Heath, E., & Howarth, O. W. (1981) *J. Chem. Soc., Dalton Trans.*, 1105-1110.
Lin, Z.-J., Konno, M., Abad-Zapatero, C., Wierenga, R., Murthy, M. N. R., Ray, W. J., Jr., & Rossmann, M. G. (1986) *J. Biol. Chem. 261*, 264-274.
Lowry, O. H., & Passonneau, J. V. (1969) *J. Biol. Chem. 244*, 910-916.
Ma, C., & Ray, W. J., Jr. (1980) *Biochemistry 19*, 751-795.
Magneson, G. R., Puvathingal, J. M., & Ray, W. J., Jr. (1987) *J. Biol. Chem. 262*, 11140-11148.
Michal, G. (1984) in *Methods of Enzymatic Analysis* (Bergmeyer, H. U., Ed.) 3rd ed., pp 191-198, Verlag, Basel.
Percival, M. D., Doherty, K., & Gresser, M. J. (1990) *Biochemistry 29*, 2764-2769.
Ray, W. J., Jr. (1967) *J. Biol. Chem. 242*, 3737-3744.
Ray, W. J., Jr. (1986) *J. Biol. Chem. 261*, 275-278.
Ray, W. J., Jr., & Long, J. W. (1976) *Biochemistry 15*, 3993-4006.
Ray, W. J., Jr., & Puvathingal, J. M. (1990) *Biochemistry 29*, 2790-2801.
Ray, W. J., Jr., Hermodson, M. A., Puvathingal, J. M., & Mahoney, W. C. (1983) *J. Biol. Chem. 258*, 9166-9174.
Ray, W. J., Jr., Post, C. B., & Puvathingal, J. M. (1989) *Biochemistry 28*, 559-569.
Ray, W. J., Jr., Post, C. B., & Burgner, J. W., II (1990) *Biochemistry 29*, 2770-2778.
Ray, W. J., Jr., Puvathingal, J. M., Bolin, J. T., Minor, W., Liu, Y., & Muchmore, S. W. (1991) *Biochemistry 30* (preceding paper in this issue).
Tracey, A. S., Galeffi, B., & Soroush, M. (1988) *Can. J. Chem. 66*, 2294-2298.

Systematic Comparison of Statistical Analyses of Electronic and Vibrational Circular Dichroism for Secondary Structure Prediction of Selected Proteins†

Petr Pancoska‡ and Timothy A. Keiderling*
Department of Chemistry, University of Illinois at Chicago, Box 4348, Chicago, Illinois 60680
Received September 26, 1990; Revised Manuscript Received April 19, 1991

ABSTRACT: The electronic (ultraviolet) circular dichroism (UVCD) and vibrational circular dichroism (VCD) of 20 proteins are systematically compared as to their relationship to the secondary structures of these proteins. The UVCD spectra are statistically treated by use of the same factor analysis methods used previously for VCD. The UVCD spectra can be reproduced as linear combinations of five subspectra. The first subspectrum reflected the expected α-helical UVCD shape, particularly at longer wavelengths, while the higher order ones had less obvious similarity to standard bandshapes. Cluster analysis on the UVCD factor analysis coefficients reflected the clustering on the basis of the fractional secondary structure parameters (from X-ray) but was less clear than VCD. Qualitative complementarity of protein VCD and UVCD spectra was demonstrated by combined cluster analysis of their respective factor analysis coefficients. Quantitative relationships between spectral coefficients and fractional secondary structure were determined by multiple regression analyses using only statistically important coefficients. These resulted in an ability to reproduce four of the structural parameters with errors for individual proteins comparable to the VCD result. In UVCD, the standard deviations of the regression fit for β-sheet were worse and for the undefined part of the structure were better than in VCD. Parallel analyses using the partial least-squares method showed UVCD in that case to have more error than VCD in reproducing the training set structural parameters. Comparison of the regression and partial least-squares methods illustrated limitations of total back-transformation of the UVCD spectra into structural parameters.

In the previous paper in this series (Pancoska et al., 1991), an analysis of the vibrational circular dichroism (VCD)[1] of proteins, which will here be referred to as paper II, we presented VCD data for 20 proteins in the amide I′ region of the infrared. Statistical analyses of these spectra using the principal component method of factor analysis (PC/FA) (Pancoska et al., 1979; Malinowski & Howery, 1980) gave us a simple way of categorizing the VCD spectra in terms of the coefficients of six subspectra into which all of the protein VCD spectra could be linearly decomposed. Since the PC/FA method sorts out correlated intensity changes in order of importance, this leads to a weighting of the importance of the characteristic coefficients for each protein. Cluster analysis (CA) (Sharaf et al., 1986) of these VCD coefficients indicated a common topology with that found from cluster analysis of

†This research was supported by NIH Grant GM30147. Partial support of the purchase of the UVCD spectrometer was provided by the National Science Foundation and the University of Illinois.
*To whom correspondence should be addressed at UIC.
‡Permanent address: Department of Chemical Physics, Charles University, Prague 2, Czechoslovakia.

[1] Abbreviations: CA, cluster analysis; FC, fractional coefficient (of secondary structure); FTIR, Fourier transform infrared (spectroscopy); KS, Kabsch and Sander (1983) protein X-ray crystal structure analysis; LG, Levitt and Greer (1977) protein X-ray crystal structure analysis; paper I, Pancoska et al. (1989); paper II, Pancoska et al. (1991); PC/FA, principal component method of factor analysis; RDI, relative dissimilarity index (in cluster analysis); S/N, signal to noise ratio; UVCD, ultraviolet circular dichroism (of electronic transitions); VCD, vibrational circular dichroism (in the infrared).

0006-2960/91/0430-6885$02.50/0

the fractional coefficients (FC) of the secondary structures of the same proteins as derived from an analysis of X-ray crystal structures by Kabsch and Sander (KS) (1983).

In the first paper of this series (Pancoska et al., 1989), here referred to as paper I, it was demonstrated that VCD could be reliably measured for proteins in D_2O. More importantly, the resultant spectra showed a qualitative enhancement over the more traditionally measured electronic (ultraviolet) circular dichroism (UVCD) in terms of sensitivity to the variation in secondary structure of the proteins studied. Since UVCD has long been used to gain quantitative insight into protein secondary structure [see, for example, Yang et al. (1986) and Manning (1989)], as was reviewed in the previous papers in this series, it is important to compare our VCD analyses (Pancoska et al., 1991) to those of UVCD. [Furthermore, UVCD data has been previously analyzed with a singular value decomposition algorithm (Hennessey & Johnson, 1981) that, to the point of determining subspectra, is mathematically equivalent to the principal component method of factor analysis that we use.] While VCD and UVCD are both chiroptical spectroscopic techniques, there is a difference in that at first level one measurement derives from dipolar excitation of nuclear motion and the other from electronic molecular excitations. Both evidence coupling of locally achiral states in the polymer to give transitions exhibiting distinct chirality that is highly correlated to the handedness of the polymer (Lal & Nafie, 1982; Sen & Keiderling, 1984). It is hoped that these differences will, in turn, lead to a complementarity of the techniques in terms of sensitivity to molecular structure. The goal of this paper is to compare VCD and UVCD data on a systematic, quantitative basis and to develop evidence for this complementarity. To this end, we chose to treat the data from each technique, UVCD and VCD, in a manner consistent with several identical statistical methods of analysis. This paper then will compare the techniques, but it explicitly does not seek to provide the "best" method for the analysis of UVCD spectra. Naturally in the process of comparison, some insight is gained about questions regarding such analyses.

Complementarity of the two techniques, UVCD and VCD, leads one to expect that their combination will result in a more sensitive and exact analysis of secondary structure. Support for this expectation follows from some contrasting features of peptide UVCD and VCD spectra. Particularly at long wavelengths, UVCD is dominated by contributions from the α-helical component of the structure which, in turn, is sensitive to the length of the helix, implying dependence on a somewhat long-range interaction (Woody & Tinoco, 1967; Chen et al., 1974; Madison & Schellman, 1972). UVCD bands of different coherent structures are highly overlapped and, for some structural types, similar in shape (Yang & Kubota, 1985). From model studies, VCD seems to be dominated by shorter range effects (Dukor et al., 1991; Yasui et al., 1986) and has spectral bandshapes of similar intensity but with a unique sign and frequency pattern for each of the major secondary structural types (Yasui & Keiderling, 1986a,b).

This paper contains first a comparison of UVCD and VCD analyses for the same proteins based on the same factor analysis and cluster analysis methods that were described in some detail in paper II. Its focus is on a set of UVCD data that was all gathered in this laboratory with one spectrometer under uniform experimental conditions. In order to carry out cluster analysis simultaneously on both types of spectral data [see, for example, Lipkus et al. (1988)], we combine the PC/FA coefficients from VCD and UVCD into one 11-component descriptor of each protein and calculate a clustering that illustrates the complementarity of the two spectroscopic methods. Next, regression relationships between the PC/FA coefficients and secondary structure fractional contributions (FC) are developed. Similarly, a parallel analysis with a partial least-squares (PLS) algorithm is carried out. Both methods are used to make quantitative as well as qualitative comparisons between UVCD and VCD for secondary structure prediction. Complementarity in predictive strength and in applicability is demonstrated through this analysis.

Experimental Procedures

The protein samples studied were the same as those detailed in paper II. Samples were prepared in D_2O (for consistency with VCD spectra) at a concentration of ~0.15 mg of protein in 1 mL of solution. Spectra were measured in a 1-mm-path-length, strain-free, UV-quartz cell (NSG Precision Cells). Electronic UVCD spectra were measured on a JASCO J-600 dichrometer under continuous dry nitrogen flushing. An average of three scans over the wavelength range of 260–180 nm, each obtained with a 0.5-s time constant, was used to obtain all spectra. The figures and subsequent calculations represent data as obtained from this averaging without smoothing. All spectra were corrected by subtraction of identically obtained UVCD base-line spectra obtained with just D_2O in the same sample cell. Data were expressed in terms of ellipticity per residue, with the assumption of an average residue molecular weight of 113 or by the use of those extinction coefficients that were available (Hennessey & Johnson, 1981). All calculations used the same software packages [SpectraCalc, Galactic; EINSIGHT, Infometrix; Statgraphics, V 2.6, Statistical Graphics Corp.; PC/FA, rewritten by us for this study and based on Pancoska et al. (1979)] as described in the previous paper.

Results

The UVCD spectra we measured were substantially in agreement with those presented previously by other authors (Hennessey & Johnson, 1981; Brahms & Brahms, 1980; Mori & Jirgensons, 1981; Manavalan & Johnson, 1983; Bullock & Myer, 1978) and thus do not warrant reproduction here. The spectra used were measured in a totally consistent manner on the same instrument in the same cell, so that any minor variations found between them and previously published UVCD spectra will not be reflected in the variance within the group measured here. It is important to realize that the relative variance of the protein spectra is the basis upon which a relationship to the variations in the respective protein structures is sought in our analyses. Any systematic deviations in our spectra from those of another laboratory, then, do not affect the UVCD analysis or its comparison to the VCD. While, in principle, noise could affect the comparison, factor analysis has a smoothing effect on the data to minimize this problem.

Factor Analysis. With use of the same procedures as those detailed in paper II, our factor analysis calculations for the set of 20 UVCD spectra yielded five significant subspectra, S_i, which are illustrated in Figure 1. The first two qualitatively resemble the most significant subspectra found by Hennessey and Johnson (1981) for their set of 15 protein (and one polypeptide) UVCD spectra, but the less significant subspectra (S_3–S_5) differ. The actual shape of the subspectra will necessarily be a function of the proteins used in the analysis, so such a deviation is not surprising. It is gratifying to note that these two independent analyses using different protein sets both required five subspectra for an adequate fit of the data. Coefficients and eigenvalues for reconstruction

Table I: Coefficients and Eigenvalues As Calculated from the Principal Component Method of Factor Analysis for UVCD Spectra

protein		PC/FA coefficients				
i	name	α_{i1}	α_{i2}	α_{i3}	α_{i4}	α_{i5}
1	trypsin	6.21	−13.12	4.01	9.00	17.75
2	trypsin inhibitor	0.92	−24.42	−2.60	−20.70	−17.98
3	triosephosphate isomerase	19.08	12.39	0.36	6.91	−0.12
4	thaumatin	10.53	5.98	−2.78	−31.86	6.47
5	ribonuclease S	8.42	−2.84	4.26	−1.84	7.47
6	ribonuclease A	9.63	−2.70	2.09	−2.35	4.71
7	papain	19.57	−9.69	−15.37	30.16	10.01
8	myoglobin	67.0	34.00	−46.95	−0.51	−31.77
9	lysozyme	20.43	1.64	−7.06	−8.3	−22.04
10	lactoferrin	23.22	9.51	−2.72	0.26	14.20
11	lactoglobulin	25.49	20.79	5.45	3.01	5.47
12	hemoglobin	44.59	32.61	2.41	−2.21	−27.64
13	elastase	4.09	−13.67	0.89	0.48	2.14
14	cytochrome c	17.38	2.89	1.74	10.82	−0.62
15	concanavalin A	18.53	51.00	77.98	47.55	−67.30
16	chymotrypsin	11.01	−19.37	−2.18	5.17	−24.77
17	chymotrypsinogen	10.43	−20.91	−5.49	1.17	2.94
18	casein	11.25	−49.31	12.26	−1.13	−47.50
19	carbonic anhydrase	4.76	−4.16	34.79	−6.95	3.84
20	albumin	68.32	21.71	−55.41	0.46	−14.83
eigenvalues		λ_1	λ_2	λ_3	λ_4	λ_5
λ_j		11.43	6.61	1.32	0.42	0.12
$\sqrt{\lambda_j}$		3.38	2.57	1.15	0.65	0.34

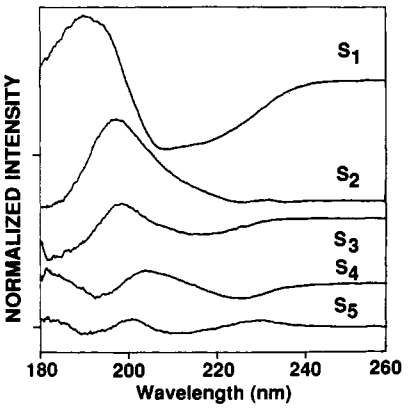

FIGURE 1: The five most significant subspectra derived from the PC/FA of the UVCD data set for 20 proteins. Since spectra are normalized, the vertical scale is relative. Spectra are offset for clarity.

of the protein UVCD from these subspectra are listed in Table I. The eigenvalues show a steep decline with decreasing subspectrum significance as expected. Since we used different proteins and 0.2-nm steps in our digitized UVCD spectra, they differ from those of Hennessey and Johnson (1981), which used only 2-nm steps. It is possible that our PC/FA routine is numerically different from the singular value decomposition program, but we feel this is minor (Haaland, 1990; W. C. Johnson, personal communication).

The first subspectrum, S_1, qualitatively resembles the standard UVCD bandshape found for predominantly α-helical molecules in that it has a strong positive band at 190 nm and overlapped strong negative bands at 208 and 217 nm. S_1 can be interpreted as a weighted average of the spectra in the training set, which contains proteins of a variety of structural types including high contributions from α-helices and β-structures (both of which are positive to short wavelength and negative to long wavelength) as well as from "random-coil" proteins, which have strong negative bands at shorter wavelengths. As a result of this averaging, the relative intensity of the positive 190-nm band and the resolution of the two negative bands differ from the UVCD spectra found for model α-helical compounds (Greenfield & Fasman, 1969) and highly helical proteins (Hennessey & Johnson, 1981; see also Figure 3a). The resemblance of S_1 to the typical α-helical UVCD spectral shape confirms the dominancy, noted above, of this contribution to protein UVCD, at least in the near-UV range.

The second subspectrum can be understood as a weighted average of residuals calculated as the difference of each UVCD spectrum, θ_j, and $\alpha_{1j}S_1$, where α_{1j} is the PC/FA coefficient of S_1 for protein j (see eq 1, paper II). Its shape is roughly the inverse of what might be expected for a "random-coil" polypeptide. That shape has been attributed to locally ordered left-handed extended helical segments in the coil (Tiffany & Krimm, 1968; Woody, 1991; Dukor & Keiderling, 1991). S_2 should represent the largest spectral variance in the UVCD data set, which is shown to be a sign reversal in the 190–200-nm region from a large positive for the predominantly α-helical proteins to a large negative for the so-called "random-coil" proteins.

Successive subspectra are progressively more oscillatory. These latter subspectra have no obvious relationship to the UVCD spectra of model polypeptides or to the protein spectra we have measured for various cluster types (see below). The effects of noise are increasingly more evident in the 180–190-nm region. In general, it can be noted that the UVCD subspectra are much smoother than those for the VCD due to the greater signal-to-noise ratio (S/N) in the UVCD spectra themselves.

Cluster Analyses. In order to compare spectral features qualitatively to protein structure, cluster analyses (Pancoska et al., 1991; Sharaf et al., 1986) of the PC/FA coefficients, α_{ij}, have been compared to the clustering of these proteins based on the X-ray-based secondary structural fractional coefficients (FC_ζ) from the Kabsch and Sander (1983) algorithm. By way of review, cluster analysis of five FC_ζ values derived from the KS descriptors, ζ = helix (α), sheet (β), turn (t), bend (b), and "other" (ρ), results in a coarse clustering of the 62 proteins studied by KS into several distinct groups that are separable at a low relative dissimilarity index (RDI) index (Pancoska et al., unpublished results). (The RDI is a measure, ranging from 0 to 1, of the relative distance between clusters in terms of the metric upon which the analysis is calculated.) As was detailed in paper II, when the cluster analysis was redone for the 13 proteins in our set of 20 that

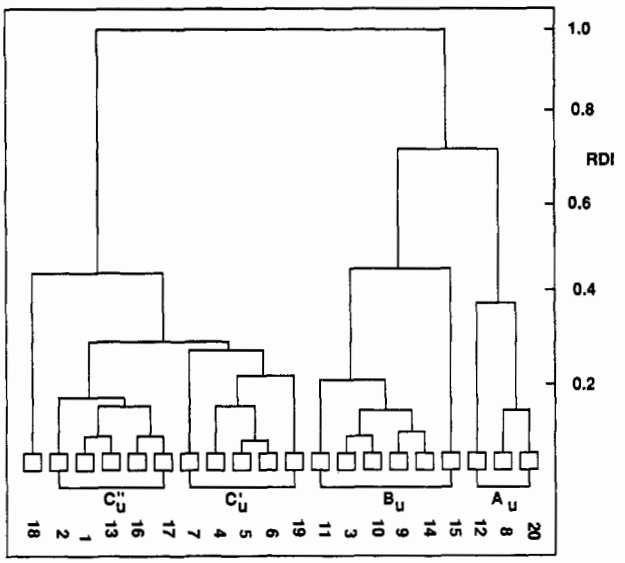

FIGURE 2: Dendrogram based on the Lance–Williams flexible CA algorithm as applied to clustering of eigenvalue-weighted coefficients of the five UVCD PC/FA subspectra for the 20 proteins studied. Proteins are identified by number corresponding to the identification in columns 1 and 2 of Table I. RDI is the relative dissimilarity index, which is a measure of the distance between spectral clusters in this measure.

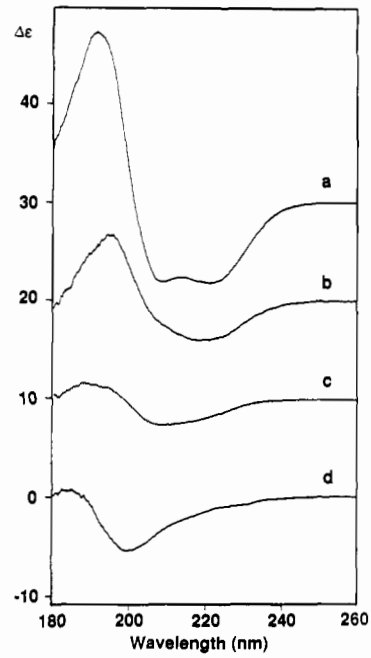

FIGURE 3: Averaged UVCD spectra for the proteins that belong to each of the primary spectral clusters: (a) A_U, (b) B_U, (c) $C_{U'}$, and (d) $C_{U''}$, which are defined in Figure 2.

overlap the KS set, representatives of a few of these clusters were found (see Figures 11 and 12a, paper II).[2] The first cluster, K1, corresponds to those proteins very high in α-helical content, having no β-sheet content, and containing myoglobin and hemoglobin for which we have measured CD. The second, K2, contains cytochrome *c*, triosephosphate isomerase, and papain, which have moderate, correlated helical (22–45%) and β-sheet (3–18%) content. Finally, the third cluster has less α-helix and more β-sheet, contains most of our studied examples, and can be divided into two dominant subclusters, K3 and K4 (Pancoska et al., 1991).

With a focus now on the UVCD spectra, the dendrogram illustrated in Figure 2 from the Lance–Williams flexible CA algorithm is a typical result and shows that the PC/FA coefficients corresponding to the UVCD spectra lead to roughly similar clustering, as was found for the coefficients of the VCD. But the degree of clustering and a few of the members in each of the main groups are different. Here the α-helical proteins, labeled as group A_U, strongly cluster and are distinct from the others to an RDI of 0.7. Aside from papain, the X-ray-determined K2 cluster of proteins also clusters in UVCD, here denoted as group B_U. In VCD, the grouping of those PC/FA coefficients was a better match to the X-ray clustering than is seen here. However, for those proteins contained in a given UVCD cluster, the similarity (judged by low RDI values) is stronger than in VCD. Concanavalin A is essentially not clustered (RDI > 0.5) with the other proteins and is the only KS-analyzed protein to change groups when different clustering algorithms are used. This reflects long-noted difficulties in using UVCD data to interpret its structure (Yang et al., 1986; Bolotina et al., 1980). Most of the X-ray-determined K3 and K4 proteins are clustered together in one subgroup of the C_U group from UVCD with RDI < 0.2. Carbonic anhydrase, thaumatin, and the ribonucleases, A and S, form another subgroup with RDI < 0.2. Both are weakly coupled to papain from K2. Casein, a "random-coil" protein, is weakly coupled (RDI > 0.4) to the rest of group C_U. In summary, the PC/FA coefficients for the UVCD spectra lead to three clusters, much as was seen for the cluster analysis of the VCD and X-ray coefficients presented in paper II. The mapping of UVCD clusters onto the X-ray clusters is not quite as good that seen for VCD. In addition, cluster analysis of both UVCD and VCD spectral coefficients have group C divided into two main subgroups, much as seen in the clusters K3 and K4 for the X-ray fractional coefficients.

To parallel our VCD study and to characterize the clusters spectrally, we averaged the UVCD spectra for the proteins for each of the main UVCD groups as illustrated in Figure 3. They were indeed distinct and reflected some of the patterns seen in the subspectra. The average over the first cluster corresponding to group A_U is a "classical" α-helical spectrum (Figure 3a) with a large positive band at 192 nm and two partially resolved, smaller negative bands at 208 and 221 nm. Group B_U and subgroup $C_{U'}$ [that part of C_U containing the ribonucleases (Figure 2)] have progressively weaker averaged UVCD spectra (Figure 3b,c) but still evidence clear positive and negative lobes at longer wavelengths. This fundamental similarity of spectral shape of most of the UVCD spectra for our set of 20 proteins is in contrast to the complete sign variation and frequency shifts seen in the VCD of these same proteins (see paper II). A qualitative description of this contrast was presented for a few of them in paper I (Pancoska, et al., 1989). The other subcluster, $C_{U''}$, containing trypsin and related proteins, has an average UVCD spectrum (Figure 3d) resembling that of casein and the second subspectrum (with the opposite sign).

The simple numerical form of the factor analysis coefficients allows us to use cluster analysis to find patterns of similarities in the descriptive vectors that have been extended to include the six PC/FA coefficients from the VCD spectra along with the five UVCD ones to form an 11-vector descriptor of the

[2] The FC_f values used are listed in Table IV and II (in part) and were derived from the KS descriptors by counting residues assigned to the respective secondary structural types. While they are not available directly in the KS paper, they are similar in spirit to those of Qian and Sejnowski (1988) but are here more complete.

Table II: Fractional Contributions to Secondary Structure As Predicted from Regression Equations (1) for UVCD Spectral Coefficients

		KS FC[a]				predicted FC[a]					
i	protein	α	β	b	ρ	α	β	b	ρ	Δ^b	Δ(UVCD–VCD)[c]
1	trypsin	8	32	15	31	8	28	16	34	2	5
2	trypsin inhibitor	14	24	17	38	14	31	13	36	3	3
3	thiosephosphate isomerase	43	17	8	24	25	21	13	26	7	11
5	ribonuclease S	18	35	12	27	15	27	14	31	4	26
7	papain	23	16	18	33	16	20	16	30	4	2
8	myoglobin	77	0	2	11	86	−6	1	12	4	6
9	lysozyme	29	8	13	29	32	20	11	28	5	5
12	hemoglobin	67	0	4	20	59	6	6	21	4	11
13	elastase	5	34	9	34	9	29	15	34	4	4
14	cytochrome *c*	26	4	22	34	21	22	14	29	9	4
15	concanavalin A	0	40	20	30	8	21	19	29	7	4
16	chymotrypsin	6	33	10	35	16	25	15	35	6	8
19	carbonic anhydrase	7	27	18	36	13	29	14	37	3	4
	standard deviation					8	10	4	3		
	standard deviation (% of dynamic range)[d]					10	25	19	12		

[a] Abbreviations of secondary structures: α, α-helix; β, β-sheet; b, bends; t, turn; ρ, other conformations. Because the fraction of turn is not reliably predicted in this model, it is not included in the table. [b] Average differences between KS and predicted data per conformation: $[\sum|\text{KS} - \text{pred}|/4]$. Note that, since regressions are independent, FC_f are not required to sum to 100%. No separate accounting for the turn fraction is needed because its effect is part of the regression on the training set. [c] The last column compares UVCD and VCD FC values as $\sum|FC_f^{UV} - FC_f^{VCD}|/4$. [d] Calculated from the differences $\Delta f = FC_f^{KS} - FC_f^{UV}$ between KS FC and UVCD-predicted FC values as the standard deviation within the given secondary structure column as $[\sum \Delta_f^2/(n-1)]^{1/2}$. Dynamic range: (max FC_f^i − min FC_f^j) within the KS column.

circular dichroism spectra for each protein. The revised clusters (shown in Figure 4 as a dendrogram determined with the Lance–Williams flexible algorithm) for the combined description fall somewhat between those for the UVCD (Figure 2) and VCD (Figure 9, paper II) results alone. Group A_{VU} is still strongly separated from the rest (RDI = 1), group B_{VU} has been "repaired" by moderately reclustering with papain and including ribonuclease A, as occurred in the VCD clustering. Similarly, aside from concanavalin A, group C_{VU} is reassembled to the X-ray K3 + K4 cluster in the combined UVCD–VCD cluster analysis. Casein and thaumatin are weakly clustered with this latter group (RDI > 0.4), and concanavalin A is characterized as being dissimilar from all of the others, primarily due to its UVCD character. To a large extent, the combined UVCD and VCD representation of the PC/FA-based clustering reflects the X-ray FC results and, consequently, the VCD clusters.

Regression Analyses. For VCD, quantitative relationships between the spectral factor analysis and the secondary structure were determined by use of regression analyses in paper II. To determine which PC/FA coefficients are related to protein secondary structure, the correlation matrix between the UVCD coefficients and the X-ray-determined FC values was constructed and analyzed as was done in paper II. At the 99% confidence level ($r > 0.684$), only the coefficient of the first UVCD subspectrum, α_{i1}, had significant correlations. However, these were with three different sets of FC_f values: α-helix ($r = 0.90$), β-sheet ($r = -0.72$), and "other" ($r = -0.90$). With a lower level of confidence, a correlation of FC_ρ with α_{i2} and of FC_α and FC_β with α_{i3} were also seen.

Next, the optimum number of α_{ij} coefficients that could be reliably used in a multiple-variable regression analysis to predict each of the FC_f values were sought. Only those correlations that are statistically significant at the 99% confidence level (unless explicitly stated otherwise) were used.[3] The form of the regression chosen depended on the following criteria. First, the fit was required to be statistically significant on the basis of critical values of the multiple-correlation coefficient ($\sqrt{R^2}$; Rohlf & Sokal, 1981). In ambiguous cases, we selected that which depended on (a) subspectra of highest importance, (b) the fewest parameters, and (c) coefficients with the highest F_s values. (F_s is a measure of the significance of the increment in $\sqrt{R^2}$ by the inclusion of the variable of interest.) The results can be summarized as

Table III: Predicted Fractional Contributions to Secondary Structures for Proteins Not in the KS Set from UVCD Regression Equations (Equation 1)

i	protein	FC_α[a]	FC_β[a]	FC_b[a]	FC_ρ[a]
20	albumin	86	−7	1	14
18	casein	19	25	14	46
17	chymotrypsinogen	17	25	14	35
11	lactoglobulin	34	17	11	25
10	lactoferrin	32	18	11	27
6	ribonuclease A	17	26	14	31
4	thaumatin	30	25	10	28

[a] Abbreviations are as in Table II.

$$\begin{vmatrix} FC_\alpha \\ FC_\beta \\ FC_b \\ FC_t \\ FC_\rho \end{vmatrix} = \begin{vmatrix} 1.20 & 0 & 0 & -0.40 & 0 \\ -0.56 & 0 & 0 & 0 & 0 \\ -0.22 & 0 & 0 & 0.14 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -0.29 & 0.18 & 0 & 0 \end{vmatrix} \otimes \begin{vmatrix} \alpha_{i1} \\ \alpha_{i2} \\ \alpha_{i3} \\ \alpha_{i4} \\ \alpha_{i5} \end{vmatrix} + \begin{vmatrix} 4.6 \\ 31.3 \\ 16.3 \\ 0.0 \\ 29.9 \end{vmatrix} \quad (1)$$

The corresponding $\sqrt{R^2}$ values are 0.93, 0.72, and 0.93 for helix (α), sheet (β), and other (ρ), respectively, (> 99% confidence), and the bend (b) correlation was worse (95% level, two variables at $\sqrt{R^2} = 0.69$). All the α_{ij} coefficients in eq 1 have a significance level of ≥98% by the F_s test except for the FC_b relationship, which is ≥92%.

The fourth row in eq 1 has been left as all zeros to emphasize that, as in the VCD analysis, the turn fractions are undetermined by the UVCD data within our stated criteria of significance. Furthermore, also as in VCD, use of multiple parameters did not provide a means of significant improvement for FC_β determination as compared to the single-parameter results. On the other hand, the reformulation of FC_ρ as dependent on α_{i2} and α_{i3} results in lower overall error than use of the single-parameter correlation to α_{i1} alone. The only "higher order" subspectrum retained in our multiple regression is S_4, which is used for the α-helix and bend determinations. This means that the S_5 subspectrum although significant in

[3] The hypothesis was tested in all cases by use of critical values for correlation coefficients corresponding to $\nu = n - 1 - m$ degrees of freedom, with n being the number of observations used for testing the correlation and m being the number of independent variables in the regression equation (Sokal & Rohlf, 1981; Rohlf & Sokal, 1981).

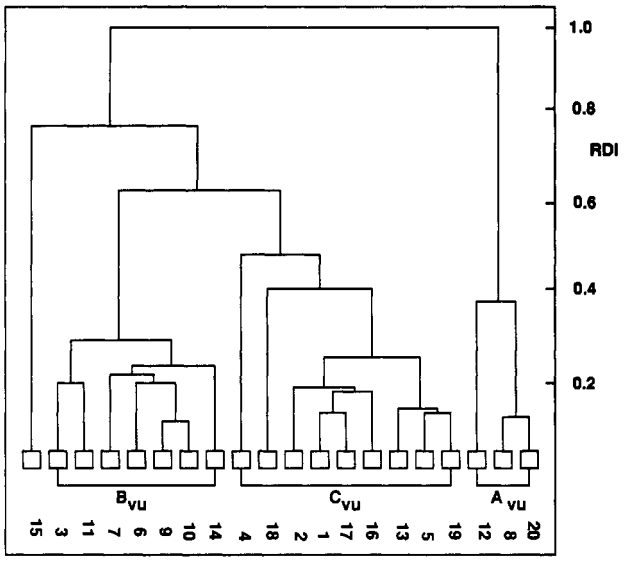

FIGURE 4: Dendrogram from the Lance–Williams CA algorithm for the combined UVCD and VCD analysis, which includes coefficients of 11 PC/FA subspectra. Protein numbering is as in Table I, columns 1 and 2.

determining the experimental UVCD bandshape is not significant for determination secondary structure, at least with this training set.

The values of four FC_ζ values fit in this regression, $\zeta = \alpha$, β, b, ρ, are listed in Table II for the 13 KS-studied proteins in our set. Those FC_ζ values predicted from eq 1 for the other seven of the 20 proteins in our factor analysis set are in Table III. An attempt to use the Varimax rotation algorithm to clarify these relationships (Davies, 1984), as outlined in paper II, resulted in relatively small changes of the first and third subspectra with larger effects seen in the others. However, the correlation to secondary structure did not improve from the original PC/FA picture.

Discussion

The goal of this study is to compare UVCD and VCD spectral analyses with respect to protein secondary structure in a manner that is as consistent as possible. Thus, we have remeasured all of the UVCD data and have used the same algorithms for the UVCD analyses as those used in paper II for the VCD analyses. Our factor analysis is equivalent to that level of analysis by Hennessey and Johnson (1981) in that mathematically equivalent methods are used to develop the eigenvalues and eigenvectors (coefficients of the subspectra) in both of our approaches. It should be made clear that factor analysis creates subspectra on the basis of their commonality and orthogonality, not on their correlation to any structural factors. To first order, these subspectra have no structural meaning other than representing the various spectral averages noted in the previous section. Thus, change in the protein set will change the subspectra. If the set is sufficiently large and flexible (i.e., contains multiple representatives of all major protein types), these subspectra become more resistant to change when the set used is slightly altered. There is no guarantee that we are at that level, but at least S_1 and S_2 are like those found before so that similar correlations would be expected to hold. As the subspectra become less significant, more variation is expected.

Our approaches to the *use of the coefficients* of these UVCD subspectra differ significantly from those of Hennessey and Johnson (1981) in that we first seek and test for relationships between the structural variables (FC_ζ) and the spectral coefficients (α_{ij}) before using them predictively. These relationships are sought in a qualitative manner by use of cluster analysis and in a quantitative manner by use of *selective regression analyses*. Here "selective" is the operative word. Only those coefficients are used that have a demonstrable correlation to some aspect of secondary structure included in the analysis. This approach is more conservative than a total inversion of UVCD spectra to yield only secondary structural parameters, as is commonly practiced by others in the field (Yang et al., 1986). At least in principle, we avoid the danger of convoluting into the analysis other factors that can affect the spectra (Manning, 1989). UVCD spectra, in particular, are very susceptible to solvent and environmental perturbations as well as are subject to considerable interference from aromatic residues. This is not a significant problem in VCD due to the inherent resolution of vibrational spectroscopy in the ground state. Use of the UVCD factor analysis coefficients for secondary structure prediction is discussed below in comparison with the results obtained with VCD by use of identical methods. They are then compared to results from a parallel application of the partial least-squares method, which is provided as an example of a standard analytical method that does make use of a total back-transformation of the spectra.

Cluster Analyses. First, it should be noted, without surprise, that the UVCD spectra when reduced in the PC/FA method and subjected to cluster analysis reflect the secondary structure of the proteins. That is, after all, a long-standing use of UVCD in protein chemistry. However, via our CA approach, it is clear that the topological match is not perfect and that the VCD match to the X-ray clusters is better than that found here with UVCD. The significance of this observation is tempered by the sensitivity of the UVCD clustering to the algorithm used.

For CA calculations using just the 13 proteins from our training set that are also in the KS set, the UVCD clusters are quite stable with respect to the various clustering methods (results not shown). Only concanavalin A (no. 15) changes classification between groups A_U and B_U, but concanavalin A only clusters to other proteins at a high RDI and has a somewhat unusual UVCD spectrum (Yang et al., 1986). As shown in Figure 5, when all 20 proteins are used in the CA calculations with the set of seven CA algorithms at our disposal, more variation is evident. The respective sensitivities of the 20-protein clustering to algorithm used are graphically illustrated for (Figure 5a) the UVCD PC/FA coefficients and (Figure 5b) the combined VCD–UVCD coefficients, following the parallel presentation of the VCD analysis presented in paper II (Figure 12). Those inside a box cluster for all algorithms and those placed outside the boxes change clusters as indicated in one or more of the CA calculations.

UVCD and VCD methods differ significantly in extent of classification variation in that more proteins change subclusters in the UVCD CA calculations (five proteins, Figure 5a) than in the respective VCD ones (two proteins, Figure 12c, paper II). More importantly, several (four) proteins change between major (A, B, C) clusters in UVCD. Even hemoglobin, certainly expected to be a member of group A_U, the helical proteins, has variable clustering in the 20-protein UVCD set. This ambiguity is a direct consequence of the dominance of the UVCD bandshape by the α-helical contribution, which is evidenced by the general similarity of protein UVCD spectra. Consequently there is relatively less differentiation between the protein groups on the basis of UVCD than VCD, which, in turn, is reflected in the dissimilarity of clustering. Similarly, there is a high level of variability in connectivity between the clusters in UVCD in contrast to the relative stability of the

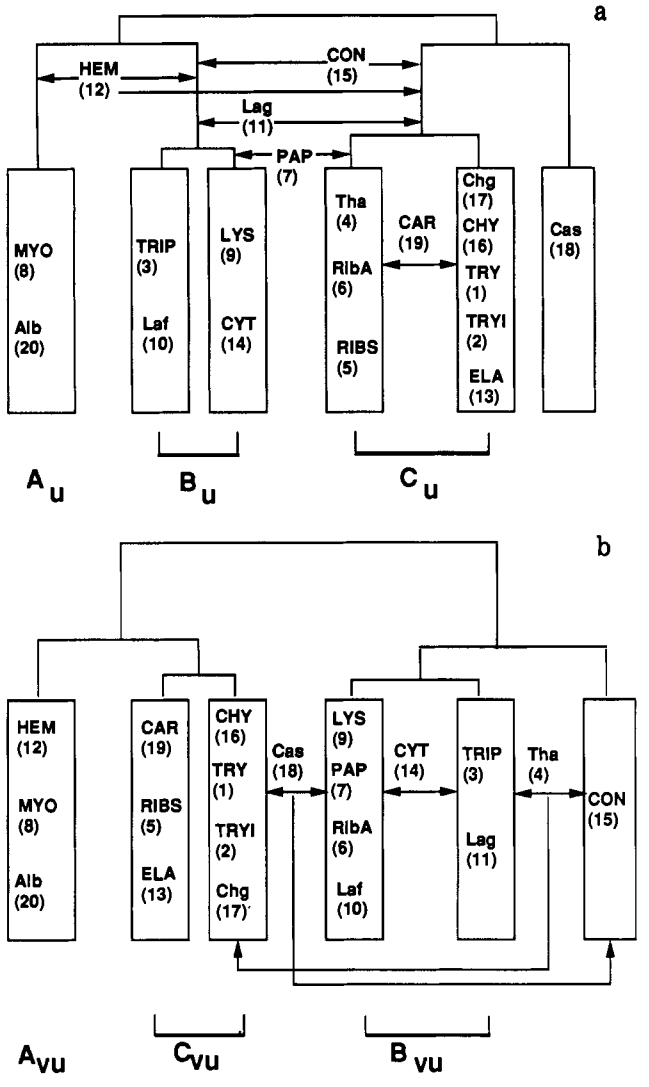

FIGURE 5: Schematic representation of the sensitivity of the clustering pattern to the CA algorithms used for analysis of all 20 proteins based on (a) the five UVCD only and (b) for the 11 (UVCD + VCD) PC/FA coefficients. Proteins placed in a box had no differences in clustering with change of algorithm. Those outside the boxes had one or more changes in clustering with different algorithms as indicated. Proteins placed between tie lines above the boxes had less clear clustering to individual subclusters in UVCD. Connections above the boxes are meant to schematically indicate the higher RDI connectivity, but for UVCD, in particular, this was quite variable and the drawing is meant to be more suggestive than quantitative. A higher level of clustering is indicated by the boxes than is discussed in the text (beyond A, B, C) in order to indicate the stability of the subcluster topology. The UVCD connectivities at high RDI were variable.

VCD clusters. Combining the UVCD and VCD sets of coefficients moves the combined clustering toward improved mapping onto the VCD cluster pattern and, consequently, onto the X-ray clusters (Figure 12a, paper II). The combined analysis, where three proteins change groupings between well-defined clusters, shows improved stability over just UVCD with change in CA algorithm (Figure 5b).

These observations lead to the conclusion that the VCD data set has a more differentiated structure than does the UVCD in terms of cluster analysis. Coupled with VCD's better mapping onto the X-ray-based clusters, this indicates that there is a clearer spectra–structure relationship for VCD than for UVCD. In summary, the cluster analyses of the factor analysis results further support the observation made in paper I that, *as compared to UVCD, VCD has an enhanced sensitivity to protein conformation* (Pancoska et al., 1989). Coupling the two sets of spectral coefficients better delineates the clustering and may give the spectral representation better sensitivity to the structural range over the protein set in terms of FC_ζ values (see discussion below).

The difference between UVCD and VCD in terms of qualitative mapping onto secondary structure is not large, but the sensitivities of the two measurements are different and, we feel, complementary. There is a clearer parallel of VCD α_{ij} coefficients with X-ray FC_ζ^i parameters via clustering. UVCD data senses "pure" α-helix and "other" contributions uniquely due to their dominance in specific regions of the UVCD, and VCD has better discriminatory powers for mixed structures, as evidenced by a clearer separation of $\alpha + \beta$ from β-rich proteins (Levitt & Chothia, 1976) with VCD. Such discriminatory capabilities have previously been claimed for UVCD on the basis of proposed subtle shape variations between protein classes (Manavalan & Johnson, 1983). The VCD variations in paper II are much more dramatic due to the roughly equivalent intensity of the contributions of each component of the structure to the VCD spectrum.

As pointed out in paper II, an important virtue of the cluster analysis method of qualitative spectra–structure correlation is its providing a guide to sensible application of more quantitative approaches to secondary structure determination from spectra. For those proteins whose PC/FA coefficients cluster well with those of the training set proteins, quantitative analyses are likely to have significance. This work shows that use of more than one type of data in the CA leads to a more stable discrimination that can give weight to the degree of confidence one should have in a given clustering.

Regression Analyses. The results of our regression analyses on the subspectral coefficients are given in eq 1 and Tables II and III, and the following comments relate to the form of the significant single-variable regressions found. The first subspectrum, S_1, represents the average of the spectral intensities; hence, α_{i1} is in some sense a measure of the overall UVCD intensity. This intensity appears to carry the information about the variation in α-helix fraction that is a consequence of both the dominance of the UVCD (at least in the near UV) by the α-helical contribution and the wide dynamic range of the FC_α values (0–77%) for the proteins in our training set (see Table II for a listing of these values). From an analysis of the proteins in the KS dictionary, but not specific to the KS set, it can be shown that FC_β and FC_ρ are correlated to FC_α at a moderate level (Pancoska et al., unpublished results). The probable source of the FC_α, FC_β, and FC_ρ correlations all to α_{i1} is thus most likely the interdependence of the FC_ζ values themselves.

The influence of the secondary structure on the electronic transitions and consequently on the UVCD is different from that acting on the vibrational ones. This is exemplified in the correlations found between the FC_ζ structural parameters and the PC/FA coefficients. In VCD, the mapping of the α-helical structure was onto the principal variance from the average spectral intensity, not onto the intensity itself as seen in UVCD. Conversely, the "other" fraction correlated to the average intensity in VCD. This overall dependence of several FC_ζ values on the average intensity in UVCD is a probable source of the past success of simple methods that used just single wavelength determined UVCD intensities or simple spectral fits for estimation of FC_α.

The PC/FA method identifies subspectra by commonality; so simple relationships of structural elements with the spectral coefficients are not required by the analysis. Furthermore, there is nothing to indicate that these structural elements are

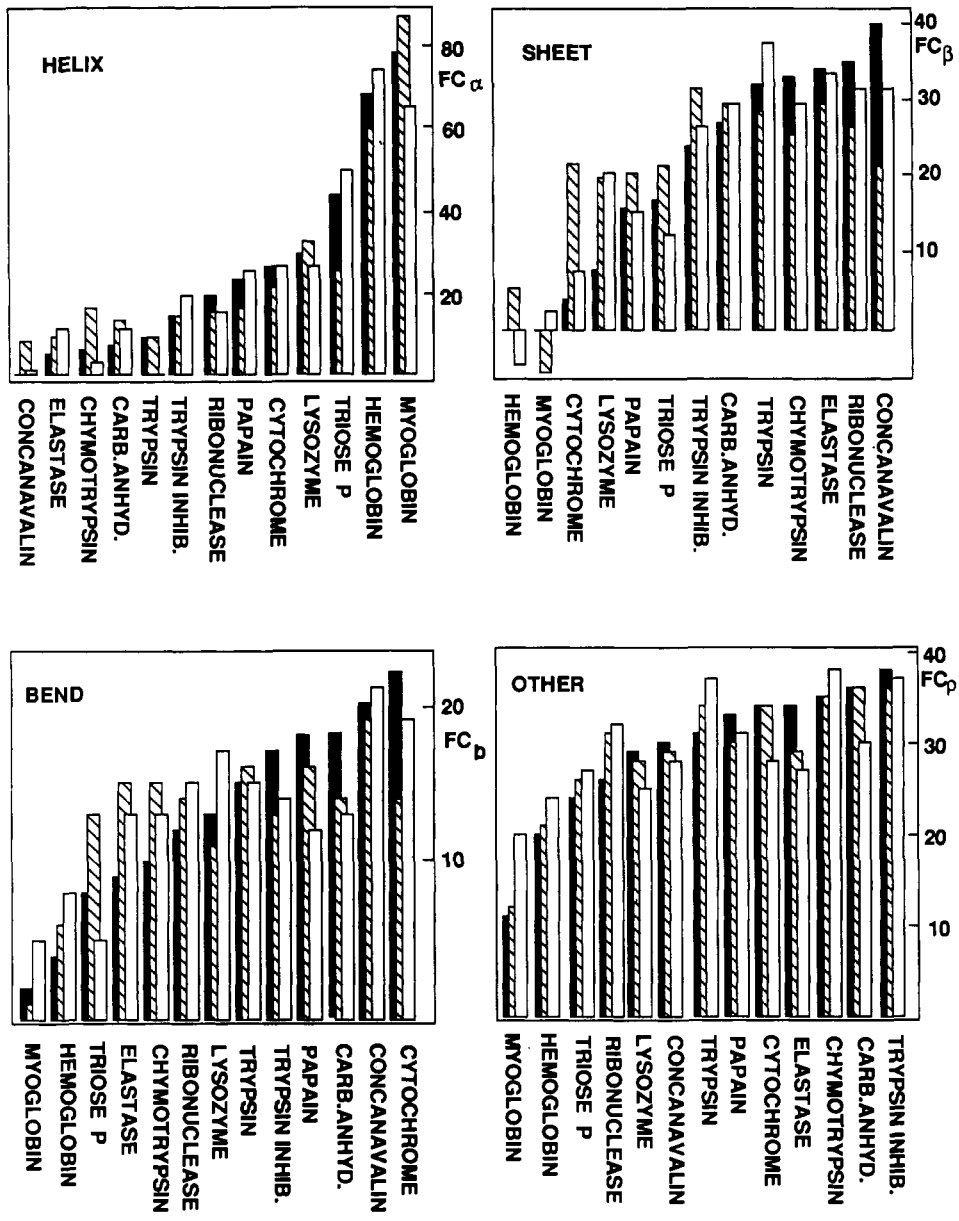

FIGURE 6: Bar graph representation of the errors in the multiple-regression fit to the X-ray-determined FC_ζ values for 13 KS-analyzed proteins in the training set. Proteins are arranged in order of increasing value of FC_ζ^j with solid bars representing the KS values, slashed bars the UVCD fit value, and open bars the VCD values.

independent, particularly for the subset of proteins we have studied; in fact, our data indicates that they are not (Pancoska et al., unpublished results). The multiple variable spectra–structure correlations in eq 1 were determined after confirmation that the regression actually showed a significant dependence on those added variables. This method avoids assigning spectral consequences of nonrelevant factors as due to the secondary structure.

As an example of our approach, the two-parameter correlation given in eq 1, which relates FC_ρ to spectral coefficients independent of α_{i1}, gives the best fit despite the single correlation noted above. The corresponding subspectra, S_2 and S_3, in large part represent an inversion of the main UVCD spectral features (Figure 1) that are traditionally assigned to the "random-coil" conformation (Yang & Kubota, 1985). As noted, the relationship of FC_ρ with α_{i1} follows from the FC_α–FC_ρ correlation itself. Our eq 1 formulation of the variance of FC_ρ being independent of that of FC_α enhances the accuracy of its prediction (see Table II) and would not be so easily done in a total back-transformation.

Comparison of UVCD and VCD. The average errors in terms of standard deviations for each FC_ζ over the 13 proteins in the training set are noted at the bottom of Table II. In terms of percent of dynamic range, σ_α is the smallest at 10% and σ_β is the highest at 25%, both of which are consistent with the degree of correlation seen in the regression equations. Compared to the VCD analysis in paper II, the UVCD standard deviations are much higher for σ_β and nearly the same for σ_α and σ_b and significantly lower for σ_ρ, perhaps due to the independent correlation noted above. The average (per conformation) error of fit for the individual proteins is relatively constant (Table II, column 11). These individual errors are comparable to those found for the VCD regressions (Table II, paper II). A comparison of the average differences (absolute values) between VCD and UVCD in terms of predicted FC values is in the last column of Table II.

Comparison of the predictive capability of the multiple-regression equations for UVCD and for VCD are presented graphically in Figure 6. The black bars represent the X-ray-determined FC_ζ values, the slashed bars the UVCD results,

Table IV: Partial Least-Squares Analysis of UVCD Spectra for Known Proteins

protein	KS FC[a]					predicted FC[a]						
	α	β	b	t	ρ	α	β	b	t	ρ	Δ[b]	Δ(UVCD–VCD)[c]
trypsin	8	32	15	14	31	11	49	3	8	29	8	8
trypsin inhibitor	14	24	17	7	38	4	29	15	16	36	5	3
triosphosphate isomerase	43	17	8	7	24	29	24	14	10	24	6	16
ribonuclease S	18	35	12	7	27	16	14	20	16	34	9	2
papain	23	16	18	8	33	23	−1	3−	12	35	7	8
myoglobin	7	0	2	10	11	93	−49	14	22	19	20	22
lysozyme	29	8	13	21	29	26	22	9	12	30	6	6
hemoglobin	67	0	4	9	20	53	−44	29	31	31	23	24
elastase	5	34	9	17	34	8	25	17	16	33	4	7
cytochrome *c*	26	4	22	14	34	27	28	8	8	29	10	16
concanavalin A	0	40	20	9	30	19	−2	11	35	36	21	11
chymotrypsin	6	33	10	16	35	8	−2	20	33	42	14	13
carbonic anhydrase	7	27	18	12	36	10	44	15	2	29	8	4
standard deviation[d]						7	23	10	10	4		
standard deviation (% of dynamic range)[d]						9	57	50	72	15		

[a] Abbreviations are as in Table II. [b] Average differences between KS and predicted data per conformation: $[\sum|KS - pred|/5]$. [c] The last column is calculated as $\sum|FC_f^{UV} - FC_f^{VCD}|/5$. [d] Calculated as in Table II.

and the open bars the VCD results. Some trends are apparent with an increase in FC_α. At low helical content, the UVCD equations overestimate FC_α and the VCD ones are generally better. Over the range studied, both spectral methods have comparable relative errors. Those proteins with little helical content are most poorly fit, in a relative sense. For example, VCD gives 0% for trypsin while UVCD gets it correctly at 8%; on the other hand, UVCD calculates concanavalin A at 8% while VCD gets 1%, very close to its value of 0%. For highly helical proteins, both techniques have more absolute error but similar relative errors, with the correct values tending to be approximately the average of the UVCD and VCD results. With the exception of myoglobin, the UVCD regression equations do not appear to have enough dynamic range to properly account for the KS FC_α values. In particular, they do not yield very low helical values, which again may be due to the strong correlation of FC_α with the overall intensity of the UVCD. While the VCD FC_α values are on the average better than the UVCD ones ($\sigma_\alpha^{VCD} = 8\%$ while $\sigma_\alpha^{UVCD} = 10\%$), it is clear that by having done analyses on both types of spectra we are in a better position to estimate the correct value.

On the other hand, the β-sheet contribution is poorly estimated by the UVCD correlation, which again does not evidence an appropriate dynamic range for the FC_β prediction. Aside from hemoglobin and myoglobin, the UVCD-based FC_β predictions are virtually flat between 20 and 30% while the VCD predictions have considerably more variation. Both spectral methods have predictions that are low at high FC_β values, but the VCD error is much smaller. For concanavalin A, the VCD prediction lies much closer to the experimental value and comes between it and the UVCD prediction, which is further evidence that the influence of the aromatic residues is not significant for amide VCD (Yasui et al., 1986a).

For bend, the behavior is similar to that with β-sheet, such that VCD on the average evidences more dynamic range than UVCD with notable exceptions. However, the standard deviations for both fits are identical. By contrast, for "other" the UVCD regression is a significantly better representation of the FC_ρ values than is the VCD one. It not only has sufficient range but the predicted values agree better with the X-ray ones for nearly all the proteins in the set.

A comparison of the predicted values in Table III for the seven "unknown" proteins to known structures indicates, in most cases, that the method is making reasonable predictions. The high FC_α component for albumin, high FC_ρ for casein, and the moderate FC_α and FC_β contributions for ribonuclease A (Levitt & Greer, 1977) and lactoferrin (Anderson et al., 1987) meet expectations based on previous data. Furthermore, the fact that only FC_β for albumin is negative is satisfying. However, the UVCD-predicted values for β-lactoglobulin deviate from the X-ray values expected for a β-barrel structure (Papiz et al., 1986) much as was seen in the VCD analysis. Similarly, thaumatin and chymotrypsinogen are predicted to have too much helix (DeVos et al., 1985; Freer et al., 1970), which follows from the UVCD overprediction of α-helix for low helix containing proteins (Figure 6). Finally, it should be clear that these regressions are obtained independently so that there is no constraint that the predicted FC_f values should sum to 100%, which was a limitation of earlier methods of calculation (Manavalan & Johnson, 1985).

Partial Least-Squares. As an alternative approach to analysis of the UVCD data set and as a continuing parallel to our VCD work, we have also used the partial least-squares (PLS) method (Haaland & Thomas, 1988a,b), as described in Paper II, to analyze the UVCD spectra. Using the set of 13 KS-analyzed proteins to provide the calibration spectra, the PLS method optimizes the estimates of the FC_f values for all five secondary structural types being considered. Our test runs indicated that a dimension of 30 (Haaland & Thomas, 1988a,b) was optimal. The FC_f values for a given protein were predicted by using the spectra of the other 12 of the 13 KS-analyzed proteins as a calibration set. These results are in Table IV along with the KS-determined FC_f values for comparison. Additionally, this test provides a means of making 13 different predictions for each unknown protein. The average and standard deviations of these calculations (Table V) can be used to assess the sensitivity of the PLS method results to the calibration set.

At the bottom of Table IV are noted the standard deviations of the predictions of each FC_f for the 13 KS proteins. In terms of percent of range, the σ_α and σ_ρ values are markedly smaller than the other three, whereby it becomes clear that the PLS method for UVCD cannot predict β-sheet, bend, and turn at all reliably. As compared to the VCD results using the PLS method, the UVCD predictions are much worse, but the difference is one of degree rather than type. Both PLS methods are not good for β-sheet, bend, and turn, with VCD being less bad. The limitations of a total transformation of the spectral variation into secondary structure factors becomes evident in comparing the PLS and regression methods. The undetermined FC_f values have a consequence on the accuracy of the more well-determined values.

As done in Table II for the regression results, the last

Table V: Partial Least-Squares Analysis for UVCD of Unknown Proteins

protein	FC_α	σ_α[a]	FC_β	σ_β[a]	FC_b	σ_b[a]	FC_t	σ_t[a]	FC_ρ	σ_ρ[a]
albumin	78	7	−27	21	18	9	12	7	20	4
casein	−10	4	7	20	13	8	44	9	46	5
chymotrypsinogen	30	3	11	7	23	4	6	4	31	2
lactoglobulin	25	5	27	19	−5	9	33	9	21	4
lactoferrin	46	3	17	7	6	3	11	3	20	1
ribonuclease A	28	2	19	6	18	3	6	3	29	2
thaumatin	27	2	22	7	13	4	7	3	31	2

[a] σ_f^i = standard deviation of FC_f calculated from 13 estimates using different 12-member training sets for each protein *i*. Abbreviations are as in Table II.

column in Table IV compares the difference in PLS-predicted FC_f values for individual proteins from UVCD with the corresponding VCD results (paper II, Table V) and gives a measure of the deviation of the two techniques. In addition, the average error in FC_f prediction (second-to-last column) for the KS-known proteins can be compared with its counterpart in paper II (Table V). This latter comparison indicates that the UVCD average errors are larger overall than the VCD errors and are larger for 8 of the 13 proteins whether the turn contribution is included or left out. The UVCD errors are also the largest errors in the set of compared numbers, particularly for myoglobin, hemoglobin, and concanavalin A. These are the proteins studied that have no contribution from either β-sheet or α-helix. For the PLS method, if one of the more "pure" spectra is removed from the training set in order to predict its FC_f values, the error in the prediction of that pure contribution will be increased since the quality of the calibration set in that region of "FC space" will decrease. VCD, by contrast, had large errors for chymotrypsin and cytochrome *c*. These are proteins with three significant ($\geq$20) FC_f values. Thus, VCD seems more sensitive to the absence of mixed structures from the calibration set while UVCD is more sensitive to missing pure structures. This sensitivity is paralleled in the relative errors of the regression equations graphically shown in Figure 6.

By contrast, when looked at as a function of the 13 training sets, the average variance for each conformational type in the unknown proteins (Table V) was slightly worse for UVCD than for VCD, with the exception of σ_β, which was significantly worse for UVCD. Given the high σ_f values in Table IV for β, b, and t, it seems unreasonable to put much faith into their determination in Table V for the unknown proteins. Since the predicted FC_f^i values are not independent and are constrained to sum to 100%, the α and ρ values may also be adversely affected (Manavalan & Johnson, 1985). As for VCD, quantitative secondary structure prediction for the unknown proteins is done better by the multiple regressions. This difference in reliability for two statistical approaches to interpretation of the spectral data in terms of structure undoubtedly relates to the underlying assumption in the PLS method that the spectral variation is totally relatable to secondary structure variations. This failure of the PLS method reinforces our reluctance to totally back-transform CD data into secondary structural parameters (Manning, 1989). Our multiple-regression approach avoids this assumption to some degree.

Despite the relative weakness of the PLS approach for this analysis, the complementarity in the two chiroptical spectroscopic approaches, VCD and UVCD, to protein conformational study is seen in the PLS analysis as it was in the regression and cluster analyses. Those proteins that are predicted the worst by one technique are often much better handled by the other. In the PLS case, myoglobin, hemoglobin, and concanavalin A are much better predicted in the VCD analysis than in the UVCD one; cytochrome *c*, triosphosphate isomerase, and lysozyme are better in UVCD, but by somewhat smaller margins in terms of average error.

Conclusions

We have shown that analysis of UVCD data in a manner totally consistent with that done for our VCD data facilitates the quantitative and qualitative comparison of these two chiroptical methods for protein structural study. The result shows that VCD has a clearer correlation than does UVCD with X-ray crystal structure derived FC_f values for given proteins. The clustering and spectra–structure correlation are much more complex for UVCD than for VCD. On the other hand, there is a strong correlation of α-helix content with the coefficients of the first UVCD subspectrum, which is the source of the historical utility of UVCD for secondary structure analyses. UVCD is dominated in magnitude by the α-helical contribution. Since all spectra heavily overlap, changes in magnitude become associated with variation in the α-helix content.

Coupling the helix-sensitive UVCD data to VCD data, where the different secondary structural types contribute on a much more even basis, broadens the sensitivity of both techniques. Our tests with various clustering algorithms showed that the most stable clusters developed when both VCD and UVCD data were used together and that these led to a reasonable correlation with X-ray data. This provides further evidence of the complementarity of the two chiroptical methods. These results agree with a neural network analysis of the same set of spectral data, which will be presented separately (Blazek et al., 1991).

In summary, we have shown that under systematic comparison the differences between the VCD and UVCD are that either one or the other is incrementally better in certain situations, that VCD is better for β-sheet prediction, and that UVCD is better for "other" prediction on average. Together they give information that acts in a mutually reinforcing manner as a guide to an improved structural analysis. But this is not the whole story. The data represented here approach the limits of UVCD. While one can push UVCD spectra further into the vacuum UV region, the changes seen in terms of secondary analysis for change in the short wavelength limit from 184 to 178 nm have been judged to be minor (Hennessey & Johnson, 1981). On the other hand, this series of papers (I, II, and the present one) are the first systematic analyses and comparison to other techniques of VCD spectra for proteins. Being relatively new, they address only one spectral transition of several that are technically accessible in the infrared spectrum. Thus, VCD has a potential for future protein structural studies that, with wider spectral coverage and improved S/N, will offer new structural insights in the future.

Another approach toward improvement of the structure predictions is the use of cluster analysis to identify those proteins that are spectrally similar. A well-defined subclass

can be used as an optimized training set for those unknown proteins which cluster with it. This more focused approach could eventually be used to create a series of subanalyses for various subclasses of proteins. Using such an approach to protein selection seems to be reasonable and avoids overextending the analysis to classes of proteins where it will fail (Yasui et al., 1990). We feel that this is an improvement over the variable selection method used by Manavalan and Johnson (1987). Development of methods to test on some physically determinable basis if the proteins are similar to select them into or out of a group for analysis is underway in our laboratories.

Acknowledgments

We thank the National Institutes of Health for support of this research and the National Science Foundation and University of Illinois for instrumentation support. Profs. Robert Woody and Mark Manning provided helpful manuscripts prior to publication for which we are most grateful. We also thank Dr. Sritana Yasui and Ms. Loretta Mickley, who provided help in obtaining the experimental data, Ms. Rina Dukor for suggestions regarding the manuscript, and Prof. W. Curtis Johnson for detailed explanations of the methods used in his analysis of UVCD.

Registry No. Trypsin, 9002-07-7; trypsin inhibitor, 9035-81-8; triosephosphate isomerase, 9023-78-3; ribonuclease S, 9001-99-4; papain, 9001-73-4; lysozyme, 9001-63-2; elastase, 9004-06-2; cytochrome c, 9007-43-6; concanavalin A, 11028-71-0; chymotrypsin, 9004-07-3; chymotrypsinogen, 9035-75-0; carbonic anhydrase, 9001-03-0.

References

Anderson, B. F., Baker, H. M., Dodson, E. J., Norris, G. E., Rumball, S. V., Waters, J. M., Baker, E. N. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 1769–1773.

Blazek, M., Pancoska, P., & Keiderling, T. A. (1991) *Proc. Neuronet* '90 (in press).

Bolotina, I. A., Chekhov, V. O., Lugauskas, V., & Ptitsyn, O. B. (1980) *Mol. Biol.* USSR (English transl.) *14*, 902; (1981) 709.

Brahms, S., & Brahms, J. (1980) *J. Mol. Biol. 138*, 149–178.

Bullock, P. A., & Myer, Y. P. (1978) *Biochemistry 17*, 3084–3091.

Chen, Y.-H., Yang, J. T., & Chau, K. H. (1974) *Biochemistry 13*, 3350–3359.

Compton, L. A., & Johnson, W. C. (1986) *Anal. Biochem. 155*, 155–167.

Davies, W. K. (1984) in *Factorial Ecology*, Chapter 5, Gower Publishing Company, Hants, England.

DeVos, A. M., Hatada, M., van der Wel, H., Krabbendam, H., Peerdeman, A. F., & Kim, S.-H. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 1406–1409.

Dukor, R. K., & Keiderling, T. A. (1991) *Biopolymers* (submitted for publication).

Dukor, R. K., Keiderling, T. A. & Gut, V. (1991) *Int. J. Pept. Protein Res.* (in press).

Greenfield, N., & Fasman, G. D. (1969) *Biochemistry 8*, 4108–4116.

Freer, S. T., Kraut, J., Robertus, J. D., Wright, H. T., & Xuong, N. H. (1970) *Biochemistry 9*, 1997–2008.

Haaland, D. M. (1990) in *Practical Fourier Transform Infrared Spectroscopy. Industrial and Laboratory Chemical Analyses* (Ferraro, J. R., & Krishnan, K., Eds.) pp 395–468, Academic, San Diego.

Haaland, D. M., & Thomas, E. V. (1988a) *Anal. Chem. 60*, 1193–1202.

Haaland, D. M., & Thomas, E. V. (1988b) *Anal. Chem. 60*, 1202–1208.

Hennessey, J. P., Jr., & Johnson, W. C., Jr. (1981) *Biochemistry 20*, 1085–1094.

Kabsch, W., & Sander, C. (1983) *Biopolymers 22*, 2577–2637.

Lal, B. B., & Nafie, L. A. (1982) *Biopolymers 21*, 2161–2183.

Levitt, M., & Chothia, C. (1976) *Nature 261*, 552–58.

Levitt, M., & Greer, J. (1977) *J. Mol. Biol. 114*, 181–239.

Lipkus, A. H., Lenk, T. J., Chittur, K. K., & Gendreau, R. M. (1988) *Biopolymers 27*, 1831–1838.

Madison, V., & Schellman, J. A. (1972) *Biopolymers 11*, 1041–1076.

Malinowski, E. R., & Howery, D. G. (1980) in *Factor Analysis in Chemistry*, Wiley, New York.

Manavalan, P., & Johnson, W. C., Jr. (1983) *Nature 305*, 831–832.

Manavalan, P., & Johnson, W. C., Jr. (1985) *J. Biosci. (Suppl.) 8*, 141–149.

Manavalan, P., & Johnson, W. C., Jr. (1987) *Anal. Biochem. 167*, 76–85.

Manning, M. (1989) *J. Pharm. Biomed. Anal. 7*, 1103–1119.

Mori, E., & Jirgensons, B. (1981) *Biochemistry 20*, 1630–1634.

Pancoska, P., Fric, I., & Blaha, K. (1979) *Collect. Czech. Chem. Commun. 44*, 1296–1312.

Pancoska, P., Yasui, S. C., & Keiderling, T. A. (1989) *Biochemistry 28*, 5917–5923.

Pancoska, P., Yasui, S. C., & Keiderling, T. A. (1991) *Biochemistry 30*, 5089–5103.

Papiz, M. Z., Sawyer, L., Eliopoulos, E. E., Northe, A. C. T., Findlay, J. B. L., Sivaprasadarao, R., Jones, T. A., Newcomer, M. E., & Kraulis, P. J. (1986) *Nature 324*, 383–385.

Qian, N., & Sejnowski, T. J. (1988) *J. Mol. Biol. 202*, 865–888.

Rohlf, F. J., & Sokal, R. R. (1981) *Statistical Tables*, W. H. Freeman, San Francisco.

Sen, A. C., & Keiderling, T. A. (1984) *Biopolymers 23*, 1519–1532.

Sharaf, M. A., Illman, D. L., & Kowalski, B. R. (1986) *Chemometrics*, John Wiley, New York.

Sokal, R. R., & Rohlf, F. J. (1981) in *Biometrika*, 2nd ed., Chapter 16, W. H. Freeman, San Francisco.

Tiffany, M. L., & Krimm, S. (1968) *Biopolymers 6*, 1379–1382.

Woody, R. W. (1991) *Advances in Biophysical Chemistry* (Bush, C. A., Ed.) Vol. 2, JAI Press, Greenwich, CT (in press).

Woody, R. W., & Tinoco, I., Jr. (1967) *J. Chem. Phys. 46*, 4927.

Yang, J. T., Wu, C.-S. C., & Martinez, H. M. (1986) *Methods Enzymol. 130*, 208-269.

Yang, J. T., & Kubota, S. (1985) in *Microdomains in Polymer Solutions* (Dubin, P. L., Ed.) p 311, Plenum, New York.

Yasui, S. C., & Keiderling, T. A. (1986a) *Biopolymers 25*, 5-15.

Yasui, S. C., & Keiderling, T. A. (1986b) *J. Am. Chem. Soc. 108*, 5576–5581.

Yasui, S. C., Keiderling, T. A., Formaggio F., Bonora, G. M., & Toniolo, C. (1986) *J. Am. Chem. Soc. 108*, 4988–4993.

Yasui, S. C., Pancoska, P., Dukor, R. K., Keiderling, T. A., Renugopalakrishnan, V., Glimcher, M. J., & Clark, R. C. (1990) *J. Biol. Chem. 265*, 3780–3788.