Statistical Analyses of the Vibrational Circular Dichroism of Selected Proteins and Relationship to Secondary Structures[†]

Petr Pancoska,[‡] Sritana C. Yasui, and Timothy A. Keiderling*

Department of Chemistry, University of Illinois at Chicago, Box 4348, Chicago, Illinois 60680

Received September 11, 1990; Revised Manuscript Received February 4, 1991

ABSTRACT: The vibrational circular dichroism (VCD) spectra of 20 proteins dissolved in D_2O are presented in the amide I′ region. These data are decomposed into a linear combination of orthogonal subspectra generated by the principal component method of factor analysis, and the results for 13 of them are compared to their secondary structures as determined from X-ray crystallography. Factor analysis of the VCD yields six statistically significant subspectra that can be used to reproduce the spectra. Their coefficients can then be used to characterize a given protein. Comparison of cluster analyses of these VCD coefficients and of the secondary structure fractional coefficients from X-ray crystallography showed that proteins clustered in the VCD analysis were also clustered in the X-ray analysis. The relative fractions of α-helix and β-sheet in the protein dominate the clustering in both data sets. Qualitative characterization of the secondary structure of a given protein is obtained from its clustering on the basis of spectral characteristics. A strong linear correlation was found between the coefficient of the second subspectrum and the α-helical fraction for the proteins studied. The second coefficient also correlated to the β-sheet fraction, and the first coefficient weakly correlated to the fraction for "other". Subsequent multiple-parameter regression analyses of the VCD factor analysis coefficients, constrained to include only significant dependencies, yielded reliable determination of the α-helix fraction and somewhat less confident determination of β-sheet, bend, and "other" components. Predictive capability for proteins not in the regression was good. Varimax rotation of the coefficients transformed the subspectra and gave simple correlations to secondary structure components but had less reliability and more restrictions than the multiple regression on the original coefficients. The partial least-squares analysis method was also used to predict fractional secondary structures for the training set proteins but resulted in somewhat higher average error, particularly for β-sheet, than the multiple regression. The turn fraction was effectively undetermined in both the regression and partial least-squares analyses. These statistical analyses represent the first determination of a quantitative relationship between VCD spectra and secondary structure in proteins.

A fundamental tenet of biochemistry is that structure and function are intimately related on the molecular scale. This has led to numerous efforts to determine the structure of biological molecules, of which the most detailed have been the X-ray crystal structural analyses. However, crystallization has always been difficult for proteins and nucleic acids so that spectroscopic techniques have been relied upon to generate more limited structural insight into the solution conformations of these molecules. NMR, due to its very high resolution, has made significant inroads into this problem in recent years. Previously, particularly for proteins, electronic (or ultraviolet) circular dichroism (UVCD)[1] and its now outmoded analogue, optical rotatory dispersion, have given needed information about aspects of the secondary structure. Since these chirooptical techniques arise in first order from an interaction of the polarized electromagnetic wave with the three-dimensional electron-density distribution in the chiral molecule, the resultant structural sensitivity of the spectra has a formal relationship to the three-dimensional sensitivity of diffraction techniques. UVCD is still broadly useful in this respect due to its ease of use and high sensitivity, which can allow one easily to obtain data on new systems.

The sensitivity of UVCD to secondary structure stems in large part from its partial dependence on through-space coupling of electric dipole transition moments associated with the repeating amide groups in the protein polymer. This is often designated as the coupled-oscillator mechanism. Electronic UVCD is also quite sensitive to other stereochemical perturbations, which leads it to have much broader use than just for the study of polymeric structures. However, this broader sensitivity can lead to complications when the prime question at hand is one of protein secondary structure analysis. For example, aromatic amino acid residues have transitions that overlap those of the amide which can obfuscate the now standard interpretations of UVCD (Woody, 1977). A recent review has highlighted the perils of assuming that the observed UVCD of proteins arises totally from secondary structure (Manning, 1989). Unfortunately this is a fundamental weakness built into many schemes of protein UVCD analysis.

We have shown recently, in the first paper of this series exploring the application of VCD to protein conformational study (Pancoska et al., 1989), that the measurement of circular dichroism in vibrational transitions of the molecular ground electronic state, termed vibrational circular dichroism (VCD), is straightforwardly accessible for proteins in D_2O solution in the amide I′ band. This experimental development is the

[†] This research was supported through a grant (GM30147) from the National Institutes of Health.

* To whom correspondence should be addressed.

[‡] 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); PC/FA, principal component method of factor analysis; RDI, relative dissimilarity index (in cluster analysis); UVCD, ultraviolet circular dichroism (of electronic transitions); VCD, vibrational circular dichroism (in the infrared).

0006-2960/91/0430-5089$02.50/0

natural outgrowth of our continuing VCD studies of polypeptide conformations (Keiderling et al., 1989). Protein VCD spectra were found to be qualitatively more sensitive to various aspects of protein conformation than are their, now standard, UVCD counterparts. In particular, differentiation of proteins by type, i.e., those having secondary structures charaterized by primarily α-helical, β-sheet, or α and β domains, is more obviously reflected by both sign pattern and frequency changes in the VCD spectra, of just the amide I′, than in UVCD spectra, even when data out to 180 nm are included. In some cases, these two chiro-optical methods are found to be complementary. This arises in principle from the dominance of the UVCD by the α-helical contribution, particularly at longer wavelengths. The VCD appears to result from relatively short-range interactions (Yasui et al., 1986; Dukor & Keiderling, 1991), which leads to all conformations contributing roughly equivalent intensity to the VCD but having varying frequencies and band shapes. Additionally, the conformational sensitivity of the amide I′ absorption generates more resolvable bands in VCD than are similarly available in UVCD.

Quantitative determination of fractions of individual secondary structure segments in a given protein from VCD or UVCD spectra is a much more complicated problem than qualitative discrimination between types. This can be demonstrated by the number of different approaches to this goal for UVCD that have appeared following the classical introductory work of Greenfield and Fasman (1969). These have been extensively reviewed (Yang et al., 1986), and the assumptions forming the basis of all of these approaches have been critically discussed (Manning, 1989). The source of the problems in such approaches can be summarized by the near universal assumption that the secondary structure composition fully determines the protein UVCD spectrum. In other words, factors like the interaction of segments of the secondary structure, side-chain contributions, aggregation or tertiary structure effects, solvation effects, or more basic (physical) factors like the differences in the sensitivity of the amide $n\pi^*$ and $\pi\pi^*$ transitions to their local environment are assumed to average out in such methods.

VCD offers an alternative experimental point of view on the spectra–structure correlation for proteins and has several advantages for studies of polypeptides and proteins. Due to the inherent resolution of the vibrational spectrum, aromatic chromophores (side chains) do not interfere with measurement of the amide I VCD (Yasui et al., 1987a; Yasui & Keiderling, 1986a). As compared to UVCD, in oligopeptide studies, the VCD signal approaches its limiting value and develops its extended form band shape for relatively short segments (Yasui et al., 1986b, 1987b; Dukor & Keiderling, 1991). The characteristic protein VCD bands (e.g., amide I′, C=O stretch) arise from normal modes with a quite well-understood infrared (IR) frequency dependence on secondary structure (Byler & Susi, 1986; Mantsch et al., 1986). These aspects led us to apply VCD to protein stereochemical analysis in an effort to extend the understanding of secondary structure now possible with infrared and UVCD spectra.

With these advantages come some difficulties, of course. Low vibrational dipolar strengths demand use of farily large amounts of protein to make the samples (>1 mg). Use of D_2O as a solvent obscures and shifts bands other than the amide I′ that might be of use in extending the protein spectral data base. Furthermore, D_2O background absorption necessitates the use of high-concentration solutions at low path lengths. [Techniques to obtain reliable preparation-independent VCD data from unrelated solid-phase samples await development (Sen & Keiderling, 1984b; Narayanan et al., 1985a,b); however, amide II VCD spectra are measurable in H_2O (Gupta and Keiderling, unpublished)]. As in UVCD, the amide I′ VCD spectra are composed of multiple overlapping bands of varying sign. Thus with the sensitivity to structural change that comes with the signed data characteristic of a chiro-optical method comes the analogous need to develop a reasonable deconvolution algorithm that does not lead to false structural analyses due to its underlying assumptions.

Despite these technical difficulties, we wish to emphasize that reliable, reproducible data are now accessible for a number of globular proteins in solution by a new technique that exhibits high sensitivity to secondary structure. The set of data we have accumulated to date is broad in terms of the types of proteins studied and encompasses a variety of secondary structural types. Its quality is high enough to lead us to attempt its systematic interpretation and comparison with data from other sources such as X-ray crystallography and UVCD. That is the nature of this and the following papers in this series. X-ray crystal structure data are available for many of the proteins that are discussed here. High- (Kabsch & Sander, 1983) and low-resolution (Levitt & Greer, 1977) analyses of the fractional composition (FC) of the secondary structures have been carried out for 13 and 15 of the proteins studied here, respectively. Further, this set significantly overlaps that used in previous analyses of UVCD spectra and of IR and Raman data, the latter two being based on frequency analyses. This gives us the possibility of comparing techniques in a systematic fashion over a self-consistent data set.

In this second paper in our series exploring the information content of the VCD data of proteins, we focus on secondary structure and its relation to the data we have observed; it is clear that other factors can also influence spectroscopic observables such as VCD. To avoid reproducing the same problems reviewed by Manning (1989) in this introduction of a new chiro-optical technique to the study of protein secondary structure, we propose to address the following question in this paper: How is information pertinent to the protein secondary structure encoded in the VCD spectra?

Our approach to answering this question has been based on the following strategy: secondary structure compositions in our broad set of protein samples provide us with a statistically important variance of the spectral properties of the molecules in the set. We seek to find which part of the spectral variance (in this case, of the VCD spectra) reflects the variations in the fractional secondary structure of the molecules in the set. This requires independent identification of correlated changes in the spectral band shapes. For ease of analysis, the variance of the "analogue" data (spectra) should then be reduced to a numerical form suitable for computer-aided comparison with secondary structural data from X-ray crystallography. For this purpose, we used the principal component method of factor analysis (PC/FA) to reduce the VCD and UVCD spectra to a small number of characteristic parameters (Malinowski & Howery, 1980; Pancoska et al. 1979). This treatment of spectral data is functionally equivalent to that used for UVCD by Hennessey and Johnson (1981). Relationships between these parameters and secondary structural characteristics of the proteins studied were then sought by using regression techniques. Here we diverge from the earlier study of UVCD. It was found most useful for us to use the techniques of cluster analysis (CA) to help identify such correlations graphically (Jardine & Sibson, 1968; Sharaf et al., 1986).

In this paper we will focus on the factor analysis of the VCD data for our protein set and its interpretation in terms of secondary structure using cluster and regression analyses. A parallel effort using partial least-squares is also presented. This whole treatment has been subsequently reapplied to the UVCD data we have remeasured for all of these proteins and to the combined UVCD–VCD data set. In the next paper in this series (Pancoska & Keiderling, 1991), we discuss the analysis of the UVCD and compare it to the VCD results using the same techniques. In a future paper, we will compare the results of a frequency study using band deconvoluted VCD and infrared absorption data (Pancoska et al., unpublished). Each of these studies is dependent on development of a consistent picture of the secondary structures for these proteins. That has been obtained in a similar manner by using PC/FA on the fractional coefficients of the secondary structure as obtained from analyses of X-ray crystallographic data (Pancoska and Keiderling, unpublished).

In the next section are discussed briefly the methods used in our factor analysis approach to the statistical treatment of spectroscopic data. Since our primary interest here is application of VCD data to the determination of protein secondary structures, we here present only a brief discussion of the theory to orient the reader to this statistically based analysis of spectral data. For more detailed reference, the basic formulas and a discussion of the concepts involved in factor analysis along with examples of its application to spectral analysis will be presented in a separate paper (Pancoska, to be published).

Theoretical Techniques. As a tool for the reduction of our spectra into a compact numerical form, we have selected the multivariant statistical approach of factor analysis (PC/FA). This mathematical treatment allows us to transform simultaneously a whole set of n experimental spectra $\{\theta_i(\nu)\}$ into a linear combination of $p \ll n$ orthogonal functions representable as $\{S_j(\nu)\}$, which we denote as subspectra. Both sets are composed of continuous functions of the frequency ν and are related by

$$\theta_i(\nu) = \sum_{j=1}^{p} \alpha_{ij} S_j(\nu) \qquad (1)$$

where the coefficients α_{ij} correspond to the fraction of each subspectrum S_j in each experimental spectrum, θ_i. These subspectra and their coefficients are simultaneously determined from the experimental spectra as a result of the PC/FA calculation using the projection matrix $\mathbf{P}$ as

$$\mathbf{S} = \theta\mathbf{P} \qquad (2)$$

The matrix $\mathbf{P}$ is obtained by diagonalizing the correlation matrix $\mathbf{R}$ of the experimental data whose elements are

$$R_{ij} = \int \theta_i(\nu)\theta_j(\nu)\, \mathrm{d}\nu \qquad (3)$$

where the experimental spectra $\theta_i(\nu)$ are amplitude normalized and

$$\mathbf{P}^{\mathrm{T}}\mathbf{R}\mathbf{P} = \Lambda \qquad (4)$$

in which Λ is the diagonal matrix of eigenvalues, λ_{ii}, of the correlation matrix. From the above, it can be seen that the coefficients, α_{ij}, relating the experimental protein spectra and the subspectra are determined from the projection matrix as

$$\alpha = \mathbf{P}^{\mathrm{T}} \qquad (5)$$

Mathematical details of the derivation of the above equations are available elsewhere (Malinowsky & Howery, 1980).

As our input, experimental spectra in the form of tables of equidistantly recorded intensities are treated as vectors in a functional space. These experimental "vectors" are normalized and all overlap integrals (correlation coefficients) between them are numerically evaluated. The resulting correlation matrix, $\mathbf{R}$, is diagonalized (eq 4), which provides us with the following: The effective dimensionality, p ($p \ll n$), of our data set is given by the number of significant eigenvalues.[2] Only the first p rows of the matrix $\mathbf{P}^{\mathrm{T}}$ corresponding to the largest λ_{ii} values are retained in α. The other subspectra generated are of lower magnitude and represent predominantly the uncorrelated noise that is present in real data. In all cases considered by us, p can be chosen to be much smaller than n without any loss of significant spectral information. This allows the dimensionality of the problem to be reduced. The matrix of eigenvectors yields the projection matrix, $\mathbf{P}$, which in turn gives the subspectra, $\mathbf{S}$ (eq 2). From $\mathbf{P}$, the values of the α coefficients of the linear combination (eq 1) are available (eq 5). The experimental spectrum for every studied sample can then be regenerated by summing over the linear combination of S_j functions after multiplying the coefficients, α_{ij}, by the corresponding norms to give back the intensity information that was initially normalized out.

The common part of the information content of the experimental spectra is thereby transformed to a common basis, $S_1(\nu)$, ..., $S_p(\nu)$. The set of coefficients α_{i1}, ..., α_{ip}, then defines the position of a given experimental spectrum θ_i in a p-dimensional space that is spanned by the truncated $\mathbf{S}$ basis.

The subspectra, $S_j(\nu)$, have no a priori relationship to any of the structural types. The first subspectrum corresponds to the weighted average at each ν value of the intensities of all the experimental spectra used to create the PC/FA. Thus, $S_1(\nu)$ clearly depends on which proteins are used to make up the data set.[3] The subsequent $(p - 1)$ subspectra, S_j, can be interpreted as the weighted average difference spectra of the experimental spectra with, first, S_1 to determine S_2 and, subsequently, with $S_1 + ... + S_{j-1}$ to determine S_j. Their statistical significance is determined by the variance of the data from the summed subspectra, which is proportional to the eigenvalue of the PC/FA algorithm. For example, in decomposing any given experimental spectrum, θ_i, the coefficient α_{i1} is a measure of the degree of typical protein VCD to be found in that spectrum. Conversely, for $p \geq j > 1$, α_{ij} are a measure of the deviation of θ_i from a typical protein spectrum as determined from the training set.

Since the subspectra, in principle, form a complete orthogonal basis set for the p-dimensional space, it is possible to transform them to another basis. To this end, we have written a program for Varimax rotation (Davies, 1984) of the PC/FA subspectra. Briefly, the method used is to modify eq 1, expressed in matrix form as

$$\theta = \mathbf{S}\alpha \qquad (6)$$

by insertion of the identity expressed as the product of a transformation matrix $\mathbf{T}$ and its inverse $\mathbf{T}^{-1}$ to give

$$\theta = [\mathbf{S}\mathbf{T}^{-1}][\mathbf{T}\alpha] = {}^{\mathrm{rot}}\mathbf{S}^{\mathrm{rot}}\alpha \qquad (7)$$

$\mathbf{T}$ is chosen in such a way that the orthogonality of the S_j is retained and that the maximum number of proteins will have

[2] "Significance" is here understood in the following way. It can be shown that dim $[\theta]$ = dim $[\mathbf{R}]$ = dim $[\Lambda]$ = p and Tr $[\mathbf{R}]$ = Tr $[\Lambda]$ = $\sum\lambda_{ii}$ = p. Thus the dim $[\theta]$ is given by the number of nonzero eigenvalues λ_{ii}. Due to the experimental noise, we determine the effective dimensionality p as the number of subspectra z for which the ratio $(\sum_{i=1}^{z}\lambda_{ii})/n$, $z = 1, 2, ..., n$, approaches 0.99. The fraction of spectral variance in the analyzed set as described by a given subspectrum is then proportional to λ_{ii}/n.

[3] This data set can be termed a "training set"; if it is large compared to unknowns added to it, the results obtained from the PC/FA on the whole set should be relatively stable.

large values of ${}^{\mathrm{rot}}(\alpha_{iz})$ for some $z \in \langle 1,p \rangle$ and minimal values for the remaining ${}^{\mathrm{rot}}(\alpha_{ik})$. We thus maximize

$$V = \sum_{j=1}^{p} \{ n \sum_{i=1}^{n} (\alpha_{ij}^2 / h_i^2)^2 - [\sum_{i=1}^{n} (\alpha_{ij}^2 / h_i^2)]^2 \} \qquad (8a)$$

where

$$h_i^2 = \sum_{j=1}^{p} \alpha_{ij}^2 \qquad (8b)$$

and p is the number of subspectra, n is the number of proteins, and α_{ij} are the coefficients from the PC/FA.

Description of Protein Secondary Structures. The basis for the factor analysis herein is the set of spectra for 20 protein samples, 13 of which are included by example in a systematic analysis of protein secondary structures (Kabsch & Sander, 1983). These 13 are globular proteins with known X-ray determined geometries that have been transformed into fractional contributions (FC) from the individual conformational types. Because this transformation can be performed at various levels, we have also compared our PC/FA results with FC values determined from other analyses of crystallographic data where possible (Levitt & Greer, 1977; Hennessey & Johnson, 1981; Chang et al., 1978; Bolotina et al., 1980).

With the nomenclature of Kabsch and Sander (KS) (1983), the X-ray secondary structural conformation associated with each residue can be expressed in terms of helices, extended β-structures, β-bends, turns (combined), and "other" conformations (i.e., all components not covered by these four specific definitions) that we lump together. Any protein sample is therefore describable by a five-component vector of fractional contributions to the secondary structure. A necessary aspect of our structure–spectra analysis was the determination of the relationships that can be found between the secondary structures of the globular proteins in the studied set. (This corresponds to four true degrees of freedom since the sum of the *X-ray determined* FC values is restricted to be 100%). Of course, there exist other studies of correlations between protein structures using various measures (Richardson, 1981; Rackovsky & Goldstein, 1988; Rackovsky, 1990). For practical reasons, we decided to use the cluster analysis methods that we used for the spectral analysis also to characterize the KS described secondary structures thereby yielding a consistent mathematical treatment of both.

In the process of categorizing the structures, some problems of using X-ray data for such an analysis came to light. Exclusive use of proteins that crystallize with facility can in principle be equivalent to selecting a certain protein structural or folding type, which would consequently reduce the generality of the whole approach. Unfortunately, this can not be avoided. Another general problem of solution-phase spectral analysis is implicit in our approach. There is no guarantee that the solution and solid-state structures are identical for these proteins. This problem is being addressed by others on many levels, but for now we will make the usual assumption that these globular proteins have sufficient stability between solid and solution forms for our test of the structure–spectra correlation. If the structures are the same, our approach is clearly reasonable. Similarly, if there is a functional relationship between the solid-state and solution structures of all the proteins in the training set, the procedure remains reasonable. If this latter case were true only for a subset of the proteins studied, we would hope to identify that subset. Cluster analysis can be used for this sort of delineation.

At this point it might be useful to summarize our approach and note what new aspects we are bringing to the analysis of protein secondary structure. We transform a spectral characterization of a set of proteins into a compact, discrete description in terms of the parameters $\{\alpha_{ij}\}$ corresponding to the data set (in this paper, the VCD data are used). Then the spectral relationships between the proteins and the informational content of the $\{\alpha_{ij}\}$ set is sought. Transformations of X-ray data into FC_i vectors as found in the literature are used to delineate their interrelationships in a parallel fashion. Finally, a comparison of the topology of the two reduced data sets is made to see if a useful spectra–structure correlation can be proposed. For this purpose, regression analyses between the two sets of vector components were found to be most useful.

Experimental Procedures

All protein samples used in this paper were purchased from Sigma and used without further purification. As suggested by a referee, these samples are listed below with species, preparation type, or product number indicated in parentheses: albumin (bovine, fatty acid and globulin free, A-0281), carbonic anhydrase (bovine erythrocyte, C-7500), casein (bovine milk, ~60% α-casein, C-7891), α-chymotrypsin (bovine pancreas, Type II, salt free, C-4129), α-chymotrypsinogen A (bovine pancreas, Type II, salt free, C-4879), concanavalin A (jack bean, Type IV, salt free, C-2010), cytochrome *c* (horse heart, Type VI, C-7752), elastase (porcine pancreas, Type III, E-0127), hemoglobin (human, H-7379), lactoferrin (human milk, iron saturated, L-5765), β-lactoglobulin A (bovine milk, L-7880), lysozyme (chicken egg white, grade I, L-6876), myoglobin (horse heart, ferric state, salt free, M-1882), papain (papaya latex, P-4762), ribonuclease A (bovine pancreas, Type III-A, protease and salt free, R-5125), ribonuclease S (bovine pancreas, Grade XII-S, R-6000), thaumatin (*Thaumacoccus daniellii*, T-7638), triosephosphate isomerase (rabbit muscle, Type X, T-6258), trypsin (bovine pancreas, Type III, salt free, T-8253), and trypsin inhibitor (soybean, Type I-S, T-9003). The samples for VCD and infrared (not shown in this paper, but needed for standardization of the intensities) spectral measurements were prepared according to a protocol described in our previous paper (Pancoska et al., 1989) that was optimized to yield high-quality VCD spectra. First, every protein was dissolved in D_2O and lyophilized three times. Then a 5% solution of each was prepared in D_2O (Aldrich), and the spectra were measured by using a sample cell composed of BaF_2 windows separated by a Teflon spacer (~25 μm) clamped between them.[4]

All the VCD spectra presented here were obtained on a dispersive instrument that has been previously described in detail (Keiderling, 1981, 1990). IR absorption data were also obtained on this instrument under identical conditions at ~ 10-cm^{-1} resolution. Some added protein VCD were run on our FTIR-VCD instrument (Keiderling, 1990; Malon & Keiderling, 1988). These data were in all cases fully compatible with the dispersive results presented here but were typically of lower signal-to-noise ratio (Pancoska et al., 1989). Every VCD spectrum presented is an average of four sample

[4] It can be noted that these sampling conditions were designed to get consistent spectra but that these may not be optimal for the handling of individual proteins. In particular, a referee has correctly noted that, due to the high concentrations used, the state of solution aggregation of the proteins studied is potentially a problem that could complicate interpretation of the spectra. It is possible, for a given protein, that aggregation could affect the secondary structure and that the result could be different from the reported crystal structure. For now we neglect such effects in this first effort to demonstrate the methods of deconvoluting structural information from VCD spectra. Future work will address problems of environmental effects, such as concentration, on structure.

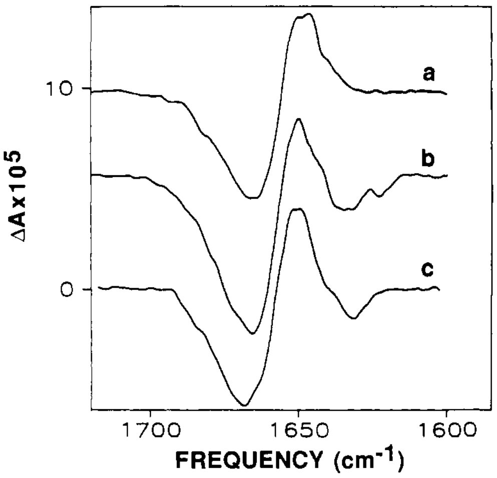

FIGURE 1: VCD spectra of three highly helical proteins in the amide I′ region. Details of the data collection are in the text. All spectra are normalized to an amide I′ absorption maximum of 1.0. Spectra are of (a) myoglobin, (b) hemoglobin, and (c) albumin.

scans collected with a 10-s time constant. These spectra were corrected for artifacts and instrument baseline by subtraction of an identically recorded spectrum of a solution of poly(L,D-lysine) (Sigma) in D_2O whose absorbance matched that of the protein sample in the amide I′ band. Calibration of the instrument was done by our standard procedures, which use a birefringent plate and polarizer assembly as a pseudosample having a known "VCD" signal (Nafie et al., 1976; Su et al., 1981; Keiderling, 1990). For the sake of comparison, the VCD in the figures are scaled so that all proteins have their amide I′ absorbance maximum equal to 1.0.

For data analyses, our spectral files were imported by and structured according to the SpectraCalc (Galactic Software Inc., Nashua, NH) protocols on a personal computer (IBM PS/2 or compatible). This allowed facile arithmetic manipulation of the data and display of the results. For presentation of the VCD and subsequent analyses, the data were smoothed via a Fourier transformation using a triangular apodization function with a breakpoint at 0.5 of the total interferogram range as implemented in SpectraCalc. Zero offsets in the baseline were automatically corrected in the PC/FA analysis. A Turbo Pascal (v 5.0, Borland International, Scotts Valley, CA) version of our previous factor analysis program (Pancoska, et al., 1979) was prepared and used on this system. The input matrices were of the order of 20 spectra by 520 spectral points, and the resulting PC/FA computational times were of the order of 10 min. The EINSIGHT (Infometrix Inc., Seattle, WA) software package was used for the cluster analyses on the parameter sets generated from FA of the VCD as well as on the FC parameters from the X-ray analyses. Regression analyses were calculated for single and multiple variables by using the Statgraphics (v 2.6) program (Statistical Graphics, Inc., Princeton, NJ). For comparison of methods (Discussion), a partial least-squares (PLS) analysis (Haaland & Thomas, 1988) on the VCD spectra was undertaken by using the PLSQuant program in the SpectraCalc package.

Results

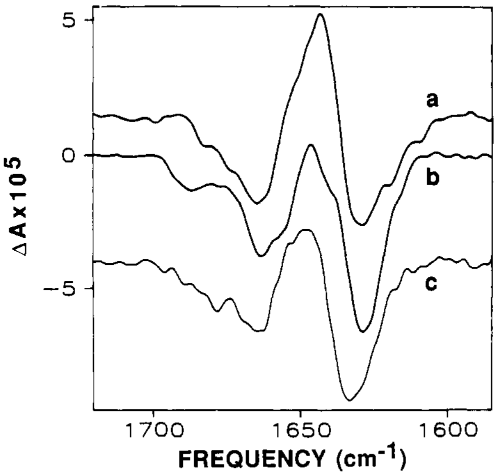

FIGURE 2: VCD spectra of three proteins with mixed structure but dominant α-helical fraction (>25%), recorded as in Figure 1. Spectra are of (a) triosephosphate isomerase, (b) lactoglobulin, and (c) papain.

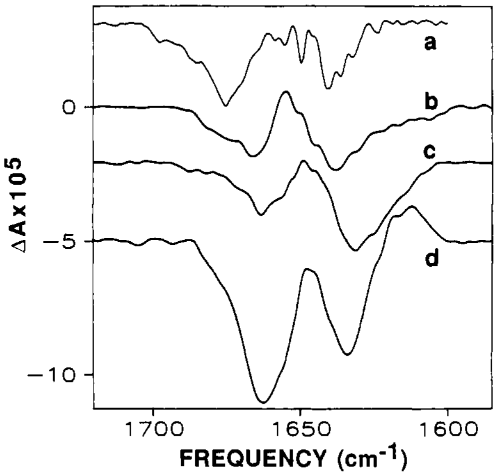

FIGURE 3: VCD spectra of four proteins with mixed structure recorded as in Figure 1. Spectra are of (a) ribonuclease A, (b) lysozyme, (c) lactoferrin, and (d) cytochrome *c*.

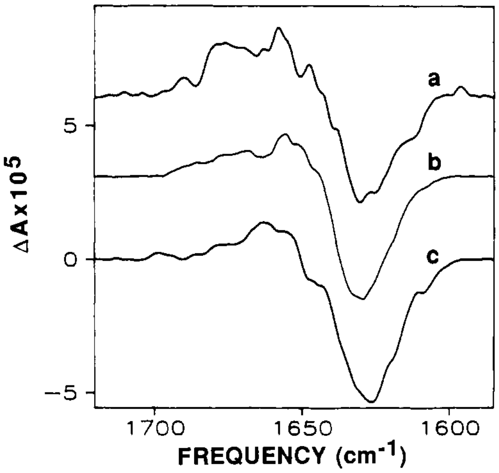

FIGURE 4: VCD spectra of three proteins with a relatively high fraction of β-sheet, recorded as in Figure 1. Spectra are of (a) elastase, (b) carbonic anhydrase, and (c) ribonuclease S.

VCD data obtained for the set of 20 proteins are presented in Figures 1–7. These VCD spectra follow the qualitative patterns noted previously (Pancoska et al., 1989) in that proteins with extensive α-helical character (Figure 1) exhibit a positive couplet to the high-frequency side of the absorption maximum, those of mixed character, such as α + β proteins (Figures 2 and 3) have two dominant negative VCD bands to either side of the absorbance maximum, and those with moderate β-sheet contributions (rarely more than 50%, Figures 4, 5, and 7), as well as "random-coil" dominated proteins (Figure 6), have a low-energy negative VCD band that is typically coupled with a small positive VCD band lying above the frequency of the absorbance maximum. The corresponding IR absorbance bands for the amide I′ are all quite similar in

Table I: Coefficients, α_{ij}, and Eigenvalues, λ_j, of the Protein VCD Spectra from the Principal Component Method of Factor Analysis

protein		PC/FA coefficients					
i	name	α_{i1}	α_{i2}	α_{i3}	α_{i4}	α_{i5}	α_{i6}
1	trypsin	26.73	−15.69	14.91	−17.86	3.34	−4.29
2	trypsin inhibitor	27.51	−1.21	8.91	−2.84	−8.19	3.44
3	triosephosphate isomerase	10.60	17.91	25.14	16.25	−15.06	3.00
4	thaumatin	22.40	−25.13	−40.17	36.16	−39.51	16.69
5	ribonuclease S	19.66	−7.08	10.41	2.62	−19.23	10.60
6	ribonuclease A	4.75	10.02	−16.24	−14.86	−22.45	−16.17
7	papain	16.97	13.39	3.05	−4.73	2.92	23.85
8	myoglobin	−1.35	29.51	13.73	−11.35	3.86	14.58
9	lysozyme	7.88	6.13	−9.52	−9.41	0.42	19.24
10	lactoferrin	12.02	6.78	−1.40	12.30	−0.80	4.33
11	lactoglobulin	20.10	15.86	0.12	34.45	−7.26	−15.66
12	hemoglobin	6.52	40.18	−4.60	−15.63	−2.58	25.77
13	elastase	13.42	−11.15	19.50	−2.23	−0.95	4.65
14	cytochrome *c*	11.60	23.16	−29.48	31.25	40.42	−18.08
15	concanavalin A	11.93	−7.29	−12.31	−12.85	17.42	0.75
16	chymotrypsin	29.18	−4.41	14.05	−30.98	20.60	−33.28
17	chymotrypsinogen	24.15	−14.91	−11.14	4.75	8.70	11.59
18	casein	23.68	−12.70	−10.37	6.10	8.82	10.94
19	carbonic anhydrase	16.27	−5.37	13.59	−4.21	2.41	−7.38
20	albumin	2.61	31.21	12.53	−2.91	12.82	6.48
	eigenvalues	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
	λ_j	11.02	5.90	1.78	0.63	0.32	0.15
	$\sqrt{\lambda_j}$	3.32	2.43	1.33	0.79	0.56	0.39

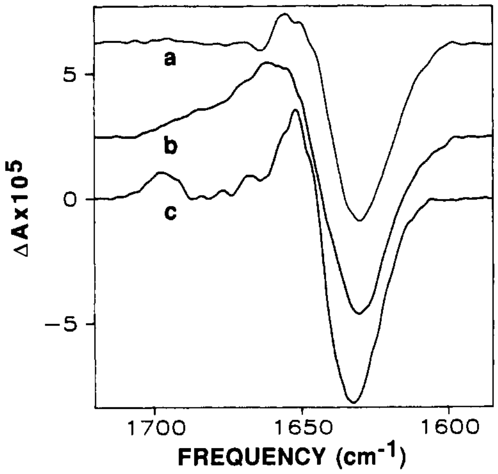

FIGURE 5: VCD spectra of three trypsin-fold proteins, recorded as in Figure 1. Spectra are of (a) trypsin inhibitor, (b) trypsin, and (c) chymotrypsin.

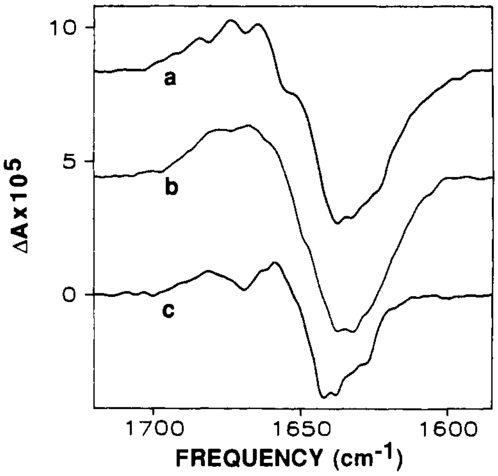

FIGURE 6: VCD spectra of three proteins with dominant extended or "other" secondary structure, recorded as in Figure 1. Spectra are of (a) casein, (b) chymotrypsinogen, and (c) concanavalin A.

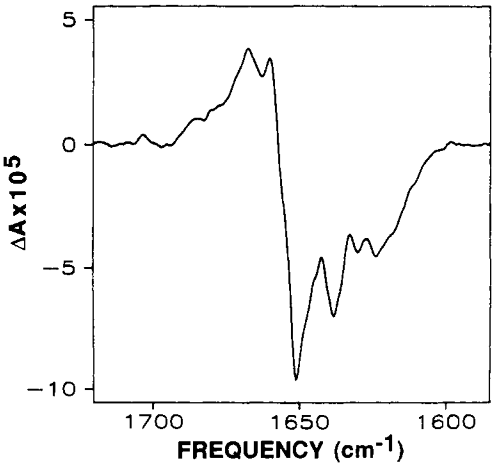

FIGURE 7: VCD spectrum of thaumatin recorded as in Figure 1.

shape and have maxima that evidence frequency shifts of up to 33 cm^{-1} from each other. These frequency aspects of our spectra will be discussed in a separate paper. The presentation of Figures 1–7 was chosen to emphasize the similarities in the VCD spectra of proteins with similar secondary structures.

Upon factor analysis of these data over the spectral region from 1590 to 1710 cm^{-1}, it was found that the spectra could be adequately reproduced with the six most significant subspectra. These respective subspectra are presented in Figure 8, and their respective coefficients for each of the 20 proteins in the training set are listed in Table I. It should be clear that, due to the nature of the factor analysis, the shapes of these subspectra are dependent on the protein set used to generate them. Given a sufficiently broad base of proteins in the set, analyses based on the subspectral coefficients should be relatively rugged; however, the information content may be more or less dispersed. Thus the correct procedure for the analysis of any unknown protein is to include it in the PC/FA with the training set of known proteins so that a sensible set of coefficients for it can be calculated.

The eigenvalues of the correlation matrix form an evaluation of the importance of each subspectrum in fitting the entire data set. The largest eigenvalue determines the most significant subspectrum, $S_1(\nu)$, which is dominated by a negative band at 1630 cm^{-1} as are the VCD spectra of several of our proteins with high β-sheet character. In fact, since most of our mea-

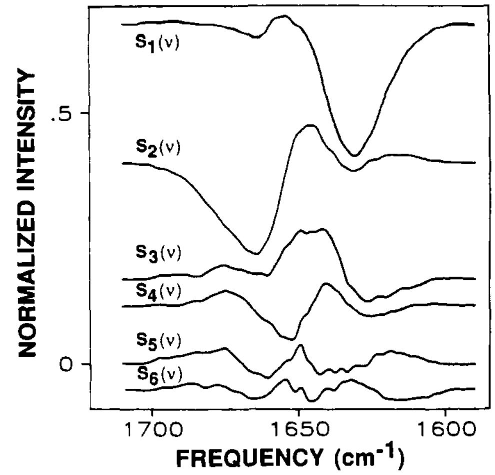

FIGURE 8: Subspectra $S_1(\nu)$ to $S_6(\nu)$ from the factor analysis of VCD spectra of 20 studied proteins (see eq 6).

sured spectra share this band to greater or lesser degree, the first subspectrum is naturally dominated by it. The second subspectrum, $S_2(\nu)$, has a large negative band at 1665 cm^{-1} and a positive band at 1650 cm^{-1}, much like the α-helical proteins in Figure 1. The higher order subspectra contribute important shifts and shoulders to the overall band shape but do so with significantly smaller magnitudes than do the first two. Moreover, they have a correspondingly larger contribution from the noise in the spectral data set and less significant information content. The analysis of the importance of each subspectrum to an individual protein spectrum is represented by the α_{ij} coefficients weighted by the square root of the eigenvalue, $\sqrt{\lambda_j}$. This multiplicative adjustment corrects the coefficients for the different intensities of individual subspectra and is based on a firm theoretical footing (Rummel, 1970).

The PC/FA coefficients (being our numerical measure of protein spectral similarity) were subjected to cluster analyses (CA) by various techniques. The dendrogram from the Lance–Williams Flexible CA algorithm (Sharaf et al., 1986), which gave results typical of the various other CA algorithms, is shown in Figure 9 for the 20 proteins studied (the numbers identifying individual proteins are defined in Table I). Two groups (G1 and G2) of 10 proteins each are separated at the 0.5 relative dissimilarity index (RDI) level. However, at the 0.3 RDI level, G_2 can be further separated into the α-helical proteins, group A ($FC_\alpha > 60\%$ and $FC_\beta \sim 0$), and six proteins with VCD dominated by two negative peaks, group B (proteins with $25 < FC_\alpha < 45\%$, $FC_\beta \sim$ 10–15%, and $FC_\rho \sim$ 25–30%). These six are connected to cytochrome *c* (14) at a slightly higher level. In G_1 at this RDI level, thaumatin (4) is separate from the other nine proteins, which form two similar subgroups (one having the trypsin-like proteins). These nine make up group C and contain our examples of low α-helix, high β-sheet containing proteins ($FC_\alpha < 15\%$ and $25 < FC_\beta < 45\%$). The exact clustering is sensitive to the algorithm used and to the number and identity of the proteins included in the analysis as is addressed further in the Discussion.

This VCD CA result shows a connection of VCD band shape with secondary structure that was already evident in the first paper in this series (Pancoska et al., 1989), which contained a qualitative study of representative proteins. To give the reader a general feeling of the typical VCD band shapes for the CA-recognized protein secondary structural types, it is reasonable to combine the spectra for a given cluster into an average representation of the VCD for the three main clusters, A, B, and C, shown in Figure 10. The average VCD spectrum A is very similar to the α-helical VCD spectrum measured in a variety of model N-deuterated polypeptides (Sen & Keiderling, 1984a; Yasui & Keiderling, 1986a,b). The average spectrum C is similar to the VCD of random-coil polypeptides but is much broader (Dukor & Keiderling, 1989; Yasui & Keiderling, 1986a,b). The average spectrum B looks like a linear combination of the other two.

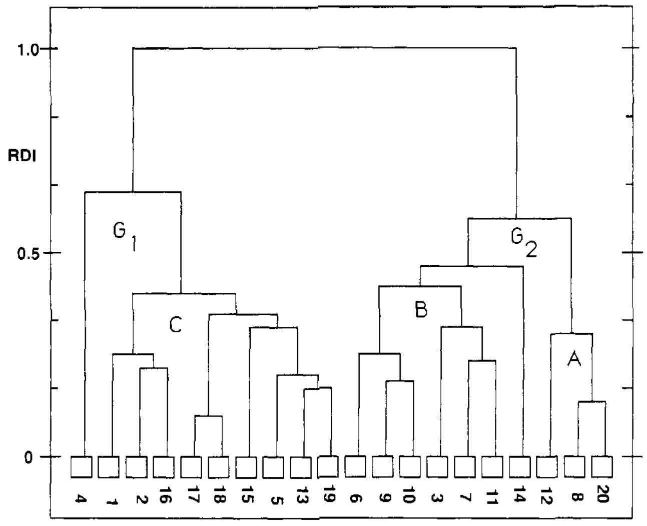

FIGURE 9: Dendrogram showing the clustering of the 20 VCD spectra based on the FA-calculated coefficients weighted by the square roots of the respective eigenvalues. The proteins are identified by numbers as in Table I. Clusters G_1, G_2, A, B, and C are discussed in the text.

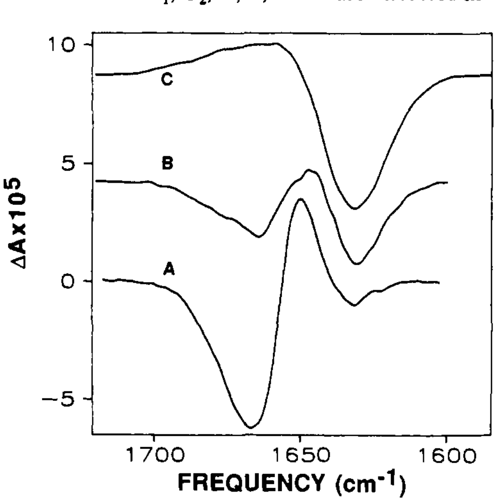

FIGURE 10: Average VCD spectra for the characteristic protein groups A, B, and C (see Figure 9) found in the cluster analysis of the VCD spectra.

Discussion

Clustering of X-ray Determined FC Values. To usefully analyze the secondary structural effects on protein VCD with the broad selection of proteins such as we have chosen, it is necessary to categorize their secondary structures in some manner. Our approach is to generate an image of the topology of these structural characteristics that is compatible with the reduced spectral information. For the sake of discussion, here we will use the "high-resolution" KS (Kabsch & Sander, 1983) categorization that evaluates the respective fractions (FC_f) of α-helix, β-sheet, turn (t), and bend (b) from the range of peptide conformations found in the respective protein structures. The other KS defined conformations not explicitly noted above are grouped with "other" (ρ) for our protein study. FC_ρ is sometimes termed the random-coil component but clearly here includes various contributions that are not at all uniformly

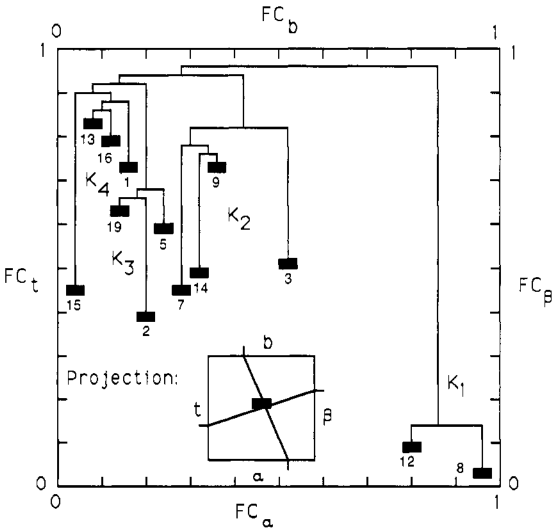

FIGURE 11: Schematic representation of relationship between X-ray determined FC_ζ of secondary structures of 13 proteins (training set used in this study). The location of each protein in the plane is a projection of its four independent FC_ζ [ζ = α-helix, β-sheet, t (turns), b (bends)]. See inset for explanation of the projection method. The points are connected according to the dendrogram resulting from use of the Lance–Williams flexible method of clustering of five X-ray fractional components for these proteins. The relative dissimilarity scale is distorted, but can be visualized by considering the distances between the points in the projection plane. The protein numbers are as in Table I.

characterized. This variation in the nature of the contributions to FC_ρ is a probable source of difficulty in relating it to the spectra. [Other analyses of secondary structure from X-ray data (noted earlier) exist in the literature and can alternatively be used for this purpose as will be discussed later.] Our approach to using the KS X-ray analysis is to create practical descriptive vectors of the secondary structure having as components the FC values for these five secondary structural types that are constrained by the X-ray analysis to sum to 100%. Then cluster analysis is used to create a measure of the similarity of these vectors and to group them by this recognition algorithm. A complete discussion of our approach to the cluster analyses of X-ray derived secondary structure data will be published separately.

Our analyses developed various schemes to view the X-ray data. For 13 of the proteins in the KS categorization, we have also measured VCD spectra.[5] As an example, in Figure 11, a combination of two methods of reducing the crystal-structure data are illustrated. First, for these 13 proteins, four independent FC values, FC_α, FC_β, FC_t, and FC_b, are used to project the five-vector describing the secondary structure onto a two-dimensional plane as shown in the inset to Figure 11. This results in grouping similar proteins on the basis of secondary structure. Secondly, the FC values for these 13 proteins have been analyzed with the Lance–Williams flexible CA algorithm. A schematic dendrogram representing this result is superimposed on the two-dimensional reduction of the FC values in Figure 11.

The CA of the FC_ζ values in this set of 13 proteins resulted in three main clusters that are primarily determined by the α-helical and β-sheet fractional components of the proteins. The reason for this follows directly from the relatively large dynamic range of the FC_α and FC_β variables (i.e., 0–85% and 0–50%, respectively) as compared to those of FC_t and FC_b (~0–26%). The first cluster (K1) corresponds to those proteins very high in α-helical content (65–77%) and with no β-sheet content. From the subset of proteins for which we have VCD, the second cluster (K2) contains cytochrome *c*, triosephosphate isomerase, papain, and lysozyme having moderate correlated α-helical (22–45%) and β-sheet (3–18%) content. The third group, K3, has less α-helix and more β-sheet content than the above two clusters and can be divided into two subgroups.

PC/FA-Based VCD Cluster Analyses. With the structural data reduced to a five-vector of FC_ζ values, it follows that use of a parallel treatment for the spectral data implies that the spectra should be reduced from relatively complex arrays of intensities at many frequencies to similar small vectors of coefficients. That is what the factor analysis (PC/FA) method provides. The coefficients of the subspectra weighted by the square root of the eigenvalues provide a set of six-vectors to represent the VCD spectra of the proteins in the PC/FA analyzed set.[6] Cluster analysis provides a unifying treatment to elucidate the correlations between the spectral data sets and the X-ray results. The Lance–Williams flexible CA algorithm, used for Figures 9 and 11, gives typical results. The VCD cluster A, as shown in Figure 9, contains only X-ray analyzed proteins from K1 (Figure 11), cluster B from K2, and cluster C from K3. Cytochrome *c*, while in K2 is, however, only weakly coupled to cluster B from our VCD analysis. The correlation between the VCD data and the X-ray derived FC values for secondary structure seems clear: at this similarity threshold for the Lance–Williams CA of our 13 proteins, there are no errors in protein clustering with both descriptors.

Other CA algorithms give very similar results. We have in fact tested this with all seven methods provided in the EINSIGHT package including single linkage (nearest neighbor), complete linkage (farthest neighbor), group average, incremental sum of squares, centroid, median, and Lance–Williams flexible. Schematic representations of the CA consistency with calculational method using these seven different algorithms are in Figure 12 panels a and b for the VCD and X-ray coefficients of the 13 KS-analyzed proteins, respectively. Here all proteins placed within a box boundary are found clustered together with all algorithms tested. Those placed between boxes have a change in clustering with one or more algorithm.

Overall the CA result is quite stable and has a clear structural correspondence. At this level, the three X-ray clusters, K1–3, have no changes with algorithm. The VCD groups A, B, and C are somewhat less stable; thus proteins whose VCD parameters lie between boxes must be considered less well-characterized by the spectra. Overall, group A and cluster K1 are α-helix rich, group C and cluster K3 are β-sheet rich and α-helix poor, while group B and cluster K2 are moderate in both, as seen in α+β proteins (Levitt & Chothia, 1976). The close correspondence of the X-ray and VCD CA results implies a strong interdependence of the two sets of coefficient vectors and hence of what they represent. This demonstrated relationship between spectra and structure is what we originally sought in validating the use of VCD as a useful new structural tool for secondary structure elucidation.

Thus viewing the protein structures only at the level of dominant type, if a new protein were added to the analysis,

[5] In some cases, we used the KS criterion for neglecting variation of proteins between species since KS provided analysis for only the best structure in a related homologous group, and we did not always have access to that protein. For myoglobin, two species (horse heart and sperm whale) gave identical VCD within the noise level. Clearly, this is too little data for establishment of a trend.

[6] By comparison, electronic CD is represented by five subspectra, thus giving rise to a set of five-vectors as shown in next paper of this series.

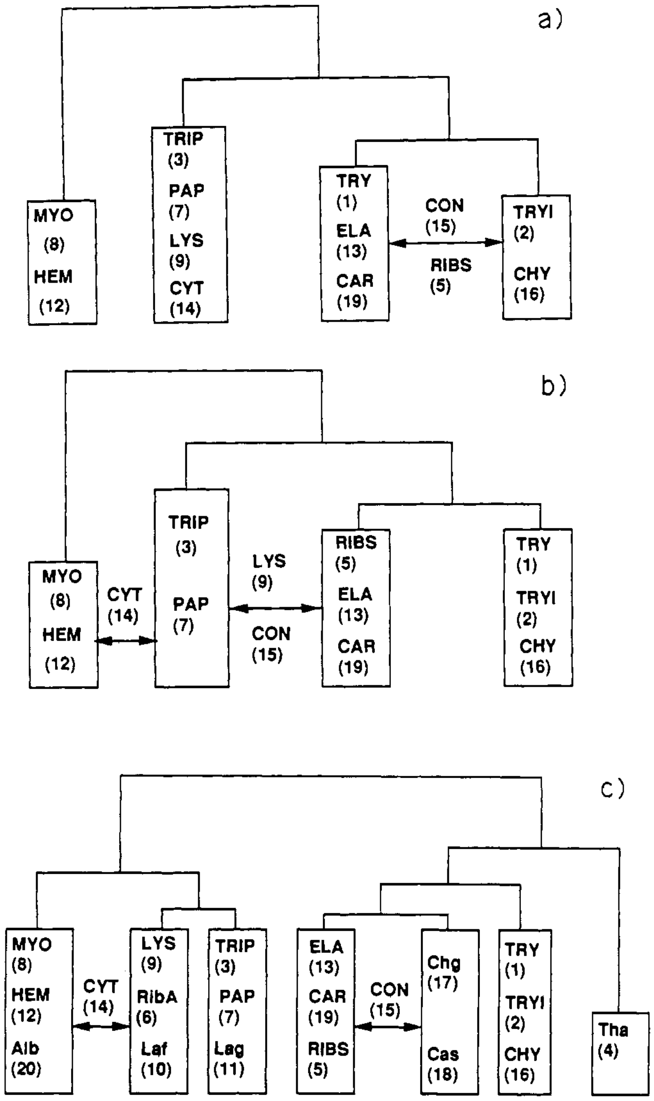

FIGURE 12: Schematic description of the cluster analyses with seven different algorithms for (a) FC values from the KS X-ray analysis, (b) PC/FA coefficients for the 13 KS-analyzed protein VCD spectra, and (c) PC/FA coefficients for all 20 proteins organized as described in the text.

it would be expected to have the general characteristics of the group into which it clustered on the basis of its VCD coefficients. Conversely, if its VCD does not cluster with those of the training set proteins, little can be said from this analysis with respect to that protein's structure. In this study, there are 7 proteins in our set of 20 for which we have VCD data but that are not included in the KS X-ray study. The result of cluster analyses with the seven CA algorithms for all 20 sets of protein VCD spectral coefficients is illustrated in Figure 12c. (The dendrogram in Figure 9 is a more detailed representation of one of these, the Lance–Williams method.) Adding proteins can alter the clustering somewhat, but it is clear that the main structure is preserved. The result with 20 proteins has actually become more stable, with only cytochrome *c* (14) having errors in clustering between groups A and B. (The added proteins are in lower case, the 13 from KS are in upper case.) From a comparison of panels a and c Figure 12, it is clear that even considering all CA algorithms, cytochrome *c* is the only mismatch in the two sets, at the discrimination level of the three main clusters. Inspection of Figure 9 shows cytochrome *c* to be only weakly clustered to group B. Multiple CA calculations may lead to the same conclusion as could be derived from more detailed study of the RDI levels for the clusters found in a "typical" algorithm.

From the VCD CA results, we expect albumin (20) in cluster A to be high in α-helix content and β-lactoglobulin A (11), lactoferrin (10), and ribonuclease A (6) from cluster B to have α and β domains. Casein (18) and chymotrypsinogen (17) have very similar spectra and are weakly coupled to cluster C; from the VCD analysis, it would be expected that they might have significant β-sheet fractions. However, it is likely that the VCD of the random-coil component, which has been determined to have some relationship to that of extended helices in polypeptides (Dukor & Keiderling, 1991), may confuse this clustering. Thaumatin is so weakly connected to cluster C that, on the basis of the above remarks, it would be unwise to use VCD to speculate on its secondary structural classification in solution, despite its known β-barrel structure from the X-ray results (DeVos et al., 1985).

Regression Analyses of PC/FA Coefficients. A major use of UVCD in biochemistry has been for estimation of fractional secondary structure. The approaches to this problem have been manifold from early model polypeptide-based determinations (Greenfield & Fasman, 1969) to more recent singular value decomposition (Hennessey & Johnson, 1981) and multivariate (Provencher & Glöckner, 1981) analyses. These have been extensively reviewed by Yang et al. (1986). A similar relationship with VCD data is expected. While we are admittedly at the beginning of this process, we have made several attempts to use these VCD data more quantitatively to find a correlation with secondary structure. Unlike the effort in UVCD, we prefer not to form a complete back transformation from the factor-analysis coefficients into *only* a description of secondary structure. This would be tantamount to assuming that *all* of the VCD band shape (or, at least, variances at a level analyzable by the first six subspectra) would be directly dependent on the secondary structure and not on any other factors. In fact, we do not believe this to be true.

One approach to interpretation of the PC/FA coefficients is to search stepwise for their correlations with the FC_ζ values Plots of $\{\alpha_{ij}\}$ versus FC_ζ^i for the i = 1–13 proteins in the KS set and for all possible pairs of j and ζ showed that the data did not indicate a need for nonlinear correlations. Therefore, first any single-parameter regressions between the VCD coefficients and the FC_ζ parameters were sought. Then more complex regressions of FC_ζ with multiple α_{ij} parameters were evaluated. In this paper we discuss only those correlations that are statistically significant at the 99% confidence level (unless explicitly stated otherwise). This hypothesis was tested in all cases by using critical values for correlation coefficients corresponding to $\nu = n - 1 - m$ degrees of freedom, 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).

From the search for simple regression analyses, only the first two coefficients, α_{i1} and α_{i2}, were found to be strongly correlated to the common measures of secondary structure. The α_{i2} parameters can be linearly fit to the FC_α and FC_β values,[7]

[7] It may seem surprising that one coefficient can correlate to two independent structural parameters. In summary (Pancoska, to be published), if one plots FC_α vs FC_β for all 62 KS proteins, nearly 80% have FC values lying within ±10% of the best fit line representing an inverse α–β relationship. The remaining 20% (typically, non-heme metalloproteins) fall outside this "galaxy" according to the notation of Rackovsky (1990) and have their FC_α and FC_β values uncorrelated. The Levitt and Greer (1977) analysis leads to a similar self-correlation on a smaller number (43) of proteins, but the roughly parallel line is substantially shifted up from the KS line.

Table II: Fractional Concentrations of Secondary Structural Types for the Training Set Proteins from the Multiple-Regression Equations

	KS X-ray FC_ζ[a]				VCD-predicted FC[a]				
protein (KS source)	α	β	b	ρ	α	β	b	ρ	Δ[b]
trypsin (bovine pancreas)	8	32	15	31	0	37	15	37	5
					−2	39	15	38	6
trypsin inhibitor (bovine pancreas)	14	24	17	38	19	26	14	37	3
					19	27	13	37	3
triosephosphate isomerase (chicken breast)	43	17	8	24	49	12	5	27	4
					52	11	3	27	6
ribonuclease S (bovine pancreas)	18	35	12	27	15	31	15	32	4
					14	30	15	33	4
papain (papaya)	23	16	18	33	25	15	12	31	3
					25	15	11	30	3
myoglobin (sperm whale)	77	0	2	11	63	3	5	20	7
					56	4	7	24	11
lysozyme (chicken egg white)	29	8	13	29	26	21	17	25	6
					25	22	18	25	7
hemoglobin (horse)	67	0	4	20	73	−5	8	24	5
					80	−8	10	25	8
elastase (porcine pancreas)	5	34	9	34	11	34	13	28	4
					13	34	14	28	5
cytochrome *c* (tuna heart)	26	4	22	34	26	8	19	27	4
					25	9	16	27	5
concanavalin A (jack bean)	0	40	20	30	1	31	21	28	4
					2	30	21	27	4
chymotrypsin (bovine pancreas)	6	33	10	35	3	29	13	38	3
					1	28	13	39	4
carbonic anhydrase (human erythrocytes)	7	27	18	36	11	29	13	30	5
					12	30	13	30	5
std dev[c]					6	6	4	5	
					9	7	5	6	
std dev in % of dynamic range[c]					8	15	19	19	
					11	17	25	24	

[a] Abbreviations of secondary structures: α, α-helix; β, β-sheet; b, bends; ρ, other conformations. In the first row are the values for eq 10, and in the second row for each protein are results for that protein based on regressions over a set reduced to 12 by eliminating that protein from the regression determination. [b] Average differences between KS and predicted data per conformation: $[\sum|FC_\zeta^i(\text{KS}) - FC_\zeta^i(\text{pred})|/4]$. [c] Calculated from the differences Δ between X-ray FC_ζ and predicted FC_ζ values as the standard deviation for a given secondary structure type $\sqrt{[\sum\Delta^2/(n-1)]}$. The dynamic range is max FC_ζ − min FC_ζ within the KS column. The second row in each case is for the reduced regressions over 12 proteins.

as determined from the KS procedure, to a high degree of correlation and with a typical error of ±10% (absolute):

$$FC_\alpha{}^i = 17.38 + 1.25 \cdot \alpha_{i2}, \qquad r = 0.90 \qquad (9a)$$

$$FC_\beta{}^i = 25.40 - 0.75 \cdot \alpha_{i2}, \qquad r = -0.91 \qquad (9b)$$

In addition, the FC_ρ values can be fit to α_{i1} with a somewhat less, but still significant (99% level), degree of confidence:

$$FC_\rho{}^i = 20.4 + 0.61 \cdot \alpha_{i1}, \qquad r = 0.70 \qquad (9c)$$

To the level needed for statistical reliability for the 13 KS proteins in our training set, no other single-parameter correlations were found.

It should be recalled that the factor analysis procedure creates subspectra based on their commonality and orthogonality but not on their correlation to particular structural aspects of the proteins. Thus, it is not surprising that single subspectra do not strongly represent individual components of the secondary structure. (Thus, in retrospect, the correlations noted above with α_{i2} are very pleasing.) Alternatively, one can seek correlations to several or all of the PC/FA parameters with the FC values. If done indiscriminately, this approach would amount to a back transformation of the data, thereby implicitly assuming that the $FC\zeta$ values and all α_{ij} coefficients are correlated. One of our early goals was to avoid imposing this assumption on the data. We have therefore optimized multiple-parameter fits to the $FC\zeta$ values by systematically testing all pairs, triples, etc. to determine which combinations of parameters give the best regressions.

To determine the form of the final multiple-regression equations (eq 10), we used the following criteria. First, the fit was required to be statistically significant based on critical values of the multiple-correlation coefficient ($\sqrt{R^2}$) (Rohlf & Sokal, 1981). In ambiguous cases where more than one relationship fit the first criterion, we selected that one 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 are summarized in matrix form as

$$\begin{bmatrix} FC_\alpha \\ FC_\beta \\ FC_b \\ FC_t \\ FC_\rho \end{bmatrix} = \begin{bmatrix} 0 & 1.25 & 0.41 & 0 & 0 & 0.37 \\ 0 & -0.75 & 0 & 0 & 0 & 0 \\ 0 & -0.23 & -0.28 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0.61 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \otimes \begin{bmatrix} \alpha_{i1} \\ \alpha_{i2} \\ \alpha_{i3} \\ \alpha_{i4} \\ \alpha_{i5} \\ \alpha_{i6} \end{bmatrix} + \begin{bmatrix} 15.4 \\ 25.4 \\ 15.8 \\ 0.0 \\ 20.4 \end{bmatrix} \qquad (10)$$

The corresponding $\sqrt{R^2}$ values are 0.96, 0.90, and 0.70 for helix (α), sheet (β), and other (ρ), respectively (≥99% level). With two parameters, the bend (b) correlation was judged to be less significant (95% level) having $\sqrt{R^2} = 0.74$. All the coefficients in eq 10 have a significance level of ≥96% by the F_S test.

We have left the fourth row in eq 10 remaining as all zeros to emphasize that, with the VCD data currently available, the turn fractions, FC_t, are undetermined within the stated criteria of significance. Furthermore, use of multiple parameters did not provide a means of significant improvement for FC_β and FC_ρ determination as compared to the single-parameter results. Thus the single-parameter eqs 9b and 9c are retained in eq 10. The only "higher order" subspectrum retained in our multiple regression is S_6, which is used for the α-helix determination. Its inclusion results in a drop of the standard

Table III: Predicted Fractional Contributions for Proteins Not Included in the Regression Training Set

protein	predicted FC[a]			
	α	β	b	ρ
albumin	62	2	5	22
casein	−1	35	22	35
chymotrypsinogen	−4	37	22	35
lactoglobulin	30	13	12	33
lactoferrin	25	20	15	28
ribonuclease A	15	18	18	23
thaumatin	−26	44	33	34

[a] Abbreviations of secondary structures: α, α-helix; β, β-sheet; b, bends; ρ, other conformations.

deviation σ (Table II) to 6 as compared to $\sigma = 9$ for a regression using just α_{i2} and α_{i3}. Furthermore the three-parameter fit results in no negative FC_α values being predicted for the KS set using eq 10.

In Table II are listed the KS-determined FC_ζ values and their VCD-predicted values (eq 10) for the 13 "known" proteins in the training set. From the last column in Table II, it can be seen that the average error in the linear fit is evenly distributed among the 13 proteins, having a range of only 3–7. Similarly, the standard deviations among the four FC values determined by eq 10 are very similar, varying from 4 to 6. However, when the dynamic range of these FC values is taken into account, it is clear that the α-helix fraction is the best determined by nearly a factor of 2.

To investigate how the regression relations depend on the training set chosen, we redetermined them (with the same form as eq 10) for 13 sets of 12 samples, systematically leaving out one of the KS proteins in each. The FC values for this protein were then calculated from the formulas and compared to the KS-determined FC_ζ values. The results are entered in Table II as the second line of values listed for each protein. The average errors showed little or no increase from the eq 10 results except for the examples with the highest helical contents, myoglobin, hemoglobin, and triosphosphate isomerase. This can be understood from the disproportionate weight these proteins have on the slope of the FC_α regression and the effect of their variance on such a small subset when one is dropped. If the 13 regression relations are treated statistically, the coefficients have standard deviations varying from 0.02 to 0.08, and the constant terms (right-hand component of eq 10) range from $\sigma = 0.5$ (α) to $\sigma = 1.4$ (ρ) in variance. Both indicate a high degree of stability in the regression relations (eq 10).

In Table III are listed the VCD (eq 10) predicted secondary structures for the "unknown" proteins, i.e., those not in the KS analysis but included in the amide I′ VCD factor analysis. Alternatively, one could use all 13 sets of regression relations (each based on 12 proteins) as derived above to predict the FC values for these seven proteins and average the results. Having done so, we found no significant difference from the Table III numbers. It should be made clear here that Table III is presented in the spirit of documentation of the method. Therefore we have not suppressed, for example, negative values as they give the reader some warning about what might come out of such a calculation. Secondly, it is useful to note that these are the results of independent regression equations and, thus, are not constrained to sum to 100%.

The estimations vary in quality. In the category of well-predicted proteins are albumin, known to be highly helical, lactoferrin, reported to be 32% α-helix and 22% β-sheet (Anderson et al., 1987), and chymotrypsinogen, which is claimed to have a structure similar to that of chymotrypsin (Freer et al., 1970). Levitt and Greer (1977)[8] evaluated the structure of ribonuclease A as having 23% helix and 47% sheet, which, accounting for the higher FC values that result from their algorithm, is also in reasonable agreement with the Table III result.

By contrast, β-lactoglobulin A has been reported to be dominated by a β-barrel conformation having 3% α-helix and 45% β-sheet (Papiz et al., 1986), which is not predicted well by eq 10. It is clear that the β-lactoglobulin VCD band shape (Figure 2) is clustered with and typical of an α+β protein (Levitt & Chothia, 1976) and is not compatible with other proteins having high β-sheet fractions such as concanavalin A (Figure 6). This mismatch could relate to a difference in crystal and solution structures or could be a result of some aspect of the β-barrel that is not well accounted for in our training set. For thaumatin, the predictions in Table III are not good. While it is also known to be a β-barrel (DeVos et al., 1985) and the VCD-predicted FC_β value is high, the large negative FC_α value predicted obviates the reliability of this prediction. As noted above, thaumatin fails to cluster well with the group C proteins, which may explain the inapplicability of the regression equations (eq 10) in this case.

Two basic questions arise regarding the above regression analysis. The first relates to how the results depend upon the algorithm used for determination of fractional secondary structural content from X-ray data. We have analyzed the single-variable fit of the PC/FA coefficients to the LG "low-resolution" parameters (Levitt & Greer, 1977), as an opposite extreme from the KS analysis. The LG data set has 15 proteins for which we have obtained VCD data. In terms of α-helix and β-sheet content evaluations, the LG algorithm for determining secondary structure also exhibits linear relationships between α_{i2} and the FC_α and FC_β:

$$FC_\alpha{}^i(\mathrm{LG}) = 23.08 + 1.60{\cdot}\alpha_{i2}, \qquad r = 0.93 \qquad (11a)$$

$$FC_\beta{}^i(\mathrm{LG}) = 39.60 - 1.11{\cdot}\alpha_{i2}, \qquad r = -0.90 \qquad (11b)$$

The LG and KS correlations are comparable with the LG slopes being larger than the respective KS ones, and the lines are shifted up in the LG correlation by 6% in FC_α and 14% in FC_β due to the coherent structure elongation effect (footnote 8). Our efforts to find a significant correlation ($r > 0.7$) of the LC $FC\zeta$ values with any of the other α_{ij} coefficients have not been successful.

We tested other algorithms for the interpretation of X-ray data by comparing the slopes of regression lines of α_{i2} coefficients as determined from the VCD spectra with FC values according to KS, LG, Bolotina et al. (1980), Chang et al. (1978), and Hennessey and Johnson (1981) (CD corrected). The linear correlation of α_{i2} with FC_α and FC_β remained valid, although some were at the lower limits of reliability due to the smaller quantity of FC values available. The regression lines again fall within a ±10% (absolute) range. Thus, in terms of applicability of the technique, the precise nature of the X-ray analysis algorithm does not seem to be significant. A set of equations can be developed for each algorithm, but the numbers so generated must be interpreted in the context of that algorithm. The predictions, even if 100% consistent, can naturally be no better than the X-ray analysis algorithm used to create the FC_ζ parameters for the training set. At this point generation of more equations seems premature, as we are only

[8] Levitt and Greer (1977) define α-helix and β-sheet in a similar way as do KS but the LG definition encompasses a wider range of dihedral angles. In essence, this elongates the coherent structures and results in uniformly larger numbers for these two FC parameters as compared to the KS values (footnote 7). While turns are still defined by LG, bends and all of the other components are grouped under the ubiquitous "other".

Table IV: Varimax Rotated PC/FA Coefficients, ${}^{rot}\alpha_{ij}$, of the Protein VCD Spectra

protein		varimax coefficients					
i	name	${}^{rot}\alpha_{i1}$	${}^{rot}\alpha_{i2}$	${}^{rot}\alpha_{i3}$	${}^{rot}\alpha_{i4}$	${}^{rot}\alpha_{i5}$	${}^{rot}\alpha_{i6}$
1	trypsin	27.34	-3.58	2.00	-7.74	-23.23	12.97
2	trypsin inhibitor	11.14	4.14	-4.32	9.51	-23.68	10.66
3	triose phosphate isomerase	-0.64	12.76	5.27	25.11	-21.22	-17.14
4	thaumatin	-42.97	-33.37	-26.43	28.10	-19.12	32.87
5	ribonuclease S	-2.98	-0.98	-3.61	7.48	-30.40	5.58
6	ribonuclease A	4.78	-0.21	-36.41	-0.29	0.44	-4.90
7	papain	27.06	3.26	-11.32	13.58	2.95	-3.55
8	myoglobin	1.45	36.29	3.56	-0.83	-2.21	-8.96
9	lysozyme	-7.43	16.29	-5.25	-5.15	-3.51	16.19
10	lactoferrin	-2.59	4.18	0.95	16.77	-3.43	6.72
11	lactoglobulin	3.26	-5.19	-3.78	45.58	-0.52	-3.22
12	hemoglobin	-8.20	48.40	-12.50	1.98	-0.47	4.75
13	elastase	9.50	-1.95	11.25	-1.09	-22.00	2.50
14	cytochrome *c*	11.21	2.38	7.50	38.63	48.49	21.13
15	concanavalin A	12.51	0.56	-0.47	-9.23	7.40	22.79
16	chymotrypsin	58.44	-1.45	-4.25	-8.15	-3.34	6.86
17	chymotrypsinogen	1.18	-6.04	4.19	5.41	-7.44	31.97
18	casein	1.24	-4.93	4.60	7.24	-6.68	30.31
19	carbonic anhydrase	18.73	-2.54	3.91	3.01	-13.05	2.18
20	albumin	8.06	32.58	7.97	8.95	5.58	-6.90

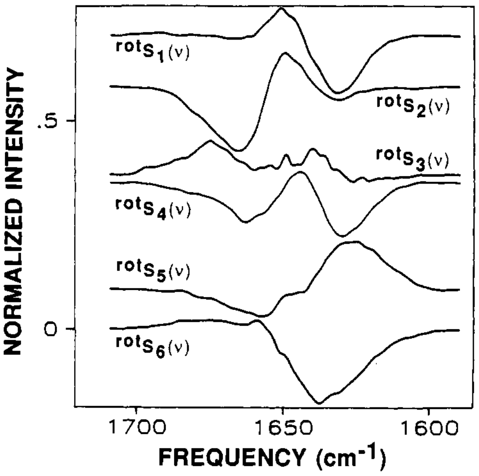

FIGURE 13: Subspectra ${}^{rot}S_1(\nu)$ to ${}^{rot}S_6(\nu)$ as calculated by the Varimax rotation of the PC/FA subspectra (for comparison, see Figure 8).

trying to develop a method of relating solution-phase spectral data with an interpretation of X-ray crystal structure data and are not trying to present the "best" final relationship, which may be dependent on refinement of both the method and the data going into it.

The second major question is the numerical stability of the result with respect to variations in the training set used for the factor analysis and in the subsequent linear correlations. As a test, we have tried to reduce the number of proteins in the PC/FA set. In those cases where the reduction did not remove completely one of the delimited structural types (A, B, or C clusters) the calculated subspectra and the coefficients for the remaining proteins were qualitatively similar to the values presented here. The linear correlation of α_{i2} with FC_α and FC_β also remained valid with the slope of the linear regression changing slightly. The area swept out by this set of line segments is encompassed by parallel lines that have a vertical offset of ±10% (absolute). This is similar to the variation noted above for different X-ray interpretation algorithms.

In summary, our PC/FA coefficients have given us a clustering that indicates proteins by type from analysis of VCD data. Additionally, we have found simple and multiple correlations of the α, β, ρ, and b FC_ζ values with a few of the α_{ij} coefficients. While the results in Table II derived from eq 10 evidence an acceptable degree of variance between the VCD-predicted FC_ζ values and the X-ray determined ones, if the factors affecting numerical stability noted above couple unfavorably with these variances, a less satisfactory situation would develop. As alluded to in the discussion of Table III, such problems can be minimized by using the results of the cluster analyses on the PC/FA coefficients of the proteins to judge the reasonability of the predictions.

Varimax Rotation of the PC/FA Coefficients. In the above analysis of the PC/FA results, the importance of a given subspectrum was used as a criterion for its inclusion in the regression. However, this results in an incomplete transformation between $[FC_\zeta]$ and $\{\alpha_{ij}\}$, and the corresponding changes in the subspectra are inaccessible.

The PC/FA subspectra can be considered to be a special transformation of the experimental data as dictated by the orthogonality condition. In an effort to go beyond this restriction and, perhaps, to give the subspectra some added meaning, use of a Varimax transformation (Rummel, 1970; Davies, 1984) of our VCD PC/FA coefficients represents an attractive possibility. Here the transformation matrix is generated analytically by minimization of interdependencies among the PC/FA coefficients but without reference to any external factors. The final transformation provides us with a mechanism to represent, with rotated subspectra, these operations on the coefficients (see eq 7).

The result of Varimax transformation on the PC/FA is in Table IV for the coefficients and Figure 13 for the subspectra. Interestingly, some of the transformed subspectra, ${}^{rot}S_i$, have shapes that reflect those that are found for the clustered groups. ${}^{rot}S_2$ looks very similar to the cluster A spectrum, ${}^{rot}S_4$ to the cluster B result and ${}^{rot}S_6$ to the cluster C result (compare Figure 13 to Figure 10). In searching for simple relationships between the rotated coefficients, ${}^{rot}\alpha_{ij}$, and the FC values, we found significant correlations for only three of the six subspectra: the ${}^{rot}\alpha_{i2}$ coefficients relate to FC_α, FC_β, and FC_ρ with the highest correlation being to FC_α ($r = 0.93$); ${}^{rot}\alpha_{i4}$ relates most to FC_β; and ${}^{rot}\alpha_{i6}$ relates to FC_b ($r = 0.61$, significant at the 95% level). Since these correlations were not quantitatively better than the multiple-regression (eq 10) results, the respective equations are not presented.

This result indicates that rotation of the subspectra may offer us an alternate view for interpretation of the VCD data. Furthermore, these results offer support for our contention

Table V: Partial Least-Square Analysis of the VCD Spectra of the Training Set Proteins

	KS X-ray FC[a]					predicted FC[a]					deviations	
protein	α	β	b	t	ρ	α	β	b	t	ρ	Δ[b]	Δ[c]
trypsin	8	32	15	14	31	2	38	20	6	34	6	4
trypsin inhibitor	14	24	17	7	38	13	27	12	12	35	3	2
triosephosphate isom	43	17	8	7	24	56	−6	4	20	26	11	9
ribonuclease S	18	35	12	7	27	16	17	18	13	36	8	6
papain	23	16	18	8	33	18	16	19	7	40	3	3
myoglobin	77	0	2	10	11	64	5	−2	18	15	7	7
lysozyme	29	8	13	21	29	30	35	9	9	19	10	7
hemoglobin	67	0	4	9	20	73	−4	11	2	19	5	4
elastase	5	34	9	17	34	1	39	18	6	36	6	4
cytochrome *c*	26	4	22	14	34	9	35	35	−14	34	18	13
concanavalin A	0	40	20	9	30	5	20	17	23	35	9	6
chymotrypsin	6	33	10	16	35	33	5	4	33	26	17	16
carbonic anhydrase	7	27	18	12	36	11	35	15	9	30	5	4
std dev[b]						8	14	5	10	5		
std dev in % of range[b]						10	34	25	49	12		

[a] Abbreviations as in Table II. [b] Calculated as in Table II. [c] Average differences calculated without predictions for turn; the sum of the absolute values of differences was divided by 4.

Table VI: Partial Least-Square Analysis of the VCD Spectra of the Unknown Proteins

	predicted FC values and standard deviations[a]									
protein	α	σ_α	β	σ_β	b	σ_b	t	σ_t	ρ	σ_ρ
albumin	58	4	12	7	−1	2	15	5	17	2
casein	20	4	20	7	18	3	14	6	27	3
chymotrypsinogen	−15	2	29	4	23	2	15	4	47	2
lactoglobulin	33	4	40	15	14	4	−10	8	23	6
lactoferrin	22	1	20	3	13	2	11	3	33	1
ribonuclease A	24	3	26	7	20	2	3	5	28	2
thaumatin	−42	5	42	12	35	4	8	8	56	5

[a] σ_i is the standard deviation of the average as calculated from 13 estimated FC values by using different 12-member training sets. Secondary structure abbreviations as in Table II.

that, although dominant, only part of the VCD information is related to these secondary structure aspects of the proteins studied.

PLS Analysis of the VCD Spectra. To provide a basis of comparison for our PC/FA method in using VCD data to determine secondary structures, we have subjected the same data set to a partial least-square (PLS) analysis (Haaland & Thomas, 1988). PLS is conceptually similar to PC/FA, but it is constructed so as to optimize the $\{\alpha_{ij}\}$ coefficient set while PC/FA optimizes the $\{S_j\}$ subspectra. In other words, the PLS technique is designed to generate subspectra that are optimal for concentration predictions using a set of known calibration spectra (Wold, 1966). In our application, the concentrations are the protein FC values obtained from the analysis of the X-ray crystal structure with the implicit assumption that the fractions are not substantially changed in solution. A dimension of 30, optimized by trial and error, was used to constrain the fit in these PLS calculations.

To test the PLS method on our VCD data, twelve spectra from the KS subset in our training set of proteins were used to predict the FC values for the remaining one "known" sample left out. This process was repeated 13 times, and an estimate of error was made by comparison with the X-ray results for the left-out protein in each calculation. The results of this study for α, β, t, b, and ρ (other) are in Table V, including average error for each protein and standard deviations for each FC as was done above for the regression calculations. The standard deviations of the FC values from the crystal structure results are $\sigma = 8$ for α, $\sigma = 5$ for b, and $\sigma = 5$ for ρ, which are very close in quality to the similarly determined multiple-regression standard deviations (Table II) for these conformations. However, the error, $\sigma = 14$, for β is much worse, and, having $\sigma = 10$, the turn fraction can be viewed as undetermined since the standard deviation encompasses the dynamic range of FC_t. The range of the average error in FC_ζ values for individual proteins is from 3 to 18 (second last column in Table V). Eliminating the turn contribution changes this range slightly (last column). Then the worst predicted FC values are found for cytochrome *c*, with chymotrypsin and triosephosphate isomerase also significantly in error, but the full range from 2 to 16 is represented.

As for the regression results, all 13 of the above calibrating sets can be used for generation of FC vectors for the seven "unknown" protein spectra. This gives 13 estimates of FC values, which allows a statistical treatment and provides a measure of the sensitivity of the quantitative results to the changes in the calibration set selection. The average PLS predicted FC values for the unknown proteins and their standard deviations are in Table VI. This study is meant to be a test of the internal consistency of the method and is *not* presented as an ultimate prediction of FC values for unknown proteins. In fact, it is clear that some proteins do not give sensible predicted FC values, presumably for reasons related to poor clustering with the training set as noted above in the PC/FA discussion. However, the predictions of the regression equations (eq 10) and the PLS for these unknown proteins differ significantly in detail. For example, chymotrypsinogen is predicted to be moderate α-helix and low β-sheet in PLS, while it has no α-helix and relatively high β-sheet in the regression result, which, in turn, agrees with expectations (Freer et al., 1970). On the other hand, the β-lactoglobulin A $\mathrm{FC}\zeta$ values look more reasonable in Table VI for the PLS method than they did in Table III for the regression method. When just the σ values are surveyed, it is clear that the α-helical predictions are quite stable while the β-sheet predictions have large relative σ values. In contrast to the regression analyses, the most consistently predicted FC_ζ parameters turn out to be the "other".

Summary. Three approaches to defining a VCD spectra–structure correlation have been described. A direct consequence of the PC/FA analysis on the simple regression level was the discovery of the FC_α and FC_β correlation in the set of proteins we have studied. This simple spectra–structure correlation, based only on the α_{i2} is unique to VCD, being not so clear for UVCD or infrared spectra as will be shown in the next papers in this series. A similar simple regression in VCD was seen with α_{i1} for FC_ρ. Multiple-regression analyses gave similar results but with better precision and offered an opportunity to estimate the bend fraction, although at a lower level of confidence. The Varimax rotation of the PC/FA solution provided a means for reformulating the subspectra but did not yield an improved quantitative view of the VCD. Finally, the PLS approach provides a means of estimating all five FC_ζ components from VCD data, but the correlation of X-ray and PLS-predicted FC_ζ values is particularly poor for sheet and turn structures. The PLS method necessarily involves the assumption that the variation in spectra used in the analysis totally results from variation in the secondary structures of the training set proteins. In general, this is an undesirable assumption.

Thus while VCD is highly sensitive to protein structure, the quantitatively reliable information content of the amide I′ spectra is limited. Lest this sound overly negative, it should be noted that many UVCD and IR analyses have not sought to carry out a similar set of tests for reliability. That is the purpose of the following papers in this series. By way of preview, we can state that VCD will be shown to have sensitivities to secondary structural elements complementary to that of UVCD and to yield somewhat better quantitative results than UVCD when treated on precisely the same analytical level (Pancoska & Keiderling, 1991). On the other hand, UVCD and IR data can currently be measured with a significantly better signal-to-noise ratio than is currently possible with VCD. This implies that for the present time, amide I′ VCD can serve as a complementary tool to those longer established techniques.

In this light it is important to establish a correct procedure to use the methods reported herein for VCD-based structural analysis of unknown proteins. The spectrum of the unknown sample, recorded under controlled experimental conditions, should be combined with the training set for the PC/FA calculation. Subsequently, cluster analysis of the coefficients so obtained should be performed to determine qualitatively the degree of similarity of new sample VCD spectrum with those in the training set [see, for example, Yasui et al. (1990)]. Dependable analyses at this stage will rely on clear clustering of the unknown with group A, B, or C. The multiple-parameter regression equations (eq 10) can be used if the α_{ij} values for the training set are effectively the same in the new PC/FA as they are in Table I. In other words, if the PC/FA is not significantly affected by the added proteins, the equations given here remain useful. If the coefficients are different, new regression equations must be determined. In either case, the level of the probable error of the estimated FC parameters can be inferred from the similarity of the unknown protein spectrum with other entries in the training set.

Conclusion

This work had demonstrated some of the possibilities for quantitative use of VCD data in the amide I′ region for secondary structural analysis of proteins. A direct relationship has been found between the coefficients of the second subspectrum as determined from factor analysis and the α-helical and β-sheet content. Thus the most important changes in the VCD spectrum are correlated to the most important coherent elements of the secondary structure. That these two structures have the clearest correlation is presumably a result of the relatively larger dynamic range of their FC values.

We feel that the validity and limitations of factor analysis of VCD in the amide I′ region are best seen by cluster analysis results on the coefficients and the correspondence of those clusters to the CA results for the X-ray determined FC_ζ values. Both sets of raw data have been transformed via systematic, albeit arbitrary, mathematical methods into concise vector descriptors of the proteins that describe seemingly independent spaces. The CA algorithms serve to create boundaries in those spaces. On the level of extended similarity discrimination, i.e., for the three main clusters, the match of the X-ray and VCD topologies is exceptional.

Regression analyses, varimax rotation of the PC/FA solutions, and PLS analysis of the training set gave us quantitative estimates of the value of the VCD spectra–structure correlation. These, plus the dependence of the clustering result on CA algorithm, imply that FC_ζ determinations from VCD data are likely to be subject to a reasonable error, perhaps as much as ±15%. Improvements may come by expansion of the training set to include examples of more structural types and variants within a type. Optimization of the VCD-sampling conditions to better represent the crystallization conditions for proteins in the training set may also improve the fit between X-ray and VCD-determined FC_ζ values. Beyond that, improvements in the signal-to-noise ratio should lower the implied error limits.

It is natural to ask the question about the difference in information content between this newer approach using VCD and the more traditional UVCD approach to protein structure analysis. We have addressed this question in detail and present a systematic analysis of UVCD and VCD in the next paper in this series (Pancoska & Keiderling, 1991). A similar comparative effort has been made with respect to the somewhat revitalized IR analyses now available with Fourier deconvolution and will be presented in a future paper. Subsequent to the submission of this paper, two reports of the analysis of protein FTIR spectra, one using factor analysis (Lee et al., 1990) and the other partial least-squares analysis (Dousseau & Pezolet, 1990), have appeared and demonstrated the utility of these band-shape analysis techniques of spectra–structure correlation for such data.

The other question that arises is the information content of the various subspectra. We feel that other factors may influence those not strongly correlated to secondary structure and that new information may result if we can determine their dependencies. Current research is underway to probe this question further by perturbation of the protein structure and subsequent analysis of the resulting VCD. Also, one can take advantage of the multiple chromophore aspect of VCD to provide added probes of structure with perhaps alternate sensitivities.

At present we are on the threshold of VCD applications to biologically relevant systems. Only through careful systematic analysis will the proper role of this new technique be clarified. It is our opinion that the real value of VCD will lie as a complementary technique to the array of spectroscopic tools already available to the protein structural chemist.

Acknowledgments

Prepublication copies of papers from Profs. Mark Manning, Shelly Rackovsky, and Robert Woody are gratefully appreciated as were the comments, in particular, of one of the referees.

Registry No. Trypsin, 9002-07-7; trypsin inhibitor, 9035-81-8; triosephosphate isomerase, 9023-78-3; ribonuclease S, 65742-22-5; ribonuclease A, 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.

Bolotina, I. A., Chekhov, V. O., Lugauskas, V., & Ptitsyn, O. B. (1980) *Mol. Biol. (Moscow) 14*, 902–909.

Byler, D. M., & Susi, H. (1986) *Biopolymers 25*, 469–487.

Chang, C. T., Wu, C. S. C., & Yang, J. T. (1978) *Anal. Biochem. 91*, 13–31.

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., and Kim, S.-H. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 1406–1409.

Dousseau, F., & Pezolet, M. (1990) *Biochemistry 29*, 8771–8779.

Draper, N. R., & Smith, H. (1966) *Applied Regression Analysis*, John Wiley, New York.

Dukor, R. K., & Keiderling, T. A. (1989) in *Peptides 1988. Proceedings of the 20th European Peptide Symposium* (Jung, G., & Bayer, E., Eds.) pp 519–521, Walter de Gruyter, Berlin.

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

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

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

Haaland, D. M., & Thomas, E. V. (1988) *Anal. Chem. 60*, 1193–1208.

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

Jardine, N., & Sibson, R. (1968) *Comput. J. 9*, 177–184.

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

Keiderling, T. A. (1981) *Appl. Spectrosc. Rev. 17*, 189–226.

Keiderling, T. A. (1990) in *Practical Fourier Transform Infrared Spectroscopy. Industrial and Laboratory Chemical Analyses* (Ferraro, J. R., & Krishnan, K., Eds.) pp 203–284, Academic, San Diego.

Keiderling, T. A., Yasui, S. C., Dukor, R. K., & Yang, L. (1989) *Polym. Prepr. 30*, 423–424.

Lee, D. C., Haris, P. I., Chapman, D., & Mitchell, R. C. (1990) *Biochemistry 29*, 9185–9193.

Levitt, M., & Chothia, C. (1976) *Nature 261*, 552–558.

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

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

Malon, P., & Keiderling, T. A. (1988) *Appl. Spectrosc. 42*, 32–38.

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

Mantsch, H. H., Casal, H. L., & Jones, R. N. (1986) in *Spectroscopy* (Clark, R. J. H., & Hester, R. E., Eds.) Vol. 13, pp 1–46, John Wiley and Sons, London.

Massart, D. L., & Kaufman, L. (1983) in *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, Chemical Analysis*, Vol. 65, John Wiley and Sons, New York.

Nafie, L. A., Keiderling, T. A., & Stephens, P. J. (1976) *J. Am. Chem. Soc. 98*, 2715–2723.

Narayanan, U., Keiderling, T. A., Bonora, G. M., & Toniolo, C. (1985a) *Biopolymers 24*, 1257–1263.

Narayanan, U., Keiderling, T. A., Bonora, G. M., & Toniolo, C. (1985b) *J. Am. Chem. Soc. 108*, 2431–2437.

Pancoska, P., & Keiderling, T. A. (1991) *Biochemistry* (in press).

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.

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.

Provencher, S. W., & Glöckner, J. (1981) *Biochemistry 20*, 33–37.

Rackovsky, S. (1990) *Proteins 7*, 378–402.

Rackovsky, S., & Goldstein, D. A. (1988) *Proc. Natl. Acad. Sci. U.S.A. 85*, 777–781.

Richardson, J. S. (1981) *Adv. Prot. Chem. 34*, 167–339.

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

Rummel, R. J. (1970) *Applied Factor Analysis*, Northwestern University Press, Evanston, IL.

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

Sen, A. C., & Keiderling, T. A. (1984b) *Biopolymers 23*, 1533–1546.

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.

Su, C. N., Heintz, V. J., & Keiderling, T. A. (1981) *Chem. Phys. Lett. 73*, 157–159.

Wold, H. (1966) in *Multivariate Analysis* (Krishnaiah, P. R., Ed.) Academic Press, New York.

Woody, R. W. (1977) *J. Polym. Sci., Macromol. Rev. 12*, 181–320.

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

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., & Keiderling, T. A., & Sisido, M. (1987a) *Macromolecules 20*, 2403–2406.

Yasui, S. C., & Keiderling, T. A., & Katakai, R. (1987b) *Biopolymers 26*, 1407–1412.

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.