structure of a protein from CD spectra without using information from X-ray structures. In this method, they introduce a third constraint, in addition to $\Sigma f_k = 1.0$ and $f_k \geq 0.0$, called "volume minimization" and optimize the vectors corresponding to the fractions f_i in a P -dimensional space, where P is the total number of components (generally $P = 5$ gives the best results). They used this method to calculate the component spectra, which they correlated with the secondary structures, and from these predicted the secondary structures. The RMS differences and the correlation coefficient between the KSH analysis of X-ray structures and the predicted secondary structural fractions for five components were: A 0.112, 0.93; B 0.095, 0.71; C 0.204, 0.73; D 0.094, 0.35; E 0.174, 0.48. We can only qualitatively compare these results with ours due to differences in the basis set. The components correspond to those used in this paper in the following fashion: A with α ; B with β ; C with U, and D + E with T. Their results show high RMS differences and moderate correlations with the X-ray structures. The self-consistent solution approach gives significantly lower RMS deviations and higher correlation coefficients. Further, Perczel *et al.* mention that their method is sensitive to the basis set used (21). The choice of basis set is not a serious factor in the self-consistent method due to the strategy of arranging the proteins in the order of increasing δ_c with the CD spectrum of the protein analyzed and including them in the analysis in a systematic manner. In fact, the self-consistent method, and the LL and VS methods, perform the analysis with different combinations of the basis set proteins and obtain a set of allowed solutions which may be averaged. The number of such allowed solutions may also be considered as an indicator of the reliability of the analysis.

We have also calculated the spectra corresponding to different secondary structure classes using the C matrix with the various acceptable solutions (differing in the number of proteins, N_p , and the number of significant

diagonal elements, N_a) and the corresponding \mathbf{F} matrix. The sets of these spectra (one set per allowed solution) differ with the test protein as the proteins included in the basis will be different for different proteins analyzed. While the proteins with high α content, e.g., myoglobin, generated component spectra with significant similarities with α -helix and unordered spectra, proteins with all secondary structures represented approximately equally generated intermediate curves with more nodes. However, the amplitudes of these component spectra were different with different proteins as test proteins, depending on the amount of information present in the basis set. The component spectra calculated were qualitatively comparable to curves A, B, and C in Fig. 2c of Perczel *et al.* (21).

A new method to analyze protein CD spectra, based on the neural network theory, has been developed by Böhm *et al.* (20). In neural network methods, the processing elements or neurons are considered in several layers. A neuron in a given layer is connected to all neurons in the next layer through numerically weighted connections; information is passed through these connections and processed in neurons. In the learning phase, the weights are adjusted iteratively until all patterns presented in the input layer are correctly projected on the output layer. Böhm *et al.* considered a hidden layer of 45 neurons connecting an input layer of 83 neurons (corresponding to 1-nm intervals over the wavelength range 178–260 nm) and an output layer of 5 neurons (corresponding to 5 secondary structural elements). The basis set, consisting of 13 proteins, was similar to that used in this paper. The weight factors determined by using 9 proteins as the learning set were used to analyze the spectra of 4 proteins and the estimates for each protein were averaged. This may be equivalent to a partial implementation of the variable selection method. Table 11 compares the differences between the X-ray and predicted secondary structural fractions from the neural network method for four proteins, for which results were available, with those from the self-consistent method. Results from the neural network method are better for two of the proteins (lactate dehydrogenase and triose-phosphate isomerase), while the self-consistent method gives better results for the other two proteins (myoglobin and glyceraldehyde-3-phosphate dehydrogenase). However, the correlation coefficients obtained from the neural network method are better than those obtained from the self-consistent method (Table 1). It should be noted that the neural network method used 13 proteins, while the self-consistent method used 16 proteins, with 10 proteins in common. We are currently exploring the possibilities of combining the neural network method with the locally linearized method to improve the accuracy of analysis.

Pancoska and Keiderling (37) have compared the analysis of CD (electronic and vibrational) spectra for

secondary structure prediction in the context of important factor analysis coefficients (38,39) and cluster analysis (40). Factor analysis brings out the important factors contained in a set of data points (39), and cluster analysis enables the classification or grouping of individual samples making use of similarity in the data (40). When applied to proteins with characteristic CD spectra, the proteins with similar spectral characteristics form separate clusters and Pancoska and Keiderling suggest using this classification with the prediction methods, creating subanalyses for different classes, as an improvement over the VS method (18). This can be included in the prediction methods by grouping the protein to be analyzed in the class to which it belongs, determined from the similarities in the CD spectra, and performing the analysis, which should ensure a more reliable prediction. The LL approach (19) followed in this paper incorporates this feature in an indirect manner by grouping the proteins with similar spectra and thus similar structures. Thus proteins belonging to the same class as the test protein will appear in the first columns of the \mathbf{C} and \mathbf{F} matrices and are included in the analysis in a systematic manner. We have used the RMS differences between the CD spectra of the basis set proteins and the test protein as a measure of similarity for the ordering of proteins. The set of solutions obtained with different numbers of proteins included in the analysis would reflect the effects of inclusion of proteins belonging to different classes. However, the number of proteins belonging to each given class is small in our basis set, and hence we are unable to perform detailed subanalyses with proteins belonging to different classes. Efforts to test the validity and performance of such subanalyses by increasing the number of proteins in the basis set and making use of cluster analysis are underway.

Other spectroscopic techniques, viz., infrared (ir), Raman, and vibrational CD, have also been used in the estimation of secondary structures of proteins (28–30,41–45). The analyses of ir spectra have proven to be superior to CD in predicting the β -sheet fraction, while CD analyses provide better estimates of the α -helix fraction. A combination of ir and CD spectra in the secondary structure analyses has been successfully carried out by Sarver and Krueger (46). They combined the CD spectra (178–240 nm) and the ir spectra (1600–1700 cm^{-1}) of 10 proteins in estimating the secondary structure by the SVD method (23). In their data matrix, the ir data were normalized by dividing each absorbance by the total area under the amide I region and the CD data were scaled down by a factor of 0.5. Larger scaling factors led to better prediction of α and smaller values improved the prediction of β , so they selected 0.5 as a compromise. The results obtained with the combined CD/ir approach were significantly superior to those obtained with either CD or ir alone. This approach of combining

TABLE 11
Deviations of the Secondary Structure Estimates from the X-Ray Structure—Comparison of the Self-Consistent Method (SC) with the Neural Network Method (NN)^a

Protein	Secondary structure				RMS
	α	β	T	U	
Lactate dehydrogenase					
SC	-0.029	0.058	-0.036	-0.046	0.044
NN	-0.014	-0.039	0.060	-0.005	0.036
Glyceraldehyde-3-phosphate dehydrogenase					
SC	0.067	-0.054	0.004	0.006	0.043
NN	0.030	-0.109	0.042	0.039	0.063
Triosephosphate isomerase					
SC	0.020	0.069	-0.056	-0.003	0.045
NN	-0.027	-0.021	0.011	0.038	0.026
Myoglobin					
SC	-0.060	0.067	0.016	0.007	0.045
NN	-0.059	-0.028	0.019	0.072	0.049

^a Abbreviations: SC, self-consistent method; NN, neural network method. The results for NN are from Böhm *et al* (20). X-ray structures used are KS for SC and HJ for NN.

two complementary optical techniques is more powerful than the individual techniques alone in predicting the secondary structure of proteins. Extending this approach to include more than two spectroscopic techniques in the analyses should further improve the quality of prediction. We are currently pursuing a similar approach with our self-consistent method.

In conclusion, we have presented a new method, called the self-consistent method, for analyzing the CD spectra of proteins for the secondary structure content and examined its performance with different resolutions of CD data and different analyses of the X-ray data. This method includes the CD spectrum of the protein analyzed in the basis set, assumes an initial guess for the unknown structure, and gives the solution which is most consistent with the basis set of proteins considered. Different initial guesses give similar results. The basic principles of this method are similar to those of the variable selection method and the locally linearized method, but the results obtained are slightly better. On comparison of different analyses of secondary structures from the X-ray data in the context of predicting secondary structure from CD, we find the Kabsch and Sander method to be most suitable. We have also suggested a new criterion for comparison of results from similar analyses which compares the RMS differences with the average fraction of the structural element present in the basis set.

The FORTRAN program for the self-consistent method is available from the authors upon request.

ACKNOWLEDGMENTS

We thank Dr. W. C. Johnson, Jr., for providing the CD spectra of the proteins used in this work and for clarifying aspects of the VS

method and Dr. S. Yu. Venyaminov for helpful discussions. This work was supported by NIH Grant GM22994.

REFERENCES

1. Yang, J. T., Wu, C-S. C. and Martinez, H. M. (1986) *Methods Enzymol.* **130**, 208-269.
2. Johnson, W. C., Jr. (1988) *Annu. Rev. Biophys. Biophys. Chem.* **17**, 145-166.
3. Johnson, W. C., Jr. (1990) *Proteins: Struct. Funct. Genet.* **7**, 205-214.
4. Woody, R. W. (1985) in *The Peptides*, Vol. 7 (Hruby, V. J., Ed.), pp. 15-114, Academic Press, New York.
5. Greenfield, N., and Fasman, G. D. (1969) *Biochemistry* **8**, 4108-4116.
6. Rosenkranz, H., and Scholten, W. (1971) *Hoppe-Seyler's Z. Physiol. Chem.* **352**, 896-904.
7. Brahm, S., and Brahm, J. (1980) *J. Mol. Biol.* **138**, 149-178.
8. Saxena, V. P., and Wetlaufer, D. B. (1971) *Proc. Natl. Acad. Sci. USA* **68**, 969-972.
9. Chen, Y.-H., and Yang, J. T. (1971) *Biochem. Biophys. Res. Commun.* **44**, 1285-1291.
10. Chen, Y.-H., Yang, J. T., and Martinez, H. M. (1972) *Biochemistry* **11**, 4120-4231.
11. Chen, Y.-H., Yang, J. T., and Chan, K. H. (1974) *Biochemistry* **13**, 3350-3359.
12. Chang, C. T., Wu, C-S. C., and Yang, J. T. (1978) *Anal. Biochem.* **91**, 13-31.
13. Bolotina, I. A., Chekhov, V. O., Lugauskas, V. Yu., Finkel'shtein, A. V., and Ptitsyn, O. B. (1980) *Mol. Biol. (Moscow)* **14**, 701-709.
14. Bolotina, I. A., Chekhov, V. O., Lugauskas, V. Yu., and Ptitsyn, O. B. (1980) *Mol. Biol. (Moscow)* **14**, 709-715.
15. Provencher, S. W., and Glöckner, J. (1981) *Biochemistry* **20**, 33-37.
16. Hennessey, J. P., Jr., and Johnson, W. C., Jr. (1981) *Biochemistry* **20**, 1085-1094.

17. Compton, L. A., and Johnson, W. C., Jr. (1986) *Anal. Biochem.* **155**, 155-167.
18. Manavalan, P., and Johnson, W. C., Jr. (1987) *Anal. Biochem.* **167**, 76-85.
19. van Stokkum, I. H. M., Spoelder, H. J. W., Bloemendal, M., van Grondelle, R., and Groen, F. C. A. (1990) *Anal. Biochem.* **191**, 110-118.
20. Böhm, G., Muhr, R., and Jaenicke, R. (1992) *Prot. Eng.* **5**, 191-195.
21. Perczel, A., Hollosi, M., Tusnady, G., and Fasman, G. D. (1991) *Prot. Eng.* **4**, 669-679.
22. Perczel, A., Park, K., and Fasman, G. D. (1992) *Anal. Biochem.* **203**, 83-93.
23. Forsythe, G. E., Malcolm, M. A. and Moler, C. B. (1977) *Computer Methods for Mathematical Computations*, Prentice-Hall, Englewood Cliffs, NJ.
24. Wollmer, A., Strassburger, W., and Glatter, U. (1983) in *Modern Methods in Protein Chemistry—Review Articles* (Tschesche, H., Ed.), pp. 361-384, DeGruyter, Berlin.
25. Venyaminov, S. Yu., Baikalov, I. A., Wu, C-S. C. and Yang, J. T. (1991) *Anal. Biochem.* **198**, 250-255.
26. Toumadje, A., Alcorn, S. W., and Johnson, W. C., Jr. (1992) *Anal. Biochem.* **200**, 321-331.
27. Provencher, S. W. (1982) Technical Report EMBL-DA05, EMBL, Heidelberg, Germany.
28. Sarver, R. W., and Krueger, W. C. (1991) *Anal. Biochem.* **194**, 89-100.
29. Williams, R. W. (1983) *J. Mol. Biol.* **166**, 581-603.
30. Pancoska, P., Yasui, S. C., and Keiderling, T. A. (1989) *Biochemistry* **28**, 5917-5923.
31. Levitt, M., and Greer, J. (1977) *J. Mol. Biol.* **114**, 181-293.
32. Kabsch, W., and Sander, C. (1983) *Biopolymers* **22**, 2577-2637.
33. Woody, A-Y. M., Sreerama, N., and Woody, R. W. (1991) *Proceedings 4th International Conference on CD*, Bochum, Germany, pp. 322-323.
34. Richards, F. M., and Kundrot, C. E. (1988) *Proteins: Struct. Funct. Genet.* **3**, 71-84.
35. Pancoska, P., Blazek, M., and Keiderling, T. A. (1992) *Biochemistry* **31**, 10250-10257.
36. Bobba, A., Cavatorta, P., Attimonelli, M., Riccio, P., Masotti, L., and Quagliariello, E. (1990) *Protein Seq. Data. Anal.* **3**, 7-10.
37. Pancoska, P., and Keiderling, T. A. (1991) *Biochemistry* **30**, 6885-6895.
38. Pancoska, P., Fric, I., and Blaha, K. (1979) *Collect. Czech. Chem. Commun.* **44**, 1296-1312.
39. Malinowski, E. R., and Howry, D. G. (1980) in *Factor Analysis in Chemistry*, Wiley, NY.
40. Sharaf, M. A., Illman, D. L. and Kowalski, B. R. (1986) in *Chemometrics, Chemical Analysis*, Vol. 82, pp. 179-295, Wiley-Interscience, New York.
41. Dousseau, F., and Pezolet, M. (1990) *Biochemistry* **29**, 8771-8779.
42. Lee, D. C., Haris, P. I., Chapman, D., and Mitchell, R. C. (1990) *Biochemistry* **29**, 9185-9193.
43. Kalnin, N. N., Baikalov, I. A., and Venyaminov, S. Yu. (1990) *Biopolymers* **30**, 1273-1280.
44. Byler, D. M., and Susi, H. (1986) *Biopolymers* **25**, 469-487.
45. Dong, A., Huang, P., and Caughey, W. S. (1990) *Biochemistry* **29**, 3303-3308.
46. Sarver, R. W., and Krueger, W. C. (1991) *Anal. Biochem.* **199**, 61-67.