not only by spectra but also by the recovery of enzymatic activity. The complete (100%) recovery of enzymatic activity was confirmed for the refolded protein in the present study (data not shown). (2) The complete unfolding of proteins in the initial condition, with use of GdmCl at high concentration (Tanford, 1968), may not always be necessary, because the unfolded molecules may anyway take a specific secondary structure characteristic of the refolding condition before the molecules start refolding.

Acknowledgments

We thank Robert L. Baldwin for critical reading of the manuscript and for valuable suggestions.

References

Baldwin, R. L. (1975) *Annu. Rev. Biochem. 44*, 453–475.
Baldwin, R. L. (1978) *Trends Biochem. Sci. 3*, 66–68.
Brandts, J. F., Halvorson, H. R., & Brennen, M. (1975) *Biochemistry 14*, 4953–4963.
Brandts, J. F., Brennen, M., & Lin, L.-N. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 4178–4181.
Canfield, R., & Liu, A. K. (1965) *J. Biol. Chem. 240*, 1977–2002.
Creighton, T. E. (1979) *J. Mol. Biol. 129*, 411–431.
Garel, J.-R., & Baldwin, R. L. (1973) *Proc. Natl. Acad. Sci. U.S.A. 70*, 3347–3351.
Garel, J.-R., & Baldwin, R. L. (1975a) *J. Mol. Biol. 94*, 611–620.
Garel, J.-R., & Baldwin, R. L. (1975b) *J. Mol. Biol. 94*, 621–632.
Hagerman, P. J. (1977) *Biochemistry 16*, 731–747.
Hagerman, P. J., & Baldwin, R. L. (1976) *Biochemistry 15*, 1462–1473.
Hammes, G. G., & Roberts, P. B. (1969) *J. Am. Chem. Soc. 91*, 1812–1816.
Hiromi, K., Ono, S., & Nagamura, T. (1968) *J. Biochem. (Tokyo) 64*, 897–900.
Karplus, M., & Weaver, D. C. (1976) *Nature (London) 260*, 404–406.
Lin, L.-N., & Brandts, J. F. (1978) *Biochemistry 17*, 4102–4181.
Manjula, B. N., Acharya, A. S., & Vithayathil, P. J. (1976) *Int. J. Protein Res. 8*, 275–282.
Nall, B. T., Garel, J.-R., & Baldwin, R. L. (1978) *J. Mol. Biol. 118*, 317–330.
Nozaki, Y. (1972) *Methods Enzymol. 26*, 43–50.
Schmid, F. X., & Baldwin, R. L. (1978) *Proc. Natl. Acad. Sci. U.S.A. 75*, 4764–4768.
Schmid, F. X., & Baldwin, R. L. (1979) *J. Mol. Biol. 135*, 199–215.
Schwarz, G. (1965) *J. Mol. Biol. 11*, 64–77.
Sophianopoulos, A. J., Rhodes, C. K., Holcomb, D. N., & Van Holde, K. E. (1962) *J. Biol. Chem. 237*, 1107–1112.
Stellwagen, E. (1979) *J. Mol. Biol. 135*, 217–229.
Tanford, C. (1968) *Adv. Protein Chem. 23*, 121–282.
Tanford, C., Aune, K. C., & Ikai, A. (1973) *J. Mol. Biol. 24*, 189–197.
Tonomura, B., Nakatani, H., Onishi, M., Yamaguchi-Ito, J., & Hiromi, K. (1978) *Anal. Biochem. 84*, 370–383.
Wetlaufer, D. B., Johnson, E. R., & Clauss, L. M. (1974) in *Lysozyme* (Osserman, E. F., Canfield, R. E., & Beychok, S., Eds.) pp 269–280, Academic Press, New York.

Information Content in the Circular Dichroism of Proteins†

John P. Hennessey, Jr., and W. Curtis Johnson, Jr.*

ABSTRACT: A method is presented for predicting the secondary structure of a protein from its circular dichroism (CD) spectrum. Eight types of secondary structure are considered: helix; parallel and antiparallel β strand; types I, II, and III β turn; all other β turns combined; and "other" structures. The method is based on mathematical calculation of orthogonal basis CD spectra from the CD spectra of proteins with known secondary structure. Five basis CD spectra are needed to reconstruct the 16 original protein CD spectra that extend into the vacuum ultraviolet region to 178 nm. Thus, one can expect to extract five independent pieces of information from the CD spectrum of a protein. Each basis CD spectrum corresponds to a known mixture of secondary structures so that the coefficients that reconstruct the protein CD spectrum can also be used to predict secondary structure. Furthermore, when the same method is applied to protein secondary structure rather than CD, it is found that only five basis secondary structure vectors are needed to reconstruct the original protein secondary structure vectors. Thus there are five independent "superstructures", consisting of a mixture of standard secondary structures, in the proteins studied. It would appear that there is enough information in the CD spectrum of a protein to predict all types of secondary structure. Our CD analyses compare favorably with the X-ray data.

It is generally accepted that the circular dichroism (CD)[1] spectrum of a protein is a direct reflection of its secondary structure. Over the past 15 years attempts have been made to correlate the two by using a variety of theoretical and empirical techniques, all finding various degrees of success. Use of CD spectra of polypeptides, in theoretically known structural conformations, as basis spectra has been one of the foremost techniques (Greenfield et al., 1967; Greenfield &

† From the Department of Biochemistry and Biophysics, Oregon State University, Corvallis, Oregon 97331. *Received August 5, 1980.* This work was supported by National Science Foundation Grant PCM76-81556 from the Biophysics Program, National Institutes of Health Biomedical Research Support Grant RR07079, and a grant from the Oregon State University Computer Center.

[1] Abbreviations used: CD, circular dichroism; vacuum UV, vacuum ultraviolet; H, helix; A, antiparallel β strand; P, parallel β strand; I, type I β turn; II, type II β turn; III, type III β turn; T, remaining types of β turn combined; O, "other" structure; *r*, correlation coefficient; rms, root mean square.

0006-2960/81/0420-1085$01.25/0

Fasman, 1969; Saxena & Wetlaufer, 1971; Brahms & Brahms, 1980). However, cross-checking of the basis CD spectra shows notable variations and brings about questions as to the actual structural conformations represented by the spectra (Baker & Isenberg, 1976).

An alternative approach is based on mathematical calculation of basis CD spectra using a set of CD spectra of proteins with known secondary structure. Saxena & Wetlaufer (1971) used three protein CD spectra to derive basis CD spectra for helix, β-sheet, and "random" structure. Chen et al. (1972, 1974) used sets of five and eight protein CD spectra to derive these three basis CD spectra. Chang et al. (1978) used 15 protein CD spectra and in turn derived spectra for helix, β-sheet, a general β-turn, and unordered structure. In all cases the CD spectra were terminated at 190 nm.

Recently, Siegel et al. (1980) used statistical methods to show that the CD of 16 proteins between 210 and 240 nm correlates only with the amount of α helix. Thus it is not surprising that workers have been unable to predict the amount of β-stand and "random" structure with CD data restricted to wavelengths longer than 200 nm.

We use an eigenvector method of multicomponent matrix analysis (Lloyd, 1969) with a set of 15 protein CD spectra and one polypeptide CD spectrum over the range 178–260 nm to generate a set of orthogonal CD spectra for use as a basis. These basis CD spectra are used to estimate the fraction of eight types of secondary structure as a function of the protein CD spectra. The secondary structures include two types of β strand and four types of β turn.

Experimental Procedures

Materials. α-Chymotrypsin (bovine), cytochrome *c* (horse heart), hemoglobin (bovine, lysozyme (egg white), myoglobin (whale), papain (papaya), and subtilisin BPN′ (*Bacillus amyloliquefaciens*) were obtained from Sigma Chemical Co. Elastase (porcine) and ribonuclease A (bovine pancreas) were obtained from Worthington Biochemical Corp. Lactate dehydrogenase (chicken heart) was the generous gift of Dr. Sonia Anderson.

Preparation. All proteins were dissolved in 0.01 M sodium phosphate buffer, pH 6.8. The solutions were then dialyzed at 5 °C against 1 L of buffer, changed twice over the course of 2 days. Dialyzed protein solutions were then clarified by using a 0.45-μm Millipore filter and stored at 5 °C.

Spectra. CD spectra were measured over the range of approximately 178–260 nm by using a vacuum ultraviolet (UV) CD spectrophotometer (Johnson, 1971). The instrument was calibrated by using (+)-10-camphorsulfonic acid, $\Delta\epsilon$ = 2.42 at 290.5 nm. Most measurements were made by using 50- and 100-μm pathlength cells and protein solutions of approximately 1 mg/mL. Pathlengths were checked by using an infrared spectrophotometer (Bree & Lyons, 1956). Each solution was measured a minimum of three times, on separate days, to assure that the solutions were not decomposing with time. When possible, our spectra were compared to published data for verification of shape and/or intensity. All spectra were measured at 25 ± 1 °C with a 10-s time constant and a scan rate of 2 nm/min. The spectral slit width was a constant 1.6 nm.

Transmission spectra were measured in the same instrument and extinction coefficients determined for each protein at 190 nm (Table I). CD spectra were terminated at the point at which the transmission spectra indicated the noise to signal ratio was greater than 0.1 (in most cases 176–178 nm). In all cases the total OD did not exceed 1.0 within the bounds of the CD spectrum.

Table I: Protein Extinction Coefficients

proteins	ϵ_{190} (L mol^{-1} cm^{-1})	proteins	ϵ_{190} (L mol^{-1} cm^{-1})
α-chymotrypsin	9690 ± 50	lysozyme	11460 ± 160
cytochrome *c*	9720 ± 50	myoglobin	9150 ± 80
elastase	10290 ± 150	papain	10100 ± 200
hemoglobin	9620 ± 140	ribonuclease	9640 ± 280
lactate dehydrogenase	8510 ± 40	subtilisin BPN′	8850 ± 130

Table II: Molar Color Yield for Proteins and Standards

sample	color yield	sample	color yield
leucine	1.00	lactate dehydrogenase	1.03
amino acid standard	0.91	lysozyme	1.05
α-chymotrypsin	1.02	myoglobin	1.01
cytochrome *c*	1.03	papain	1.04
elastase	1.00	ribonuclease	1.06
hemoglobin	1.00	subtilisin BPN′	1.02

Additional CD spectra of flavodoxin (*Desulfovibrio vulgaris*), glyceraldehyde-3-phosphate dehydrogenase (rabbit muscle), prealbumin, subtilisin Novo (*Bacillus amyloliquifaciens*), and triosephosphate isomerase (rabbit muscle) were obtained from Brahms & Brahms (1980). The spectrum of poly(L-glutamic acid) was obtained from Johnson & Tinoco (1972). These spectra are redrawn here for completeness and for the convenience of the reader.

Analysis of Protein Concentrations. Protein concentrations were determined by taking duplicate 60-μL aliquots from the protein solution on 2 successive days. Each sample was hydrolyzed by using 2.0 mL of constant-boiling HCl in separate, nitrogen-flushed, evacuated, 5-mL drying ampules, and heating to 110 °C in a boiling toluene bath for 22 h. Sample ampules were then rotoevaporated to dryness and the residue was redissolved in 1.00 mL of 0.20 N sodium citrate buffer (with 0.5% thiodiglycol solution, 0.01% octanoic acid, and 0.06% BRIJ solution), pH 2.20. From each ampule a 0.50-mL aliquot was extracted and rediluted with an additional 2.00 mL of the 0.20 N sodium citrate buffer. To the resulting solution was added 1.25 mL of a ninhydrin solution (Moore, 1968). The final solution was mixed by inversion, heated to 100 °C in a boiling water bath for 20 min, and cooled to room temperature in a water bath. Standards were run in a similar manner (deleting hydrolysis) with each set of samples. Standards were prepared by using amino acid standard H, No. 20089, from Pierce Biochemicals. Processed samples and standards were then measured for OD at 570 nm on a Cary 14 spectrophotometer. We measured the total amino acid content of each protein by using the ninhydrin reaction common to amino acid analysis, but without separating the residues. The molar color yield ratio for each protein (Table II) was calculated by using Table III as modified from Moore & Stein (1948, 1954).

From the OD_{570} readings of the standards and the samples, a concentration can be computed as moles of amino acid equivalent. Multiplication of each protein concentration (in molar equivalents) by its corresponding molar color yield ratio will yield its true concentration in moles of amino acid per volume. Reproducibility by this method is ±2%. Results were checked and verified by using a modified Beckman 120B amino acid analyzer.

Protein Secondary Structure. Secondary structure data for each of the proteins were obtained from analysis of X-ray diffraction data found in the "Atlas of Macromolecular Structure on Microfiche" (AMSOM), (Feldman, 1976). Published X-ray data were interpreted by using a set of criteria

Table III: Molar Color Yield for Amino Acids

amino acid	color yield	amino acid	color yield
Asp	0.94	Met	1.02
Asn[a]	1.92	Ile	1.00
Thr[b]	0.94	Leu	1.00
Ser[b]	0.95	Tyr	1.00
Glu	0.99	Phe	1.00
Gln[a]	1.96	Lys	1.10
Pro	0.05	His	1.02
Gly	0.95	Trp[b]	0.97
Ala	0.97	Arg	1.01
Val	0.97	Cys^+	0.99
Cys/2[b]	0.55	NH_3	0.97

[a] Asn and Gln release an amino group upon hydrolysis, thus adding an NH_3 equivalent to their color yield. [b] Amino acids that are destroyed during hydrolysis leave an intact α-amino group to react with ninhydrin. These figures are adjusted to reflect the average color yield after hydrolysis.

designed to create a structural data matrix, **S**, that was both internally consistent and oriented toward contributions of the secondary structure to the CD of the amide chromophore. As part of these criteria we used the amide groups of the protein molecules as our basic unit rather than the amino acid residues.

Helical structure was identified by visual inspection of molecular stereo drawings with confirmation of end points of the helices from "possible β bend" and "possible hydrogen bond" tables in AMSOM. It was usually possible to differentiate α- and 3_{10}-type helices on the basis of "possible hydrogen bond" data. However, for the present we have chosen to combine all helical structures (H) because there is little 3_{10}-type helix in our data set. Designation of helical structure was confined to those segments that met all definitions of being helical and contained at least four amide groups (five residues).

β strands were identified in much the same manner as were helices with a similar confirmation of endpoints. Of the various types of protein secondary structure, β strand shows the greatest dependence of CD signal on strand length (Woody, 1969). This indicates a strong reinforcement pattern in the intrastrand amide–amide interactions with the strength of the signal coming from the center of the strand. For this reason, terminal amides of β strands, when adjacent to "other" structures, were counted as half β strand and half "other" structure. Additionally, a minimum size of three amide groups (four residues) was maintained for β-strand designation. The strands were divided into the parallel (P) and the antiparallel (A) conformations.

β turns were identified primarily by the "possible β bend" table and confirmed by visual inspection of the stereo drawings. β turns were classified into types as per the criteria of Chou & Fasman (1977). However, for our purposes all β turns other than types I, II, and III were grouped into a general category of turns (T) because of the scarcity of the remaining turn types in our data set. Special treatment was given in designating structure in the case of hairpin turns involved in a β sheet in that half of each end amide of the β turn was allocated to the β-strand designation while the remainder counted as β turn.

All amides unaccounted for under the designation of helix, β strand, or β turn were placed in a category designed "other" (O), that is sometimes called random, aperiodic, or irregular.

Method of Analysis

Eigenvector Method. A number of methods of component analysis have been used in the past for elucidation of the number and composition of significant components found in a data matrix. The most multidisciplinary approach has tended to be factor analysis (Harman, 1976); however, matrix rank analysis has been used successfully in spectroscopic work (McMullen et al., 1967). A related technique using modern matrix algebra allows for the quick processing of large data sets to generate orthogonal components in the form of eigenvectors with associated eigenvalues. This eigenvector method of multicomponent analysis (Lloyd, 1969) allows for either a qualitative or quantitative evaluation of the significance of each component and suits itself very well to vectoral data.

In our application of this technique we started with a 16 × 42 data matrix, **C**, containing the CD spectra of 15 proteins and a helical polypeptide in the range 178–260 nm at 2-nm increments. A symmetric square matrix containing all our knowledge of the system is constructed by multiplying **C** by its transpose, $\mathbf{C}^T$. Diagonalization of this square matrix produces a matrix of 16 eigenvectors, **U**, and a diagonal matrix of 16 eigenvalues, **E**, such that

$$(\mathbf{C}\mathbf{C}^T)\mathbf{U} = \mathbf{U}\mathbf{E}$$

The matrix of basis CD spectra, **B**, is generated by the operation

$$\mathbf{B} = \mathbf{U}^T\mathbf{C}$$

These 16 orthogonal basis CD spectra are linear combinations of the original protein CD spectra. They are valuable because the relative importance of each basis CD spectrum, as a component of the original protein CD spectra, goes as the square root of its corresponding eigenvalue. They are unique because any limited number of basis CD spectra with the largest corresponding eigenvalues will give the best possible reproduction of the original protein CD spectra for that limited number of bases.

Within the framework of the eigenvector method of multicomponent analysis, the variance in the data set unaccounted for by the μ most important basis CD spectra is given by (Lloyd, 1969)

$$\sigma_\mu^2 = \frac{1}{n(m-\mu)} \sum_{i=\mu+1}^{m} e_i$$

Here σ is the standard deviation, n is the number of data points in each original vector, m is the number of original vectors, μ is the number of basis vectors in the truncated set used for reconstruction, and the e_i are the eigenvalues arranged in decreasing magnitude. The random error in intensity from noise and repeatability is about 0.3 $\Delta\epsilon$ unit for our measured CD spectra. From the equation above, σ for our reconstruction of the protein CD spectra is 0.38 for $\mu = 3$, 0.24 for $\mu = 4$, 0.17 for $\mu = 5$, and 0.12 for $\mu = 6$. Four basis CD spectra bring σ within our error level, but shapes are poor, as might be expected for a single standard deviation. Five basis CD spectra give a reconstruction within two standard deviations of the measured values and provide significant fine structure in the shape of the original protein CD spectra. A sixth basis CD spectrum provides little improvement.

Errors in $\Delta\epsilon$ units for a root mean square (rms) error comparison between the measured CD spectrum for each individual protein and its reconstructed CD spectrum using five basis CD spectra are given in Table IV. The values range from 0.25 for ribonuclease A to 0.08 for elastase. Correlation coefficients are not helpful here since they are all 0.99 or 1.00. The five basis CD spectra are graphed in Figure 1 and tabulated in Table V. Figure 2 shows the reconstruction of the papain CD spectrum using the five basis CD spectra.

Protein Analysis. Having established the truncation of our basis set at $\mu = 5$, we can reconstruct our original protein CD spectra minus the insignificant components (i.e., noise). When

Table IV: Error in Proteins[a]

protein	rms CD meas − recon	r structure XRAY − CD	r structure XRAY − PRE
α-chymotrypsin	0.20	0.95	0.91
cytochrome *c*	0.14	0.98	0.95
elastase	0.08	1.00	0.99
flavodoxin	0.12	0.74	0.66
glyceraldehyde-3-phosphate dehydrogenase	0.11	0.89	0.87
hemoglobin	0.12	0.99	0.95
lactate dehydrogenase	0.10	0.99	0.99
lysozyme	0.19	0.83	0.76
myoglobin	0.15	1.00	0.99
papain	0.11	0.99	0.99
prealbumin	0.17	0.92	0.61
ribonuclease	0.25	0.94	0.91
subtilisin BPN'	0.13	0.91	0.86
subtilisin Novo	0.10	0.96	0.89
triosephosphate isomerase	0.13	0.99	0.98
poly(L-glutamic acid)	0.09	1.00	0.96

[a] Abbreviations: rms, root mean square; *r*, correlation coefficient; XRAY, structure from X-ray data; CD, structure from analysis of CD using basis CD spectra; PRE, structure from analysis of CD using basis CD spectra constructed without protein being analyzed.

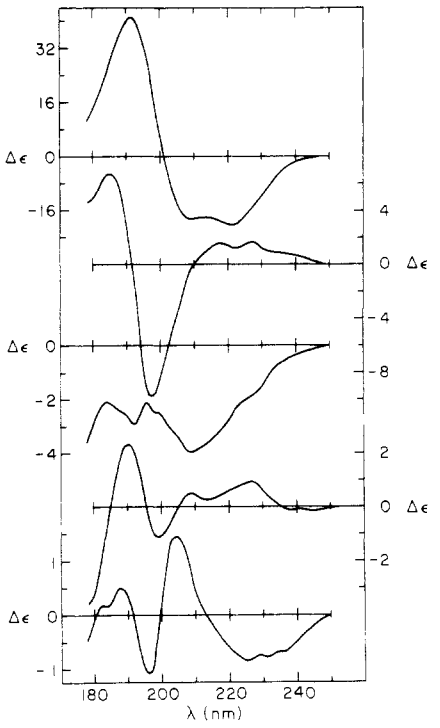

FIGURE 1: Five most significant basis CD spectra generated by the eigenvector method of multicomponent matrix analysis in descending order of significance starting at the top.

the subscript μ is used, to denote the truncation, the operation is $\mathbf{C}_\mu = \mathbf{U}_\mu\mathbf{B}_\mu$. This establishes each protein CD spectrum as a function of $\mathbf{B}_\mu$, so that

$$C_i = u_{i1}B_1 + u_{i2}B_2 + u_{i3}B_3 + u_{i4}B_4 + u_{i5}B_5$$

where *i* indexes the individual protein vectors and their corresponding coefficients in the **U** matrix.

By maintaining our premise that protein CD spectra and protein secondary structures are highly correlated, it is possible to carry the eigenvector method of analysis one step further. Each basis CD spectrum, B_i, corresponds to a mixture of secondary structures (in analogy to a protein) represented by

Table V: Basis CD Spectra

λ (nm)	basis spectra (Δε) 1	2	3	4	5
178	10.4	4.6	−3.6	−3.6	−0.5
180	13.5	4.9	−2.9	−3.3	−0.1
	18.6	6.0	−2.3	−2.1	0.2
	24.4	6.7	−2.1	−0.7	0.2
	30.8	6.6	−2.3	0.7	0.3
	36.4	5.3	−2.5	1.9	0.5
190	40.0	2.3	−2.6	2.3	0.4
	41.2	−1.2	−2.9	2.0	−0.2
	36.3	−6.0	−2.5	0.9	−0.8
	28.0	−9.6	−2.1	−0.5	−1.1
	17.4	−9.6	−2.5	−1.1	−0.9
200	5.4	−8.5	−2.5	−1.1	0.2
	−3.8	−6.1	−3.0	−0.8	1.1
	−11.7	−4.4	−3.3	−0.2	1.4
	−16.8	−2.6	−3.6	0.2	1.4
	−18.3	−0.9	−4.0	0.5	1.0
210	−18.5	0.2	−3.9	0.5	0.5
	−18.3	0.6	−3.8	0.3	0.1
	−18.4	1.0	−3.6	0.3	−0.1
	−18.9	1.4	−3.4	0.3	−0.3
	−19.4	1.6	−3.1	0.5	−0.5
220	−20.1	1.4	−2.8	0.5	−0.6
	−20.3	1.2	−2.3	0.7	−0.7
	−19.3	1.3	−2.1	0.8	−0.8
	−17.2	1.7	−2.0	0.9	−0.8
	−14.8	1.7	−1.8	0.9	−0.7
230	−11.7	1.2	−1.5	0.5	−0.8
	−9.1	1.0	−1.1	0.3	−0.7
	−6.8	0.9	−0.8	0.1	−0.7
	−4.4	0.8	−0.6	0	−0.7
	−2.6	0.7	−0.5	−0.1	−0.5
240	−1.4	0.6	−0.4	−0.1	−0.4
	−0.6	0.5	−0.3	−0.1	−0.3
	−0.3	0.3	−0.2	−0.1	−0.2
	−0.2	0.2	−0.1	−0.1	−0.1
	0	0	0	0	0
250	0	0	0	0	0

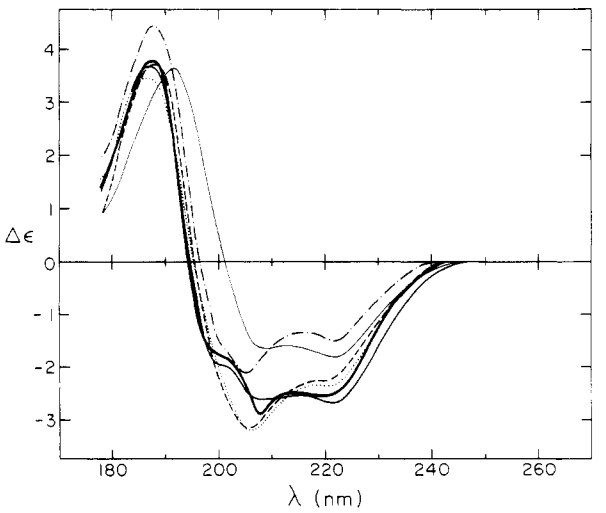

FIGURE 2: Reconstruction of the CD spectrum of papain using one to five basis CD spectra. One basis spectrum (—); two basis spectra (---); three basis spectra (- - -); four basis spectra (···); five basis spectra (—); measured spectrum (▬).

the vector D_i', as given in Table VI. The structure matrix **D'** is the same linear combination of protein data used to generate **B**; however, the matrix of protein secondary structures, **S**, is substituted for the CD spectra matrix **C** such that $\mathbf{D}' = \mathbf{U}^T\mathbf{S}$. Protein secondary structure can be estimated from

Table VI: Secondary Structures Corresponding to Basis CD Spectra[a]

basis spectra	H	A	P	I	II	III	T	O
1	1.78	0.11	0.12	0.10	0.03	0.09	0.08	0.57
2	−0.29	0.02	−0.17	−0.05	0.00	−0.04	−0.05	−0.31
3	0.36	0.55	0.09	0.20	0.11	0.07	0.14	1.02
4	−0.14	0.21	0.13	0.04	0.00	0.00	0.02	0.11
5	0.07	−0.12	0.14	0.00	0.00	−0.01	0.02	0.46

[a] Abbreviations: H, α- and 3_{10}-helix combined; A, antiparallel β sheet; P, parallel β sheet; I, type I β turn; II, type II β turn; III, type III β turn; T, remaining β turns combined; O, other or "random" structures.

Table VII: Secondary Superstructures

superstructures	H	A	P	I	II	III	T	O
XRAY, i	1.77	0.30	0.20	0.19	0.07	0.12	0.14	1.06
ii	0.56	−0.47	−0.06	−0.11	−0.07	−0.01	−0.09	−0.76
iii	0.06	0.38	−0.12	0.01	0.02	0.01	0.01	−0.18
iv	0.00	0.06	0.27	−0.04	−0.02	0.00	0.03	−0.06
v	−0.01	−0.01	0.02	0.16	0.02	0.05	0.00	−0.03
CD, i	1.77	0.31	0.20	0.19	0.07	0.11	0.14	1.07
ii	0.54	−0.41	−0.11	−0.13	−0.07	−0.02	−0.09	−0.72
iii	0.04	0.31	−0.13	0.05	0.03	0.02	0.01	−0.15
iv	0.00	0.04	0.16	0.01	−0.03	0.01	0.01	−0.05
v	0.00	−0.01	0.00	0.02	0.01	0.03	0.03	−0.01

[a] Abbreviations as in Table VII.

the truncated set of eigenvectors by the operation $\mathbf{S}_\mu' = \mathbf{U}_\mu\mathbf{D}_\mu'$. This establishes each protein structure from its CD spectrum as a function of the secondary structures, D_μ', such that

$$S_i' = u_{i1}D_1' + u_{i2}D_2' + u_{i3}D_3' + u_{i4}D_4' + u_{i5}D_5'$$

We find that while five eigenvectors are significantly better than four in reconstructing the X-ray structural data, a sixth eigenvector provides little improvement. This is consistent with our choice of five eigenvectors from consideration of the protein CD spectra.

Secondary Superstructures. The eigenvector method can also be applied to our X-ray structural data. The symmetric square matrix $\mathbf{SS}^T$ that contains all our structural knowledge of this system is diagonalized in an analogous manner. The five most important eigenvectors (independent secondary superstructures) given in Table VII reconstruct the original X-ray structural data with an rms error for each type of structure of less than 0.001. The correlation, r(XRAY-SS), for each type of secondary structure is given in Table VIII. In spite of the low rms error, r is somewhat less than 1.00 for the β turns because of the small fractions of I, II, III, and T in each protein (Table IX). Four eigenvectors also give a low rms error, but give significantly lower r values for the β turns.

Since only five basis CD spectra will reconstruct the original CD spectra of the 16 proteins, we can learn about no more than five independent types of secondary structure from a CD spectrum. Fortunately, the eigenvector analysis of secondary structure shows that only five independent superstructures are necessary to describe the eight standard secondary structures considered in this paper. Thus we are justified in considering the eight standard secondary structures (which are not independent to within a small error) in our protein analyses.

CD Intensities. We produced internal consistency in the CD intensities of our protein data set by assuming that the sum of the structural contributions found by CD should come to 1.0 for each protein. In some proteins it may be argued that, theoretically, structural contributions sum to 1.0, but because of cancellation caused by mirror image β turns, structural contributions found by CD should be less than 1.0. However, few proteins actually have enough in the way of canceling β turns to show a significant effect within the realm of CD measurements. Furthermore, when mirror image β turns cancel, only that portion of the CD corresponding to the β-turn structure would be canceled, thereby leaving the remaining contribution due to the intrinsic asymmetry of the α carbons. This remainder would be equivalent in most respects to the amide groups already found in the "other" category. Therefore, for the purpose of further discussion it will be assumed that each protein will have structural contributions found by CD that sum to 1.0.

Table VIII: Error in Structures[a]

structures	r XRAY − SS	rms XRAY − CD	r XRAY − CD	rms XRAY − PRE	r XRAY − PRE
H	1.00	0.06	0.98	0.08	0.95
A	1.00	0.07	0.83	0.10	0.66
P	1.00	0.05	0.71	0.07	0.51
I	0.99	0.04	0.53	0.05	−0.07
II	0.63	0.01	0.77	0.02	0.51
III	0.50	0.02	0.36	0.05	−0.44
T	0.65	0.02	0.71	0.02	0.38
O	1.00	0.07	0.88	0.10	0.72
I + II + III + T	0.90	0.05	0.78	0.08	0.25

[a] Abbreviations as in Tables IV and VI; SS, secondary superstructures.

As long as the sample is pure, we expect the shape of our CD spectra to be reliable. However, a variety of systematic and random errors cause error in the intensity measured. Using the eigenvector method of multicomponent analysis, it is possible to combine the concepts of internally consistent CD intensities and generation of structural components that sum to 1.0.

Ideally, the eigenvector method is able to reconstruct input vectors by using a limited number of basis vectors. This will work best if all of the input vectors are in some way normalized to one another to bring about some internal consistency. Here we chose to normalize the CD spectra by requiring that all reconstructed CD spectra represent 100% of the secondary structure of their respective proteins. Examination of original CD spectra showed that, upon reconstruction using five basis CD spectra, all proteins except α-chymotrypsin reproduced structural contributions of 100 ± 10%, indicating that these

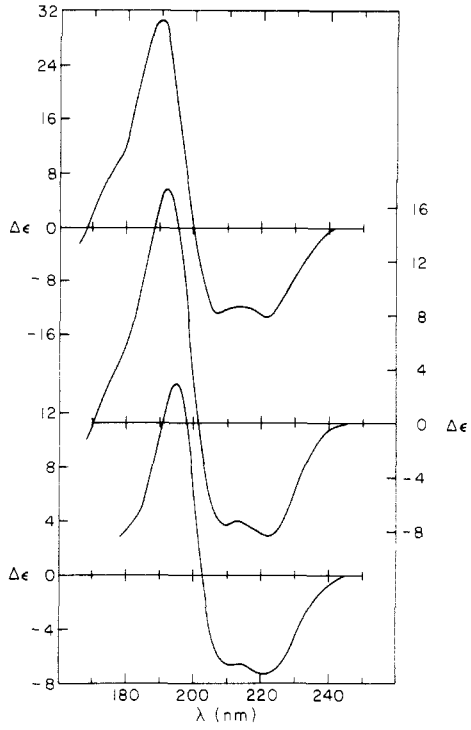

FIGURE 3: Three CD spectra from our original data set. From top to bottom: poly(L-glutamic acid) [from Johnson & Tinoco (1972)], myoglobin, and hemoglobin.

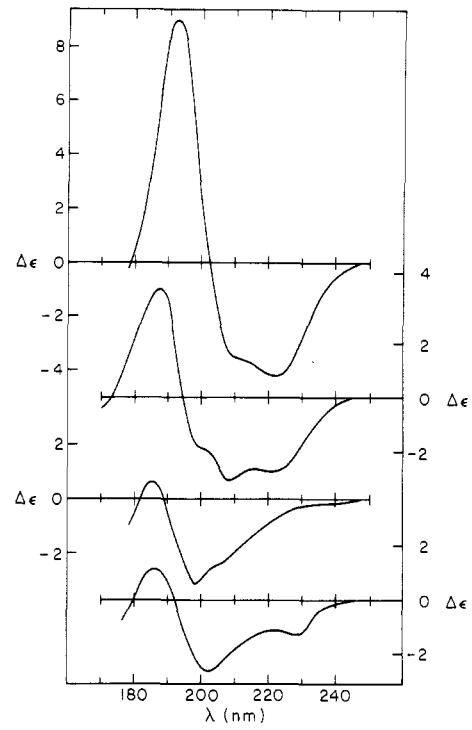

FIGURE 4: Four CD spectra from our original data set. From top to bottom: lactate dehydrogenase, papain, elastase, and α-chymotrypsin.

CD spectra were all within their expected limits of error. This discrepancy for α-chymotrypsin can be viewed as marginal, but acceptable, since the 190-nm band has exceptionally low intensity. From this premise the intensities of each of the 16 CD spectra were adjusted so that, upon eigenvector analysis, the reconstructed spectra represented the desired 100% of structure.

Under our assumptions that (1) five basis CD spectra should provide 100% of the secondary structure (the remaining basis CD spectra provide structure from our analysis of the noise) and (2) any deviation from 100% structure is due to a multiplicative error in the intensity, the method provides a way of averaging random intensity errors in the CD spectra of *different* proteins. This multiplicative adjustment is very different from constraining the sum of structures to equal 1.0 during least-squares fitting.

Results and Discussion

CD Spectra. CD spectra measured in this laboratory and not yet adjusted for internal consistency of intensity are given in Figures 3–5. Five other CD spectra from Brahms & Brahms (1980) that are used in this work are redrawn for convenience in Figure 6. Examination of the 16 CD spectra reveals a number of features that are commonly shared. Most obvious is the presence of three major bands at approximately 190, 208, and 222 nm. Equally apparent is the positive nature of the 190-nm band and the negative nature of the 208- and 222-nm bands. This generalized picture of a protein CD spectrum, while consistent with all samples presented here, shows wide individuality. The 190-nm band can be seen to shift from 185 (in elastase) to 195 nm (in cytochrome *c*) with a 50-fold fluctuation in intensity. The 208- and 222-nm bands likewise show considerable individuality with shifts of up to ±8 nm in either band. However, intensity shifts are limited to 4-fold. The band shifts, combined with independent intensity fluctuations, allow for a variety of patterns in this region of the spectrum. Moreover, lesser peripheral bands are apparent in papain, cytochrome *c*, ribonuclease A, myoglobin, hemoglobin, and poly(L-glutamic acid).

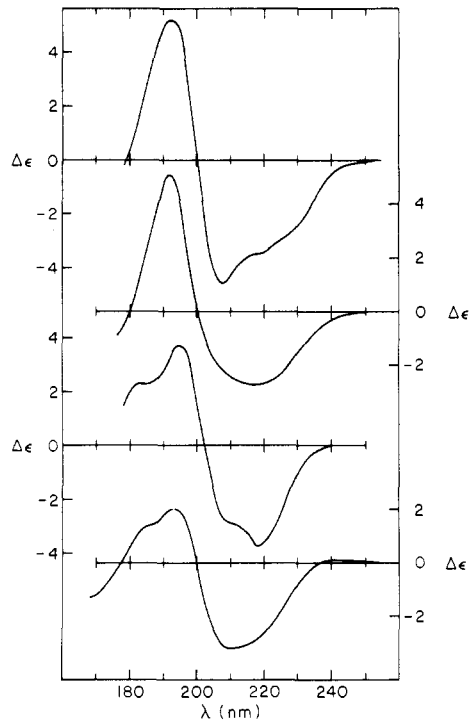

FIGURE 5: Four CD spectra from our original data set. From top to bottom: lysozyme, subtilisin BPN′, cytochrome *c*, and ribonuclease A.

The frequency of band shifting indicates that there are more than three distinct bands in the vacuum-UV region. The prevalence of peripheral bands would seem to bear this out as well as account for the frequency of band shifting.

The presence of a broad peripheral band centered at approximately 185 nm is indicated in the spectra of poly(L-glutamic acid), myoglobin, and hemoglobin. This would ef-

Table IX: Protein Secondary Structures[a]

protein	method	H	A	P	I	II	III	T	O	total
α-chymotrypsin	XRAY	0.10	0.34	0.00	0.11	0.03	0.02	0.04	0.36	1.00
	CD	0.15	0.24	0.02	0.08	0.05	0.02	0.05	0.39	1.00
	PRE	0.16	0.21	0.02	0.07	0.05	0.03	0.06	0.39	0.99
cytochrome *c*	XRAY	0.38	0.00	0.00	0.03	0.06	0.01	0.07	0.45	1.00
	CD	0.42	0.01	0.02	0.05	0.03	0.03	0.04	0.40	1.00
	PRE	0.44	0.03	0.02	0.06	0.01	0.04	0.03	0.34	0.97
elastase	XRAY	0.10	0.27	0.00	0.09	0.07	0.03	0.03	0.41	1.00
	CD	0.10	0.28	0.01	0.09	0.05	0.03	0.05	0.39	1.00
	PRE	0.10	0.28	0.01	0.09	0.04	0.03	0.07	0.38	1.00
flavodoxin	XRAY	0.38	0.00	0.24	0.06	0.02	0.03	0.05	0.22	1.00
	CD	0.27	0.09	0.10	0.06	0.02	0.03	0.05	0.38	1.00
	PRE	0.26	0.11	0.08	0.06	0.02	0.03	0.05	0.40	1.01
glyceraldehyde-3-phosphate dehydrogenase	XRAY	0.30	0.14	0.16	0.01	0.02	0.04	0.07	0.26	1.00
	CD	0.27	0.12	0.08	0.06	0.02	0.02	0.05	0.38	1.00
	PRE	0.27	0.12	0.08	0.06	0.02	0.02	0.04	0.39	1.00
hemoglobin	XRAY	0.75	0.00	0.00	0.08	0.01	0.04	0.01	0.11	1.00
	CD	0.68	−0.01	0.02	0.04	0.02	0.05	0.04	0.16	1.00
	PRE	0.62	0.00	0.05	0.00	0.02	0.05	0.06	0.26	1.06
lactate dehydrogenase	XRAY	0.41	0.06	0.11	0.01	0.02	0.05	0.03	0.31	1.00
	CD	0.39	0.07	0.12	0.05	0.01	0.03	0.04	0.29	1.00
	PRE	0.39	0.08	0.12	0.06	0.01	0.02	0.04	0.28	1.00
lysozyme	XRAY	0.36	0.09	0.00	0.14	0.06	0.07	0.05	0.23	1.00
	CD	0.29	0.24	0.01	0.08	0.04	0.04	0.05	0.25	1.00
	PRE	0.28	0.26	0.03	0.06	0.03	0.03	0.05	0.30	1.04
myoglobin	XRAY	0.78	0.00	0.00	0.02	0.00	0.08	0.02	0.10	1.00
	CD	0.83	−0.03	0.03	0.02	0.00	0.04	0.02	0.09	1.00
	PRE	0.85	−0.05	0.04	0.02	0.00	0.02	0.02	0.09	0.99
papain	XRAY	0.28	0.09	0.00	0.05	0.02	0.03	0.04	0.49	1.00
	CD	0.26	0.08	0.04	0.05	0.03	0.02	0.04	0.48	1.00
	PRE	0.24	0.08	0.07	0.06	0.04	0.01	0.04	0.44	0.98
prealbumin	XRAY	0.07	0.38	0.07	0.04	0.03	0.00	0.07	0.34	1.00
	CD	0.15	0.24	0.12	0.08	0.02	0.03	0.06	0.30	1.00
	PRE	0.21	0.14	0.14	0.10	0.02	0.05	0.04	0.28	0.98
ribonuclease	XRAY	0.24	0.27	0.00	0.02	0.02	0.02	0.08	0.35	1.00
	CD	0.30	0.24	−0.03	0.08	0.05	0.04	0.05	0.27	1.00
	PRE	0.33	0.20	−0.01	0.11	0.05	0.05	0.05	0.31	1.09
subtilisin BPN′	XRAY	0.30	0.02	0.07	0.12	0.01	0.03	0.05	0.40	1.00
	CD	0.26	0.16	0.10	0.06	0.02	0.03	0.04	0.33	1.00
	PRE	0.26	0.18	0.11	0.05	0.02	0.02	0.04	0.32	1.00
subtilisin Novo	XRAY	0.31	0.02	0.08	0.07	0.00	0.01	0.03	0.48	1.00
	CD	0.38	−0.03	0.14	0.04	0.00	0.02	0.04	0.41	1.00
	PRE	0.41	−0.05	0.16	0.02	0.01	0.03	0.04	0.34	0.96
triosephosphate isomerase	XRAY	0.52	0.00	0.14	0.01	0.00	0.00	0.03	0.30	1.00
	CD	0.54	−0.01	0.10	0.04	0.00	0.03	0.03	0.27	1.00
	PRE	0.55	−0.01	0.09	0.04	0.01	0.04	0.03	0.25	1.00
poly(L-glutamic acid)	XRAY	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
	CD	0.99	0.00	−0.02	0.00	0.00	0.02	0.00	0.00	1.00
	PRE	0.94	−0.03	−0.13	0.06	−0.01	−0.17	−0.04	0.13	1.09

[a] Abbreviations as in Tables IV and VI. The CD for five of the proteins were taken from Brahms & Brahms (1980).

fectively explain the drastic blue shifting of the far-vacuum-UV band in α-chymotrypsin, elastase, and papain as well as the split bands apparent in cytochrome *c* and ribonuclease A. All of this considered, it would appear that the other major positive band would actually be centered at approximately 193 nm.

The proteins of high helical value all show the presence of distinct bands at approximately 208 and 222 nm. Furthermore, it appears that two other bands can be elucidated. Elastase, α-chymotrypsin, and papain all have a peripheral band at 198–200 nm, and α-chymotrypsin shows a distinct band at 228 nm.

Analysis of Secondary Structure. All 15 proteins in our data set have been analyzed by X-ray diffraction to determine their atomic coordinates and secondary structure as given in AMSOM (Feldman, 1976). Six proteins have been analyzed to a resolution of 2.0 Å or better, and the rest have been analyzed at a resolution of better than 3.0 Å. The structural coefficients that we present in Table IX as "XRAY" may differ somewhat from coefficients found in the literature because of our consistent method of structural elucidation discussed above.

Two sets of calculated secondary structures are also presented in Table IX. Those labeled "CD" are an analysis of each protein CD spectrum with the five basis CD spectra that can be obtained from matrix **B**. As such, the structural contributions of each protein have been normalized to 1.00, but no restriction has been placed on negative coefficients. A few small negative coefficients appear in the antiparallel β-strand contribution for proteins of fairly high helical contribution, but they are within the margin of error.

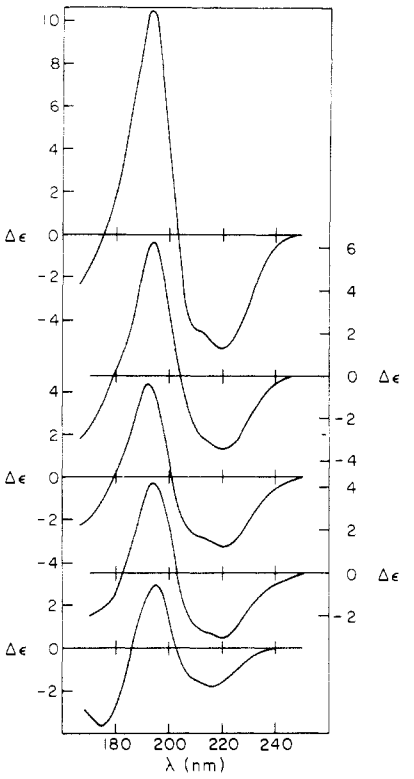

FIGURE 6: Five additional CD spectra, taken from Brahms & Brahms (1980). From top to bottom: triosephosphate isomerase, subtilisin Novo, glyceraldehyde-3-phosphate dehydrogenase, flavodoxin, and prealbumin.

Structures labeled "PRE" are predicted coefficients calculated from an unconstrained least-squares analysis (Baker & Isenberg, 1976) of each CD spectrum using the five basis CD spectra generated from diagonalization of the protein CD spectra remaining when the predicted protein is not included. Negative coefficients are again generated, but they are small. The sums of the "PRE" structures for each protein were unconstrained but are still close to 1.0.

Inspection of Table IX shows good agreement in all eight classifications for the three sets of secondary structures. Table IV provides a more quantitative picture of the agreement. There is a high correlation between the XRAY structure and CD structure for all 16 proteins. The correlation between XRAY and PRE structure is lower, but still quite good.

The correlation of each structural type among the set of proteins is given in Table VIII. There is good correlation between the XRAY structures and the structures obtained from an analysis of the CD. We see that although the rms error is smaller for the β turns, the correlation is somewhat poorer. Here again the small fractions of β turn lead to smaller values for r, as we found for the comparison of XRAY and superstructures. The correlation between XRAY and PRE is surprisingly poor for the β turns in view of the good agreement evident in Table IX. Actually, it is poly(L-glutamic acid) that causes most of the damage. Apparently the basis set provided by the 15 remaining proteins is not sufficient to describe this polypeptide. It is evident that considering all the β turns as a single class improves the correlation for XRAY-CD and XRAY-PRE with the limited number of proteins used in this basis set.

When the limited scope of their origin is noted, the proteins used can be viewed as defining a vector space whose dimensions are defined by asymmetric characteristics of the proteins that contribute to their CD spectra. The space contains various secondary structural types with auxiliary contributions from aromatic side chains, tertiary and quaternary structure, and other sources of asymmetry. Other proteins may contain asymmetric contributions to the CD that are outside the vector space.

Attempts were made to expand our vector space by inclusion of CD spectra of model polypeptides reportedly representing specific types of secondary structure. However, results showed that, for the most part, CD spectra representative of a given type of secondary structure were actually comprised of sizable amounts of other structural types and they adversely affected our analysis. The one exception was poly(L-glutamic acid), representative of an "infinite" helix. Inclusion of this spectrum made a positive contribution to our analysis, particularly in predicting secondary structure for highly helical proteins, although the polypeptie did destroy r(XRAY-PRE) for the β turns as noted above.

Attempts to analyze CD spectra that are outside the space result in large negative coefficients and structure sums markedly different from 1.0. This is a benefit of an unconstrained analysis in that it gives a clear indication of problems in the analysis rather than masking them through constraints.

Basis CD Spectra. Application of eigenvector analysis to our spectral data provides us with a set of 16 orthogonal basis CD spectra describing all of the independent components found in our 16 protein CD spectra. The five basis CD spectra used to reconstruct the original protein CD spectra are given in Figure 1. Each of the five significant components has a unique role to play in formation of a protein CD spectrum. Basis spectrum 1 imparts an obviously helical nature to any spectrum in which it is included (Table VI). Basis spectrum 2 serves mainly as a shifting mechanism, when combined with (1), to allow either red or blue shifting of the 190-nm band and contributes mainly helix, parallel β-strand, and "other" structures. Contributions from (2) in the range 210–250 nm allow the intensity of the 222-nm band to be shifted more or less independently from that of the 208-nm band. Basis spectrum 3 is complementary to (2) in that its main contribution is a fairly independent intensity shifting mechanism for the 208-nm band. Subtle shifts to the 190-nm band are also inherent in (3). It produces interchanges among the eight structures. The fourth basis spectrum does little in the way of band shifting except for a minor influence in red-shifting the 222-nm band. Both (4) and (5) serve mainly to impart subtle changes to the shape of the composite spectrum to account for some of the idiosyncrasies of the various protein spectra, as can be seen in Figure 1. These two basis spectra do not affect the β-turn structures significantly.

The structural contributions of the five basis CD spectra (Table VI) can be viewed in much the same manner as those of protein CD spectra. The one important difference is that, for the basis CD spectra, the sum of the structures is not limited to 1.0. The individual spectra could be normalized to allow the sum to be 1.0, but this would be reflected in the coefficients and would cancel the unitary characteristics of the eigenvector matrix (**U**).

Data Truncation. We truncated our data to determine the wavelength dependence of our analysis and to show the effects of a reduced information set. Analyses were made by using cut-offs of 178, 184, 190, and 200 nm. This scheme systematically eliminates most of the information content of the 185- and 193-nm bands and is eventually equivalent to early studies of this type that were done without the use of vacuum-UV spectroscopy.

The 184-nm cut-off made little differnece in the analysis of the proteins. As long as we do not eliminate any major CD band, our method is fairly insensitive to the wavelength range. Changes that did occur were primarily in the "other" category and were reflected in the structure totals. Truncation to 190 nm resulted in striking changes, reflected almost solely in the β strand and "other" categories, as well as the totals. Further truncation of 200 nm accentuated some of these changes while actually correcting others. These truncations eliminated too much of the information content by eliminating most of the 185- and 193-nm bands.

Interestingly, the helix estimates changed very little no matter which method of truncation was used. This indicates that helical content is very strongly expressed in the near-UV and needs little reinforcement from bands in the vacuum-UV. Simultaneously, it is also evident that β-strand and "other" categories are significantly expressed at the shorter wavelengths and proper analysis must draw on the information in the vacuum-UV.

It appears that there are six CD bands in the wavelength range studied. Negative bands at 200, 208, 222, and 228 nm essentially give a complete set of information in themselves in that they comprise all (negative) CD contributions found in protein spectra from 250 nm down to approximately 195 nm. In many cases a scan to 195 nm may not reach the crossover point of the spectrum, but it will include all negative inflection points in this region. Similarly, the range from 178 nm to approximately 200 nm encompasses another body of information, notably the positive peaks at 185 and 193 nm. As can be seen in the case of poly(L-glutamic acid) this range takes in all major positive inflection points in this spectral region.

Previous studies of the correlation between protein CD spectra and secondary structure have usually involved most of the first body of information (195–250 nm). Our analysis shows this region to be too general, by itself, to provide enough information for proper evaluation of all secondary structure contributions. Inclusion of the second body of information (178–200 nm) increases the specificity of structure determination tremendously, allowing a much more usable structure estimation scheme.

As can be seen in protein CD spectra extended further into the vacuum-UV to 165 nm, a third body of information begins to appear below approximately 180 nm. Extension of protein spectra to 165 nm shows very little in the way of additional inflection points (possibly one more at approximately 174 nm) and would seem to provide, more importantly, an extension of our second body of information to allow its region to extend from 200 nm down to 170 nm. This would include most, but not quite all, of the far-vacuum-UV crossover points in this region. A useful excursion into the third body of information would necessitate facilities and methods to reach approximately 150 nm in order to make substantial inroads into this region.

Secondary Superstructures. Our purpose here was to determine the number of independent secondary superstructures in our set of proteins. We find good correlation (Table VIII) with five superstructures. Thus there appears to be a pattern in the way that various secondary structural types occur in the presence of each other. This would seem reasonable in the case of functional globular proteins where the complex folding required to maintain functionality would maintain some consistency within the generalized requirements of protein folding.

Diagonalization of the protein structure matrix can be approached from essentially two viewpoints. The first, and most forthright, is to simply diagonalize the matrix of X-ray diffraction data, **S**. This has the advantage of working with original unaltered data. The second is to diagonalize the matrix of CD structural data, $\mathbf{S}_\mu'$. However, this makes the assumption that the CD structure matrix is more reliable than the original XRAY data. In practice, it makes little difference which alternative is used.

When the results of the two alternatives in Table VII are compared, a strong similarity between the first four superstructures is apparent. In both bases the first superstructure stresses a helical–"other" relationship of about 2:1, with other structures included in smaller amounts. The second superstructure is supplementary to the first in allowing a further change in helical content while allowing an opposite change in both "other" and antiparallel β strand with an inclusion of parallel β strand. The fourth superstructure brings in a larger share of parallel β strand with a reverse of the polarity of "other" and antiparallel β strand found in the third superstructure. The fifth contributes largely to β turns.

Using the five superstructures to reconstruct their respective parent matrices, we find (the expected) very good agreement between each parent matrix and its reconstruction. From this we can see that the differences in the XRAY and CD structure matrices are reflected in the differences in the two sets of superstructures. Upon reexamination of Table VII, it is apparent that the differences lie mainly in the β-strand structures. This indicates that our greatest source of error for structural measurement lies in our β-strand designation. This in turn must affect other structures, the most likely choice being the "other" category.

This is consistent with the problems of elucidating secondary structure by X-ray analysis. β strands are less distinct than the helical or β-turn structures and, therefore, for those proteins analyzed at lower resolution, the β-strand designations would be of the greatest potential error. The fact that so many of our proteins were analyzed at lower resolutions reflects on all of the proteins since the matrix method of multicomponent analysis averages the errors among them.

Error Analysis. An overall view of the structures estimated from CD shows that, as the contribution of a structural type to the sample set increases, so does the deviation of the estimate by CD spectra analysis. This is indicative of random error in our estimates rather than absolute error. The error appears to be a function of both the amount of the structural type in the data set and the relative assurance that each structural type can be properly identified by X-ray analysis. This accounts for both the error inherent to increased estimates of helical and "other" structure and the increased error of identifying β-strand structure (through X-ray analysis) over other structural types, even though its frequency of occurrence is less.

Analyzing Other Proteins. As part of our research for other projects we have measured vacuum-UV CD for some other proteins. In all cases, the fractions of each secondary structure are reasonable and the sum of all secondary structures is close to 1.0. However, the reader should understand that the basis CD spectra generated from this limited set of 16 proteins will apply only when the structural characteristics of the protein to be analyzed are well represented in the basis set. Furthermore, while this method does not ignore CD contributions from aromatic side chains, the method must handle these contributions within the limits imposed by the use of only five basis CD spectra. Certainly a protein with unusual CD contributions from the aromatics will not be well analyzed. However, it should be obvious when the analysis fails. If an

unconstrained analysis is used, then large negative coefficients or large deviations from 1.0 for the sum of all secondary structures will indicate that the basis set has been overextended. Adding contraints only hides to fact that the analysis has failed.

Acknowledgments

We thank Professor R. R. Becker for the use of his Beckman 120B amino acid analyzer.

References

Baker, C. C., & Isenberg, I. (1976) *Biochemistry 15*, 629–634.
Brahms, S., & Brahms, J. (1980) *J. Mol. Biol. 138*, 149–178.
Bree, A., & Lyons, L. E. (1956) *J. Chem. Soc.*, 2658–2670.
Chang, C. T., Wu, C.-S. C., & Yang, J. T. (1978) *Anal. Biochem. 91*, 13–31.
Chen, Y.-H., Yang, J. T., & Martinez, H. M. (1972) *Biochemistry 11*, 4120–4131.
Chen, Y.-H., Yang, J. T., & Chau, K. H. (1974) *Biochemistry 13*, 3350–3359.
Chou, P. Y., & Fasman, G. D. (1977) *J. Mol. Biol. 115*, 135–175.
Feldman, R. J. (1976) *Atlas of Macromolecular Structure on Microfiche*, Tracor Jitco, Inc., Rockville, MD.
Greenfield, N., & Fasman, G. D. (1969) *Biochemistry 8*, 4108–4116.
Greenfield, N., Davidson, B., & Fasman, G. D. (1967) *Biochemistry 6*, 1630.
Harman, H. H. (1976) *Modern Factor Analysis*, 3rd ed., The University of Chicago Press, Chicago.
Johnson, W. C., Jr. (1971) *Rev. Sci. Instrum. 42*, 1283–1286.
Johnson, W. C., Jr., & Tinoco, I., Jr. (1972) *J. Am. Chem. Soc. 94*, 4389–4390.
Lloyd, D. (1969) Ph.D. Thesis, University of California, Berkeley.
McMullen, D. W., Jaskunas, S. R., & Tinoco, I., Jr. (1967) *Biopolymers 5*, 589–613.
Moore, S. (1968) *J. Biol. Chem. 243*, 6281–6283.
Moore, S., & Stein, W. H. (1948) *J. Biol. Chem. 176*, 367–388.
Moore, S., & Stein, W. H. (1954) *J. Biol. Chem. 211*, 907–913.
Saxena, V. P., & Wetlaufer, D. B. (1971) *Proc. Natl. Acad. Sci. U.S.A. 68*, 969–972.
Siegel, J. B., Steinmetz, W. E., & Long, G. L. (1980) *Anal. Biochem. 104*, 160–167.
Woody, R. W. (1969) *Biopolymers 8*, 669–683.

Nonhistone Proteins Cross-Linked by Disulfide Bonds to Histone H3 in Nuclei from Friend Erythroleukemia Cells†

Alice E. Grebanier‡ and A. Oscar Pogo*

ABSTRACT: Nonhistone proteins in close proximity to histone H3 were detected in cross-linking experiments. Disulfide bonds were formed in nuclei between histone H3 and nonhistone proteins containing sulfhydryl groups by oxidation with H_2O_2. The histones were solubilized either as nucleosomes extracted by 1 mM EDTA from micrococcal nuclease treated nuclei or else as DNA-free proteins extracted by a buffer containing 1 M KCl from nuclei that had been extensively digested with DNAase I. The extracts were examined for disulfide cross-linked proteins by two-dimensional electrophoresis. The nonhistone proteins appeared to be extracted selectively by the two extraction procedures; neither procedure was completely efficient in releasing the histones from the nuclei. One cross-linked nonhistone protein (M_r 46 000) was extracted in 1 mM EDTA along with the nucleosomes. The high salt buffer extracted many nonhistone proteins, including four cross-linked nonhistones which were not found in the EDTA extracts; three of the cross-linked proteins were about M_r 50 000 and one was about M_r 36 000. Four of the five cross-linked, nonhistone proteins which are extracted either by EDTA or high salt are found cross-linked to histone H3 in nuclei prepared from cells lysed in the presence of potassium iodoacetate. Only the disulfide bond to the M_r 36 000 protein requires highly oxidizing conditions for its formation. Thus, many of the disulfide bonds formed between H3 and nonhistone proteins may occur naturally. These experiments disclose a set of nonhistone proteins which are close enough to histone H3 that a disulfide bond can be formed between the histone and nonhistone. The close association implies that these proteins may have significant structural or functional roles in the nucleus.

Previous work from this laboratory has been concerned with preparation and characterization of the nuclear matrix (Miller et al., 1978), an ordered, skeletal structure within the nucleus (Berezney & Coffey, 1974). In the course of experiments with matrix prepared from nuclei of Friend erythroleukemia cells, the discovery was made that the morphology of the nucleus is preserved after almost all of the DNA has been removed by DNAase I digestion (Long et al., 1979). The intranucleosomal histones remain with the nuclear matrix after digestion of the DNA, and the interactions between the histones, as monitored by the use of cross-linking reagents, appear to be undisturbed (Long et al., 1979; Grebanier & Pogo, 1979). The histones are known to be organized into octameric units, the nucleosome cores, around which the DNA is wrapped (Kornberg, 1974; van Holde et al., 1974). However, the histones normally do not maintain such large structures at physiological ionic strength in the absence of DNA. Therefore, the interactions between the histones and the other

† From the L.F. Kimball Research Institute of the New York Blood Center, New York, New York 10021. *Received August 18, 1980.* This work was supported by grants from the National Institutes of Health and the National Science Foundation.

‡ Present address: 35–18 21 Avenue, Astoria, NY 11105.

0006-2960/81/0420-1094$01.25/0