Predictions of Secondary Structure Using Statistical Analyses of Electronic and Vibrational Circular Dichroism and Fourier Transform Infrared Spectra of Proteins in H2O

ARTICLE *in* JOURNAL OF MOLECULAR BIOLOGY · JULY 1996

Impact Factor: 4.33 · DOI: 10.1006/jmbi.1996.0357 · Source: PubMed

3 AUTHORS:

Vladimir Baumruk
Charles University in Prague
88 PUBLICATIONS **1,901** CITATIONS
SEE PROFILE

Petr Pancoska
University of Pittsburgh
101 PUBLICATIONS **3,078** CITATIONS
SEE PROFILE

Timothy A Keiderling
University of Illinois at Chicago
276 PUBLICATIONS **7,305** CITATIONS
SEE PROFILE

J. Mol. Biol. (1996) **259**, 774–791

JMB

Predictions of Secondary Structure using Statistical Analyses of Electronic and Vibrational Circular Dichroism and Fourier Transform Infrared Spectra of Proteins in H_2O

Vladimir Baumruk, Petr Pancoska and Timothy A. Keiderling*

Department of Chemistry
University of Illinois at
Chicago, 845 W. Taylor St.
Chicago, IL 60607-7061 USA

Vibrational circular dichroism (VCD) and Fourier transform IR (FTIR) methods for prediction of protein secondary structure are systematically compared using selective regression analysis. VCD and FTIR spectra over the amide I and II bands of 23 proteins dissolved in H_2O were analyzed using the principal component method of factor analysis (PC/FA) and regression fits to fractional components (FC) of secondary structure. Predictive capability was 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 (RMR) of PC/FA coefficients with FC values, both independently for each spectral data set as well as for the VCD and FTIR sets grouped together and with similarly obtained electronic CD (ECD) data. The distribution of predictive error for a set of the best RMR relationships that use a given number of spectral coefficients was used to select the optimal prediction algorithm. Minimum predictive error resulted for a small subset (three to six) of spectral coefficients, which is consistent with our earlier findings using VCD measured for proteins in ^2H_2O and ECD data. Subtracting the average absorption spectrum from all the training set FTIR spectra before analysis yields more variance in the FTIR band shape and improves the predictive ability of the best PC/FA RMR to near that for the VCD. Both methods (FTIR and VCD) using data for proteins in H_2O are somewhat better predictors than amide I' (in ^2H_2O) VCD alone and, for helix, worse than ECD alone. Combining FTIR and VCD data did not dramatically change the prediction results. Predictions are improved by combining both with ECD data, indicating that the improvement is due to using their very different structural sensitivities. The coupled H_2O-based spectral analyses and the mixed amide I' + II VCD plus ECD analysis are comparable for the helix and sheet components, indicating that partial deuteration is not a major source of prediction error.

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

**Corresponding author*

Permanent addresses: V. Baumruk, Institute of Physics, Charles University, Ke Karlovu 5, Prague, Czech Republic; P. Pancoska, Department of Chemical Physics, Charles University, Ke Karlovu 3, Prague 2, Czech Republic.

Abbreviations used: FTIR, Fourier transform IR; DF, difference FTIR; ECD, UV electronic circular dichroism; FC, fractional contributions to secondary structure; H, E, S, T, and C, helix, sheet, bend, turn and ''other'', respectively; RMR, restricted multiple regression; *k*, number of coefficients considered in RMR; KS, 1983, Kabsch & Sander (1983); MCT, $Hg_x\ Cd_{1-x}\ Te$ IR detector; PC/FA, principal component method of factor analysis; PDB, Protein Data Bank; *rr*, regression coefficients; σ and σ^{rel}, standard and relative standard deviation; S/N, signal-to-noise ratio; VCD, vibrational circular dichroism.

0022–2836/96/240774–18 $18.00/0

Introduction

In recent years there has been a rebirth of various statistical techniques to derive secondary structural parameters from optical spectroscopic data, such as UV electronic circular dichroism (ECD), Fourier transform IR (FTIR), vibrational CD (VCD) in the IR, and Raman spectra (Manavalan & Johnson, 1987; Perczel *et al*., 1991; van Stokkum *et al*., 1990; Pancoska & Keiderling, 1991; Toumadje *et al*., 1992; Sreerama & Woody, 1993, 1994; Lee *et al*., 1990; Dousseau & Pezolet, 1990; Venyaminov & Kalnin, 1990; Sarver & Krueger, 1991a,b; Pribic *et al*., 1993; Pancoska *et al*., 1991, 1994, 1995; Williams, 1986; Berjot *et al*., 1987; Bussian & Sander, 1989). These optical techniques are particularly useful for fast, universal, qualitative estimation of average secondary structure and determination of structural integrity during some biochemical process or environmental change in proteins. The ability of such spectral techniques to sense small changes in protein conformation is well established (Keiderling *et al*., 1995; Keiderling & Pancoska, 1993). However, quantification of the structural distribution in the protein in terms of fractional contributions (FC) of specific secondary structural types to the overall conformation is less reliable. We have recently reported an extensive study of the reliability of such methods and an in-depth analysis of our own approach, which there used a restricted multiple regression (RMR) method as applied to ECD and VCD data (Pancoska *et al*., 1994, 1995). An important part of our method was its facility for combining data from different techniques by first processing their spectral representation with the principal component method of factor analysis (PC/FA) to effect a simplification of the mathematical form of these spectra, thus facilitating further empirical calculations used to extract the FC parameters.

The major conclusion of our previous work was that, while using data from more than one technique and from more than one transition (in the case of VCD) led to enhanced prediction accuracy, it was at the same time necessary to reduce the number of spectral components used from each spectral data set in order to achieve the most reliable predictions. Using variations of these analysis methods (which are related to many, now standard, methods used for protein ECD analyses (Hennessey & Johnson, 1981; Provencher & Glöckner, 1981; Johnson, 1985, 1988; Manning, 1989)), other groups have reported similar findings in both respects. Sarver & Krueger (1991b) have combined FTIR and ECD data for improved fits and van Stokkum and co-workers (van Stokkum *et al*., 1990; Pribic *et al*.; 1993) have found that a reduction in dimension aids their single and combined spectral analyses. This improvement in quality with a reduction in dimension implies that the typical structure-spectral analyses under-utilize the information available from optical spectroscopic data. By extension, these results suggest that optical spectroscopic data could yield added detail about the protein structure, if it were used more optimally.

Going further in this direction, the lack of dramatic improvement in structure prediction, despite all this method development, for example as recently compared by Sreerama & Woody (1994) for ECD analyses, has led us to propose that such methods must face an intrinsic limitation (Pancoska *et al*., 1994, 1995; Keiderling, 1996). One aspect of this limitation should be fundamentally due to the variance in detailed structure for those parts of the protein traditionally assigned to a given secondary structural type. Such structural variations could alter the spectral response without affecting the structural descriptor (FC), leading to error in the analyses. In other words, the problem in these analyses seems to be endemic to the question itself (e.g. average fractional secondary structure) that is typically being posed. Our previous studies were based on ECD and VCD (in ^2H_2O) data, but we here compare FTIR and VCD data for proteins in H_2O to determine if the same qualities of limitations of the improvement in prediction and its dependence on only a few components of the spectra still hold. Since all the needed data and methods are available, we also test which combination of techniques, i.e. VCD, FTIR and/or ECD, gives the best, even if limited, predictions. More generally, this broader-based analysis can test the contention that the errors found in predicting fractional components of secondary structure are intrinsic to the protein structures rather than being properties of the data measured or the technique used. Finally, we address the problems (or benefits) of partial deuteration of the amide groups when proteins are studied in ^2H_2O rather than H_2O.

The RMR techniques based on the PC/FA that we have developed for VCD and ECD are fully adaptable to any type of spectrum provided some facet of the resulting band shapes has a dependence on the structure. This flexibility in our methods is due, in part, to its use of the numerical part of the PC/FA spectral decomposition (the coefficients) rather than the component band shapes (the subspectra) to establish a structural correlation. FTIR data have been shown to be highly correlatable to secondary structure and have a resolution and signal-to-noise ratio (S/N) advantage over ECD and VCD (Mantsch *et al*., 1986; Surewicz *et al*., 1993). There is also some enhancement for FTIR and VCD over ECD in terms of sensitivity to sheet structure, while ECD seems to be superior to both FTIR and VCD in terms of helix prediction. This naturally suggests that combining data from the two techniques may have an advantage, as was seen for VCD and ECD (Pancoska *et al*., 1995). The analyses of vibrational spectroscopies (FTIR or Raman) in terms of structure have often depended on assignment of frequencies to components of a broad transition band, which 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). The band shape analysis approach is less sensitive to the frequencies than to the relative intensity distribution in the spectrum (which, of course, depends on frequency to some extent). However, the FTIR spectrum measured for a protein in the amide I and amide II bands has a smooth band profile with little variance as compared with that characteristic of protein ECD and VCD spectra. The latter are intrinsically differential spectroscopies, which gives them remarkable sensitivity to small structural changes. By contrast, FTIR is a ''summed'' dipolar spectroscopy that, when used for study of the amide I and II bands in particular, effectively adds up the contribution of a large number of very similar dipole oscillators. We present here a modified approach using a differential representation of the experimental spectra to optimize the application of our PC/FA methods to single-signed and thus less variable band shapes such as seen in FTIR and Raman spectra.

Most early FTIR and VCD studies of proteins were done in ^2H_2O solution in order to enhance the signal to noise ratio (S/N) and allow the use of cells with convenient path-lengths. They consequently focused on spectra of the amide I' band (primarily the C = O stretch of the amide group when N-H is deuterium-substituted), but not all the amide groups are deuterated if the protein is just dissolved in ^2H_2O or even H-^{2}H exchanged to an equilibrium state, as has been our past practice (Pancoska *et al*., 1989). The water bending deformation mode occurs at $\sim$1650 cm^{-1} and has an absorbance >1 for paths >10 μm, which would effectively block reliable measurement of the amide I band shape for proteins in H_2O. However, with modern instrumentation it is possible to obtain FTIR data routinely with a high level of accuracy for proteins at sufficiently high concentrations in H_2O when sampled using a 6 μm path-length cell. We have demonstrated our ability to determine VCD spectra under similar conditions of high concentration and short path-lengths (Baumruk & Keiderling, 1993). Under such conditions the amide I and II bands (as well as lower-energy transitions, which will not be addressed further here) can be measured simultaneously on the same sample and under the same conditions. Furthermore, proteins measured under these conditions do not have the potential difficulties that might result from partial H-^{2}H isotopic exchange. However, the required concentrations are quite high, the signal averaging times required are long, and the interference and possible FTIR band shape distortion from the water absorbance (which must be subtracted) is high. Comparison of the results of our RMR PC/FA structural analysis methods on such H_2O-based protein data with those obtained previously should directly address the cost to the analysis of the convenience of measuring spectra in ^2H_2O or, conversely, the costs for using the seemingly simpler, single sample/single scan approach possible for proteins in H_2O. Furthermore, comparison of FTIR and VCD data for the same samples should test the effects of both improved resolution and structural sensitivity for predictions of FC secondary structure parameters.

Here, we present a systematic comparison of the prediction of protein secondary structure with amide I and II dispersive VCD and FTIR absorption data collected systematically on the same samples of 23 proteins in H_2O. Since 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, we systematically test the relative predictive ability of these multi-component regression models for proteins not included in the RMR development. An important aspect of our work is the evaluation of a series of restricted regressions that demonstrate that the optimal predictive ability of these analyses is obtained with only part of the spectral data. While these aspects of our study have been addressed before, this is their first application to H_2O VCD data and the first comparison of VCD and FTIR data as well as the combination of the two. Furthermore, some new method enhancements were implemented in addressing some computational problems involved in extending our methods to this study. The prime effort here addresses the problem of choosing the most reliable regression relationships, or a set of them, for purposes of structure prediction from the multitude of possible choices available, especially when using data from different techniques. Our focus is to understand, or at least highlight, the main sources of error in these analyses so that we might support any conclusions drawn about the general methods of quantitative spectra-structure correlation. Our comparative approach is designed to identify the aspects of prediction that are technique-related and those intrinsic to the protein.

Results

Factor analysis results

PC/FA decompositions were done independently for the VCD and FTIR spectra of the amide I + II bands for 23 proteins. From comparison of reconstructed spectra to the experimental spectra, we retained ten orthogonal components for the VCD and ten for the FTIR spectra to reproduce the respective experimental spectra. These subspectral components are given in Figures 1 and 2. In the DF analysis, we retained 11 subspectra (shown in Figure 3). We feel that with our selective RMR methods, it is better to choose too many rather than too few coefficients, particularly for the VCD spectra. Those with no structural significance will not be retained in the optimal RMR relationships. Thus we have not striven to find the absolute minimum number of subspectra needed to match the experimental data, but rather have made sure that we encompassed enough of them in our analysis to reproduce all the important spectral

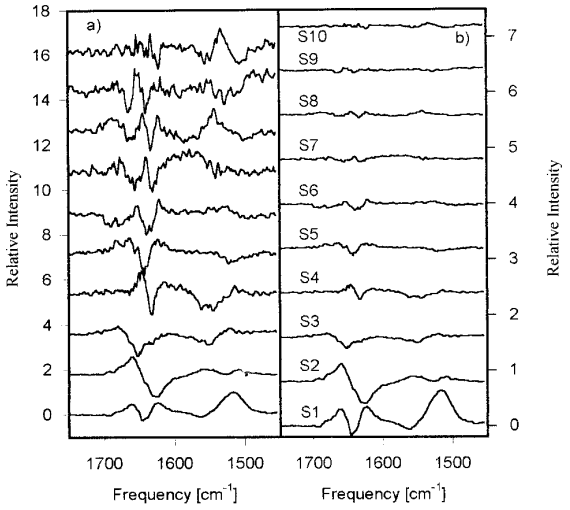

Figure 1. PC/FA protein amide I + II (H_2O) VCD subspectra represented as (a) normalized to the maximum amplitude (relative units, to illustrate shape) and (b) constant amplitude (to illustrate relative importance). Note the decrease in S/N for higher subspectra.

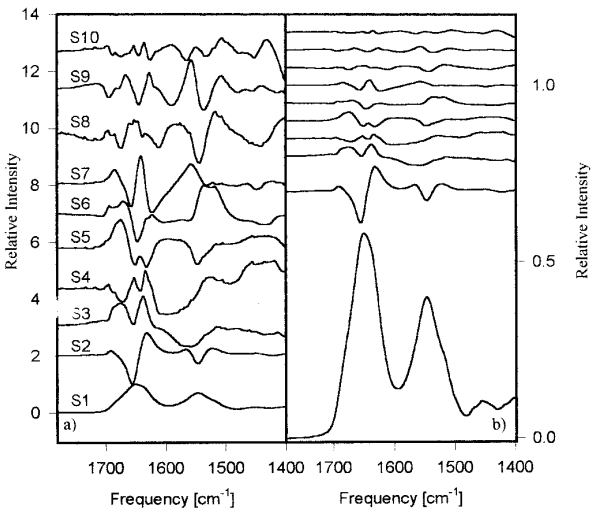

Figure 2. PC/FA protein amide I + II (H_2O) FTIR subspectra represented as (a) normalized to the maximum amplitude (relative units, to illustrate shape) and (b) constant amplitude (to illustrate relative importance). Note the very large relative intensity of S1, which is of little structural consequence.

features for each technique. In our previous study, we retained six subspectra for the amide I' PC/FA analysis and six more (which were not as significant due to S/N limitations) for the amide II, while Lee *et al*. (1990) and Pribic *et al*. (1993) have retained as many as 11 for their FTIR spectral analyses. These studies could be viewed as being consistent with the number of subspectra retained here, but their significance and effects on the analyses will be addressed below. For the sake of comparison, PC/FA reductions of our previously obtained ECD spectra and amide I' VCD spectra in ^2H_2O (Pancoska *et al*., 1995) for the same 19 training set proteins used here were re-analyzed so that the comparison between techniques (ECD) and environments (amide I') would be on the same basis.

The VCD subspectral intensities can be roughly divided into three groups with the first two (S1, S2) being most significant (having the largest contributions to the reconstructed spectra), the next four (S3 to S6) having reasonable intensities, and the final four (S7 to S10) being quite weak. This is shown in Figure 1 (right side), where the subspectra are all plotted on the same scale. This grouping is naturally reflected in the noise content of each subspectrum, with the first two being relatively smooth while the last four are fairly noisy, as illustrated in the left side of Figure 1, where the subspectra are plotted all with the same amplitude to better illustrate their variance in shape. The first subspectrum is dominated by a three-feature pattern (–, +, –) in the amide I and a single signed negative† in the amide II with a maximum excursion at 1517 cm^{-1}. While the former band shape is qualitatively consistent with a mixed helix-sheet structure, the latter is more associated with proteins with a high helical content proteins. However, since loss of helix typically involves an overall loss of amide II intensity, the helix contribution dominates the amide II intensity of proteins and thus dominates the first subspectrum, much like the situation for ECD (Hennessey & Johnson, 1981; Pancoska & Keiderling, 1991). On the other hand, the second subspectrum has a dominant couplet in the amide I with much less contribution to the amide II, the shape being somewhat like one expects for proteins with a high sheet content. Since the second subspectrum represents the major variance in the set from the average, contributions of the second subspectrum, when combined with the first subspectrum, shift the pattern from the average, mixed helix/sheet band shape to one characteristic of less structured and more β-form-containing proteins.

As is obvious from the plot of the FTIR subspectra in Figure 2 (right side), the FTIR subspectra are totally dominated in intensity by the first subspectrum, which looks like a typical protein absorbance spectrum, as should be expected, since there is so little variation among the FTIR absorbance spectra for globular proteins in H_2O. The correlation matrix coefficients (overlap integrals) used to carry out the PC/FA (Malinowski & Howery, 1980; Pancoska *et al*., 1979) reflect this similarity, being all large, close to 0.9. Since this might have introduced some numerical problems in the diagonalization required to generate the subspectra and coefficients needed for the regression tests, the difference FTIR (DF) calculations

† It should be clear that the sign of a particular subspectrum is arbitrary. Its contribution to reconstruction of the experimental spectrum is controlled by the sign and magnitude of the respective coefficient.

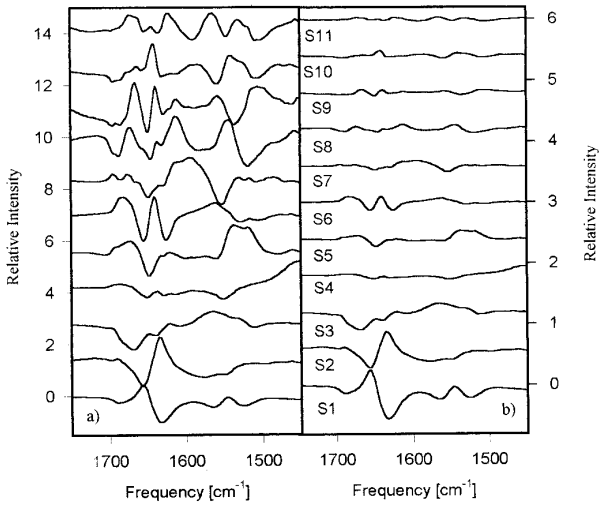

Figure 3. PC/FA protein amide I + II (H_2O) DF subspectra (obtained from the FTIR data set with the average spectrum presubtracted) represented as (a) normalized to the maximum amplitude (relative units, to illustrate shape) and (b) constant amplitude (to illustrate relative importance). Note the similarity of S1 to S2 of Figure 2.

were carried out with the average absorbance spectrum pre-subtracted from each training set spectrum. This DF representation is mathematically more like the VCD, which is a true difference spectroscopy, and consequently has much more variation in its correlation matrix. While other algorithms could be chosen to transform the FTIR data, subtraction of the average intensity seemed consistent with the PC/FA method and avoids reduction of the order of the training set used in the analysis for purposes of comparison.

In these DF calculations, there is some similarity of the subspectra (Figure 3) to those with different indices (based on relative magnitude of the corresponding eigenvalues) in the FTIR subspectra. Thus the first subspectrum in the DF set (S1, Figure 3) and the second in the FTIR (S2, Figure 2) are dominated by a derivative shape centered over the amide I and three features of alternating sign pattern over the amide II. These separately serve to shift the frequency of the amide I band center and to change the band-width of the amide II. The second DF subspectrum is opposite in phase to the first for the amide I and more of a single signed broad feature on the high-energy side of the amide II, while the third FTIR subspectrum (S3, Figure 2) is similar for the amide II but has two of the same signed peaks for the amide I, thus resembling the third DF subspectrum (S3, Figure 3). For the fourth and further subspectra, the intensities are very low even for the DF set, but there are further correspondences between the two sets of subspectra. At this level, it seems the DF approach has divided the intense differential aspects of the spectra into more components than does the unmodified FTIR analysis. While with the FTIR analysis the highest subspectra have a vanishingly small intensity compared with the first one, in the DF analysis the change is less than an order of magnitude. But even for the higher-order DF subspectra, the variations seen with wave number are significantly above the noise level. This is in contrast to the higher-order VCD subspectra, where the S/N level deteriorates more rapidly. Even with ECD data, if subspectra as high as ten were considered, a substantial noise contribution would be evident, which is consistent with the observations regarding ECD analyses reported by Hennessey & Johnson (1981), that only about five subspectra above the noise level are significant. The continuing significance of these higher-order subspectra reflects the extremely good S/N obtainable in modern FTIR spectra using these short-path, high-concentration measurement techniques.

Correlation of spectra and structure and development of RMR equations

In our approach, the next step is to systematically search through all possible combinations of spectral coefficients from this complete analysis of our training set to identify the best fitting RMR of each order k with each type of data. It is important to realize that this search for correlations is complete; for example, for $k = 2$, we choose all possible pairs with no restriction as to ordering; and we do this independently for each structural type, helix, sheet, etc. For a given number, k, of coefficients considered, several regressions based on different combinations of coefficients were found to have virtually equal *rr* values. These combinations often had in common the particular coefficient that was found to be the most important when only one coefficient was used to determine a fit, i.e. the best $k = 1$ regression. For example, the restricted regression fits for α-helix using the VCD data set all tend to depend on the coefficient of the second (S2) subspectrum and, using FTIR, also on the S2 subspectrum, but with the DF representation the S1 subspectrum was more important, as would be expected from its relationship to the FTIR. Since for $k > 1$, several ''better'' RMR relationships containing these coefficients, or their counterparts for other structural types, were nearly equivalent, we retained four or five of the best ones to test prediction capability.

The complete RMR analyses were next recalculated for 19 different protein data sets, each encompassing the same type of spectra for 18 of the training set proteins, one protein having been sequentially omitted to create the 19 sets. The input to all of these calculations were the same, systematically obtained PC/FA coefficients; the difference from the initial calculation being only that a smaller training set (18 proteins) was used for the RMR. Each of these sets was tested for prediction error of each of the structural parameters for the protein left out by optimizing just the four or five best regression forms for a given number of

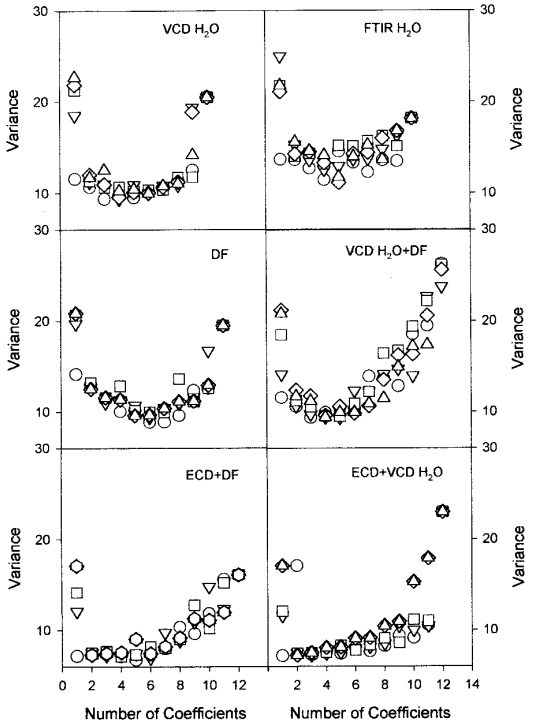

Figure 4. Effect on the standard deviation or variance of prediction, for the protein left out (calculated successively over the entire training set), of increasing the number, *k*, of subspectral coefficients in the restricted regression tests for the helix (H) fractional component (FC) with six different data sets, as indicated in the Figure. The combined data sets were constrained to 12 for simplification of the calculations. The different symbols represent the σ values obtained with different regression relations of the same *k* value. The circle in each case indicates σ for the regression relation having the highest *rr* for the entire set.

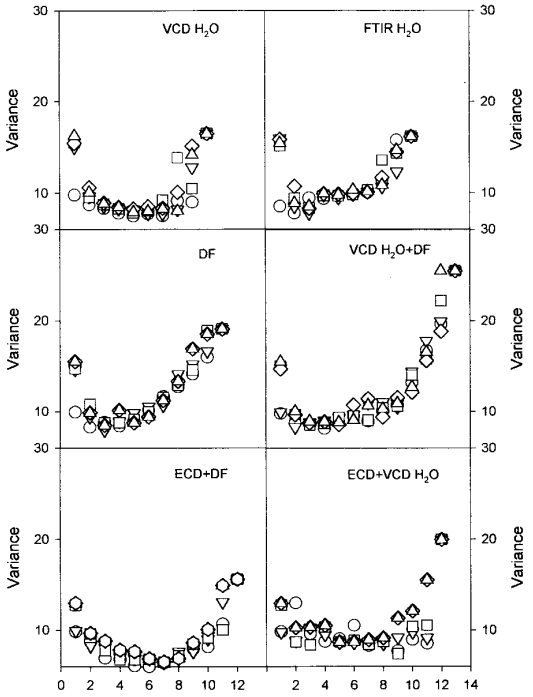

Figure 5. Effect on the standard deviation or variance of prediction, for the protein left out (calculated successively over the entire training set), of increasing the number, *k*, of subspectral coefficients in the restricted regression tests for the sheet (E) fractional component (FC) with six different data sets, as indicated in the Figure. Details as in Figure 4.

spectral coefficients, i.e. those that had the highest *rr* value for that structural type in the previous series of full training set structure-spectra fit optimizations. As in our previous study, an enormous amount of numbers requiring comparison resulted. We have found that a graphical representation of the variation of the prediction error with *k* is the most useful way of comprehending the relative reliability of these predictions. Examples of these are shown in Figures 4 and 5, wherein, for each *k*, the variance averaged over the 19 protein structure predictions for a specific structural type is indicated with different symbols representing the best, second best etc., regression form based on the full training set *rr* value. The minima found in the variance with *k* indicate the optimally predicting *k* and the tightness of the variance dispersion for a given *k* indicates the reliability or stability of the predictions. The optimal predictions obtained with the data from each technique are then summarized in Table 1 in terms of the standard (σ) and relative (σ^{rel}) deviations (as compared with the dynamic range of each structural parameter) of what we judge to be the best predictions of the 19 protein secondary structures from actual X-ray values derived from the KS algorithm.

An important characteristic of our method is the ability to combine different data sets with a goal of improving the analysis by inclusion of a broader variety of spectral transitions or sensitivities. In the VCD plus ECD study (Pancoska *et al*., 1995), this proved to provide the analysis with a significant self-correcting ability so that the excursions of error were greatly restricted. In Table 2 are summarized the same sort of average errors for combined analyses using the VCD + FTIR or DF sets and the ECD + DF or VCD sets. To carry out these combinations we needed to restrict the number of coefficients considered due to computer limitations. While in principle the method is not restricted, combining sets leads to an enormous growth in the possible combinations that we should consider. Truncating the coefficient vectors is somewhat arbitrary, but we are considering the most significant aspects of the data.

Again, as seen previously (Pancoska *et al*., 1991, 1994, 1995), two classes of prediction evolve: helix and sheet as compared with the rest. For the three new data sets reported here (Table 1), VCD, FTIR and DF, all but the FTIR calculations have helix (H) and sheet (E) predictions of comparable quality (∼12% and ∼16% of the dynamic range, respectively). By comparing relative standard deviations, it

Table 1. Standard and relative deviations for predictions of fractional secondary structures using individual different spectroscopic data sets

Technique	Helix	Sheet	Turn	Bend	Other
FTIR full					
σ	11.2	7.8	3.2	2.5	4.8
σ_{rel} [%][a]	14.5	16.3	22.9	13.6	24.7
k[b]	5	3	3	8	4
DF					
σ	8.9	8.0	3.2	2.8	4.8
σ_{rel} [%][a]	11.5	16.7	23.0	14.9	25.0
k[b]	6	3	5	8	3
VCD in H_2O					
σ	9.4	7.5	3.8	3.9	3.7
σ_{rel} [%][a]	12.1	15.7	27.1	20.8	19.3
k[b]	4	5	2	5	5
VCD H_2O AI only					
σ	10.8	8.4	4.0	3.7	3.1
σ_{rel} [%][a]	13.9	17.7	28.4	19.7	16.2
k[b]	4	4	2	2	2
VCD H_2O AII only					
σ	10.2	8.5	3.5	3.4	4.8
σ_{rel} [%][a]	13.2	17.8	24.8	18.4	24.7
k[b]	3	4	3	1	3
ECD					
σ	7.2	8.2	3.5	3.4	3.8
σ_{rel} [%][a]	9.3	17.1	24.9	18.6	19.5
k[b]	1	3	2	2	3
VCD in ^2H_2O					
σ	12.1	7.7	4.1	4.0	5.4
σ_{rel} [%][a]	15.7	16.1	29.2	21.7	28.2
k[b]	2	3	1	1	2

[a] Relative standard deviations (calculated as standard deviations compared with the dynamic range of each structural parameter).

[b] Number of coefficients used in the optimal regression function for prediction.

Table 2. Standard and relative deviations for optimal predictions of secondary structures using combinations of different spectroscopic data

Techniques	Helix	Sheet	Turn	Bend	Other
VCD H_2O + FTIR					
σ	9.3	8.2	2.8	3.1	3.7
σ_{rel} [%]	12.1	17.2	20.0	16.8	19.2
k	5	3	6	8	7
VCD H_2O + DF					
σ	9.4	8.2	3.0	2.9	3.7
σ_{rel} [%]	12.2	17.1	21.2	15.6	19.0
k	5	4	6	6	4
VCD H_2O AI + AII					
σ	9.4	7.0	3.3	2.9	3.0
σ_{rel} [%]	12.2	14.6	23.4	15.6	15.6
k	3	5	4	5	3
ECD + VCD H_2O					
σ	7.2	7.4	3.2	3.1	2.5
σ_{rel} [%]	9.3	15.5	22.9	17.0	13.0
k	1	9	3	3	9
ECD + DF					
σ	6.7	6.0	3.2	3.1	3.8
σ_{rel} [%]	8.7	12.6	22.9	17.0	19.5
k	5	6	5	6	3

See Table 1 for an explanation of the symbols.

is seen that the bend (S) is also roughly predicted in FTIR and DF analyses while the ''other'' (C) is reasonably predicted in VCD analyses. Surprisingly, if the VCD of the amide I and II bands are analyzed separately, the helix and sheet errors are about identical with each other and about 2% above the combined analysis σ^{rel} results. By comparison, for this smaller training set (Table 1), our previous amide I' VCD yields prediction errors somewhat higher for the helix while the ECD has much lower errors than seen in the newer H_2O data (VCD or FTIR). In both these cases, the errors for the sheet component are about the same as the newer data.

When the VCD and FTIR analyses were combined (Table 2), the difference between the FTIR and DF approaches was diminished, giving both about the same level of accuracy for helix prediction. For this combination of techniques, the bend and other fractions were both reasonably predicted. The turn error was improved but remained high, >20% when viewed relative to its dynamic range. As in our previous study, further improvement comes when ECD data are coupled to the vibrational spectra. In this case, coupling ECD to either the VCD (amide I + II in H_2O) or the DF representation of the FTIR of the same transitions gives a clear improvement for the helix and sheet. They are about the same for helix, the ECD + DF is a bit better for sheet and the ECD + VCD is better for ''other'. Neither of these combinations is as good at predicting helix and sheet as is the combination used in our previous work (Pancoska *et al*., 1995) of ECD + VCD (amide I' in ^2H_2O) + VCD (amide II in H_2O). Since helix and sheet are of prime interest in prediction methods, we list in Tables 3 and 4 the errors in predicting these two structural types for each protein with each individual model and with the combined models, respectively. Scanning down the columns, one is immediately struck by the smaller values (with a few notable exceptions) in Table 4 when ECD is combined with VCD or FTIR data, much as we reported earlier with ^2H_2O-based data (Pancoska *et al*., 1995).

In Figures 4 and 5 the error in prediction of helix and sheet is plotted as a function of the number of coefficients used in the RMR with the circles plotted at each value of *k* indicating the prediction error found with the RMR model having the highest *rr*. While in most cases the best-fitting RMR relations gave the best prediction, that was not always true. This extension of our algorithm avoids missing those cases where an alternative relation is a better predictor. More importantly, it is clear again, as was seen previously for the ECD and VCD in ^2H_2O data (Pancoska *et al*., 1995), that these errors do not in general decrease as arbitrarily more coefficients are added to the scheme; but, in fact, the error always goes through a minimum and then increases with increasing *k*. Often substantial prediction errors are obtained when as many as ten or more coefficients are used for structure prediction. This is consistent with not only our findings for VCD and ECD but also with those of van Stokkum and co-workers for ECD and FTIR data (van Stokkum *et al*., 1990; Pribic *et al*., 1993). The difference in our approach is the selectivity in our RMR, which is not at all restricted as to which coefficients are kept in the final relation. In comparing our different data sets, it is seen that

Table 3. Errors in predictions of α-helix and β-sheet fractions from optimal models for individual spectroscopic methods

	Helix							Sheet						
Protein name	FTIR full	DF	VCD H_2O	VCD AI	VCD AII	VCD ^2H_2O	ECD	FTIR full	DF	VCD H_2O	VCD AI	VCD AII	VCD ^2H_2O	ECD
α-Chymotrypsinogen A	2.4	1.9	1.2	2.7	4.9	11.4	1.4	0.6	2.1	3.8	8.8	0.7	0.2	0.2
Alcohol dehydrogenase	7.3	3.5	9.7	14.0	0.9	1.6	2.0	8.9	12.5	7.2	7.9	1.7	3.7	4.5
α-Chymotrypsin type II	1.7	0.5	7.9	6.6	6.5	4.3	1.0	1.3	0.4	1.3	3.7	2.1	6.0	1.0
Concanavalin A	16.8	5.2	3.9	9.2	12.7	16.3	12.7	4.3	4.1	9.6	4.4	12.2	13.0	10.4
Carbonic anhydrase	5.1	2.3	13.9	11.3	13.1	2.4	8.7	6.5	5.3	11.9	13.5	7.9	1.0	10.6
Cytochrome *c*	6.6	10.8	1.8	7.3	2.7	7.3	12.2	13.8	15.7	1.9	11.9	4.1	1.9	19.2
Glutathione reductase	1.5	9.5	2.5	6.3	2.4	4.2	1.9	1.2	3.7	6.4	6.0	3.2	6.6	0.9
Hemoglobin	1.9	0.8	2.5	2.9	2.0	0.1	1.1	2.9	0.8	10.3	11.3	10.3	4.8	10.0
λ-Immunoglobulin	13.5	13.1	2.0	3.1	3.4	14.9	2.8	13.8	17.2	10.0	12.9	7.4	15.0	2.5
Lactate dehydrogenase	5.6	13.7	4.5	9.5	5.6	23.3	4.1	4.6	3.8	3.9	11.7	0.7	7.3	2.2
Lysozyme	22.4	12.1	0.2	23.3	1.0	12.1	4.5	2.4	5.0	1.8	9.2	3.3	14.1	0.7
Myoglobin	18.6	16.5	15.8	18.2	17.4	28.5	5.3	3.3	0.5	3.5	0.9	2.9	4.7	11.8
Ribonuclease A	1.6	13.4	7.6	3.4	10.1	9.6	1.8	8.1	9.3	9.6	10.6	9.3	4.1	10.1
Ribonuclease S	9.7	2.1	6.9	2.7	13.6	10.7	2.2	0.1	7.7	5.6	3.2	12.8	1.4	6.0
Subtilisin BPN'	0.7	3.8	7.6	4.7	13.7	0.8	7.4	3.3	0.7	5.9	1.8	17.4	13.0	4.3
Superoxide dismutase	25.2	6.0	12.7	15.0	3.1	2.3	17.3	16.5	10.2	6.0	3.8	8.3	3.9	5.2
Triose phosph.isomerase	1.1	4.8	0.1	1.1	1.8	9.8	1.0	10.7	2.9	4.5	1.7	6.5	2.8	10.5
Trypsin inhibitor	0.9	7.1	21.4	16.6	22.9	6.7	9.9	7.0	8.6	14.2	7.2	14.5	4.5	0.2
Trypsin	5.0	8.3	11.8	1.9	9.6	8.1	0.5	0.2	2.5	3.4	6.3	2.0	5.8	6.9

Table 4. Errors in predictions of α-helix and β-sheet fractions from optimal models for combinations of spectroscopic methods

	Helix					Sheet				
Protein name	ECD+ VCD H_2O	ECD+DF	VCD H_2O +DF	VCD ^2H_2O +DF	VCD H_2O AI+AII	ECD+ VCD H_2O	ECD+DF	ECD+ FTIR	VCD ^2H_2O+ DF	VCD H_2O AI+AII
α-Chymotrypsinogen A	1.4	0.8	1.5	3.1	5.5	1.7	2.8	1.7	7.2	6.5
Alcohol dehydrogenase	2.0	1.5	2.9	5.6	2.6	2.7	7.2	0.2	1.8	8.3
α-Chymotrypsin type II	1.0	3.5	1.6	2.2	4.2	2.8	6.6	1.2	2.5	0.2
Concanavalin A	12.7	7.3	1.8	8.9	4.6	9.9	3.9	2.3	2.2	3.9
Carbonic anhydrase	8.7	10.4	10.9	9.7	16.4	12.4	8.5	9.8	10.1	11.9
Cytochrome *c*	12.2	5.7	4.8	5.9	6.6	3.1	8.4	6.9	8.0	0.2
Glutathione reductase	1.9	5.1	0.0	2.5	3.3	4.0	6.8	4.4	2.9	0.6
Hemoglobin	1.1	3.6	1.1	4.8	3.3	4.6	10.5	3.4	5.2	10.1
λ-Immunoglobulin	2.8	8.0	1.4	7.8	0.6	4.0	9.5	16.8	17.0	1.6
Lactate dehydrogenase	4.1	0.0	4.1	13.6	2.1	6.2	3.5	5.1	5.2	4.7
Lysozyme	4.5	1.9	0.4	9.5	6.2	6.2	2.3	2.4	2.1	8.2
Myoglobin	5.3	1.2	15.8	12.7	17.3	0.8	2.2	1.9	1.5	0.0
Ribonuclease A	1.8	2.9	10.8	6.1	9.0	11.9	6.8	9.6	8.3	12.9
Ribonuclease S	2.2	3.7	6.0	7.4	11.9	9.1	3.9	11.6	8.4	4.1
Subtilisin BPN'	7.4	7.2	8.8	9.6	5.3	11.8	0.9	5.2	1.1	1.7
Superoxide dismutase	17.3	19.2	22.6	17.5	15.1	11.3	7.1	9.9	14.6	5.1
Triose phosph.isomerase	1.0	5.8	3.7	2.2	1.7	3.8	1.7	8.2	4.1	4.4
Trypsin inhibitor	9.9	3.8	19.4	17.3	18.5	7.1	2.9	17.0	15.4	13.4
Trypsin	0.5	2.8	1.4	2.5	3.4	4.0	0.5	0.6	0.3	2.5

the DF transformation of the FTIR data set results in an improvement for helix prediction and that the definition of the *k* value of minimum error is more precise for the DF than for the FTIR set which tends to have a flatter distribution. The same general characteristic is seen for sheet and coil prediction. The VCD predictions and the DF predictions are nearly the same, but the most reliable VCD RMR predictions encompass more spectral coefficients (minimize for higher *k* values).

Going beyond our previous analyses, which were aimed solely at finding the minimum error, the clustering (or dispersion) of these errors for the set of different RMRs considered at a given *k* value is indicative of stability of the analysis to the specific coefficients chosen for making the prediction. The largest divergence of predictive error between RMR models is typically found for $k = 1$, which indicates that one of the coefficients has a much stronger predictive capability than all the others for that particular structural parameter. This new dimension to our analyses provides a new criterion for selection of the RMR that should be employed in the predictive analyses of unknown proteins. The lowest ultimate error for the set of proteins considered must be balanced with the need for stability to inclusion of different subspectral coefficients. Such a compromise is advantageous since, in general, one does not know how an unknown protein might deviate from the test set of proteins. For such an unknown, one could use plots of its predicted FC values with different predicting models as a function of *k* to evaluate prediction stability for that protein. Here, with effectively 19 unknowns, we have chosen to focus on plots of σ-variability for compactness of discussion. With this in mind, it seems the best and most reliable predictors will have from three to six coefficients for the spectral sets analyzed here. That is somewhat more than found for ECD (typically one) and amide I' (^2H_2O) VCD (typically two to three) spectral analyses (Pancoska *et al*., 1995) and indicates a dispersal of structural sensitivity over the subspectra. This dispersion may indeed be a result of analyzing two independent transitions in a single uniform analysis. Probably what is most striking about all of these plots is their ''sameness'' rather than the clear differentiation with technique that we found in our previous VCD-ECD comparative study.

As can be seen by scanning Table 3, there is no significant overall correlation between the individual protein errors found for predictions with one set of data and those with another, but some proteins are less well predicted with several sets and some are relatively good. It seems that for the helix and sheet prediction, those proteins with roughly equivalent fractions of each type of structure are more consistently predicted by the FTIR or DF sets. On the other hand, those more extreme proteins (having very high helix or sheet and almost none of the other type) are not so well predicted, but the difference is small. In this latter case, the VCD does a bit better, but tends to have higher errors for the mixed structure proteins.

Discussion

While this work uses a methodology that we have explored in depth separately, particularly with regard to its fundamental limitations, the application of the RMR approach as carried out here on VCD and FTIR data for proteins in H_2O offers useful tests of other aspects of the method not previously addressed and answers some questions posed by our previous work. Since this work encompasses ECD, FTIR and VCD data for all the main transitions studied in typical separate reports utilizing each technique, it provides a perspective not previously available and exemplifies the benefits of combining data from various techniques that become possible with the filtering of spectral information inherent in the PC/FA method. The approach is general such that data obtained systematically with another technique from such a training set can be straightforwardly combined with this analysis. Yet the patterns of minimizing predictive error with relatively few spectral coefficients, as noted in previous studies (Pancoska *et al*., 1994, 1995; van Stokkum *et al*., 1990; Pribic *et al*., 1993) still hold here. In brief, we have shown that deuteration is not an obstacle to protein spectral analysis in terms of fractional secondary structure (FC values), that differential FTIR representation of the data has some advantages, and that FTIR as well as VCD and ECD evidence a loss of predictive capability as the number of spectral coefficients used for FC prediction is increased. Finally, we have developed an improvement on our RMR method by using the variance in prediction achieved with a cluster of the best fit regressions to select the order, *k*, of regression to best use for stable prediction. These points are discussed below in some more detail.

Spectral technique comparison

We have demonstrated that the amide I + II VCD as well as the FTIR spectra of proteins in H_2O can be decomposed into a series of subspectra and coefficients and that these spectral coefficients can be used to predict the fractional secondary structure of globular proteins. Thus these H_2O VCD spectra are as fully accessible and analyzable as were our previously published ^2H_2O data for most proteins, despite the somewhat increased experimental difficulty in making the measurements in H_2O. While this should have been expected, it has been contested. However, it is more interesting to see that the quality of prediction for the VCD and FTIR (if analyzed as differential data, as in our DF set) are similar. While both of these H_2O-based analyses are better than VCD predictions based solely on the amide I' for proteins in ^2H_2O, particularly for helix component as shown in Table I, a more proper comparison would be to the combined data set,

amide I' + II, used in our previous study (Pancoska *et al*., 1994, 1995). It is important to realize that this H_2O-based analysis settles a lingering question about conventional VCD analyses of protein secondary structure. The partial deuteration effects that inevitably accompany dissolving native proteins in ^2H_2O do not impose a detrimental effect on these analyses. A different perspective on the equivalence of information in H_2O and ^2H_2O-based VCD spectra can be found in our recently reported neural network study demonstrating the interconvertablity of these spectra (Pancoska *et al*., 1996). In fact, comparison of our previous results (Pancoska *et al*., 1994, 1995) with those reported here indicate that the amide I' (^2H_2O) + amide II (H_2O) predictions are, when averaged over all the structural components, somewhat better than the amide I + II (H_2O) VCD predictions presented here. Such a comparison then implies that the deuteration process might actually help the analyses. In principle, partial deuteration can be used to sense the least exchangeable residues characteristic of the buried parts of the structure. These often involve helices, whereas turns, which might contribute to spectral overlap with helices, are often on the surface of a globular protein and more easily exchanged. We are currently pursuing the implications of such partial deuteration effects on spectral analyses, under more controlled conditions.

On the other hand, in contrast to our previous coupling of ECD and VCD data (Pancoska *et al*., 1995), no dramatic increase in the quality of the fit was obtained by coupling the VCD spectra with the FTIR results in a combined analysis of spectra for proteins in H_2O. In this respect, even for the reduced 19 member protein set, the ECD predictions of helix are consistently better than those of the VCD (H_2O or ^2H_2O) or of the FTIR. The joining of ECD and FTIR data was previously reported by others (Sarver & Krueger, 1991b; Pribic *et al*., 1993), and we have repeated such a process but by using the RMR methods to allow a systematic comparison with coupling H_2O-based VCD to ECD. While the predictions definitely improve, especially for the helix fraction, the improvement is not so dramatic as seen with our ^2H_2O-based VCD data. In all three instances, clearly it is the combination of sensitivities between the ECD and vibrational spectra that is the key to the success seen for combining data sets. We assume that sensing qualitatively different spectral transitions is a key to the difference in sensitivity of the two types of combined spectral analyses. Since the individual FTIR (DF) and VCD analyses were so similar, one can infer that they both sense the important aspects of secondary structure needed to complement the ECD. We view the combination of vibrational and electronic spectroscopies as a compensating analysis, each technique adding its strengths but, through the process of the RMR, damping the wide variations in predictions that can be found with a single technique. On the other hand, combining VCD and FTIR (DF) does not bring new transitions into the analysis. Thus the errors in Table 4 for the combined DF + VCD or the FTIR + VCD analyses are more consistent than those found in Table 3 for single technique analyses. This improvement of the FTIR for structural correlation by combination with ECD has been noted before (Sarver & Krueger, 1991b; Pribic *et al*., 1993) but our approach using the selective RMR following a complete search over all significant spectral components and emphasizing their predictive ability is unique and permits us to explore the physical origins of the improvement. It is important in this respect to note that all the optimally predicting combined technique models (except ECD + VCD, FC_{helix}) depend on a mixture of spectral coefficients from both techniques. This development is not accidental, since we completely search all possible combinations (including those representing only one technique) for optimal regressions.

This comparison of prediction quality would imply that the variances in the VCD and FTIR are sensitive to similar structural parameters. That seems to be true when the FTIR are represented differentially (DF set). Our initial choice of such a representation of the FTIR data was driven by a desire to have better numerical behavior for the diagonalization of the correlation matrix, a property we have found useful in other analyses. In the end, that did not seem so critical for FC determination, since both data sets, FTIR and DF, seemed to be well behaved in the PC/FA process and to yield reasonable regression relations and predictions. That the DF did better at helix prediction is probably due to the added differentiation in the significant subspectra, those most important to the overall band shape. One sign that this may be the cause for improvement in helix prediction is that the first two subspectra for the DF set are similar in magnitude, intense and of similar shape for the amide I but are quite different for the amide II. Both of these contribute to the best RMR for helix and sheet determination giving it added flexibility to represent the training set†. For the FTIR set, coefficients of S2, S3 and S4 dominate the optimal prediction relation, but only those of S2 are large, S2 being the FTIR subspectrum having the greatest overlap with S1 of the DF set. It is interesting that the shapes of the more important (in terms of variance) FTIR and DF subspectra, S2 and S1, respectively, are similar to that of the VCD (H_2O) subspectrum S2. Each of these is respectively the most important subspectral coefficient for helix prediction using that data set.

† The reader should note that the dependence of two structural descriptors (helix and sheet) on the same spectral components is consistent with our earlier results and is characteristic of the helix-sheet interdependence in the training set (Pancoska *et al*., 1992, 1995, 1995; Keiderling, 1996). The fact that the DF set has two spectral parameters correlated to these structural parameters gives it more accuracy than if it depended essentially on just one, as we have seen for ECD analyses.

One can qualitatively appreciate this similarity in the FTIR and VCD analyses by noting that the second subspectra represent the major band shape deviations from the average for the VCD and FTIR cases, and in the DF case, due to the pre-subtraction of the average FTIR spectrum, the first subspectrum represents this dominant variance. At the same time, from a structural point of view, the helix fraction is the dominant secondary structural variance, having a dynamic range from 0 to 70% in our training set. That these subspectra should both be important for helix prediction is thus reasonable. That they have similar shapes implies that the underlying transitions and their frequencies are important. This is consistent with the established observations (Mantsch *et al*., 1986, Byler & Susi, 1986; Surewicz *et al*., 1993) that helices give rise to higher IR frequencies than those characteristic of sheets and ''coil'', the two alternative dominant secondary structural forms. It can be seen that the effect of the second subspectrum is to shift the dominant amide I frequency up or down, depending on the sign of its contribution. Since both techniques measure the same transitions, it is reasonable that the most important subspectrum for helix prediction represents this frequency shift and has the same shape in both analyses.

The method of complete search for the optimal RMR is the key to developing this first correlation between the band shape and frequency determination methods of FTIR-based secondary structure analysis. Put another way, the coincidence of this ''mathematically'' derived band shape representing the major spectral variance with the commonly accepted structural dependence of the frequency dispersion of protein IR spectra lends further confidence to the RMR method. Its success justifies the method of complete search through all combinations of spectral coefficients to optimize the structural predictions. If we compare the regression coefficients for the optimal prediction regression relations with different values of *k*, the number of coefficients used in the relation, the amplitudes for the second FTIR subspectral coefficient clearly dominates both the helix and sheet predictions. As we have discussed earlier (Pancoska *et al*., 1992, 1995), the correlation of helix and sheet fractions in the training set (as well as in the PDB) leads in part to this dependence of both prediction relations on the same spectral coefficient. For the DF set, the first coefficient dominates the helix relations, but the second is more important for the sheet prediction relations. Here we can see that the differential representation has helped separate the helix and sheet dependencies of the spectra, giving it a better response to structural variation which, consequently, is evidenced in the lower helix errors for the DF set in Tables 1 and 3.

RMR prediction accuracy

The prediction error as a function of *k* (results in Figures 4 and 5) convincingly demonstrate that although increasing the number of spectral components improves the fit of any given regression, eventually it leads to a degradation in the prediction capability. As compared with our previous studies, the $k = 3$ to 6 coefficients found to be optimal for prediction using VCD or FTIR of proteins in H_2O is about the same as found with our amide I' (^2H_2O) + amide II results (Pancoska *et al*., 1994). Converting to the DF set of data led to a sharper delineation of the optimal *k* value (size of the regression relation) by imposing a sharper fall and rise, toward and away from, respectively, the minimum prediction error as compared to the unaltered FTIR set.

The new part of the RMR analysis, employed here for the first time, is the comparison of the predictive ability of several regressions of the same order, *k*, that have similarly high regression coefficients, *rr*, in the original fitting process. When these predictions have a low dispersion in terms of their error for the one-out strategy employed here then we know the result will not be very sensitive to the form of the regression relationship chosen for the prediction. That dispersion also gives us a realistic idea of whether the improvement seen with modification of the regression is real or illusory. Thus we now suggest that one should not just choose the regression of highest *rr* and minimal prediction error, but one should look at the other regression relations with low prediction error and choose the particular set of relations whose errors are tightly clustered. (Obviously this will not apply where the $k = 1$ case is optimal.) This revision of the RMR approach might tend to favor larger numbers of coefficients, but that impact is not large. Successive values of *k* yield very similar predictions making the choice among them almost arbitrary. That is a desirable quality of the method, it is relatively robust.

It might be surprising that the amide I + II VCD is not better or even as good as the amide I' + II VCD for predicting secondary structure. One possibility we looked at was the independence of the amide I and II analyses. In a separate analysis, we split the amide I and II spectra into two separate data sets, carried out a new PC/FA reduction to get a new set of subspectra and then executed a separate selective RMR analysis on each separately and both combined, but from independent PC/FA analyses. The best prediction errors for these reconfigured VCD data sets are summarized in Table 1 for the separate bands and in Table 2 for their recombination, but as independently varying data sets. All three can be compared with the results from Table 1 for the original, naturally combined amide I + II VCD analysis. The results from our previous study for amide I' + II are listed for comparison. It is clear that, as compared to either alone, combining amide I and amide II data helps the prediction just as seen in our previous ^2H_2O-based study (Pancoska *et al*., 1994). Both helix and sheet predictions improve on combination, and this independent combination of amide I and II VCD

leads to a lower average error than their ''natural'' combination. However, for the helix descriptor, the change from that found by treating the spectra as one data set is not significant, while for the sheet the improvement is small. The errors for the minor components (bend, turn and other) showed the largest decrease. This may mean that the extra effort involved in separating the data sets is not justified by the final modest improvement obtained. For each structural type, combining the sets yielded their best predictions when spectral components from both bands were in the RMR, much as we saw in the ^2H_2O-based study. The end result of this independent combination of amide I and II (H_2O) VCD spectra is slightly better, but virtually identical, in terms of average error as compared with the amide I' (^2H_2O) + amide II spectral prediction errors for a larger training set (Pancoska *et al*., 1994). Thus the apparent virtue of the ^2H_2O data set noted above may well in part be due to the independence of the spectral component analyses implicit in that earlier calculation. Again, even at this level, there is no loss of analytical applicability by using VCD of partially deuterated proteins dissolved in ^2H_2O.

In the end though, as we have noted before, only part of the spectrally reliable information content is used for these secondary structure predictions. Since the studies reported here use data from different spectral techniques, for proteins in different solvents and at different concentrations yet still get the same behavior as found in our previous study, the recurring observation of under-utilization of optical spectral data in optimal structural predictions has a wide impact and implies we are sampling a fundamental aspect of the problem of protein structure-spectra correlation. Since many spectra-structure prediction methods utilize the entire spectral representation developed with the principal component method to create structural predictions for unknown proteins, they are most probably not optimal predictions in terms of error. It is important for the wider biochemical community of users to recognize this potential problem in their use of now-standard ECD and FTIR spectral-structure prediction techniques.

As noted before, the RMR method is a different philosophical approach from that underlying the variable selection method of 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. We instead use all the proteins but select the most structurally sensitive components of the spectra. The present study confirms our previous hypothesis (Pancoska *et al*., 1995) that, even with very high S/N FTIR data, use of large numbers of coefficients in the regression analyses is dangerous. The results of van Stokkum *et al*. (1990) and Pribic *et al*. (1993) can be viewed as supporting this. The difference between these latter studies and this one is that the selective RMR method chooses the best set of coefficients from all those available, rather than taking them in rank order. As clear from our results for FTIR, the first subspectrum is of no use for secondary structure prediction so that its inclusion is not profitable, at best, and is counterproductive, at worst. The physical basis for ignoring such subspectra is perhaps most evident in FTIR, where the observed protein spectra result from the overlap of many, near-identical amide residue spectral contributions that are perturbed only by their spatial relationships with each other. Thus we would expect that the dominant feature of the spectrum would reflect the presence of amides themselves and not secondary structure, just as we see. Subtraction of the average spectrum as done for the DF set sharpens the dependence of the spectrum on the difference from the isolated oscillators. In VCD, as in ECD, the spectra primarily result from the interaction of the oscillators. This interaction basis makes the subspectral coefficients more characteristic of the secondary structure. Carrying out the DF pre-treatment of the FTIR data then puts the VCD and FTIR data sets on a more equivalent footing, and, perhaps surprisingly, leads to a near equivalence of their predictive ability.

Conclusions

Our work shows that the fundamental limitations found earlier in secondary structure analyses using VCD and ECD of proteins in ^2H_2O apply to VCD and FTIR analyses of proteins in H_2O. Our conclusion is that these errors are more characteristic of the proteins themselves than of the methods or of the spectral techniques. Putting aside the questions that might have arisen from the use of protein in ^2H_2O data, from use of new techniques (such as VCD), from limitations of ECD, etc., as we have tried to do with this systematic comparative study of FTIR and VCD data, points to the fundamental nature of these observations as regards the relationship of spectra and structure. It is clear that describing the protein structure in terms of FC values alone is too limiting to fully use the structurally sensitive data available from optical spectra. Our analyses provide a basis for future protein structural studies that can go beyond such limitations through use of a different protein structural description that can be correlated to optical spectra. Development of such a descriptor (Pancoska *et al*., 1994) and tests of its predictability are topics of an ongoing research effort in this laboratory.

Materials and Methods

Samples

The 19 proteins of known structure used as a training set for this study are summarized in Table 5, 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 & Sander (KS, 1983) algorithm. (The values in Table 5 reflect the best possible match between crystal forms in the PDB and the actual

Table 5. Training set proteins, sources, and X-ray determined secondary structures

Protein	Species	Source	PDB file[a]	KS helix	KS sheet	KS turn	KS bend	KS other
Alcohol dehydrogenase	Horse liver	Fluka 05648	4ADH	24.9	20.6	14.7	13.6	26.2
Carbonic anhydrase	Bovine erythrocytes	Sigma C-7500	1CA2	16.0	28.9	12.9	15.2	27.0
α-Chymotrypsinogen A	Bovine pancreas	Sigma C-4879	2CGA	14.3	32.2	14.3	12.7	26.5
α-Chymotrypsin type II	Bovine pancreas	Sigma C-4129	5CHA	11.8	32.1	11.4	14.4	30.4
Concanavalin A	Jack bean	Sigma C-2010	3CNA	0.0	40.5	9.3	19.8	30.4
Cytochrome *c*	Tuna	Sigma C-2011	1CYT	42.7	0.0	15.5	8.7	33.0
Glutathione reductase	Wheat germ	Sigma G-6004	2GRS	29.3	18.7	10.4	19.3	22.3
Hemoglobin	Human	Sigma H-7379	1HCO	62.7	0.0	18.8	6.6	11.9
λ-Immunoglobulin	Human	Fluka 56834	1REI	2.8	47.7	14.0	11.2	24.3
Lactate dehydrogenase	Rabbit	Calbio. 427217	4LDH	36.8	11.3	14.3	13.1	24.6
Lysozyme	Hen egg-white	Sigma L-6876	7LYZ	38.8	7.8	20.9	16.3	16.3
Myoglobin	Horse heart	Sigma M-1882	1MBN	77.1	0.0	9.8	1.9	11.1
Ribonuclease A	Bovine	Sigma R-5125	1RN3	21.0	34.7	11.3	14.5	18.6
Ribonuclease S	Bovine pancreas	Sigma R-6000	1RNS	20.8	35.2	7.2	14.4	22.4
Subtilisin BPN'	Bacterial	Sigma P-8038	1SBT	30.2	17.8	15.3	12.0	24.7
Superoxide dismutase	Bovine erythrocytes	Fluka 86200	2SOD	2.0	38.4	14.6	20.5	24.5
Triose phosph.isomerase	Yeast	Sigma T-2507	1TIM	45.8	17.0	7.3	8.9	21.1
Trypsin inhibitor	Soybean	Sigma T-9003	3PTI	20.7	24.1	6.9	19.0	29.3
Trypsin	Bovine pancreas	Sigma T-8253	3PTN	9.9	32.3	14.8	17.9	25.1
Albumin	Bovine serum	Sigma A-0281	—					
Lactoferrin	Human milk	Sigma L-5665	—					
α-Casein	Bovine milk	Sigma C-8032	—					
β-Lactoglobulin A	Bovine milk	Sigma L-7880	—					

FC values determined using the Kabsch & Sander (1983) algorithm with helix being the sum of α-helix and 3_{10}-helix, sheet being both parallel and antiparallel components, and ''other'' being all components not specifically selected.

[a] Protein Data Bank file name for reference structure.

proteins used for our spectral determinations as regards species and general conditions as we have discussed previously (Pancoska *et al*., 1995).) As an independent, qualitative monitor for the behavior of the results, we include in the data set four extra proteins that were not in the PDB and consequently had no structural input to the spectral analysis. (Five of the total of 28 proteins used in our previous study (Pancoska *et al*., 1995) were not sufficiently soluble to prepare the higher concentration samples needed for this H_2O-based study.) Data from all 23 proteins are decomposed by our PC/FA together in a single calculation for each technique into sets of subspectra and their coefficients that are needed for linear reconstruction of the experimental spectra†. This consistent manner of spectral reduction to coefficients provides a unified set of parameters for the diverse sets of spectral data that are needed for the subsequent regression steps.

All proteins were used as obtained without further purification and were directly dissolved in a small amount (∼40 μl) of double-distilled water at ∼200 mg/ml. Solutions to be studied were placed in refillable cells (Graseby Specac) consisting of two CaF_2 windows separated by a 6 μm mylar spacer. Although the cell has a demountable design, it was assembled before being filled and used as a sealed sample cell, but it was cleaned by disassembling after each protein measurement. Samples were drawn into the cell by placing a drop in one filling port and applying gentle suction to the opposite port with a syringe. The cell was initially filled with water and used to determine both the VCD and absorbance baseline. Then it was dried by evacuation and the protein sample was inserted. It was also possible to change samples without disassembling the cell, provided extensive rinsing was used to adequately clean it between samples.

† Programs for carrying out these or similar analyses on a DOS-compatible PC are available on request from the authors.

Spectroscopy

The amide I and II VCD and IR absorbance spectra at room temperature were measured on our dispersive VCD instrument, which has been thoroughly described elsewhere (Keiderling, 1981, 1990). Our sampling and data collection methods for obtaining VCD of proteins in H_2O were just as described (Baumruk & Keiderling, 1993). In summary, the VCD were obtained with ∼10 cm^{-1} resolution as the average of typically ten sample and ten baseline scans, each obtained with a ten second time-constant, taking a total of ∼20 hours. VCD spectra were calibrated with the usual methods (Nafie *et al*., 1976; Keiderling, 1981, 1990) and were baseline-corrected with identical scans of H_2O as before. The VCD spectra for the amide I and II regions were normalized to the amide I absorbance maximum in this region (i.e. scaled so that $A = 1.0$). It should be noted that this and other possible normalization schemes are potentially the source of some error in the subsequent analyses but that we have treated the FTIR and VCD consistently here (Bitto, 1993).

FTIR absorption spectra of all the polypeptides and proteins studied were measured on the same samples over the 4000 to 400 cm^{-1} spectral region on a Digilab FTS-40 FTIR spectrometer using an MCT detector, 4 cm^{-1} resolution and an average of 1024 scans. These were corrected for water interference by subtraction of a separately collected water vapor spectrum and, for the liquid H_2O absorption, by use of the automated solvent subtraction method of Dousseau *et al*. (1989). The FTIR spectra also provided a check on 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 then using the same shift to correct

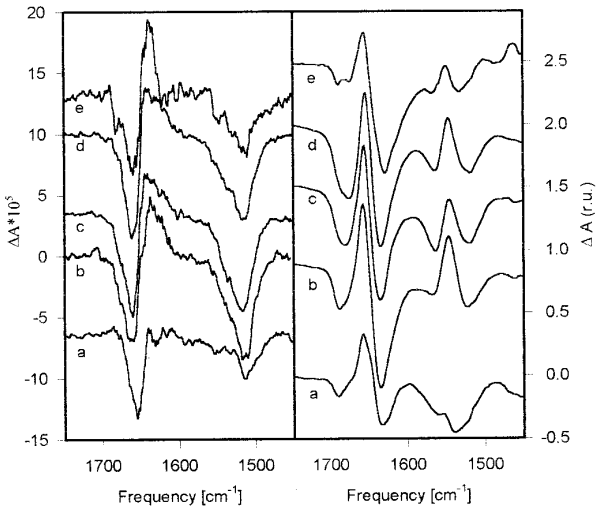

Figure 6. VCD (left) and DF (right) spectra for the group 1, ''helical'' cluster of proteins in H_2O in the training set: a, cytochrome *c*; b, albumin; c, hemoglobin; d, myoglobin; e, lactate dehydrogenase. Spectra are offset for clarity by an amount indicated by the shift of the intercept with the ΔA axes. Baselines would correspond to a horizontal line from that intercept point but are omitted for clarity.

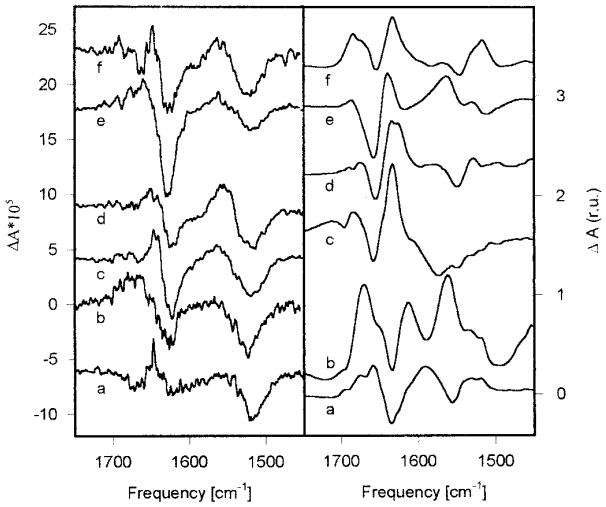

Figure 8. VCD (left) and DF (right) spectra for the group 3, ''sheet/other'' cluster of proteins in H_2O in the training set: a, subtilisin; b, casein; c, chymotrypsinogen; d, carbonic anhydrase; e, trypsin inhibitor; f, trypsin. Baselines and offsets as in Figure 6.

the VCD. Since concentrations are inexact, at best, we have also normalized all the FTIR spectra to an absorbance of 1.0 at the amide I maximum. This means that our spectra are all standardized to the same norm, whatever its reliability.

Sample spectra obtained with the VCD and FTIR techniques for some of these proteins have been presented (Baumruk & Keiderling, 1993). The FTIR spectra used agree with those widely available in the literature. Data for the entire set, as were used in this study, are shown in graphical form for the VCD in Figures 6 to 9 (left half); and, since we have found it useful to use difference FTIR spectra (see below) for these comparative analyses, that representation of those spectra is presented in the right half of Figures 6 to 9. The groupings of proteins presented in these four Figures are somewhat arbitrary, but stem from cluster analyses of the VCD band shapes. These tend, on average, with exceptions, to reflect the predominant fold type as noted in the legends. Data of either type (VCD, FTIR or difference FTIR plus the amide I' and II VCD and the ECD data from our previous report (Pancoska *et al*., 1995)) in the digital form (ASCII format) that was used for these computations as well as the programs for PC/FA and RMR are available from the authors upon request.

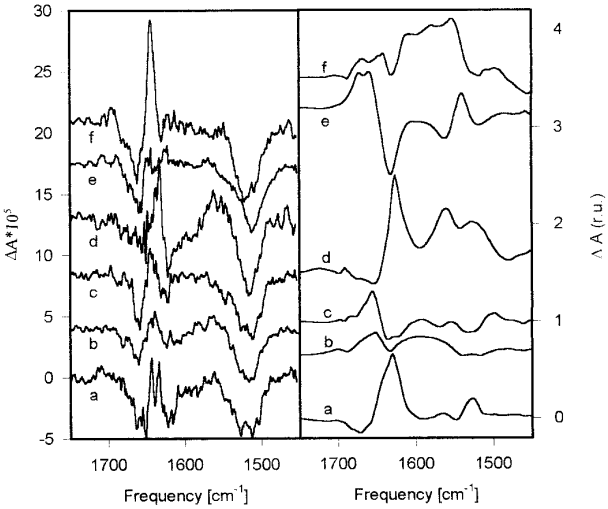

Figure 7. Subspectra. VCD (left) and DF (right) spectra for the group 2, mixed ''helical/sheet'' cluster of proteins in H_2O in the training set: a, alcohol dehydrogenase; b, glutathione reductase; c, lactoferrin; d, β-lactoglobulin A; e, lysozyme; f, triose phosphate isomerase. Baselines and offsets as in Figure 6.

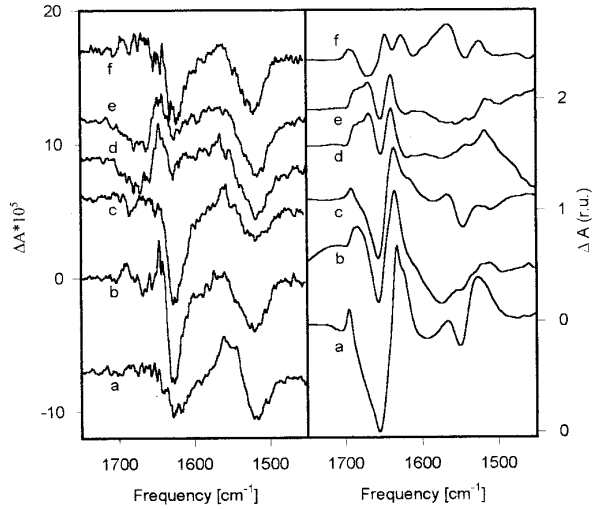

Figure 9. VCD (left) and DF (right) spectra for the group 4 ''sheet/other'' cluster of proteins in H_2O in the training set: a, concanavalin A; b, chymotrypsin; c, immunoglobulin; d, ribonuclease S; e, ribonuclease A; f, superoxide dismutase. Baselines and offsets as in Figure 6.

Calculations

The PC/FA method (Malinowski & Howery, 1980; Pancoska *et al*., 1979) that we have applied to the decomposition of the VCD and FTIR spectra into a linear combination of orthogonal subspectra has been described in detail (Pancoska *et al*., 1991, 1994, 1995). It is used here without change other than for the data sets to which it is applied. Since the FTIR data are inherently all positive (being an absorbance) and consequently less variable in band shape as compared with the VCD, we did a separate analysis on the FTIR data set by first subtracting the average of all the spectra in the training set from each individual spectrum before computing the subspectra with the PC/FA. This data set will be referred to as the Difference FTIR (DF) set. Pre-treating the FTIR data in this way enhances the differentiation between spectra (see Figures 6 to 9, right side), improves the numerical reliability of the diagonalization of the correlation matrix, and is consistent with our findings (*vide supra*) that the coefficients of the first subspectrum (effectively the average spectrum) of the conventional FTIR representation are not important for secondary structural analyses. This method also emphasizes the relative frequency-intensity variations between the FTIR spectra of different proteins, which is the primary spectral property after the overall intensity is removed. It should be noted that, since the VCD and FTIR measure the same spectral transitions, some comparison of subspectra and the structural correlations found for their coefficients with these two different physical techniques becomes possible.

While our RMR methods are also fundamentally the same as previously described (Pancoska *et al*., 1991, 1994, 1995) we have here extended the method to allow testing of some of the assumptions implicit in our previous work. Only the aspects that are new or characteristic for this study are summarized below. To establish the best regression model, we again carried out several tests using all the PC/FA coefficients derived above. The p_V coefficients obtained from the amide I + II VCD analysis and the p_F coefficients from the amide I + II FTIR analysis were used first separately, to allow comparison of the fits with each data set, and then combined into sets of $(p_V + p_F)$ coefficient vectors for combined analyses. The same sort of test was done with the DF coefficients. For comparison with our previous work and that of others, we also combined the FTIR and VCD coefficients separately with those obtained from a revised PC/FA for this training set of the ECD spectra (as obtained previously; Pancoska *et al*., 1995). Optimal RMR 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 (fractional component) values of the secondary structure (Kabsch & Sander, 1983), where ζ represents helix, sheet, bend, turn and ''other'' crystallographic secondary structures (here encoded as H, E, S, T, C, respectively, reflecting the KS definitions, with H further being the sum of α-helix and 3_{10}-helix, E being both parallel and anti-parallel components, and C being all the other components not specifically noted in this list). Restricting the regression to a selection of k coefficients and then testing the dependence of the quality of both the RMR and prediction (see below) on k is the key to our gaining insight into the sensitivity of the method to specific spectral components. In an extension of the previous approach, for each set of coefficients of a given size k, several RMR relationships corresponding to the sets of combined coefficients providing the highest group of regression coefficients, *rr*, are retained in this version of our software (typically four or five here). Previously we kept only the single RMR with the highest *rr*, i.e. the best fitting one, but here we test the assumption that this relationship would be best for prediction.

To evaluate the capability of the derived RMR relationships to predict the FC values, we repeated the above procedures 19 times, each time eliminating one protein of known structure from the training set. The set of k coefficient regression equations retained as above, $k = 1$ to p for each reduced, 18 member, training set, were then used to predict the FC values for the protein left out. Each prediction was compared with the actual values in Table 5, and the average deviation for the 19 predictions was used to characterize the prediction capability of that regression model. Choosing between the models was done graphically by plotting the error *versus k*, much as in our previous studies (Pancoska *et al*., 1994, 1995), but this time a cluster of values is plotted for each k corresponding to the set of regression relationships retained. The minimum error† is easy to determine in this manner, but, in addition, the stability of the RMR predictions for unknown proteins is indicated by the degree of clustering of these errors for the set of relationships at any given k value. Examples of this procedure are presented in the Results section to clarify this distributive plus minimum error criterion.

† We use the following error characteristics to compare spectral, FC^s_{ij}, and X-ray, FC^x_{ij}, secondary structure parameters for protein i and structural type j:

Average error, δ_i, FC values of individual proteins i over n_s types of secondary structures j:

$$\delta_i = (1/n_s)\Sigma_j|FC^s_{ij} - FC^x_{ij}|$$

Standard deviation of FC values for given type of secondary structure, for all N known proteins from the training set:

$$\sigma_j = [(\Sigma_i(FC^s_{ij} - FC^x_{ij})^2)/(N - 1)]^{1/2}$$

Relative standard deviation with respect to the range of values in the training set for an FC value j:

$$\sigma^{rel}_j = (100\sigma j)/(FC^x_{max,j} - FC^x_{min,j})$$

Acknowledgements

This work was primarily supported by a grant from the National Institutes of Health (GM-30147) for which we are most grateful. Cooperation between Charles University and UIC was 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. The research in Prague was supported in part by grants GACR 203-93-0714 and GAUK 302 (to P.P.). Equipment grants from the NSF, NIH and University of Illinois supported purchase of instrumentation used.

References

Baumruk, V. & Keiderling, T. A. (1993). Vibrational circular dichroism of proteins in H_2O solution. *J. Amer. Chem. Soc.* **115**, 6939–6942.

Berjot, M., Marx, J. & Alix, A. J. P. (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. Diploma thesis, Charles University, Prague, Czech Republic.
Bussian, B. M. & 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, D. M. & Susi, H. (1986). Examination of the secondary structure of proteins by deconvolved FTIR spectra. *Biopolymers*, **25**, 469–487.
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.
Dousseau, F., Therrien, M. & Pezolet, M. (1989). On the spectral subtraction of water from the FTIR spectra of aqueous solutions of proteins. *Appl. Spectrosc.* **43**, 538–542.
Dukor, R. K., Pancoska, P., Prestrelski, S., Arakawa, T. & Keiderling, T. A. (1992). Comparison of FTIR and vibrational circular dichroism results for two protein hormones. Implications regarding secondary structure. *Arch. Biochem. Biophys.* **298**, 678–681.
Hennessey, J. P., Jr & Johnson, W. C., Jr (1981). Information content in the circular dichroism of proteins. *Biochemistry*, **20**, 1085–1094.
Johnson, W. C., Jr (1985). Circular dichroism and its empirical application to biopolymers. *Methods Biochem. Anal.* **31**, 61–163.
Johnson, W. C., Jr. (1988). Secondary structure of proteins through circular dichroism spectroscopy. *Annu. Rev. Biophys. Chem.* **17**, 145–166.
Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
Keiderling, T. A. (1981). Vibrational circular dichroism. *Appl. Spectr. Rev.* **17**, 189–226.
Keiderling, T. A. (1990). Vibrational circular dichroism. comparison of technique and practical considerations. In *Practical Fourier Transform Infrared Spectroscopy. Industrial and Laboratory Chemical Analyses* (Ferraro, J. R. & Krishnan, K., eds), pp. 203–284, Academic Press, San Diego.
Keiderling, T. A. (1996). Vibrational circular dichroism. Applications to conformational analysis of biomolecules. In *Circular Dichroism and the Conformational Analysis of Biomolecules* (Fasman, G. D., ed.), pp. 555–598, Plenum Press, New York.
Keiderling, T. A. & Pancoska, P. (1993). Structural studies of macromolecules using vibrational circular dichroism. In *Biomolecular Spectroscopy Part B* (Clark, R. J. H. & Hester, R. E., eds), pp. 267–315, Wiley, Chichester.
Keiderling, T. A., Wang, B., Urbanova, M., Pancoska, P. & Dukor, R. K. (1995). Empirical studies of protein secondary structure by vibrational circular dichroism and related techniques. α-Lactalbumin and lysozyme examples. *Faraday Disc.* **99**, 263–285.
Lee, D. C., Haris, P. I., Chapman, D. & Mitchell, R. C. (1990). Determination of protein secondary structure using factor analysis of infrared spectra. *Biochemistry*, **29**, 9185–9193.
Malinowski, E. R. & Howery, D. G. (1980). *Factor Analysis in Chemistry*, Wiley, New York.
Manavalan, P. & Johnson, W. C., 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.
Mantsch, H. H., Casal, H. L. & Jones, R. N. (1986). Resolution enhancement of infrared spectra of biological systems. In *Spectroscopy* (Clark, R. J. H. & Hester, R. E., eds). vol. 13, pp. 1–46, Wiley & Sons, London.
Nafie, L. A., Keiderling, T. A. & Stephens, P. J. (1976). Vibrational circular dichroism. *J. Am. Chem. Soc.* **98**, 2715–2723.
Pancoska, P. & Keiderling, T. A. (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., 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., Yasui, S. C. & Keiderling, T. A. (1989). Enhanced sensitivity to conformation in various proteins. Vibrational circular dichroism results. *Biochemistry*, **28**, 5917–5923.
Pancoska, P., Yasui, S. C. & Keiderling, T. A. (1991). Statistical analyses of the vibrational circular dichroism of selected proteins and relationship to secondary structures. *Biochemistry*, **30**, 5089–5103.
Pancoska, P., Blazek, M. & Keiderling, T. A. (1992). Relationships between secondary structure fractions for globular proteins. Neural network analyses of crystallographic data sets. *Biochemistry*, **31**, 10250–10257.
Pancoska, P., Wang, L. & Keiderling, T. A. (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., Bitto, E., Janota, V. & Keiderling, T. A. (1994). Quantitative analysis of vibrational circular dichroism spectra of proteins. Problems and perspectives. *Faraday Disc.* **99**, 287–310.
Pancoska, P., Bitto, E., Janota, V., Urbanova, M., Gupta, V. P. & Keiderling, T. A. (1995). 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. *Protein Sci.* **4**, 1384–1401.
Pancoska, P., Janota, V., & Keiderling, T. A. (1996). Interconvertability of electronic and vibrational circular dichroism spectra of proteins. A test of principle using neural network mapping. *Appl. Spectrosc.* In the press.
Perczel, A., Hollosi, M., Tusnady, G. & Fasman, G. D. (1991). Convex constraint analysis: a natural deconvolution of circular dichroism curves of proteins. *Protein Eng.* **4**, 669–679.
Pribic, R., von Stokkum, I. H. M., Chapman, D., Haris, P. T. & Bloemendal, M. (1993). Protein secondary structure from Fourier transform infrared and/or circular dichroism spectra. *Anal. Biochem.* **214**, 366–378.
Provencher, S. W. & Glöckner, J. (1981). Estimation of globular protein secondary structure from circular dichroism. *Biochemistry*, **20**, 33–37.

Sarver, R. W. & Krueger, W. C. (1991a). Protein secondary structure from Fourier transform infrared spectroscopy. *Anal. Biochem.* **194**, 89–100.

Sarver, R. W. & Krueger, W. C. (1991b). An infrared and circular dichroism combined approach to the analysis of protein secondary structure. *Anal. Biochem.* **199**, 61–67.

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

Sreerama, N. & Woody, R. W. (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.

Surewicz, W., Mantsch, H. H. & Chapman, D. (1993). Determination of protein secondary structure by Fourier transform infrared spectroscopy: a critical assessment. *Biochemistry*, **32**, 389–394.

Toumadje, A., Alcorn, S. W. & Johnson, W. C., Jr (1992). Extending CD spectra of proteins to 168 nm improves the analysis for secondary structures. *Anal. Biochem.* **200**, 321–331.

van Stokkum, I. H. M., Spoelder, H. J. W., Bloemendal, M., van Grundelle, R. & Groen, F. C. A. (1990). Estimation of protein secondary structure and error analysis from circular dichroism spectra. *Anal. Biochem.* **191**, 110–118.

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

Williams, R. W. (1986). Protein secondary structure analysis using Raman amide I and amide III Spectra. *Methods Enzymol.* **130**, 311–331.

Edited by P. E. Wright

(Received 15 January 1996; received in revised form 13 March 1996; accepted 21 March 1996)