Comparison of and limits of accuracy for statistical analyses of vibrational and electronic circular dichroism spectra in terms of correlations to and predictions of protein second…

ARTICLE *in* PROTEIN SCIENCE · JULY 1995

Impact Factor: 2.85 · DOI: 10.1002/pro.5560040713 · Source: PubMed

Protein Science (1995), *4*:1384–1401. Cambridge University Press. Printed in the USA.

Comparison of and limits of accuracy for statistical analyses of vibrational and electronic circular dichroism spectra in terms of correlations to and predictions of protein secondary structure

PETR PANCOSKA,[1,2] EDUARD BITTO,[2] VIT JANOTA,[2] MARIE URBANOVA,[1,3] VIJAI P. GUPTA,[1] AND TIMOTHY A. KEIDERLING[1]

[1] Department of Chemistry, University of Illinois at Chicago, Chicago, Illinois 60607-7061
[2] Department of Chemical Physics, Charles University, Ke Karlovu 3, 121 16 Prague 2, Czech Republic
[3] Institute of Chemical Technology, Department of Physics and Measurements, Technicka 5, 166 23 Prague 6, Czech Republic

(RECEIVED January 27, 1995; ACCEPTED April 18, 1995)

Abstract

This work provides a systematic comparison of vibrational CD (VCD) and electronic CD (ECD) methods for spectral prediction of secondary structure. The VCD and ECD data are simplified to a small set of spectral parameters using the principal component method of factor analysis (PC/FA). Regression fits of these parameters are made to the X-ray-determined fractional components (*FC*) of secondary structure. Predictive capability is determined by computing structures for proteins sequentially left out of the regression. All possible combinations of PC/FA spectral parameters (coefficients) were used to form a full set of restricted multiple regressions with the *FC* values, both independently for each spectral data set as well as for the two VCD sets and all the data grouped together. The complete search over all possible combinations of spectral parameters for different types of spectral data is a new feature of this study, and the focus on prediction is the strength of this approach. The PC/FA method was found to be stable in detail to expansion of the training set. Coupling amide II to amide I′ parameters reduced the standard deviations of the VCD regression relationships, and combining VCD and ECD data led to the best fits. Prediction results had a minimum error when dependent on relatively few spectral coefficients. Such a limited dependence on spectral variation is the key finding of this work, which has ramifications for previous studies as well as suggests future directions for spectral analysis of structure. The best ECD prediction for helix and sheet uses only one parameter, the coefficient of the first subspectrum. With VCD, the best predictions sample coefficients of both the amide I′ and II bands, but error is optimized using only a few coefficients. In this respect, ECD is more accurate than VCD for α-helix, and the combined VCD (amide I′+II) predicts the β-sheet component better than does ECD. Combining VCD and ECD data sets yields exceptionally good predictions by utilizing the strengths of each. However, the residual error, its distribution, and, most importantly, the lack of dependence of the method on many of the significant components derived from the spectra leads to the conclusion that the heterogeneity of protein structure is a fundamental limitation to the use of such spectral analysis methods. The underutilization of these data for prediction of secondary structure suggests spectral data could predict a more detailed descriptor.

Keywords: circular dichroism; secondary structure prediction; spectra-structure correlation; vibrational circular dichroism

Optical spectroscopic studies are particularly important for fast, universal, qualitative estimation of average secondary structure in proteins. The ability of various such techniques to sense small changes in conformation is well established. Many attempts have been made to quantify the measured spectra in terms of global conformational characteristics, such as the fractional components (*FC*) of the secondary structure that are typically drawn from X-ray crystal structure studies for comparison to spectra (Greenfield & Fasman, 1969; Chen & Yang, 1971; Chen et al., 1972; Chang et al., 1978; Siegel et al., 1980; Hennessey & Johnson, 1981; Provencher & Glöckner, 1981; Byler & Susi, 1986; Compton & Johnson, 1986; Mantsch et al., 1986; Manavalan

Reprint requests to: Timothy A. Keiderling, Department of Chemistry, m/c 111, University of Illinois at Chicago, 845 W. Taylor Street, Chicago, Illinois 60607-7061; e-mail: tak@uic.edu.

& Johnson, 1987; Dousseau & Pezolet, 1990; Lee et al., 1990; van Stokkum et al., 1990; Venyaminov & Kalnin, 1990; Pancoska et al., 1991; Pancoska & Keiderling, 1991; Perczel et al., 1991; Sarver & Krueger, 1991a, 1991b; Toumadje et al., 1992; Pribic et al., 1993; Sreerama & Woody, 1993, 1994). However, despite utilizing different optical techniques and analytical methods of increasing mathematical complexity, the quantitative accuracy of such determinations never matches the qualitative sensitivity of the methods to conformational change (Keiderling et al., 1995).

Many previous spectral-structure correlation studies have focused on electronic circular dichroism (ECD) due to its high sensitivity to structural variation and ease of measurement (for reviews see Johnson, 1985, 1988, 1990; Yang et al., 1986; Manning, 1989; Sreerama & Woody, 1994). Most approaches for the interpretation of ECD now use band shape or pattern recognition analyses to derive structural correlations. A number of protein structural studies have utilized vibrational spectroscopic techniques, mostly Fourier transform (FT) IR but also Raman, which have a resolution and signal-to-noise ratio (S/N) advantage (for FTIR) over ECD (Byler & Susi, 1986; Mantsch et al., 1986; Williams, 1986; Berjot et al., 1987; Bussian & Sander, 1989; Dousseau & Pezolet, 1990; Lee et al., 1990; Venyaminov & Kalnin, 1990; Sarver & Krueger, 1991a, 1991b; Pribic et al., 1993). The analyses of these latter spectroscopies in terms of structure has primarily depended on assignment of transition frequencies that we have earlier shown to be inherently ambiguous (Pancoska et al., 1993) and potentially misleading if used in isolation without reference to data from other techniques (Dukor et al., 1992). Some more recent studies have utilized band-shape methods for FTIR spectra (Dousseau & Pezolet, 1990; Lee et al., 1990; Sarver & Krueger, 1991a, 1991b; Pribic et al., 1993). Somewhat between these two approaches is the use of vibrational CD (VCD), a hybrid of CD and IR, to predict protein conformation (Keiderling, 1993, 1994; Keiderling & Pancoska, 1993). VCD, like IR or Raman, can be used to sense several different spectrally resolved transitions involving different localized vibrations of the molecule, which, like ECD, have a distinct dependence on molecular stereochemistry. VCD spectra will have shapes that represent relatively local interactions of the amides, but these shapes for local structural components will be distributed over the same frequencies as in IR spectra. The overall VCD spectral information derives from both the vibrational frequencies of the component transitions and the band shapes resulting from their interaction. Hence, VCD has the potential to yield spectral-structure correlation on two levels.

Previously, we have demonstrated that VCD measured for the amide I′ transition (primarily amide C=O stretch for N-deuterated proteins) has a qualitatively enhanced sensitivity to structure as compared to IR and ECD (Pancoska et al., 1989; Keiderling et al., 1995). This sensitivity arises from both the relatively higher resolution and conformational sign dependence of the VCD as compared to ECD and IR, respectively, and the relatively short-range interaction distances that characterize VCD (Yasui et al., 1986b; Dukor & Keiderling, 1991; Keiderling & Pancoska, 1993; Yoder et al., 1995). A quantitative scheme (Pancoska et al., 1991) to derive *FC* values from the VCD spectral band shapes was developed based on the principal component method of factor analysis (PC/FA) (Malinowski & Howery, 1980). Similar methods have been used in previous ECD data analyses that build on the seminal work of Johnson's group (Hennessey & Johnson, 1981; Johnson, 1985) and have been extended to analysis of FTIR spectra of proteins (Lee et al., 1990). Systematic comparison of amide I′ VCD for proteins in D_2O with those of an exactly parallel analysis of ECD for the same small set of proteins indicated that, indeed, VCD also had a quantitative advantage for determination of at least the β-sheet fraction, whereas ECD was better for the α-helix (Pancoska & Keiderling, 1991). Similar deductions were derived from comparative analyses of ECD and FTIR spectra (Pribic et al., 1993).

It is now feasible to measure VCD for proteins in H_2O as well (Gupta & Keiderling, 1992; Baumruk & Keiderling, 1993), particularly for the amide II (N-H bending and C-N stretch), which is accessible for all soluble globular proteins. Amide II data are also characteristic of protein secondary structure but are not as qualitatively differentiated as are the amide I′ VCD of proteins in D_2O. Because one of the major features of the VCD technique is its ability to probe several resolved transitions, it is important to see if coupling the amide II data to that of the amide I′ will improve the quantitative spectra-structure analysis. Such a coupling has proven beneficial in previous protein IR analyses (Lee et al., 1990). Although amide I VCD (N-H protonated) is also measurable in H_2O, the set of proteins adaptable to the higher concentrations required is more limited.

In this paper we present a systematic comparison of the correlation of protein secondary structure with amide I′ and II VCD data and with ECD spectra. The overall goal of this work is to probe the limits of CD spectra, vibrational and electronic, in *predicting* the average fractional secondary structure of proteins. Furthermore, it is a well-known aspect of spectral analyses of protein structure that correlations are sometimes dependent on the training set chosen (Yang et al., 1986; Manavalan & Johnson, 1987; van Stokkum et al., 1990). It is important to demonstrate that our analyses are stable upon inclusion of both more proteins and proteins of various types in the training set. The effects of expansion of the protein data base in terms of spectral transitions used and numbers of proteins with known crystal structures included in the training set are both examined here.[4]

Although correlation of observables with structure can be intellectually stimulating, the real justification for development of empirical optical spectra-structure relationships is to *predict* the structure of proteins for which other structural reference data are unknown. Therefore, the focus of this paper is a systematic comparison of the relative predictive ability of multicomponent regression models for proteins not included in the regression development. As employed in previous analyses (Hennessey & Johnson, 1981; Pancoska et al., 1991; Pribic et al., 1993), sets of data with one protein systematically left out provide the basis for testing prediction. An important aspect of our work is the evaluation of a series of *restricted regressions* that demonstrate that the optimal predictability of these analyses is obtained with only selected components of the spectral data for each protein. In contrast to the variable selection method (VSM; Manavalan & Johnson, 1987), no proteins are left out of the regression analysis except for that one selected for prediction; and, in contrast to the spectral component fitting approach (Pribic et al., 1993), all proteins are included in the PC/FA. Our meth-

[4] Due to sampling constraints, VCD for the amide I in H_2O cannot be measured over the full set of proteins studied here. Therefore, the analysis of combined H_2O amide I and II VCD and FTIR data will be presented separately (V. Baumruk, P. Pancoska, & T.A. Keiderling, manuscript in prep.).

ods are an alternate, complementary approach to the VSM analysis whereby we select those spectral elements that best predict secondary structure for all the proteins rather than selecting those proteins for which the best analysis of a particular unknown results.

Restricting the spectral components led to some surprising predictive results. For ECD the optimal prediction algorithm uses only one spectral component. The full analysis of the impact on prediction of all possible combinations of independent spectral components as done in this study is generally new. The analysis of separate and combined ECD and VCD data provides insight into the full potential of using coupled optical spectral data sets to probe protein structure. Together these aspects of our study lead to interpretations of the structural prediction limitations of optical spectral data that are the central issue of the paper.

In summary, the methodology employed in this paper is not to develop a "best method" with reduction of error at any cost. Rather, this paper focuses on the use of a systematic comparison of methods to understand the sources of error in optical spectral analysis of protein structure. Consequently, we do not undertake to compare error limits with other computational methods of spectral analysis. Nonetheless, the combination of ECD with VCD data is systematically shown to have superior predictive capabilities to either alone, such that each technique compensates for the weaknesses of the other, resulting in a more rugged method of spectra-structure correlation.

Materials and methods

Samples

The proteins chosen for our expanded training set of 23 known protein structures are summarized in Table 1, together with their source, species origin, Protein Data Bank (PDB) codes, and secondary structure composition in terms of *FC* values as derived from the Kabsch and Sander (1983) (KS) and Levitt and Greer (1977) (LG) secondary structure analysis algorithms. This protein subset (which we refer to as the training set) is the source of the reference structural information for our quantitative analyses. Following our earlier study, we also include five extra proteins in the data set that are not in the PDB and consequently have no structural input to the spectral analysis as an independent, qualitative monitor for the behavior of the results. These five are treated in the same way one should treat real unknowns, which in our approach are processed by the PC/FA together with the training set proteins to generate subspectral coefficients for all in a consistent manner. Thus, the subsequent regression steps are based on a unified set of parameters. This approach avoids some potential problems inherent in other variants of PC/FA where the coefficients for unknowns are determined by fitting spectra to the subspectra obtained from only the training set.

We made an effort to select the best match possible between crystal forms in the PDB and the actual proteins used for our

Table 1. *Training set proteins, sources, and X-ray-determined secondary structures*

Protein name	Species	Source	PDB file	KS helix	KS sheet	KS turn	KS bend	KS other	LG helix	LG sheet	LG LT	LG RT	LG other
Alcohol dehydrogenase	Horse liver	Fluka 05648	4ADH	24.9	20.6	14.7	13.6	26.2	31.2	41.3	12.0	5.6	9.9
Carbonic anhydrase	Bovine erythrocytes	Sigma C-7500	1CA2	16.0	28.9	12.9	15.2	27.0	19.1	42.8	15.2	9.3	13.6
α-Chymotrypsinogen A	Bovine pancreas	Sigma C-4879	2CGA	14.3	32.2	14.3	12.7	26.5	14.6	45.1	14.6	8.5	17.1
α-Chymotrypsin type II	Bovine pancreas	Sigma C-4129	5CHA	11.8	32.1	11.4	14.4	30.4	10.6	50.4	16.1	8.5	14.4
Concanavalin A	Jack bean	Sigma C-2010	3CNA	0.0	40.5	9.3	19.8	30.4	2.5	63.5	16.8	7.1	10.1
Cytochrome *c*	Tuna	Sigma C-2011	1CYT	42.7	0.0	15.5	8.7	33.0	46.2	5.8	15.4	17.3	15.3
Tosyl elastase	Pig porcine pancreas	Sigma E-0127	3EST	10.8	34.2	17.1	10.4	27.5	11.6	45.2	17.0	9.5	16.6
Glutathione reductase	Wheat germ	Sigma G-6004	2GRS	29.3	18.7	10.4	19.3	22.3	39.0	33.8	10.6	4.8	11.9
Hemoglobin	Human	Sigma H-7379	1HCO	62.7	0.0	18.8	6.6	11.9	81.7	0.0	12.0	0.0	6.3
λ-Immunoglobulin	Human	Fluka 56834	1REI	2.8	47.7	14.0	11.2	24.3	0.0	67.6	8.3	14.8	9.3
Lactate dehydrogenase	Rabbit	Calbio. 427217	4LDH	36.8	11.3	14.3	13.1	24.6	45.8	23.9	10.3	6.7	13.3
Lysozyme	Hen egg white	Sigma L-6876	7LYZ	38.8	7.8	20.9	16.3	16.3	50.0	12.3	12.3	13.1	12.3
Myoglobin	Horse heart	Sigma M-1882	1MBN	77.1	0.0	9.8	1.9	11.1	90.3	0.0	3.9	0.0	5.8
Papain	Papaya latex	Sigma P-4762	8PAP	24.5	16.5	8.5	20.3	30.2	28.2	24.9	15.5	12.2	19.3
Rhodanese	Bovine liver	Sigma R-4125	1RHD	29.7	10.9	16.4	13.3	29.7	41.2	20.1	11.9	9.2	17.7
Ribonuclease A	Bovine	Sigma R-5125	1RN3	21.0	34.7	11.3	14.5	18.6	29.6	44.8	10.4	6.4	8.8
Ribonuclease S	Bovine pancreas	Sigma R-6000	1RNS	20.8	35.2	7.2	14.4	22.4	30.4	42.4	4.8	11.2	11.2
Subtilisin BPN′	Bacterial	Sigma P-8038	1SBT	30.2	17.8	15.3	12.0	24.7	37.0	31.5	14.1	6.5	10.9
Superoxide dismutase	Bovine erythrocytes	Fluka 86200	2SOD	2.0	38.4	14.6	20.5	24.5	3.3	59.2	15.1	9.2	13.2
Thermolysin	Bacterial	Sigma P-1512	2TLN	33.5	15.2	14.6	16.5	20.3	43.8	28.7	10.1	5.1	12.3
Triose phosphate isomerase	Yeast	Sigma T-2507	1TIM	45.8	17.0	7.3	8.9	21.1	54.4	22.6	6.85	4.8	11.3
Trypsin inhibitor	Soybean	Sigma T-9003	3PTI	20.7	24.1	6.9	19.0	29.3	25.4	40.7	8.5	10.2	15.3
Trypsin	Bovine pancreas	Sigma T-8253	3PTN	9.9	32.3	14.8	17.9	25.1	11.2	52.7	17.0	8.9	10.3
Albumin	Bovine serum	Sigma A-0281	–										
Lactoferrin	Human milk	Sigma L-5665	–										
Casein	Bovine milk	Sigma C-7891	–										
β-Lactoglobulin A	Bovine milk	Sigma L-7880	–										
Thaumatin	Bacterial	Sigma T-7638	–										

spectral determinations as regards species and general conditions, which in some cases required recalculation of the KS parameters from revised or alternate crystal structure data. It is important to note that the original KS paper assumed that knowledge of the characteristics of only a selected example of a given protein type was sufficient. We have shown that the small structural variations between species are sensed by ECD and VCD, and, consequently, mismatches would be reflected in the quality of fit obtained with our methods. Thus, it is important, as far as is possible, to compute the reference structural parameters for the specific proteins used to obtain the spectral training set.

All proteins were used as obtained without further purification. For VCD and IR studies of the amide I′ band, these proteins were typically exchanged in D_2O and lyophilized three times before small volume (~20 μL) solutions of each sample were prepared to a concentration of ~50 mg/mL in D_2O (Aldrich). For the amide II, proteins were directly dissolved in a similarly small amount of double-distilled H_2O at ~200 mg/mL. Solutions to be studied were placed in cells consisting of two CaF_2 windows separated by a 0.025- or 0.015-mm Teflon spacer for measurement of the amide I′ or amide II spectra, respectively. For ECD, more dilute solutions, ~1 mg/mL, were prepared in doubly distilled H_2O and placed in strain-free quartz cells (Precision Cells) of 0.05 or 0.1 cm pathlength (allowing measurement to ~180 nm).

Spectroscopy

For purposes of this work, an important factor in the evaluation of the use of spectroscopic data for structural correlation is self-consistency. We did not want any of our residual errors to be attributable to different instrumental or measurement conditions. Consequently, all spectra were measured in our laboratory as described below and none were taken from the literature. To be sure that we were not including any gross error due to protein preparation or instrument failure, extensive checks were made to assure that our measured spectra were consistent with whatever data were available in the literature. A separate discussion of effects of experimental error in the VCD experiments is available (Pancoska et al., 1995).

The amide I′ and II VCD and IR absorbance spectra at room temperature were measured on the UIC dispersive VCD instrument, which has been thoroughly described elsewhere (Keiderling, 1981, 1990). Details of our sampling and data collection methods for amide I′ (Pancoska et al., 1989, 1991) and II (Gupta & Keiderling, 1992) VCD have been previously reported. In summary, the VCD were obtained with ~10 cm^{-1} resolution as the average of four scans. VCD spectra were calibrated with the usual methods (Keiderling, 1981, 1990) and were baseline corrected with identical scans of poly-D,L-lysine as before. Although the amide I′ spectra were used as measured, the lower S/N amide II VCD spectra were smoothed using the FT method in the SpectraCalc package (Galactic Industries, Nashua, New Hampshire). Amide I′ spectra were normalized to the absorbance maximum in this region (i.e., scaled so that $A = 1.0$). This normalization is subject to some error if the spectra do not reach baseline inside the spectral range studied. For the amide II, due to overlapping band contours, it was necessary to develop an alternative approach using band-shape fitting. The absorbance band shape over the 1,800–1,370-cm^{-1} range was fit with several Gaussian band components (typically 6–9). Of these, we selected those having a frequency within the 1,580–1,510-cm^{-1} interval. The areas of these components were summed and arbitrarily divided by 30 to obtain a normalization of the VCD in units comparable to those used in the amide I′ region. These normalization schemes are potentially the source of some error (Bitto, 1993).

ECD spectra were measured on a JASCO J-600 spectropolarimeter at room temperature and are averages of up to 12 scans to get a smooth overall high S/N representation of the spectrum. Identical scans of pure solvent in the same cell were used for a baseline. The protein concentrations of the solutions were determined from known extinction coefficients or absorbance at 190 nm (assuming a molar absorptivity of 10,000 M^{-1} cm^{-1} $residue^{-1}$) (Johnson, 1988, 1990). The resulting ECD intensities were critically checked against available published data (Brahms & Brahms, 1980; Hennessey & Johnson, 1981). When spectra for the same protein form and species were available, our data were found to be fully consistent with the published spectra in both shape and intensity.

FTIR absorption spectra of all the proteins studied were remeasured over the entire spectral region on a Digilab FTS-60 FTIR spectrometer using a TGS detector, 4 cm^{-1} resolution, and an average of 1,024 scans. Spectra were corrected for water interference by subtraction of separately collected water vapor and solvent absorbance spectra. These spectra were used for the normalization of the amide II VCD spectra as described above and for checking the frequency consistency of the dispersive data. Frequency errors in our dispersive data were computationally corrected by shifting the dispersive absorbance spectrum to overlay the FTIR band shapes and using the same correction shift for the VCD.

Sample spectra obtained with each of the techniques for the proteins studied have been presented in our previous papers (Pancoska et al., 1991; Pancoska & Keiderling, 1991; Gupta & Keiderling, 1992). Data for any particular proteins, or for the entire set, as were used in this study, all in the SpectraCalc format that was used as input for the computations or in ASCII tables, are available in digital form from the authors and in graphical form as Figures S1–S21 in the Electronic Appendix.

Calculational methods

The description of the principal component method of factor analysis (Pancoska et al., 1979; Malinowski & Howery, 1980) as applied to the decomposition of the amide I′ and amide II VCD spectra into a linear combination of orthogonal subspectra can be found in our previous paper (Pancoska et al., 1991). Here we summarize only those aspects that are new or characteristic for this study.

The first steps of the calculation were performed separately for amide I′ and amide II VCD and for the ECD spectra. The matrix of the correlation coefficients, which are measures of the spectral similarity, was calculated numerically by a spline function integration algorithm as the overlap integrals of all pairs of spectra in the set being analyzed (VCD I′, II, or ECD). The resulting matrix was diagonalized using the TQL and TRED2 algorithms (Press, 1992). The matrix of the resulting eigenvectors transforms the experimental data matrix into a set of orthogonal subspectra. The transpose of the eigenvector matrix represents the set of coefficients necessary for regeneration of

any experimental spectrum in the subspectral basis. By ordering the subspectra according to their associated decreasing eigenvalue, the most significant subspectra can be identified, a linear combination of which can be used to reproduce the experimental spectra to a level commensurate with their experimental reliability. This provides a simplification represented by:

$$[\theta_j] = [\phi_i] \cdot [C_{ij}], \quad (1)$$

where $[\theta_j]$ is a matrix whose columns are all the n equidistantly digitized experimental spectra being analyzed, $[\phi_i]$ is a matrix whose columns are the p significant subspectra that are retained for the analysis, and $[C_{ij}]$ is an $n \times p$ matrix of coefficients for the contribution of the ith subspectrum to the jth experimental protein spectrum. Thus, each continuously varying spectrum is transformed into a compact numerical form representable as a vector of a few (p) coefficients. Although p theoretically could be taken to be as large as n, this number is considerably reduced by focusing only on the significant subspectra. For each spectral data set, p was quantitatively determined to be large enough so that at least 98% of the variance in the data set was included in the spectra, θ_j, as reconstructed from the p subspectra, ϕ_i. On a qualitative basis, the residuals due to subspectra $i > p$ yield no identifiable band shapes (other than noise).

The structural data used for correlation to the spectra are derived from X-ray crystal structures as found in the PDB. Atomic coordinates from PDB files were input into the KS DSSP program (Kabsch & Sander, 1983; Bitto, 1993) to obtain the structural parameters, FC_ζ values, used for spectral correlation. From the standard output of this program, α-helical and 3_{10}-helical structures were grouped into one class, denoted as helical (H), and all β-sheet structures form a single class (S). For testing the prediction capability for minor components, we further selected bends (B, no hydrogen bond, Cα-Cα dihedral angle <74°) and turns (T, roughly helical dihedral angles for <4 amino acids). All other amino acid residues were assigned to a class we prefer to refer to as "other" (C) because it lacks any common descriptor. For *FC* values corresponding to the Levitt and Greer (1977) definition we wrote our own program following their algorithm and chose the categories: helix (H), sheet (S), left-turn (LT), right-turn (RT), and "other" (C). It should be noted that C is simply what is left over and is *not* equivalent to a "random coil" though it may have spectral characteristics in common with that structure (Yasui & Keiderling, 1986; Dukor & Keiderling, 1991; Baumruk et al., 1994).

For the regression tests, the p_I coefficients obtained from the amide I′ VCD analysis, the p_{II} ones from the amide II, and the p_E coefficients from the ECD spectra were used first separately, to allow comparison of the fits with each data set, and then combined into sets of ($p_I + p_{II}$) and ($p_I + p_{II} + p_E$) coefficient vectors for combined analyses. Optimal restricted multiple linear regression relationships at each level were sought by testing all possible combinations of k spectral coefficients, where $k = 1, 2, \ldots, p$, for the significance of the regression obtained for each of the FC_ζ values, where ζ represents α-helix, β-sheet, bend, turn, and "other" crystallographic secondary structures. To our knowledge, restricting the regression to a selection of k coefficients and then testing the dependence of the quality of both the regression and prediction (see below) for all possible combinations of coefficients from the set of p possibilities is not a method used before in spectral–structure analyses. It is the key to our gaining insight into the sensitivity of the method to specific spectral components. For the purposes of this comparative study, the standard multiple correlation coefficient *rr* is calculated (Sharaf et al., 1986) and used to rank these sets of regressions according to the goodness of fit. For each set of coefficients of a given size, k, the combination providing the highest *rr* is retained and used for further calculations.

Although such a fitting exercise of known structural data is definitely not our ultimate goal in this study, it is a necessary first step in the development of a prediction algorithm. As such, it was required that we systematically test the statistical significance of the corresponding multiple linear regression model. We have calculated the Z-value for a given regression as

$$Z = [(n - k - 1)/k] \cdot [rr^2/(1 - rr^2)], \quad (2)$$

where n is the number of proteins in the training set and k is the number of subspectral coefficients considered in the regression. This value was then compared with the critical values of the corresponding F-distribution (Sharaf et al., 1986) to determine significance at the 99% confidence level, unless stated differently. To get a feel for which coefficients had the major impact on the fit, we plotted the *rr* values vs. k. Qualitatively, we view the most significant regression as that one containing the largest combination of coefficients, which still leads to a substantial improvement in *rr*. Such a restricted regression has more physical consequence as we will demonstrate in our prediction tests. These results will be denoted as *fit* results, as they demonstrate the capability of spectra as encoded into C_{ij} values to be transformed into the *FC* values of each secondary structure type.

By contrast, to test the predictive capability of the derived relationships between spectral features into the *FC* values, we repeated the above procedures 23 times, each time eliminating from the training set 1 of the 23 proteins having a known structure. The k-coefficient regression equations of highest correlation coefficient, *rr*, for $k = 1$ to p for each reduced, 22-member training set were then used to predict the *FC* values for the protein left out. After each cycle the predicted *FC* values are compared to the actual values in Table 1, and the average deviation is used to characterize the *prediction* capability of our method depending on the data set being tested.[5]

Error parameters

We use the following error characteristics of the fit and prediction to compare spectrally determined, FC_{ij}^s, and X-ray derived, FC_{ij}^x, secondary structure parameters for protein i and structural type j:

1. Average error, δ_i ($i = 1$ to 23), of *FC* values of individual proteins over n_s ($n_s = 5$, normally) types of secondary structures considered in the analysis:

$$\delta_i = (1/n_s) \cdot \sum_j |FC_{ij}^s - FC_{ij}^x|. \quad (3)$$

[5] To give some idea of the scope of this process, for the combined amide I′ + II and ECD analysis, this required 470,810 regression calculations, resulting from the same procedure of 94,162 determinations, being independently repeated for all five secondary structure types considered here.

2. Standard deviation of *FC* values for given type of secondary structure, calculated from deviations of spectral from X-ray *FC* values for all N (normally = 23) proteins from the training set:

$$\sigma_j = \left\{ \left[\sum_i (FC_{ij}^s - FC_{ij}^x) \right]^2 \Big/ (N-1) \right\}^{1/2}. \quad (4)$$

It should be noted that this formula is used to evaluate standard deviation for *fit* as well as for *prediction*.

3. Relative standard deviation in the percentage of the dynamic range of *FC* values for the respective secondary structure type:

$$\sigma_j^{rel} = (100\sigma_j)/(FC_{max,j}^x - FC_{min,j}^x). \quad (5)$$

4. We also monitored the non-negativity of the predicted *FC* values and the difference of their sum from 100%. Nevertheless, these "natural" conditions were not directly included in our calculations. Therefore, they act as independent measures of the reliability of the resultant regression equations much as used in other spectral-structure fitting routines (Manavalan & Johnson, 1987). Due to the nature of our algorithm, the sum of predicted *FC* values is composed of five fully independent values because the individual secondary structure types are treated separately with regard to development of a regression relation and testing of its statistical significance.

We have shown recently that the *FC* values for globular proteins as obtained from the PDB are not mutually independent (Pancoska et al., 1992). This fact can cause substantial confusion in the interpretation of the reliability of spectral-structure correlations and, to our knowledge, was not considered explicitly in any previous spectral-structure analyses. Therefore, for comparative purposes, we have also calculated the predicted values for the nonhelical secondary structure types, FC_ξ^i, as derived from the X-ray-based helix fractions for each protein, i, using those previously published relationships that were derived from a neural network analysis of a large set of crystal structure data for globular proteins.

Results

Factor analysis results

PC/FA decompositions were done independently for the amide I′ and amide II VCD and ECD spectra of the 28 proteins, 23 from our training set along with the 5 "unknowns." From the criteria described above and by comparison of reconstructed spectra to the experimental spectra, we retained six orthogonal components for both the amide I′ and amide II VCD spectral regions and five for the ECD spectra to reproduce the experimental spectra. Six is the same number of significant subspectra found in our previous study of the amide I′ VCD for a smaller number (20 total) of proteins (Pancoska et al., 1991). The ECD result is in agreement both with our previous (Pancoska & Keiderling, 1991) and other ECD spectral analyses (Hennessey & Johnson, 1981; van Stokkum et al., 1990) using similar projection techniques. The significance of these numbers

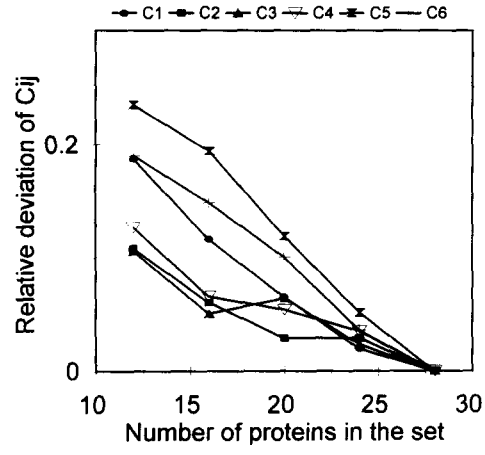

Fig. 1. Relative average deviation of the subspectral coefficients for the amide I′ VCD obtained with various protein sets of reduced size as compared to those obtained for the full set of 28 proteins.

of coefficients and their effects on the analyses will be addressed in the next section.

The band shapes of both the ECD and VCD amide I′ subspectra found in this analysis are not significantly different from those calculated on the reduced protein set in our previous study, implying that the decomposition method is relatively stable with respect to the data set. Actually, having such behavior is dependent on the subset of proteins reflecting the same distribution of protein types as in the whole set. If a subset is biased toward one type, the subspectra will be different, resulting in an emphasis of different spectral features in each. To test the generality of our set, we subdivided the 28 proteins into four clusters that gave rise to similar amide I′ VCD spectra as found using the Lance-Williams flexible cluster analysis algorithm (Einsight package). Then four subsets of protein spectra containing 12, 16, 20, and 24 proteins were assembled that contained examples from each of the clusters in a roughly equal proportion. These were subjected to an independent factor analysis and the coefficients for each protein were compared to the results found with the entire set of 28 proteins. A steady drop in the difference from the final coefficients for the full set was found with the largest deviations, as might be expected, for the fifth and sixth coefficients. These trends for the amide I′ data are illustrated in Figure 1, where relative differences are plotted versus the number of proteins in the set. The final changes from 24 to 28 proteins involve <5% modifications in the coefficients. These results give us confidence that addition of more proteins to the set will not substantially change our results. Of course, addition of proteins to the set that do not reflect the distribution of protein conformations we have selected may have significant effects. To obtain a "general" set, in the complete set of 28 proteins, we tried to include the major types of folds (Levitt & Chothia, 1976) and to overlap the proteins used as bases for other, independent spectral analyses.

The subspectral band shapes obtained by the PC/FA method reflect the most common elements of the 28 experimental spectra and their most significant variances. Because the subspectra are stable with shapes reflecting those presented earlier (Pancoska et al., 1991), they are provided graphically only in

Table 2. *Standard deviations and correlation coefficients for regression fits of spectral coefficients to secondary structure types*

	Helix		Sheet		Bend		Turn		Other	
	σ	rr	σ	rr	σ	rr	σ	rr	σ	rr
AI′ restr.[a]	9.9	0.850	7.1	0.857	3.6	0.283	3.7	0.597	4.5	0.621
Full[b]	9.7	0.858	6.8	0.866	3.4	0.399	3.6	0.618	4.3	0.670
AII restr.	11.1	0.808	7.9	0.814	3.3	0.485	3.4	0.688	4.0	0.722
Full	10.4	0.835	7.6	0.832	3.1	0.585	3.3	0.699	3.7	0.771
ECD restr.	6.2	0.943	8.3	0.793	3.4	0.424	3.5	0.657	3.3	0.815
Full	5.9	0.949	8.0	0.812	3.2	0.508	3.4	0.662	3.2	0.823
AI′+AII restr.	7.5	0.918	5.9	0.903	3.3	0.485	2.6	0.829	2.5	0.903
Full	7.1	0.970	5.5	0.967	2.6	0.626	2.5	0.862	1.7	0.941
AI′+AII+ECD restr.	4.0	0.977	4.4	0.948	2.2	0.871	2.3	0.874	1.5	0.965
Full	3.1	0.993	3.9	0.976	2.1	0.946	2.1	0.923	1.4	0.972
σ_{rel}[c]	4.0		8.1		14.9		11.1		6.4	

[a] Best fit parameters for the restricted regression fit giving the best prediction (see Table 3 for number of coefficients used). AI′, amide I′; AII, amide II.
[b] Fit parameters using complete sets of subspectral coefficients, i.e., 6 for amide I′ and amide II, 5 for ECD, 12 for amide I′+II, and 17 for amide I′+II+ECD calculation.
[c] Relative standard deviation for the full AI′+AII+ECD calculation.

the Electronic Appendix as Figures S22–S24. Briefly, the first amide I′ subspectrum is dominated by a band at 1,627 cm^{-1} because proteins with that feature dominate our training set, and the second subspectrum has major oppositely signed features at 1,659 and 1,641 cm^{-1}, representing the most common change from the average.[6] In the amide II, the first subspectrum has an intense band at 1,520 cm^{-1}, whereas the second subspectrum peaks at 1,560 cm^{-1}. Linear combinations of these two encompass most of the spectral change seen in the amide II region; the remaining four are much less significant (Gupta & Keiderling, 1992; Baumruk & Keiderling, 1993). The ECD decompositions are also virtually the same as for our previous report (Pancoska & Keiderling, 1991) with the first subspectrum resembling the dominant contribution to the overall ECD, which is oppositely signed at 209 nm and 190 nm.

Correlation of spectra and structure via regression equations

Table 2 summarizes the results of the restricted fits and those with the maximum number of coefficients (total spectral variation, labeled "full") in terms of the standard deviations and regression coefficients obtained for the amide I′, II, ECD, I′+II, and I′+II+ECD spectra, respectively. Errors are listed in the table in terms of the standard deviations on the actual FC_ζ values for each secondary structure subtype, ζ, as determined from the KS analysis (Kabsch & Sander, 1983). The relative standard deviation, as determined with Equation 5, is also listed as the final entry of the table, corresponding to the maximum number of coefficients. The relative standard deviations enable more facile comparison of the error for different structural parameters, which are determined completely independently in our method. As detailed in Supplementary Tables S1a–S1e (Electronic Appendix), the regressions stop showing large improvements after only a few coefficients are considered. This is illustrated graphically by plotting rr versus k for the helix and sheet component in Figure 2. All fits improve with increasing numbers of coefficients used, as they must, but the degree of improvement with added coefficients is generally small beyond the first few. For those regressions that meet the 99% confidence level on the F test, we have taken the break in the rise of correlation coefficient as indicative of there being no significant improvement in the regression with addition of more coefficients. In general, these break points are not precise and typically encompass the set found (see next section) to be the best predicting. Thus, in Table 2 we have used that "best predicting" set (next section) as an example of the restricted regression set. These are listed in Table 2 as "restricted regression" σ and rr values. It is clear from comparing these results that only a few coefficients led to almost the entire fit. That means that the apparent regression improvements for the fits that use the total variance of the spectral data set are not physically meaningful. The fits for the VCD coefficients for these 23 proteins were degraded in precision as compared to our earlier results with only 13 "known" proteins, but the corresponding ECD change between these protein data sets was much smaller. Furthermore the dependence on only a few coefficients is also consistent with our earlier analysis of the smaller set of protein data (Pancoska et al., 1991).

In general we found two classes of fit. Helix and sheet gave statistically significant regressions to all data sets, whereas bend and turn gave generally poorer regressions that were judged to

[6] Note that the absolute sign of the subspectrum is not important, because its contributions to the reconstruction of the experimental spectra are controlled by the sign as well as magnitude of the linear coefficients, C_{ij}.

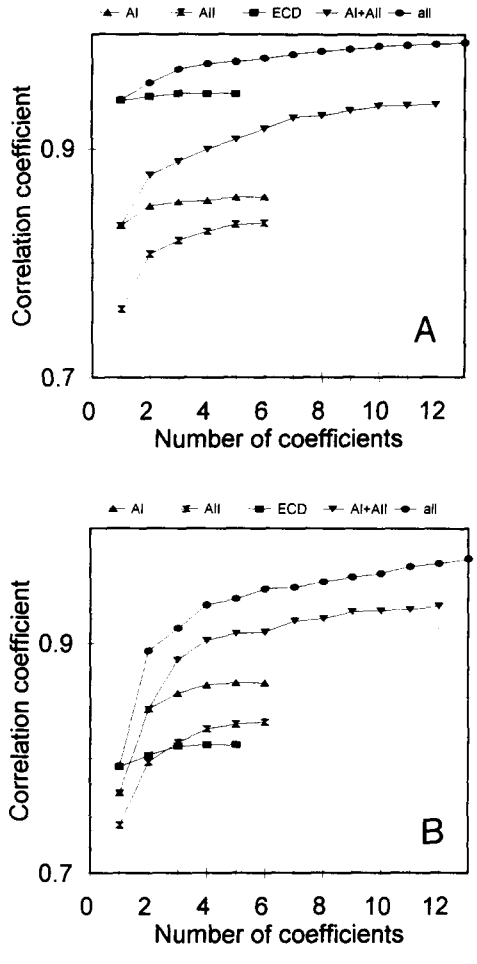

Fig. 2. Effect on standard deviation, *rr*, of increasing the number, *k*, of subspectral coefficients in the restricted regression tests for (**A**) helix and (**B**) sheet fractions. Plot for the combined amide I′+II VCD and ECD data sets was truncated at 12 for simplicity. Final value is indicated by the dashed extension of the curve.

be statistically unreliable for the turn component, except for the combined amide I′+II VCD plus ECD computation. The "other" component was in some cases fit well and in others less reliably. The regression coefficients for the α-helix and β-sheet fits are quite high (>0.84 in most cases, except the amide II and the ECD β-sheet), the standout being the ECD correlation to helix content ($rr > 0.94$). Although this numerically confirms the well-appreciated high sensitivity of ECD to the helical fraction, it also indicates that such a dependence is based on the first subspectrum, as *rr* does not improve much for added ECD coefficients. The same flatness with *k*, but a much worse *rr* value, is found for the ECD correlation to β-sheet. In contrast to ECD, the amide I′ VCD regression coefficients are virtually equivalent for helix and sheet and achieve their stable values by improving significantly with addition of at least a second coefficient. This reflects the inherently different sensitivity of these spectral methods to structural variations in peptides (and, consequently, proteins) (Keiderling et al., 1989; Freedman et al., 1995).

When spectral results from these three methods were used independently, the amide I′ VCD coefficients typically gave more precise fits than did the amide II VCD coefficients, but the ECD, other than for β-sheet, gave even higher *rr* values. This loss of precision may relate to amide II VCD spectra having lower S/N ratios and less band shape variability over the training set than do the amide I′ VCD, whereas the ECD have higher S/N than either VCD spectral type. However, fits based on the amide I′ and II VCD combined are generally better than fits to coefficients from the individual regions, including ECD, with the exception of the α-helical component. In the amide I′+II fits, the best restricted regressions contained coefficients from both amide I′ and amide II sets. This pattern is evident in all fits for the I′+II data set except the single coefficient ones. In the case of combining VCD and ECD data, again the regressions improved for all structural types. The best selected regressions generally included coefficients from both ECD and VCD.

We re-emphasize here that our calculations were complete, which means that by testing all possible combinations of coefficients in this composite case we necessarily include sets of coefficients from only one individual spectral type. It is important to know that the specific subspectral coefficients used in the multiparameter fit that has the best *rr* value are not as important as the trends found upon addition of coefficients. Determining which coefficients are sampled for the collection of better fits leads to a pattern establishing the structural relationship to spectra. For a given number of coefficients, *k*, a number of regressions based on different combinations of coefficients were found to have virtually equal *rr* values. These combinations usually had in common that particular coefficient found to be most important when only one coefficient was used to determine a fit. The form of these restricted regressions apparently impacts what we find below to give the best prediction of structure for proteins left out of the regression. The restricted regression fits for α-helix using the amide I′ VCD data set all depend on the coefficient of the second subspectrum and, using ECD, on the first subspectrum. The same pattern is seen for β-sheet. This reliance of two (or more) structural types on one subspectral contribution implies an interdependence. We have addressed this issue in detail separately (Pancoska et al., 1992, 1995) and will return to it in the Discussion.

Although the absolute errors for the structural parameters other than helix and sheet appear to be small, when one accounts for the dynamic range, i.e., using σ_{rel}, they are demonstrated to be much less well determined. In particular, except for computations involving all three data sets, there was no statistically significant correlation found with the turn fraction to even below the 95% confidence level, in agreement with our previously reported VCD-based observation (Pancoska et al., 1991). Similarly, considering just the amide I′ VCD data, for the bend and "other" secondary structure types there is only one equation for each structure (one-coefficient and two-coefficient, respectively), which surpasses the 99% confidence level. Both have remarkably smaller correlation coefficients than found for helix and sheet structures. However for the amide II, the "other" fraction is determined on a level of confidence comparable to those found for regressions to helix and sheet fractions. When both amide I′ and II are combined, the bend error drops only slightly, but the "other" error is almost halved, clearly showing the effect of adding the amide II sensitivities to the amide I′ set. Adding the ECD data to the VCD data again marginally improves all the *FC rr* values as compared to either the combined VCD or the ECD results.

Table 3. *Standard deviations for prediction of KS FC values for one protein left out of the training set*

	Helix			Sheet			Turn			Bend			Other		
	σ	σ_{rel}	No.	σ	σ_{rel}	No.	σ	σ_{rel}	No.	σ	σ_{rel}	No.	σ	σ_{rel}	No.
AI′ restr.[a]	11.6	15.1	2	8.2	17.3	3	4.0	28.8	1	4.3	23.3	2	5.2	23.8	2
Full[b]	13.9	18.0		10.2	21.3		5.1	36.2		5.3	28.7		6.4	29.4	
AII restr.	13.1	16.9	2	9.3	19.4	3	3.8	26.8	2	3.8	20.5	2	4.7	21.6	3
Full	15.9	20.6		10.8	22.6		5.3	37.9		4.3	23.3		6.2	28.4	
ECD restr.	6.7	8.7	1	9.5	19.9	1	3.8	26.9	1	3.9	20.8	1	3.7	17.1	2
Full	8.5	11.0		12.5	26.3		4.7	33.5		5.0	27.2		4.3	19.6	
AI′+AII restr.	10.1	13.1	6	7.4	15.5	4	3.8	26.8	2	3.3	17.9	5	3.3	15.3	4
Full	19.0	24.6		17.8	37.2		8.6	61.6		6.3	33.9		4.7	21.6	
AI′+AII+ECD	5.6	7.3	5	6.9	14.4	6	2.6	18.7	7	3.7	19.8	8	2.8	12.6	6
Full	15.9	20.6		21.7	45.4		8.6	61.7		14.3	77.2		9.2	42.0	

[a] Standard deviations, relative standard deviations, and the number of subspectra for the best restricted regression prediction. AI′, amide I′; AII, amide II.

[b] Standard deviations and relative standard deviations for the regression predictions using complete sets of subspectral coefficients, as in Table 2.

Use of regression relations for prediction of structure

The complete factor analyses were next recalculated for 23 different protein data sets, each encompassing spectra for 22 training set proteins, one systematically omitted. To make the computations manageable, the regression forms with the highest *rr* value for a given number of spectral coefficients were then used for testing predictions.[7] This series of regression types was then used repeatedly to predict the structure of the 23 proteins left out, and the results of those 23 predictions were compared to the X-ray values. This, of course, was done for each regression model k, $k = 1$ to p, for each data type: I′, II, ECD, I′+II, and I′+II+ECD. As should be clear, an enormous amount of numbers resulted requiring comparison. Tables S2a–S2e in the Electronic Appendix encompass some of these data, which are then summarized in Table 3 in terms of the standard and relative deviations of the best predictions of the 23 protein secondary structures from actual X-ray values (KS). For comparison, prediction errors for the full data sets are also included in Table 3.

Overall in comparing Tables 2 and 3, one can see that prediction standard deviations are higher than for the fits. Again we see two classes of prediction: helix and sheet as compared to the rest. In the ECD and both combined calculations, the "other" prediction is of the same quality as that of the sheet. This is best seen by comparing relative standard deviations. As might be expected from the quality of the fits obtained above, the ECD prediction of helix was much better than that for any of the VCD data sets, even when combined. Similarly the VCD data, particularly the amide I′ and the I′+II combination, did a better job predicting the sheet content than was possible with the ECD data. Thus, the trends seen in the regression are again evident here for prediction.

From another point of view, our prediction results can be seen to be fully compatible with the independent criteria that the predicted *FC* values should be >0 and that they should sum to 100%. For any given analysis based on a specific set of spectral data, at most one predicted coefficient was found to be negative. Averages of the summed *FC* values for each protein were found to be 100, 99.9, and 99.9% for the amide I′, amide II, and ECD-based predictions, with standard deviations of 2.2, 4.5, and 1.9%, respectively. In the combined calculations these numbers were even better.

As shown in Figure 3, by plotting the prediction error as a function of the number of coefficients used to predict helix, sheet, and "other" fractions from the different data sets, it is clear that these errors do not in general decrease as more coefficients are added to the scheme but in fact go through a minimum and then increase, in some cases to large prediction errors. An optimal predictive capability is generally found with just a few parameters for all the data sets. The most extreme case is for ECD where just *one* parameter yields the best prediction relationship for everything but "other." This demonstration that the regression with the minimal predictive error depends on just a few parameters directly parallels our experience in determining improvement for regression fits as noted above. That was exactly our original reasoning for using a restricted regression analysis for VCD and ECD spectra (Pancoska et al., 1991; Pancoska & Keiderling, 1991). We felt then that such an approach would enhance the reliability of prediction, and the results illustrated in Figure 3 bear out that hypothesis. Thus, the steady, but not significant, reduction in error noted above for the *fit* regressions as obtained with increasing numbers of spectral coefficients (Fig. 2) is *not* realizable for *prediction* (Fig. 3). This has ramifications on the real predictive ability of many other meth-

[7] To confirm that using only the best regression for prediction testing causes no significant error, we selected the 6–10 best regressions of each type for the amide I′ set and computed the prediction quality for each. In all cases, the regression of highest *rr* value gave the best prediction. For multiple coefficient-based predictions, there were often several combinations with near equivalent predictive ability, but these all had common elements. For example all the best amide I′ VCD predictors of the helix fraction depended on the coefficient of the second subspectrum.

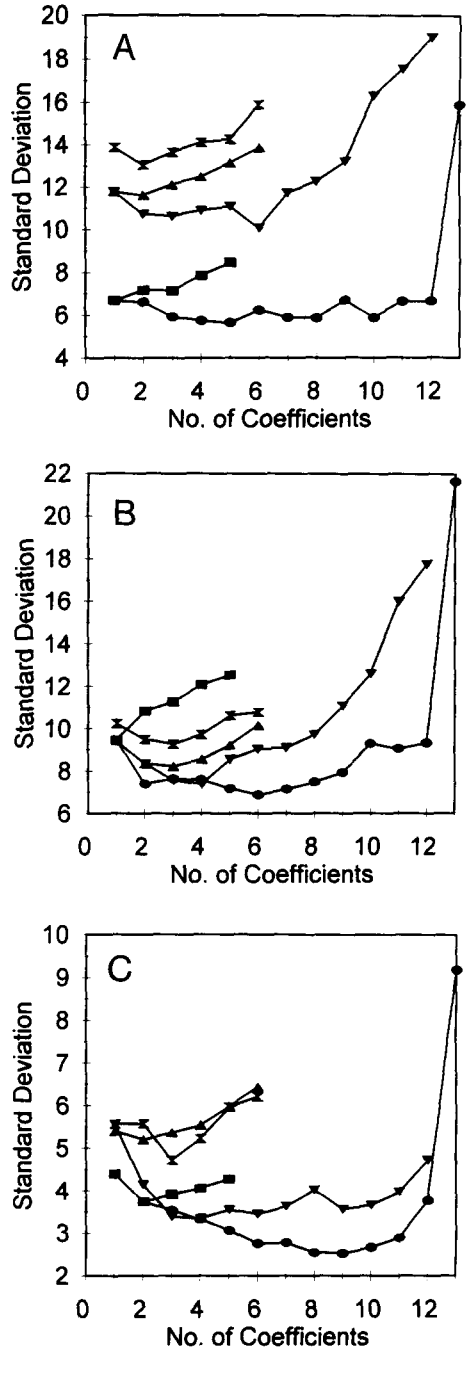

Fig. 3. Standard deviations for the best predictions of (**A**) helix, (**B**) sheet, and (**C**) "other" secondary structure components with the amide I′, II VCD and ECD, and combined VCD and VCD+ECD data sets as a function of increasing the number of spectral coefficients (symbols as in Fig. 2).

ods now used to interpret spectra data in terms of structure as discussed in the next section.

For ECD, the optimal function for prediction of helix and sheet has one coefficient, that of the first subspectrum, bend also has just one, whereas for "other" it has just two, and turn predictions from just ECD are unreliable. Incorporation of any other subspectral coefficient leads to an increase of the prediction error for ECD analysis. For example, prediction of helix with a one-coefficient function has 1.3 times smaller error than by using all five coefficients such as would be implemented in a conventional full transformation of the spectral data into structure.

For the VCD from the amide I′ and amide II regions analyzed separately, two-coefficient equations are optimal in prediction of helix, and three-coefficient ones for prediction of sheet *FC* values. The rest of the components are not predicted nearly so well but also optimize with 2–3 coefficients. Combination of amide I′ and amide II information in all cases improves the prediction performance of the VCD analyses. The optimal combined regressions always involve coefficients from both regions, paralleling the fit results, and in most cases involve only a few coefficients. The improvement obtained by combining the spectra is up to a factor of 1.4 in standard deviation reduction. For all structural components but helix, the combined VCD analyses provide better prediction than does the ECD analysis. For helix, the prediction ability of the optimal (six-coefficient) amide I′+II equation still does not exceed that of the single coefficient ECD prediction, being a factor of 1.5 worse in terms of error. The improvement factor of the combined VCD prediction over that of ECD is less, being largest for sheet, at 1.28, and smaller yet for "other," at 1.12, and for bend, 1.16.

Another dimension of the performance of various regression results can be discovered by inspection of the individual errors for *prediction* of *FC* values for specific proteins using different data sets (see Table 4). There is no significant overall correlation between the errors found in prediction with one set of data and those with another, but some proteins are poorly predicted with several sets and some are relatively good. For example, concanavalin A has a large helix error using VCD and ECD data sets as well as with the combined set, though it improved there, but its sheet error is relatively low for the amide II and combined sets. Hemoglobin is predicted well with both techniques, but the myoglobin helix content is poorly predicted in the VCD calculations while having only an average error with ECD. However, for the sheet content the myoglobin situation is reversed, with VCD doing a much better job than ECD. The spectra cannot seem to encompass myoglobin's very high degree of helicity. Cytochrome *c* and superoxide dismutase have anomalously high errors with ECD but are relatively well predicted with VCD. On combining sets, they both yield reasonable errors. On the other end of the scale, several proteins, for example, alcohol dehydrogenase, glutathione reductase, and trypsin, are relatively low in average error for most of the data sets. Several proteins are very well predicted using the combined data sets, which compensate for the limitations of each CD type as is evident by the much lower errors in the fourth column under each secondary structural type.

Despite the large number of possible prediction algorithms tested, there are more possible. To determine if our selection of the best regression for any given number of parameters biased our prediction results, for the amide I′ we tested the next best regressions for each k where the rr values of the fit were comparable for prediction error. This proved to be very enlightening in terms of finding the internal interdependencies of our prediction algorithms. For one parameter, the helix and sheet fraction were best predicted with the coefficient of the second amide I′ VCD subspectrum, and only marginally as well by the first subspectral coefficient. The others were significantly worse.

Table 4. *Errors in predicted FC values of individual proteins for helix, sheet, and "other" for best predicting models*

	Helix				Sheet				Other				Averages[b]			
Protein name	AI'[a]	AII[a]	ECD	Comb.[a]	AI'	AII	ECD	Comb.	AI'	AII	ECD	Comb.	AI'	AII	ECD	Comb.
α-Chymotrypsinogen A	10.9	8.8	1.5	2.0	0.6	3.8	3.4	1.8	1.7	1.1	2.0	2.8	4.4	4.6	2.3	2.2
Alcohol dehydrogenase	0.1	2.3	1.4	4.1	2.9	2.0	3.7	5.4	1.8	0.3	2.1	1.1	1.6	1.5	2.4	3.5
α-Chymotrypsin type II	5.9	12.7	1.3	3.7	6.5	1.9	1.6	2.9	0.4	5.9	3.5	2.8	4.3	6.8	2.1	3.1
Concanavalin A	16.9	12.7	13.4	7.4	12.2	3.1	10.3	3.4	6.9	2.5	0.9	2.7	12.0	6.1	8.2	4.5
Carbonic anhydrase	0.9	6.9	8.1	3.2	1.8	5.5	4.8	0.1	1.4	3.7	0.2	1.9	1.4	5.4	4.4	1.7
Cytochrome *c*	8.3	5.6	11.8	0.6	0.1	5.9	20.3	2.4	9.3	17.0	6.7	1.3	5.9	9.5	12.9	1.4
Tosyl elastase	15.3	3.0	1.5	2.3	14.9	3.5	1.5	7.2	1.8	0.5	0.2	0.8	10.7	2.3	1.1	3.4
Glutathione reductase	5.1	2.2	1.6	0.6	5.4	2.5	3.0	0.9	2.5	3.2	0.4	3.7	4.3	2.6	1.7	1.7
Hemoglobin	0.4	3.5	1.1	4.2	5.0	13.4	0.9	1.7	11.1	5.5	5.6	2.3	5.5	7.5	2.5	2.7
λ-Immunoglobulin	12.5	2.3	3.5	1.6	14.7	15.5	14.2	8.8	2.4	5.3	5.6	1.5	9.9	7.7	7.8	4.0
Lactate dehydrogenase	19.7	12.8	4.2	6.3	9.1	6.6	2.5	0.2	7.9	9.4	5.3	3.5	12.2	9.6	4.0	3.3
Lysozyme	11.9	3.3	4.2	0.2	14.7	4.8	10.1	0.8	6.6	7.7	3.1	0.4	11.1	5.3	5.8	0.5
Myoglobin	28.6	29.1	5.1	3.8	2.9	7.9	19.0	8.2	10.9	7.6	1.2	0.2	14.1	14.9	8.4	4.1
Papain	7.9	5.1	0.9	2.0	2.7	15.8	8.0	4.9	6.2	6.2	0.4	0.1	5.6	9.0	3.1	2.3
Rhodanese	11.7	7.3	3.6	2.4	16.5	9.4	7.8	8.5	1.2	4.9	8.8	5.6	9.8	7.2	6.7	5.5
Ribonuclease A	5.3	1.3	2.0	1.0	4.9	11.6	10.6	2.6	1.2	4.5	5.4	4.7	3.8	5.8	6.0	2.8
Ribonuclease S	13.2	7.3	2.4	1.8	0.1	12.9	11.3	1.8	0.8	3.0	1.2	2.3	4.7	7.7	4.9	2.0
Subtilisin BPN'	4.0	19.4	6.7	2.9	8.9	9.0	6.7	2.8	0.7	4.2	0.1	1.8	4.5	10.9	4.5	2.5
Superoxide dismutase	4.3	19.7	17.3	6.7	1.2	13.4	11.7	5.0	1.5	0.6	2.3	2.4	2.3	11.2	10.4	4.7
Thermolysin	4.0	14.2	7.2	9.2	6.2	3.0	7.4	3.0	5.7	4.8	2.1	2.2	5.3	7.3	5.6	4.8
Triose phosphate isomerase	12.6	18.2	1.2	4.1	2.1	7.4	8.0	2.6	0.8	0.3	2.4	0.4	5.2	8.6	3.9	2.4
Trypsin inhibitor	6.5	23.2	9.4	1.6	6.2	23.4	7.7	2.1	1.1	2.6	0.3	2.2	4.6	16.4	5.8	2.0
Trypsin	7.1	15.5	1.0	2.9	4.8	9.5	0.4	1.9	3.4	1.2	4.2	0.4	5.1	8.7	1.9	1.7

[a] AI', amide I; AII, amide II; Comb., combined amide I', amide II, and ECD data sets.
[b] Average error calculated as the arithmetic mean of the helix, sheet, and "other" prediction errors listed here.

For two coefficients the five best combinations all involved the coefficient of the second subspectrum, which was the dominant single predictor. The combination of two coefficients that had the best prediction ability for the proteins successively left out in fact turned out to be that with the highest regression coefficient for the entire set. This pattern was maintained for predictions with other sets of multiple coefficients to a level where predictions based on another set of coefficients were either insignificantly different from that with the best regression coefficient or were definitely worse. None were significantly better. Thus, our method of selecting out the regression with the best *rr* value is justified by example. The only caveat is that although the major contributors to the predictions, as identified above, are indeed significant, one should not put excessive weight on specifically which coefficients are used as minor contributors in the multiple coefficient-based predictions.

Discussion

Fits and predictions

We have demonstrated that the amide I' and II VCD as well as the ECD spectra can be decomposed into a series of subspectra and coefficients and that it is possible to develop a correlation between these spectral coefficients and secondary structure. Furthermore, the amide II correlation is determined to be less significant than the amide I' as might have been predicted from qualitative observation of the reduced variance of the amide II VCD band shape with structural change (Gupta & Keiderling, 1992; Baumruk & Keiderling, 1993). Combining the components of the two vibrational transitions, amide I'+II, did improve the VCD correlation with structure, much as was seen previously for FTIR spectral analyses (Dousseau & Pezolet, 1990; Lee et al., 1990; Pribic et al., 1993), but aside from the "other" component, the change was not dramatic. On the other hand, a dramatic increase in the quality of the fit was obtained by coupling the two VCD spectra with the ECD results in a combined analysis. All structural types show an improvement in fit with this more flexible approach, and even the turn fraction is fit to a statistically significant level. In terms of the dynamic range, errors drop to 10% or below in most cases. Each type of spectrum is shown to have its own strengths and weaknesses for determining average secondary structure so that their combination gives the fitting routine the most flexibility. Clearly, the fits improve with added coefficients, as they must, yet this improvement is small beyond a few coefficients for each structural type.

If fitting known structures were our goal, or even if it were useful, we would be very pleased with the combined approach to spectral–structure regressions. However, accurate prediction of the secondary structure of proteins not in the training set was and remains a more important and useful goal. In this regard, errors were higher and less strikingly improved in the total combined set. In fact, only a few coefficients are used to yield the most reliable predictions for proteins left out of the analysis. This is most striking in the ECD analysis, where the helix and sheet determinations depend significantly on only one coefficient. Furthermore we have shown these predictions to worsen significantly as the number of coefficients used is increased. It

is important to realize that our methods of testing all possible combinations of spectral components and comparing them explicitly for their predictive capability is a new and, we think, important approach to the use of optical spectra for secondary structure analyses. It might be noted that the VSM method (Manavalan & Johnson, 1987) is complementary in this respect by systematically searching for the most important proteins to eliminate from the training set while keeping all their spectral components that were previously judged to be significant. This aspect combined with the coupling of data from different techniques is central to the advance offered in this work. It should be noted that Sarver and Krueger (1991b) and Pribic et al. (1993) previously reported combining ECD and FTIR data for improved analyses with a linear model. From their analyses, the improvements followed a similar pattern, ECD yielding the better helix and FTIR better sheet predictions, but the improvement was not as dramatic as seen here.

It is precisely due to this lack of dimensionality in ECD structural response, first noted in our qualitative comparisons (Pancoska et al., 1989; Keiderling, 1993, 1994; Pancoska & Keiderling, 1993), that the VCD studies were designed. However, despite all of the improvement found in the regressions by inclusion of both amide I′ and II VCD data, the predictions do not show the hoped-for large improvement in accuracy when aimed at determining just fractional secondary structure. The predictive reliability for the best set of coefficients from the combined amide I′ + II VCD data set was 10–20% better than with just the amide I′. Although, in principle, adding data from amide III or other transitions yet to be sampled could help, our experience here (see below) argues that it will not be significant.

On the other hand, combining VCD with ECD data has resulted in a significantly improved analysis and an ability to predict structure that is far superior to all our analyses based on single techniques using the same systematic restricted regression methods. Improvements for the dual, ECD plus VCD, technique predictions correspond to a reduction in average error on the order of 30% (Table 3) over the single data set predictions. However, the real impact of the improvement is seen in Table 4, where the individual protein prediction errors for the combined data set are more equivalent. All errors are less than 10% for both helix and sheet with the combined set (Table 4, columns 5, 9), and the largest average error (Table 4, column 17) for helix, sheet, and other is 5.5%. The relative strengths of ECD and VCD prediction capabilities reflect the individual strengths of the components being used in the combined prediction. Although combining the ECD and VCD data sets does not result in a revolutionary decrease of predictive error, it must be recognized that the combined ECD–VCD prediction is definitely stronger than either alone; each method compensates for the weaknesses of the other. The end result is really excellent: helix, sheet, and even "other" are predicted with small error margins (7–15% of their dynamic range in the test set); and, for the first time, in our opinion, turn and bend predictions have some predictive reliability (~20%).

Limitations of the method

The incremental improvement in the prediction error with added spectral information is intimately coupled to our observation that prediction error not only fails to improve but, in fact, worsens upon addition of coefficients to the regression beyond the few of most significance to secondary structure. Such a behavior was implied in the results of van Stokkum et al. (1990) and Pribic et al. (1993) who saw degraded predictions with matrices of higher rank. However, their methods do not seem to encompass selection of the components based on significance for secondary structure, but rather take them in order of significance in terms of spectral residue. The degradation of prediction with increase of spectral components (or rank of the matrix for inversion methods) speaks directly to the method used or to the question being posed. We now feel it is the latter that is in most need of revision. Our observations also lead to a corollary observation that, if one uses the total spectral variation available to predict structure in terms of *FC* values, the results could be dramatically bad, leading to real qualitative error in interpretation. An approach of totally correlating the spectral variation with secondary structure is a common underlying assumption of many spectral band shape interpretive methods (Manning, 1989; Sreerama & Woody 1993, 1994). We feel this is too limiting and not warranted by other observations. This conflict over goals of the structural biologist and limits of the approach leads to the core of this work. What is the fundamental limit for secondary structure prediction that we are seeing, and what can be done to address it so that spectral data might be better used for determining structures for globular proteins in solution?

Before addressing these fundamental issues directly, it is worthwhile to briefly investigate other possibilities for the apparent limitations of these analyses. As shown in Figure 1, the analysis is reasonably stable to increase in the number of the proteins used. In fact, the training set used is about the same size as used in many protein spectral analyses that have been reported in recent years. However, increasing the size of the basis set of proteins did decrease the accuracy of prediction for the amide I′ VCD as compared to our previous study (Pancoska et al., 1991). The effect was less for ECD (Pancoska & Keiderling, 1991). In comparing predictions made with the full set of proteins as compared to one of the abbreviated sets used for the Figure 1 stability test, we found the difference in prediction accuracy to be not very large. The parallel analyses of ECD also point out that the limitations of this approach are not a problem of using VCD data but are more general. A similar decrease in accuracy was noted by Pribic et al. (1993) in comparing their 21 protein set to the 10 protein set of Sarver and Krueger (1990).

One might have thought the problem could arise from the noise level inherent in VCD spectra. There is no doubt that the VCD spectra are in some sense limited by S/N considerations. The corollary techniques of ECD and FTIR can both generate spectra of higher S/N. However, those spectra do not have the intrinsic band-shape variations of VCD. Thus, even burdened by reduced S/N, VCD can exhibit distinct reproducible and identifiable spectral variations with structure. This property actually manifests itself in the number of subspectra identified in the factor analysis methods. Although ECD with relatively good S/N yields five significant subspectra, VCD in the amide I′ band has six.[8] In this light it is interesting to note that parallel style analyses of very high S/N FTIR spectra of the amide I band (Lee et al., 1990) have indicated as many as 11 subspectra can be iden-

[8] The amide II also has six, but it is doubtful that these all arise from real variation in the amide II VCD bandshape. Interferences with sidechain vibrations and lower intrinsic S/N may be the root source of some of the higher subspectra for the amide II.

tified, yet subsequent tests (Pribic et al., 1993) show that only a few of these are needed for optimal prediction. Because VCD samples the very same transitions as FTIR, one might expect to obtain at least that many subspectra in VCD. There are three main differences between VCD and FTIR: the VCD spectra are obtained at lower resolution and lower S/N than the FTIR, and the two spectroscopies have different intensity mechanisms. The first will cause higher resolution components to decrease in relative contribution to the band shape, and the second will convolute some of the high-resolution subspectral components with noise. The difference in intensity mechanism means that some transitions could be observable with FTIR and not VCD, or vice versa, leading to spectral variation differences between the two techniques. Hence, all three of these differences can result in subspectra of lower significance being unimportant in our study. We have tested this contention regarding noise effects by doing a repeat analysis on the amide I′ VCD data after smoothing the spectra to reduce the noise contribution. Still, six significant spectra were found, and the errors in prediction were virtually indistinguishable from those in Table 3. Thus, in practice, the factor analysis does segregate out the largest part of the noise contribution, as it should.

Although it might seem serious that some of our spectral information is not being isolated and identified due to noise or resolution problems, this is not really cause for significant concern because the overriding issue here is really why so much of the spectral data that we do isolate is not used for prediction in the first place. Our results suggest that, if more spectral data were used in the analyses, the quality of the prediction would actually deteriorate. For the problem originally laid out in this paper, i.e., determination of fractional secondary structure, inclusion of more spectral components is definitely not the solution.

The problem of concentration determination, functionally equivalent to overall intensity variation, clearly should be a source of error. We have separately tested this and have shown that up to about 20% random error in intensity, the fitting error for the amide I′ or for the ECD is not significantly impacted (Bitto, 1993; Pancoska et al., 1995). However, if one had a systematic concentration error, the results could be affected more severely because that would skew the calibration (in an analytical sense) provided by the training set. We have used three different means of concentration normalization. For ECD, we tried to determine concentration spectrophotometrically. Although these concentrations are possibly only good to ~10%, these errors are most likely random and the tests above indicate that prediction should be preserved. In the VCD, both normalizations implicitly assume that all conformations have a common molar absorbance. This is not correct, but the band-shape analysis can partially offset such an error in that the correlation sought is to a component of the band shape and its relative contribution to the whole. Our algorithm does not force the correlation to be simple, and non-zero offsets in the linear equations are the rule. These can act as a first-order correction for absorbance variations by structural type.

Another potential problem is interferants, components of the molecule (or sample preparation) that lead to spectral overlap in the region of interest that we wish to assign fully to amide transitions. A common problem relates to side-chain absorbances. In the near UV, the aromatic group and disulfide near UV transitions cause some problem with analysis (Manning, 1989; Manning & Woody, 1989; Manning et al., 1992; Grishina & Woody, 1995). Similarly, glycosylation can be a problem in the far UV (Urbanova et al., 1994). Both remain somewhat difficult to overcome for ECD analyses. For VCD, side-chain absorbances are more of a problem in the amide II than amide I; but in both cases they are generally weak due to their arising from specific, well-separated groups that give rise to sharp transitions at relatively fixed frequencies, resulting in small increments to the integrated absorbance with consequently little impact on the normalization. If such groups give rise to VCD, they would contribute to higher-order subspectra and be reduced in impact due to their lack of correlation with the amide spectral variance. In practice, we have not seen significant side-chain VCD even in our model studies of homo-oligo- and polypeptides (Keiderling et al., 1989; Freedman et al., 1995). Thus, we do not expect that the side chains are the root cause of the problem of fractional secondary structure determination.

A different type of interferant is potentially more pernicious. We do not discriminate between α-helices and 3_{10}-helices in our analyses, nor do we distinguish parallel and antiparallel sheets. To test for effects of the former approximation, using both the ECD and the amide I′ VCD data sets, we ran a series of regression and prediction tests with the 3_{10}-helix fraction moved from the H to the C component. Inclusion or separation of the 3_{10} component in the helix fraction had virtually no effect on the prediction accuracy, in terms of relative error, of the α-helix. Tested separately, the 3_{10}-helix could not be predicted with any statistical reliability at all over its limited range, which is less than 15% for our training set. Having values so close to the prediction error for total helix, the 3_{10} component, though a significant structural feature, proves to have too little impact on the spectral correlation to fractional structure. It is important to realize that 3_{10}-helices do give qualitatively identifiable VCD spectra in oligopeptides (Yasui et al., 1986a), but for the amide I′ they are very weak and discrimination is dependent on amide II measurement. The ECD spectra of 3_{10}-helices are poorly distinguished from those of α-helices (Toniolo et al., 1991). These aspects contribute to the problem of identifying 3_{10}-helical contributions in protein spectra. Despite the qualitative issues, the real issue for this paper is that the method of treating 3_{10}-helices in the regression makes no significant difference to the fundamental prediction error.

A question that might arise regarding the errors in our prediction methods is the method used to determine the reference structures. It is clear that crystallographers do not have a uniform method of allocating residues to specific secondary structural types. Thus, we and many others have used more "objective" routines, at least not biased by our usage, to obtain such descriptors (Sreerama & Woody, 1993). We have relied most on the KS data set and the DSSP routine for abstracting parameters from structures in the PDB (Kabsch & Sander, 1983). This description can be viewed as somewhat conservative for helix and sheet, emphasizing residue conformations compatible with hydrogen bond formation. Perhaps a broader description is more appropriate for spectra that are in large part determined by the through-space interactions afforded by dipole coupling. To test this possibility, we developed a code to download PDB structural parameters and generate fractional secondary structure representations in the LG (Levitt & Greer, 1977) algorithm. This was tested for predictability using the amide I′ and ECD data sets in the manner described above for the KS structures with the results summarized in Table 5. The net prediction relative

Table 5. *Standard deviations for prediction of LG FC values for one protein left out of the training set*

	Helix			Sheet			RT			LT			Other		
	σ	σ_{rel}	No.	σ	σ_{rel}	No.	σ	σ_{rel}	No.	σ	σ_{rel}	No.	σ	σ_{rel}	No.
AI′[a]	14.7	16.3	1	11.7	17.3	3	3.9	29.8	2	4.1	23.8	1	3.4	25.6	3
ECD	8.5	9.5	2	11.8	17.5	1	3.6	27.5	1	3.5	20.4	1	3.0	22.3	1

[a] Standard deviations, relative standard deviations, and the number of subspectra for the best restricted regression prediction. AI′, amide I′.

errors found for amide I′ VCD were a bit worse than the results in Table 3 for the KS structure descriptors. For ECD, the error for the sheet fraction slightly improved and that for the "other" category worsened. Both are probably related to the redistribution of contribution from "other" to better defined structures using the LG definitions; but neither change was large enough to merit further consideration. Given this and the similar finding by Sreerama and Woody (1993) that the KS set gave the best results for their self-consistent correlation of structure and ECD, it is reasonable to continue using the KS values as a basis for our analyses.

It might be noted that, considering just the amide I′ data set, the regression fit for the α-helix is not as good with this expanded training set of 23 proteins as found before with a smaller set (13 of these proteins) (Pancoska et al., 1991). The errors for the other components are smaller in this larger set, though those differences between the two calculations are small. This implies that expansion of the set has introduced error in terms of the helical content. A better comparison in this sense would be to predictions made with our reduced training sets whose factor analyses were summarized in Figure 1. For five separate combinations of proteins from our training set, ranging from 11 to 19 members, the prediction error for helix and sheet varied both above and below what were found for the 23 protein set as reported in Table 3. The larger sets from these subsets did have errors somewhat higher than for the full set, and one smaller set (12 proteins) did have better prediction errors. Such an improvement in prediction by using reduced training sets was found previously by Johnson and coworkers and led to development of the VSM for ECD analysis (Manavalan & Johnson, 1987). These results suggest that the original data set did not adequately represent the variance of protein VCD that might be observed or conversely that the original set was more uniform in structure or in the relationship of its secondary structures to its spectra than is the case for the larger set studied here. It is important to realize that these subsets of proteins were selected on the basis of maintaining representation from each of the dominant VCD spectral types as identified by cluster analysis of the spectra. Thus, the variation seen in prediction capability should be optimal for VCD and must reflect the sensitivity of the regression surface to inclusion of individual proteins (see below).

One possible variation in the VCD spectrum that might not have been fully represented in the smaller set could be a difference in the degree of deuteration of the helices in the two sets. Deuteration has definite effects on the helix VCD, changing the amide I from a couplet to a three-feature band shape (Sen & Keiderling, 1984; Baumruk & Keiderling, 1993). It is reasonable to assume that some, perhaps smaller, changes will occur for the VCD of other secondary structure types on deuteration. Our method implicitly assumes that the proteins are fully or at least uniformly exchanged. However, some proteins do not exchange as well under our conditions of repeated dissolution and standing at ambient temperature in D_2O. This could cause a variance in the spectral coefficient distribution that would not fit the regression well. Such effects would be very minor for ECD and could be one molecular factor behind the observation that its analysis seems less sensitive to expansion of the training set (Pancoska & Keiderling, 1991). We plan to address this issue directly in a study now underway in our laboratory that uses amide I and II VCD data exclusively from H_2O solution, thereby avoiding the issue of deuteration (Baumruk et al., in prep.).

Restricted regression method and prediction accuracy

The key to appreciating, if not understanding, the difference in the prediction accuracy from the regression accuracy is to consider how the regression responds to removal of one protein from the set. Although the original linear relationship was optimized to all the data, with one protein removed from the set, the regression no longer needs to compensate for errors involving that selected protein. Thus, the regression surface will always tend to move away from the protein left out and the prediction error will always be larger than the fit error when it is included. That, of course, does not explain why the change is so large. The surface must respond to all contributions to the error, and at present we cannot conclusively determine which source is the cause of this response. However, our tests of induced intensity fluctuation and reduced noise by smoothing as described above imply that random error is not the source of this problem. If, following the above discussion, that original set was more homogeneous, then removal of one protein would have less effect. If the error was random and only a few coefficients were used in the fit, removal of a protein would have little impact. That is why the more limited regressions give better predictions: they are more stable. The more parameters used to fit the set, the more sensitive it will be to a given protein's removal. If there is a systematic error due to a nonuniform spectral response to a structural variation, removal of one of these "outlier" proteins will have a large impact on the regression surface determined with the set and the prediction error for that protein taken out will be large.

Our new approach, reported here, which generates complete searches of the multiple coefficient fits and predictions, exposes these sensitivities. Thus, for a large number of coefficients, the regression surface has flexibility to conform to the individual variations in the set with modest error. Because these spec-

tral variations are, in principle, independent, they could form a separate coordinate on which the surface varied (as emphasized by the PC/FA approach). If a given protein is removed, the regression surface does not need to adjust to its contribution to the "independent coordinate" any more and can thereby move significantly away from the removed protein's spectral–structural relationship. That is the source of the continued worsening of the predictions with increase in the number of parameters, some of which represent these "independent" contributions to the spectra.

In summary, although our factor analysis is apparently stable to addition of proteins judged at the level of band-shape comparison, the regression analyses with multiple parameters are sensitive to the nature of the training set used. What is different in our regressions, which make use of only a few coefficients? With fewer parameters, the regression must compensate for its lack of flexibility and in effect become more stable to change in the training set. Thus, although we can gain precision in terms of fitting by expanding the number of spectral coefficients used, we do not increase the accuracy of prediction. Prediction accuracy is enhanced by having the most encompassing relationship; that means one of less precision but more stability.

It is important to see that these observations have a wider impact than this particular study. Because in most other spectra–structure prediction methods the entire spectral representation developed with the principal component method are merely fit in some sort of regression scheme or are totally projected onto the structural descriptor, those results are equivalent to our "full set" fits. In terms of prediction, the results presented here imply that this must be unwise at best. Johnson and others (Hennessey & Johnson, 1981; Manavalan & Johnson, 1987; van Stokkum et al., 1990; Pribic et al., 1993; Sreerama & Woody, 1993, 1994) have been aware of the need to look at predictive capability of their analyses for some time. One response was the development of the variable selection method (Manavalan & Johnson, 1987). That method still uses the entire spectral data set to develop a relationship between spectra and structure but eliminates contributions from selected proteins. Because the VSM selects a set of proteins most similar to the one of interest, it tends to minimize the hypersensitivity of the fitted surface to any proteins left out, such as the unknown protein for which one wishes to predict structure. Alternatively, Pribic et al. (1993) have restricted the rank of the spectral matrix, which effectively reduces the dimension of the fitting surface, and also restricted the proteins used with a locally linearized (LL) model, a version of the VSM. A combined approach is the self-consistent method of Sreerama and Woody (1993, 1994) that uses the LL with inclusion of the unknown in the basis set and iteration to a self-consistent prediction of its structural parameters. Studies using FTIR data suggest that use of large numbers of coefficients in the regression analyses (Lee et al., 1990) can reduce the predictive ability of those results, which is consistent with our conclusions. We go one step further by selecting just those spectral components that most strongly relate to secondary structure, rather than those that most strongly contribute to the observed intensity, but we keep all proteins. Future extension to a VSM or LL approach with our restricted regression methods could be interesting, but we feel that is not the right direction to pursue now.

Going beyond how many subspectra are important for prediction, we should look at which ones are most influential. Here, a perhaps surprising pattern emerges. For ECD, the single coefficient important for both helix and sheet prediction corresponds to the first subspectrum. Inspection of the first ECD subspectrum shows it to resemble the expected band shape for an α-helix. That the first subspectrum is so shaped should not be surprising because the α-helix contribution effectively dominates globular protein ECD. This is presumably why the methods of Perczel et al. (1991) and the older, simpler method of using the ellipticity at 222 nm to determine helix content work at all (Yang et al., 1986; Johnson, 1988). In the amide I′ VCD, the second subspectrum is essential to predicting both helix and sheet, but the best predictions make use of other coefficients also. In a striking parallel to the ECD situation, the second subspectrum of the amide I′ VCD resembles the VCD found for highly helical proteins. The amide II situation is less clear. The first amide II subspectrum mostly resembles spectra found for α-β proteins. The second subspectrum peaks at 1,560 cm^{-1} and, with the first, can encompass most of the change seen in the amide II. Both of these are important in predicting helix and sheet. Thus, a consistent pattern emerges between ECD and VCD where the coefficient of the subspectrum most like that expected for an α-helix is the best predictor for helix (naturally) but is also best for predicting sheet.

Given our earlier development of a relationship between the helical and sheet fractions in a very large subset of the proteins in the PDB, this interrelation of subspectral dependencies is understandable (Pancoska et al., 1992). However, the lack of any other contributions on any significant level was surprising. Our method has the ECD, for example, predicting the helix well due to its high sensitivity to the helical ECD signature and predicting the sheet component only from the interrelationship of the helix and sheet. Although β-sheets give a qualitatively identifiable ECD spectrum, the relationship of that spectral contribution to variation in the *FC* value representative of the structure must be less well defined than the correlation between α-helix and β-sheet content. Our ECD regression selects the predictor of sheet with the highest correlation to its *FC* value. This appears to be the helical component, and therefore the qualitative spectral manifestation of sheet is not utilized directly in the quantitative scheme. In fact, including more spectral components so as to encompass it degrades the sheet prediction. One could say that at our level of predictor, sheet is not directly measured using ECD but is simply inferred from the excellent determination of helix. These observations have fundamental repercussions in the array of various structure-predicting algorithms in the literature that rely only on ECD spectral input. One reason that one might have problems analyzing the structure of some unknown protein with ECD is that if its ratio of helix to sheet does not reflect the pattern in the PDB, or more precisely in the training set used for a particular analysis, the method cannot predict the sheet contribution because, in general, it does not sense it.

On the other hand, VCD, although exhibiting the same problem, does utilize more of the spectral data for prediction so it can sense the sheet contribution marginally better. To explore this, we predicted the sheet components for the proteins in our training set from their crystal structure helix fractions, from the ECD-predicted helix fractions, and from the VCD helix-predicted fractions and, as summarized in Table 6, compared them to the values found with our spectral prediction algorithms and to the crystal structure values in Table 1. It is clear from Ta-

Table 6. *Comparison of deviations in FC_β determined from the crystal structure derived FC_β–FC_α relationship with errors in FC_β predictions from spectra*

	X-ray	Amide I′	Amide II	ECD
σ[a]	5.8	9.5	10.2	8.3
σ_{rel}[a]	12.2	19.9	21.3	17.4
σ_{rel} Spectra[b]	–	17.3	19.4	19.9

[a] Standard deviations of FC_β values for training set proteins calculated from FC_α taken from X-ray or predicted from spectra as indicated in the column headings.

[b] Relative standard deviation of best predictions of FC_β from spectra as taken from Table 3 for comparison.

ble 6 that if one knew the helix fraction exactly, the helix–sheet relationships derived previously from neural network analysis of the PDB (Pancoska et al., 1992) would provide an estimation of the sheet content superior to any of those available from the spectroscopic regressions. However, that is not the case for any real unknown. ECD gives us the best estimation of helix content. Using it as input to the helix–sheet relation in fact gives a better prediction of sheet than does our spectral prediction. On the other hand, use of the VCD-generated helix fractions with the helix–sheet relation does worse than the relative error of the spectral predictions as compared to the values from Table 3. Admittedly the differences are not huge, but the patterns are totally consistent, lending more significance to this discussion. This demonstrates that the VCD prediction method senses the sheet content beyond its being a counterpoint to helix. Thus, although the ECD predictions of sheet content are dominated by the relationships following from interdependence of helix and sheet in the training set proteins, the VCD-predicted values are less so. However, it must be recognized that interdependence still exists in the training set, so that any algorithm will reflect it to some degree. That is evidenced by the "perfect" helix values giving better predictions of sheet than any spectral method.

Conclusions

Our work shows that, although spectral analyses in terms of average secondary structure as represented by *FC* values can be improved to an excellent level of accuracy by combining data from ECD and VCD, overall, such methods are subject to fundamental limitations that yield errors characteristic of the proteins themselves and not of the methods. It is projected that such understanding will provide a means for future analyses, which can go beyond such limitations, through use of a different basis for protein structural description as regards optical spectra.

Given the intrinsic limits to accuracy in predicting fractional secondary structure that we have found, one might ask: what is the logical next step? Our work demonstrates that there is a wealth of spectral information available that is not now being used in trying to address the relatively narrow question of fractional secondary structure, represented by a vector of a few numbers summing to 1.0. Describing the proteins in terms of fractional secondary structure is a major part of the problem in utilizing all of the spectral information. By no means are the segments uniform as assigned to various categories by any algorithm. It is this distortion from ideality that we feel is the biggest problem with the standard approaches taken to interpret spectra in terms of structure. One approach that we have suggested (Pancoska et al., 1995) and are now testing is to represent the segments and their interconnectivities by a more detailed descriptor that would have a matrix rather than a vector form. Due to the conformational distortions at the junctions, the residues involved will certainly have a different spectral response than the segments they connect. Similarly, the distribution of lengths of coherent segments can affect the overall spectral response for VCD (Dousseau & Pezolet, 1990; LaBrake et al., 1993) as well as ECD (Yang et al., 1986). Both of these issues affect the descriptors currently being tested (Pancoska et al., 1995), but the optimal design to fit the various types of spectral data available in the IR and UV regions of the spectrum is far from settled.

Supplementary material in the Electronic Appendix

Included in subdirectory Pancoska.SUP of the SUPLEMNT directory of the Electronic Appendix are *x*, *y*-formatted ASCII data to create graphical representations of the original VCD (amide I′ and II) and ECD spectra for the 28 proteins as well as of the PC/FA subspectra derived from them. More complete tables of the fit and prediction errors with regard to crystal structure values for different numbers of coefficients based on the various data sets, amide I′, II, I′+II, ECD, and I′+II+ECD are also provided.

Acknowledgments

This work was primarily supported by a grant from the National Institutes of Health (GM30147), for which we are most grateful. Cooperation between Charles University and UIC is supported in part by a National Science Foundation grant (to P.P. and T.A.K., INT 91-07588) and a University Scholar Award to T.A.K. from the University of Illinois. Equipment grants from the NSF, NIH, and University of Illinois supported purchase of instrumentation used. V.P.G. thanks the Council for the International Exchange of Scholars for a Fulbright Travel Grant. We thank Dr. Rina Dukor for help downloading PDB files.

References

Baumruk V, Huo D, Dukor RK, Keiderling TA, LeLeivre D, Brack A. 1994. Conformational study of sequential Lys-Leu based polymers and oligomers using vibrational and electronic circular dichroism spectra. *Biopolymers 34*:1115–1121.

Baumruk V, Keiderling TA. 1993. Vibrational circular dichroism of proteins in H_2O solution. *J Am Chem Soc 115*:6939–6942.

Berjot M, Marx J, Alix AJP. 1987. Determination of the secondary structure of proteins from the Raman amide I band: The reference intensity profiles method. *J Raman Spectrosc 18*:289–300.

Bitto E. 1993. Study of protein conformation by mathematical analysis of spectroscopic data [thesis]. Prague: Charles University.

Brahms S, Brahms J. 1980. Determination of protein secondary structure in solution by vacuum ultraviolet circular dichroism. *J Mol Biol 138*: 149–178.

Bussian BM, Sander C. 1989. How to determine protein secondary structure in solution by Raman spectroscopy: Practical guide and test case DNase I. *Biochemistry 28*:4271–4277.

Byler DM, Susi H. 1986. Examination of the secondary structure of proteins by deconvolved FTIR spectra. *Biopolymers 25*:469–487.

Chang CT, Wu CSC, Yang JT. 1978. Circular dichroic analysis of protein conformation: Inclusion of the β-turns. *Anal Biochem 91*:13–31.

Chen YH, Yang JT. 1971. Secondary structure of proteins by circular dichroism. *Biochem Biophys Res Commun 44*:1285–1291.

Chen YH, Yang JT, Martinez HM. 1972. Determination of the secondary structure of proteins by circular dichroism and optical rotatory dispersion. *Biochemistry 11*:4120–4131.

Compton LA, Johnson WC Jr. 1986. Analysis of protein circular dichroism spectra for secondary structure using a simple matrix multiplication. *Anal Biochem 155*:155–167.

Dousseau F, Pezolet M. 1990. Determination of the secondary structure content of protein in aqueous solutions from their amide I and amide II infrared bands. Comparison between classical and partial least-squares methods. *Biochemistry 29*:8771–8779.

Dukor RK, Keiderling TA. 1991. Reassessment of the random coil conformation. Vibrational circular dichroism study of proline oligopeptides and related polypeptides. *Biopolymers 31*:1747–1761.

Dukor RK, Pancoska P, Prestrelski S, Arakawa T, Keiderling TA. 1992. Comparison of FTIR and vibrational circular dichroism results for two protein hormones. Implications regarding secondary structure. *Arch Biochem Biophys 298*:678–681.

Freedman TB, Nafie LA, Keiderling TA. 1995. Vibrational optical activity of oligopeptides. *Biopolymers (Pept Sci) 37*. Forthcoming.

Greenfield N, Fasman GD. 1969. Computed circular dichroism spectra for evaluation of protein conformation. *Biochemistry 8*:4108–4116.

Grishina IB, Woody RW. 1995. Contribution of tryptophan side chains to the circular dichroism of globular proteins: Exciton couplets and coupled oscillators. *Faraday Discuss Chem Soc 99*. Forthcoming.

Gupta VP, Keiderling TA. 1992. Vibrational circular dichroism of the amide II band of proteins and model polypeptides in H_2O. *Biopolymers 32*:239–248.

Hennessey JP Jr, Johnson WC Jr. 1981. Information content in the circular dichroism of proteins. *Biochemistry 20*:1085–1094.

Johnson WC Jr. 1985. Circular dichroism and its empirical application to biopolymers. *Methods Biochem Anal 31*:61–163.

Johnson WC Jr. 1988. Secondary structure of proteins through circular dichroism spectroscopy. *Annu Rev Biophys Chem 17*:145–166.

Johnson WC Jr. 1990. Protein secondary structure and circular dichroism: A practical guide. *Proteins Struct Funct Genet 7*:205–214.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.

Keiderling TA. 1981. Vibrational circular dichroism. *Appl Spectrosc Rev 17*:189–226.

Keiderling TA. 1990. Vibrational circular dichroism comparison of technique and practical considerations. In: Ferraro JR, Krishnan K, eds. *Practical Fourier transform infrared spectroscopy industrial and laboratory chemical analyses*. San Diego, California: Academic Press. pp 203–284.

Keiderling TA. 1993. Vibrational circular dichroism of proteins, polysaccharides, and nucleic acids. In: Baianu IC, Pessen H, Kumosinski TF, eds. *Physical chemistry of food processes, advanced techniques, structures, and applications*. New York: van Nostrand Reinhold. pp 307–337.

Keiderling TA. 1994. Vibrational circular dichroism spectroscopy of peptides and proteins. In: Nakanishi K, Berova N, Woody RW, eds. *Circular dichroism principles and applications*. New York: VCH Publishers. pp 497–521.

Keiderling TA, Pancoska P. 1993. Structural studies of macromolecules using vibrational circular dichroism. In: Clark RJH, Hester RE, eds. *Biomolecular spectroscopy, part B*. Chichester, UK: Wiley. pp 267–315.

Keiderling TA, Wang B, Urbanova M, Pancoska P, Dukor RK. 1995. Empirical studies of protein secondary structure by vibrational circular dichroism and related techniques: α-Lactalbumin and lysozyme examples. *Faraday Discuss Chem Soc 99*. Forthcoming.

Keiderling TA, Yasui SC, Dukor RK, Yang L. 1989. Application of vibrational circular dichroism to synthetic polypeptides and polynucleic acids. *Polymer Preprints 30*:423–424.

LaBrake CC, Wang L, Keiderling TA, Fung LWM. 1993. Fourier transform infrared spectroscopic studies of the secondary structure of spectrin under different ionic strengths. *Biochemistry 32*:10296–10302.

Lee DC, Haris PI, Chapman D, Mitchell RC. 1990. Determination of protein secondary structure using factor analysis of infrared spectra. *Biochemistry 29*:9185–9193.

Levitt M, Chothia C. 1976. Structural patterns in globular proteins. *Nature 261*:552–558.

Levitt M, Greer J. 1977. Automatic identification of secondary structures in globular proteins. *J Mol Biol 114*:181–239.

Malinowski ER, Howery DG. 1980. *Factor analysis in chemistry*. New York: Wiley.

Manavalan P, Johnson WC Jr. 1987. Variable selection method improves the prediction of protein secondary structure from circular dichroism spectra. *Anal Biochem 167*:76–85.

Manning M. 1989. Underlying assumptions in the estimation of secondary structure content in proteins by circular dichroism spectroscopy—A critical review. *J Pharmaceut Biomed Anal 7*:1103–1119.

Manning M, Sreerama N, Woody RW. 1992. Improved models for tyrosine and phenylalanine contributions to the protein circular dichroism spectra. *Biophys J 61*:A347.

Manning M, Woody RW. 1989. Theoretical circular dichroism studies of polypeptide helices: Examination of important electronic and geometric factors. *Biopolymers 31*:561–586.

Mantsch HH, Casal HL, Jones RN. 1986. Resolution enhancement of infrared spectra of biological systems. In: Clark RJH, Hester RE, eds. *Biomolecular spectroscopy, vol 13*. London: Wiley & Sons. pp 1–46.

Pancoska P, Bitto E, Janota V, Keiderling TA. 1995. Quantitative analysis of vibrational circular dichroism spectra of proteins. Problems and perspectives. *Faraday Discuss Chem Soc 99*. Forthcoming.

Pancoska P, Blazek M, Keiderling TA. 1992. Relationships between secondary structure fractions for globular proteins. Neural network analyses of crystallographic data sets. *Biochemistry 31*:10250–10257.

Pancoska P, Fric I, Blaha K. 1979. Modified factor analysis of the circular dichroism spectra, applied to a series of cyclodipeptides containing L-proline. *Collect Czech Chem Commun 44*:1296–1312.

Pancoska P, Keiderling TA. 1991. Systematic comparison of statistical analyses of electronic and vibrational circular dichroism for secondary structure prediction of selected proteins. *Biochemistry 30*:6885–6895.

Pancoska P, Wang L, Keiderling TA. 1993. Comparison of protein FTIR absorption and vibrational circular dichroism. VCD frequency analyses in terms of secondary structure. *Protein Sci 2*:411–419.

Pancoska P, Yasui SC, Keiderling TA. 1989. Enhanced sensitivity to conformation in various proteins. Vibrational circular dichroism results. *Biochemistry 28*:5917–5923.

Pancoska P, Yasui SC, Keiderling TA. 1991. Statistical analyses of the vibrational circular dichroism of selected proteins and relationship to secondary structures. *Biochemistry 30*:5089–5103.

Perczel A, Hollosi M, Tusnady G, Fasman GD. 1991. Convex constraint analysis: A natural deconvolution of circular dichroism curves of proteins. *Protein Eng 4*:669–679.

Press WH. 1992. *Numerical recipes in C: The art of scientific computing, 2nd ed, version 20*. New York: Cambridge University Press.

Pribic R, van Stokkum IHM, Chapman D, Haris PI, Bloemendal M. 1993. Protein secondary structure from Fourier transform infrared and/or circular dichroism spectra. *Anal Biochem 214*:366–378.

Provencher SW, Glöckner J. 1981. Estimation of globular protein secondary structure from circular dichroism. *Biochemistry 20*:33–37.

Sarver RW, Krueger WC. 1991a. Protein secondary structure from Fourier transform infrared spectroscopy. *Anal Biochem 194*:89–100.

Sarver RW, Krueger WC. 1991b. An infrared and circular dichroism combined approach to the analysis of protein secondary structure. *Anal Biochem 199*:61–67.

Sen AC, Keiderling TA. 1984. Vibrational circular dichroism of polypeptides II. Solution amide II and deuteration results. *Biopolymers 23*: 1519–1532.

Sharaf MA, Illman DL, Kowalski BR. 1986. *Chemometrics*. New York: John Wiley.

Siegel JB, Steinmetz WE, Long GL. 1980. A computer assisted model for estimating protein secondary structure from circular dichroism spectra: Comparison of animal lactate dehydrogenases. *Anal Biochem 104*:160–167.

Sreerama N, Woody RW. 1993. A self-consistent method for the analysis of protein secondary structure from circular dichroism. *Anal Biochem 209*:32–44.

Sreerama N, Woody RW. 1994. Protein secondary structure from circular dichroism spectroscopy—Combining variable selection principle and cluster analysis with neural network, ridge regression and self-consistent methods. *J Mol Biol 242*:497–507.

Toniolo C, Crisma M, Bonora GM, Klajc B, Lelj F, Grimaldi P, Rosa A, Polinelli S, Boesten WHJ, Meijer EM, Schoemaker HE, Kamphuis J. 1991. Structure of peptides from α-amino acids methylated at the α-carbon. *Int J Pept Protein Res 38*:242–256.

Toumadje A, Alcorn SW, Johnson WC Jr. 1992. Extending CD spectra of proteins to 168 nm improves the analysis for secondary structures. *Anal Biochem 200*:321–331.

Urbanova M, Pancoska P, Keiderling TA. 1993. Spectroscopic study of the temperature-dependent conformation of glucoamylase. *Biochim Biophys Acta 1203*:290–294.

van Stokkum IHM, Spoelder HJW, Bloemendal M, van Grundelle R, Groen FCA. 1990. Estimation of protein secondary structure and error analysis from circular dichroism spectra. *Anal Biochem 191*:110–118.

Venyaminov SYu, Kalnin NN. 1990. Quantitative infrared spectroscopy of peptide compounds in water (H_2O) III: Estimation of the protein secondary structure. *Biopolymers 30*:1273–1280.

Williams RW. 1986. Protein secondary structure analysis using Raman amide I and amide III spectra. *Methods Enzymol 130*:311–331.

Yang JT, Wu CSC, Martinez HM. 1986. Calculation of protein conformation from circular dichroism spectra. *Methods Enzymol 130*:208–269.

Yasui SC, Keiderling TA. 1986. Vibrational circular dichroism of polypeptides VII. Film and solution studies of B-forming homo-oligopeptides. *J Am Chem Soc 108*:5576–5581.

Yasui SC, Keiderling TA, Bonora GM, Toniolo C. 1986a. Vibrational circular dichroism of polypeptides V. A study of 3_{10} helical octapeptides. *Biopolymers 25*:79–89.

Yasui SC, Keiderling TA, Formaggio F, Bonora GM, Toniolo C. 1986b. Thermolysis of 1R,2R-1,2-dideuteriocyclobutane. An application of vibrational circular dichroism to kinetic analysis. *J Am Chem Soc 108*:4988–4993.

Yoder G, Keiderling TA, Crisma M, Formaggio F, Toniolo C. 1995. Characterization of β-bend ribbon spiral forming peptides using electronic and vibrational circular dichroism. *Biopolymers 35*:103–111.