The optimization of protein secondary structure determination with infrared and circular dichroism spectra

Keith A. Oberg, Jean-Marie Ruyschaert and Erik Goormaghtigh

Center for Structural Biology and Bioinformatics, Laboratory for the Structure and Function of Biological Membranes, Free University of Brussels (ULB), Belgium

We have used the circular dichroism and infrared spectra of a specially designed 50 protein database [Oberg, K.A., Ruyschaert, J.M. & Goormaghtigh, E. (2003) Protein Sci. 12, 2015–2031] in order to optimize the accuracy of spectroscopic protein secondary structure determination using multivariate statistical analysis methods. The results demonstrate that when the proteins are carefully selected for the diversity in their structure, no smaller subset of the database contains the necessary information to describe the entire set. One conclusion of the paper is therefore that large protein databases, observing stringent selection criteria, are necessary for the prediction of unknown proteins. A second important conclusion is that only the comparison of analyses run on circular dichroism and infrared spectra independently is able to identify failed solutions in the absence of known structure. Interestingly, it was also found in the course of this study that the amide II band has high information content and could be used alone for secondary structure prediction in place of amide I.

Keywords: circular dichroism; FTIR; PLS; protein secondary structure.

Multivariate statistical analysis methods have proved to be powerful tools for the analysis of component concentrations in chemical mixtures. Because of their effectiveness in systems where there are strongly overlapping bands, these chemometric methods have also proved effective in the analysis of protein spectra [1–12]. In contrast to most typical applications of statistical analysis methods, the reported accuracy in protein studies, especially those that treat infrared (IR) spectra, varies widely. It is often assumed that differences result primarily from the analytical methods applied, and so new analytical methods have appeared continuously over the last two decades (see references above, and [5,6,13–18]). In addition to the reported margins of error, there are a number of discrepancies to be found in IR studies, including the optimal data regions [8,19,20] and spectral preprocessing methods [20]. It has not been resolved whether these differences arise primarily from the analysis methods or the protein basis sets that have been used. Thus there is a need for a systematic evaluation of the steps involved in protein secondary structure analysis where the dependence of the results on the protein basis set is minimized.

The key to the effectiveness of statistical methods is concentration-dependent changes in spectra that are directly related to the concentrations of the chemical species being determined. In simple chemical quantification systems, this is typically an increase in signal intensity at certain positions in a spectrum that depends linearly on the analyte concentration. In some cases, interactions between the components of a mixture may result in additional bands or changes in the signal. In these cases the concentration dependence becomes more complex. However, statistical analysis algorithms can usually model these complexities with linear systems, and thus provide accurate analysis results.

The situation encountered in the analysis of protein spectra is less straightforward. This comes from a certain amount of independence in the variation of protein spectra on the structure content. Such behavior arises in part from the way secondary structure is assigned from crystal structures. Assignment algorithms necessarily involve the simplification of crystal structure data in the form of combining residues with somewhat different conformations into a single structure assignment.

There are several possible ways to handle structure content-independent variations in protein spectra. The primary example of such an analysis is that given in [21] in which protein amide I bands were analyzed by fitting with a series of Gaussian curves. The success reported in the original paper was spectacular: the rms errors for \( \alpha \)-helix
and β-sheet were in the order of ± 2.5%. The curve-fitting method compensates for band position variation by assigning all component bands found in given regions of the spectrum to a particular structure. This method can be highly effective when applied by one experienced in its use; however, curve-fitting requires a series of subjective decisions that can dramatically affect both the results and the interpretation [22–25]. Furthermore, curve-fitting has a tendency to overestimate the β-sheet content of primarily helical proteins, and routinely finds 15–20% β-sheet for proteins that actually have none [21,26–30].

Statistical analyses are generally accepted as being the best way to analyze protein CD spectra, but curve-fitting is still widely used for determining protein structure from infrared spectra. From reviewing the literature and considering just the results for α-helix determinations with IR data, it can be seen that the reported accuracy (rms determination error) ranges from 3.9 to 10.3%, but the 3.9% value was obtained from a 17 protein set, and two sets with nine and 21 proteins obtained rms errors of around 10%. The different algorithms used in these studies may indeed be responsible for the differences, but as statistical analyses depend both on the algorithm and reference spectra used, the source of the discrepancies cannot be unambiguously identified. It remains possible that these differences reflect the internal consistency of the spectra in each respective basis set rather than the expected general accuracy of the methods. In a recent paper, Sreerama & Woody [6] investigated the effect of the number of reference proteins (29–48) on the accuracy of the prediction obtained by three publicly available CD analysis software programs.

In this paper we propose to extend this analysis to IR and combined CD/IR spectra. We took great care to use a protein database that presents the largest possible structure variety. We constructed a protein database that covers, as far as possible, the α/β space, the fold space as described by CATH (class, architecture, topology and homology classification of proteins [31]) as well as other structural features such as helix length, and the number of chains in a sheet. We identified 50 commercially available proteins that can be obtained with sufficient purity and for which we assessed the quality of the crystal-derived structure. We call this set of 50 rationally selected reference proteins RaSP50 and details of this database have been published recently [32].

We report in this study the application of the RaSP50 set to spectroscopic protein structure determination. We have attempted here to establish an optimal approach to using existing methods. This was achieved by focusing on the input for statistical analysis algorithms, such as data types, spectral preprocessing, and secondary structure assignments. Because IR and CD are the most widely used spectroscopic tools for the determination of protein secondary structure, both were applied in this study. They were tested alone and in combination in an effort to evaluate their respective strengths, weaknesses and complementarities. It was found that the quality of the reference protein database rather than the algorithms used determines the efficiency of the secondary structure prediction. Clear complementarities between IR and CD spectra allow a further enhancement of the secondary structure accuracy.

Experimental procedures

Input data for analysis

The set of reference proteins used for this study is an ‘optimal’ basis set, and has been described elsewhere [32]. It represents a wide range of helix and sheet FC values as well as 60 different protein domain folds. The final set of 50 proteins fully spans several different ‘conformational spaces’, and has distributions of structures that reflect the natural abundances found in the PDB. The spectra of the 50 proteins are available on request from the authors.

Protein secondary structure tabulation from oessp output

The secondary structures of the RaSP proteins were determined with the oessp program [33]. There are eight assignments made by oessp. Six are familiar to protein chemists: α-helix (denoted by H), 3 _{10}-helix (G), π-helix (I), β-sheet (E), turn (T), and unassigned structure (indicated by a blank space in the oessp program output, but which we denote with C). Unassigned structure has been referred to by many names, such as irregular, other, disordered or coil. The fractional composition of their secondary structures (FCs) were tabulated from the oessp output. The α-helix assignment can be tabulated as a simple count of the residues assigned H by the oessp program, or it can be divided into ‘ordered’ (denoted by O) and ‘disordered’ (do) helix by giving the disordered helix assignment to the two residues at each helix end (ends) [10] or to helical residues with less than one or two hydrogen bonds within the helix (denoted by < 1 or < 2), and giving all other α-helical residues the ordered assignment. We also separated parallel and antiparallel β-sheet.

Spectroscopic data collection and processing

All protein preparations were desalted by dialysis or size-exclusion chromatography. CD spectra were collected on a JASCO J-710 CD spectrometer using filtered protein solutions in 2 mM Hepes pH 7.2 with an absorbance of ≈ 0.5–0.8 at 192 nm (≈ 0.1 mg mL⁻¹) in a 0.1 cm cell. Each CD spectrum was the accumulation of eight scans at 50 nm⁻¹ with a 1 nm slit width and a time constant of 0.5 s for a nominal resolution of 1.7 nm. Data was collected from 185 to 260 nm. CD spectra were background corrected and scaled to mean residue ellipticity based on the absorbance at 205 nm. The extinction coefficient used was ε₂₀₅ = 5167 per peptide bond; this was determined using a combination of data from Scopes [34] and Hennessey and Johnson [35].

Infrared measurements were made on a dry-air purged Bruker IFS-55 FTIR spectrometer with an MCT detector. Data were collected at 2 cm⁻¹ resolution; 512 scans were accumulated for each spectrum. Transmission IR spectra were collected from ≈ 3% (w/v) solutions sandwiched between CaF₂ windows with a 5 μm Teflon spacer in a demountable cell. The protein signal was extracted from IR spectra by subtracting a buffer spectrum with a scaling factor determined by the method of Powell et al. [36]. The contribution of water vapor from infrared spectra was subtracted using a scaling factor determined from the integrated absorbance of the 1717 or 1772 cm⁻¹ bands.
Preprocessing of spectra for analysis

When indicated, the contributions from amino acid side chains were subtracted for some analyses using data from Venyaminov [37]. For these subtractions, a synthetic side chain spectrum was generated using the amino acid composition of the protein listed in the SWISS-PROT database [38,39]. The side chain spectrum was then subtracted from the protein spectrum using a scaling factor determined by Fourier self-deconvolving the spectra, determining the tyrosine band area ratios at 1518 cm\(^{-1}\), and then using this as the scaling factor for subtracting the original spectra [40]. Due to low tyrosine signal, the proteins BTE, FTN, MTH, PAB, SOD and TRO could not be processed in this manner, so the relative extinction coefficients were used to estimate the subtraction scaling factor [40].

Spectral scaling, baseline corrections and normalizations were carried out by custom routines added to the PLSPLUS analysis software (discussed below). In the following discussion, intensity or point normalization refers to the scaling of spectral regions to a constant maximum intensity of 1, area normalization refers to adjusting the intensity so that all spectra had the same integrated area in a chosen region (0.1 absorbance units cm\(^{-1}\)). Before normalization, IR spectral regions were either baseline-corrected to bring both endpoints to zero, or if baseline-correction was not used, the first value in the band was subtracted from all other data points in order to bring the minimum to zero.

Combination of CD and IR spectra

To analyze combined CD and IR data, hybrid spectra were made by placing CD and IR data in a single array. In these spectra, one unit on the x axis corresponds to one of the native units for each data type (nm for CD and cm\(^{-1}\) for IR), and the data point spacing is 1 per x unit. It was necessary to scale the CD spectra to be consistent with the intensities of the IR spectra (each spectrum was multiplied by 0.0015), but the exact type of unit in each region is less important than the fact that both data types were of similar intensity in the hybrid spectrum. This ensured that they had similar contributions in the model building process. The limits of the data regions used were 1720–1600 cm\(^{-1}\) for the IR amide I band, 1590–1500 cm\(^{-1}\) for the amide II band, and 185–260 nm for far-UV CD data.

Analysis methods

The bulk of the analyses performed in this report were made with PLSPLUS version 2.1 (Galactic Industries, Salem, NH). PLSPLUS is integrated into GRAMS (Galactic Industries Corporation, Woburn, MA, USA), which uses the Array Basic\textsuperscript{\textregistered} programming language. It was therefore possible to create all other necessary software in a single environment with a common data format.

Although the PLS-1 algorithm used here extracts factors in order of their relevance to the structure being quantified, the problem of selecting the total number of factors to use for each structure type remains. It was found that the first two-to-five factors for \(\alpha\)-helix or \(\beta\)-sheet typically accounted for 95% of the variation in the RaSP50 spectra. For turn and other structures, up to 10 factors were required to reach the 95% level, but the first six usually accounted for 90% or more of the total variation. The factors themselves typically began to show significant contributions from noise at the 6th or 7th factor. For automatic selection of the optimal number of factors for each model, the maximum was set at 10. However, if there was a smaller set of factors that provided similar accuracy (< 1.05 times the minimum rms), the algorithm would select this as the optimum. For each analysis, cross validation was performed and the rms deviation between the experimental and calculated set was determined for all possible numbers of factors.

Unless otherwise indicated, all analysis results reported here are from cross validations of the full RaSP50 set. Cross validation is performed by removing each spectrum, in turn, from the reference set. The remaining spectra (49 in this case) are then used to generate a statistical model, which is used to determine the structure of the eliminated protein. Finally the calculated FC for each protein is compared with its actual structure, and the determination error is evaluated and stored. After the cross validation is complete, the rms error of all analyses are determined with \(n – 1\) degrees of freedom in PLS-1, and \(n – (s – 1)\) degrees of freedom for simultaneous methods, where \(n\) is the number of spectra and \(s\) is the number of structures being determined simultaneously.

Results and Discussion

Methods for evaluating analysis performance

Cross validation, as explained in the Materials and methods, treats each protein as an unknown and evaluates its structure using the remaining proteins as a training set. For the RaSP50 proteins, great care was taken to eliminate protein preparations with impurities and to use only high-quality spectra. Because of this it can be assumed that a high cross validation error for any of the RaSP50 proteins arises not from a problem with the sample, but from an inconsistency between its secondary structure assignment and the actual variations of structure that give rise to the protein spectrum.

Because every sample in the basis set is analyzed as an unknown during cross validation, this procedure provides data that can be used to estimate the expected accuracy for analyses of true unknowns by calculating the cross validation errors (determined FC – actual FC) for all basis samples and then taking the rms of these errors is obtained. The rms is potentially a good estimate of the overall performance of an analysis method because it is a summary of many unknown analyses (50 in this case). The rms represents the error bounds for \(\approx 2/3\) of the training samples in cross validation. Thus, it can be expected that there is a 67% chance that FC determination results for an unknown protein will be within \(\pm\) rms of the ‘true’ value.

The rms, if presented alone, is also uninformative. Assume that a hypothetical cross validation of the RaSP50 basis set was performed and an rms for FC\(_T\) of \(\pm 5.0\%\) T was reported. It would be natural to conclude that the turn determination accuracy for this method was quite good, but this is not the case. If we look at the crystal structures of the RaSP50 proteins and examine the distribution of their actual FC\(_{TS}\), we find a mean of
Preprocessing of protein spectra for statistical analysis

CD spectra of proteins were scaled based on the protein (amino acid residue) concentration of the sample used to collect a spectrum as explained in the Experimental procedures. For IR spectra the possible processing steps include normalization, baseline correction, the subtraction of side chain contributions and artificial band narrowing (using Fourier self-deconvolution or differentiation). There is consensus in the literature that band narrowing does not improve statistical analysis results [20] therefore such procedures were not re-evaluated here. The subtraction of side chain spectra before analysis has been used in only one study [10]. Normalization of IR spectra is required before analysis. Typical normalization methods that have been used for IR spectra include scaling to the same maximum intensity [19,20] or area [8,19,20]. In essence, these normalizations are intended to increase the correlation of IR band shape with protein secondary structure, and thus improve analysis accuracy.

The effect of spectral normalization on analysis results. For the infrared spectra (Fig. 1), various normalization procedures were followed: normalization of the band maximum to a constant intensity; normalization to a constant area; normalization of the combined amide I and II bands, with separate analyses of each band afterward; and separate normalization of the amide I and II bands. All these normalizations were tested with and without baseline correction. For the CD spectra (Fig. 1), the mean residue ellipticity was used. The errors of cross validation (rms) of each spectroscopic ‘data region’ used in this study (IR amide I, IR amide II, and CD) were obtained (not shown). It was immediately apparent that most of the normalizations do not radically change the analysis accuracy. Subtracting a baseline and normalizing separately on amide I and II was close to the best solution for every structure.

Side chain signal subtraction. IR and CD protein spectra can contain significant contributions from amino acid side chain bands. In CD spectra, these contributions arise from aromatic groups or disulfides and are dictated by the local environment of each side chain [42,43]. It is therefore impossible to determine the exact nature of their contribution to a given protein spectrum a priori. For IR spectra, side chain contributions are consistent enough for the generation of synthetic spectra based on data from model compound studies [23,37,44,45]. There is much debate over the usefulness of subtracting side chain spectra, as they are typically broad, relatively featureless and may be affected to a small extent by the local environment of each residue. In fact, simple baseline correction often has an effect similar to subtraction [23]. Our data (not shown) indicate that side chain subtractions only moderately improved the rms values for some, but not all, secondary structures. It was not used further in this study.

The sensitivity of different spectroscopic methods to protein secondary structure

The relative sensitivities of different spectroscopic methods to protein secondary structure. It is usually accepted that CD measurements provide more accurate estimations of protein α-helix content whereas IR is thought to be more sensitive to β-sheets. The data in Table 1 support this, and provide some quantitative information about the extent to which it is true. In comparing the optimal rms values for the IR amide I and CD data types it was found that determinations of the α-helix content are (relatively) 18% better for CD than IR, and IR is (relatively) 30% more accurate than CD in β-sheet determination. There is also a difference in sensitivity to the other structural classifications listed in the table. CD determination of the C + B + G + S assignment proved to be about (relatively) 10% more accurate than IR, but the IR rms for FC_{I} analysis is lower than the rms for CD.

Also noteworthy are the results from the amide II band normalization results. It has long been recognized that the amide II band is conformationally sensitive. However, the dependence of amide II band shape on secondary structure is complex, so it has not been considered systematically for qualitative analyses. The results in Table 1 indicate that the PLS-1 algorithm was able to extract information from the isolated amide II band. In fact, amide II cross validation results for α-helix and turn were more accurate than analysis using only the amide I region. This finding suggests, for the first time, that the amide II band could be used alone for protein FC_{II} determination.

Combining data regions for analysis. While the spectroscopic signals in the amide I and II regions arise from γ(C = O), δ(N − H) and γ(C − N), the CD data arise from electronic transitions. It is therefore probable that they contain different independent structural information. Consequently, if they are combined into a single hybrid spectrum (Fig. 1), an analysis algorithm should be able to extract complementary information from each region and thus provide more accurate results. Such a combination has been tested before [8,19,20], including for vibrational and electronic CD spectra [41] or vibrational CD and IR [1] but the conclusions reached in these studies are contradictory. This is presumably because different basis sets and/or different mathematical methods were used. To resolve this conflict, cross validations with different data region
combinations were performed with the RaSP50 protein spectra. The results presented in Table 1 are given with the optimal normalization strategy for each structure type. It was found that for the IR data, combining the amide I and II bands substantially improved determinations of the α-helix and C + B + G + S structure assignments. Combined IR and CD data was more accurate than either of these methods alone. The relative improvements in determination accuracy using all three data regions compared to IR amide I band alone (Table 1) are 48% for α-helix, 5% for β-sheet (39% compared to CD alone), 12% for turn and 9% for the sum of the remaining assignment.

![Fig. 1. Concatenated CD (186–260 nm) and IR (1720–1500 cm⁻¹) spectra of the 50 proteins described in the Experimental procedures. Spectra have been rescaled and offset along the y-axis for a better readability. Proteins are sorted according to their α-helix content.](image-url)
The correspondence between secondary structure definitions and spectral features

Now that optimum spectral processing methods have been established, the other major input for statistical analysis, structure assignments, can be explored. Bond angles, H-bonds, tertiary structure, resonance, exciton coupling and side chain signals are all factors that contribute to protein spectral band shapes. Reducing this natural complexity to a small set of secondary structure assignments involves a simplification that obviously cannot accurately describe all aspects of IR or CD spectra.

The DSSP program [33] is currently the most widely used method for assigning secondary structure types to individual residues. DSSP makes eight structure assignments, each identified by a single letter code, including three types of helix (α-helix, H; 3_10-helix, G; π-helix, I), β-sheet (E), turns (T), and what we will refer to as irregular structures (C). While the first four structures are periodic, the DSSP program assigns two aperiodic structures in addition to T and C; these are typically found within stretches of C. They are isolated–extended (B), a residue with φψ angles in the β-sheet range that does not participate in a β-sheet, and bend (S), a sharp turn in the protein chain that does not meet all the criteria for a T assignment. From the descriptions of these assignments, the question naturally arises: where would the signals from these structures be expected to appear? For example, should the B structure give rise to a band characteristic of β-sheet, irregular structure, or will it have a unique signal? Similar questions apply to other assignments as well. Optimal determination accuracy can only be achieved by placing such residues in an appropriate assignment group. A few structure assignment combinations have appeared in the literature [9,35]. To tackle these questions in a systematic manner, the performance of PLS-1 analysis for various DSSP assignment combinations was evaluated.

Note that, due to their natural abundances, the FC variation (σCS) of some structure types (G, S, B and I) in the RaSP50 set is small. Accordingly, statistical analysis cannot be accurate for these structures unless their FC/signal correlation is strong. Most proteins in the RaSP50 set have 12–13% turn, and the 3_10-helix content is below 10% for all proteins except lysozyme and α-lactalbumin, with the majority having 3–6%. The π-helix (I) was essentially nonexistent in the RaSP50 proteins, and was therefore not considered a quantifiable structure.

Individual secondary structure assignments and their combinations. First, let us consider the performance of the individual assignments made by DSSP. The ζ scores given in Table 2 indicate how much information the analysis algorithm could extract from the reference spectra. This analysis was performed for all the structure types, alone or combined. A summary of the results appears in Table 2. The results suggest that the DSSP program overclassifies some secondary structures, at least as far as IR and CD spectra are concerned. That is, there are several structure types, which apparently do not give rise to unique, detectable spectral characteristics (for example, ζ ≈ 1 for the G and B assignments). Note however, that the FC distributions (σCS) for these structures are also narrow in the RaSP50 set. The only individual secondary structure assignments made by DSSP that are determined by the PLS algorithm with ζ ≥ 2 are α-helix (H) and β-sheet (E), none of the ζ scores for other individual assigned structures are higher than 1.25 (not shown).

Residues with different secondary structure assignments can also be combined into a single assignment class. Such grouping has a possible advantage in secondary structure determination. Grouping residues with different assignments that may have similar spectroscopic signals may also increase the sensitivity of analysis. The results show that some ζ scores could be improved by combining structures, such as C + S + B (compared to each of these structures alone), but none of the combinations tested had ζ scores comparable to ζH or ζE. In addition, combining other structures with α-helix (H + G or H + C) or β-sheet (E + B) did not improve the ζ scores compared to H and E determinations alone. The structure combination that was found to have the strongest FC/signal correspondence is C + T + G + S + B. It can be hypothesized that grouping all the structures but the α-helix and β-sheet must yield a prediction correlated with the α-helix and β-sheet prediction, as the value for C + T + G + S + B is simply 100 − (H − E).

The correspondence between secondary structure definitions and spectral features

Now that optimum spectral processing methods have been established, the other major input for statistical analysis, structure assignments, can be explored. Bond angles, H-bonds, tertiary structure, resonance, exciton coupling and side chain signals are all factors that contribute to protein spectral band shapes. Reducing this natural complexity to a small set of secondary structure assignments involves a simplification that obviously cannot accurately describe all aspects of IR or CD spectra.

The DSSP program [33] is currently the most widely used method for assigning secondary structure types to individual residues. DSSP makes eight structure assignments, each identified by a single letter code, including three types of helix (α-helix, H; 3_10-helix, G; π-helix, I), β-sheet (E), turns (T), and what we will refer to as irregular structures (C). While the first four structures are periodic, the DSSP program assigns two aperiodic structures in addition to T and C; these are typically found within stretches of C. They are isolated–extended (B), a residue with φψ angles in the β-sheet range that does not participate in a β-sheet, and bend (S), a sharp turn in the protein chain that does not meet all the criteria for a T assignment. From the descriptions of these assignments, the question naturally arises: where would the signals from these structures be expected to appear? For example, should the B structure give rise to a band characteristic of β-sheet, irregular structure, or will it have a unique signal? Similar questions apply to other assignments as well. Optimal determination accuracy can only be achieved by placing such residues in an appropriate assignment group. A few structure assignment combinations have appeared in the literature [9,35]. To tackle these questions in a systematic manner, the performance of PLS-1 analysis for various DSSP assignment combinations and subclassifications was evaluated.

Note that, due to their natural abundances, the FC variation (σCS) of some structure types (G, S, B and I) in the RaSP50 set is small. Accordingly, statistical analysis cannot be accurate for these structures unless their FC/signal correlation is strong. Most proteins in the RaSP50 set have 12–13% turn, and the 3_10-helix content is below 10% for all proteins except lysozyme and α-lactalbumin, with the majority having 3–6%. The π-helix (I) was essentially nonexistent in the RaSP50 proteins, and was therefore not considered a quantifiable structure.

Individual secondary structure assignments and their combinations. First, let us consider the performance of the individual assignments made by DSSP. The ζ scores given in Table 2 indicate how much information the analysis algorithm could extract from the reference spectra. This analysis was performed for all the structure types, alone or combined. A summary of the results appears in Table 2. The results suggest that the DSSP program overclassifies some secondary structures, at least as far as IR and CD spectra are concerned. That is, there are several structure types, which apparently do not give rise to unique, detectable spectral characteristics (for example, ζ ≈ 1 for the G and B assignments). Note however, that the FC distributions (σCS) for these structures are also narrow in the RaSP50 set. The only individual secondary structure assignments made by DSSP that are determined by the PLS algorithm with ζ ≥ 2 are α-helix (H) and β-sheet (E), none of the ζ scores for other individual assigned structures are higher than 1.25 (not shown).

Residues with different secondary structure assignments can also be combined into a single assignment class. Such grouping has a possible advantage in secondary structure determination. Grouping residues with different assignments that may have similar spectroscopic signals may also increase the sensitivity of analysis. The results show that some ζ scores could be improved by combining structures, such as C + S + B (compared to each of these structures alone), but none of the combinations tested had ζ scores comparable to ζH or ζE. In addition, combining other structures with α-helix (H + G or H + C) or β-sheet (E + B) did not improve the ζ scores compared to H and E determinations alone. The structure combination that was found to have the strongest FC/signal correspondence is C + T + G + S + B. It can be hypothesized that grouping all the structures but the α-helix and β-sheet must yield a prediction correlated with the α-helix and β-sheet prediction, as the value for C + T + G + S + B is simply 100 − (H − E).
Subclassification of helix and sheet assignments. Several authors have attempted to improve the accuracy of structure determination by dividing helix into ordered, H(O), or disordered, H(do), subclassifications or by discriminating between antiparallel, E(AP), and parallel, E(par), β-sheet [6, 8, 19]. This is motivated by the idea that the differences between the geometries, H-bonding, etc. of these structures should be sufficient to produce differing band shapes that could be modeled by statistical analysis algorithms. Typically, this practice results in lower rms values for the segregated structures (Table 2). A reduction of rms values for segregated structures was also observed in this study, but subdividing α-helix and β-sheet classifications also reduces the corresponding σCS values. The consequence is usually lower ς scores for both substructures. Therefore segregation actually degrades analysis performance in general. The only exceptions were found to be the removal of kinks from α-helices, H(O), which increased the determination accuracy for one of the subassignments (ςHIR = 3.29 → ςIHO < 1, IR = 3.49, ςLCD = 2.90 → ςG1CD < 1, CD = 2.98), and β-sheet segregation which improved the accuracy of IR determinations for antiparallel β-sheet (ςEIR = 2.44 → ςEAPIR < 2.8). The ς scores for the counterparts of these structures, ςH(do < 1) and ςE(par), were close to one.

As for the remaining DSSP structures, a ς score larger than 1.4 is obtained only when C and T are combined for analysis (C + T + G + S + B, and other combinations that include C + T). If more detailed structural information is desired, grouping sharp turns in the amide backbone (T + S or T + G + S) may provide some information if FC(T+S) or FC(T+G+S) are unusually high in the protein. Similarly, the C + B + S + G or the C + B (for IR) structure combinations also show moderate correlation to protein spectra band shapes (ς ≈ 1.3).

The complementarities of information in CD and IR spectra. By comparing the ς scores in the amide I + II and CD columns of Table 2, it can be seen that their information contents are comparable for all structure types. However, the change in the ς scores between cross validations from separate CD or IR spectra and the hybrid spectra (IR + CD) reveal that there can be complementarities between the information contained in each data region. This point is investigated further in the next section of the paper.

Estimating accuracy in the secondary structure determination of an unknown

The rms, ς scores and correlation coefficients presented in the tables above are all summaries of the overall performance characteristics of protein structure statistical analyses. While these values allow different methods to be compared, the question asked during the analysis of an unknown is typically not ‘How accurate is the method in general?’, but rather ‘How accurate was the analysis for this protein?’. In theory, there are several quantities that can be derived from a statistical analysis that can assist in answering this question. An accuracy evaluation procedure would be most useful if it could define an expected margin of error for each unknown analysis. Presumably, it should be possible to derive additional information from the way that a statistical model reconstructs the spectrum of the unknown protein.

Quality of the fit

The match between the original and reconstructed spectra can be evaluated by taking the difference between the two. The sum of the absolute values of the differences between the original and fit spectra at each point is used here to
characterize the residuals of the fit. By comparing the residuals of an unknown with the residuals of the reference spectra obtained during cross validation, an estimation of reliability could be made. A small residual for a given analysis is usually considered as a good indication of an accurate, reliable analysis. Such statistical properties of chemometric analyses are meaningful in systems where spectra vary in a purely concentration-dependent manner, but because protein spectra do not necessarily follow this rule, the quality of the fit may not be related to the accuracy of the analysis. In fact, the fit residuals provide what is perhaps the most convincing indication that statistical analysis of protein spectra is fundamentally different than simpler quantitation systems. The situation is illustrated in Fig. 2A, in which the error for FC det determination in cross validation of the RaSP50 set is plotted against the spectral residual for the fit to each protein. Similar plots were obtained for all other structure assignments (not shown). The correlation coefficient (a linear regression R2) for the data plotted here is 0.024, indicating that the ability of the algorithm to reconstruct the protein spectrum has essentially no relationship to the accuracy of the analysis.

The Mahalanobis distance of the factors

Another accuracy validation criterion, based on values derived in the model-construction step, has appeared in protein structure analysis method reports. When using the Mahalanobis distance as a reliability evaluation criterion, the set of factor scores for each spectrum is treated as a vector in a coordinate system defined by the factors in a statistical model (f-space). That is, each axis in f-space is one of the factors, and the coordinates of a spectrum in f-space are its factor scores. Typically the score vectors for the reference spectra form an ellipsoid in f-space. The Mahalanobis distance is a measure of how far from the center of this ellipsoid the score vector for a given spectrum lies. If the score vector for an unknown falls significantly outside the ellipsoid formed by the reference spectra score vectors, the scores for the unknown therefore follow a different pattern than those of the basis set. It can be suggested that a large Mahalanobis distance for an unknown indicates that the statistical analysis algorithm was unable to properly evaluate the structure of the unknown protein.

The Mahalanobis distances vs structure determination error for β-sheet for the RaSP50 set cross validation is shown in Fig. 2B. It is clear that the Mahalanobis distance is also not a useful validation method for protein spectra, at least within the proteins of the RaSP50 basis set. The finding that the Mahalanobis distance does not correlate with analysis accuracy for the RaSP50 set has important consequences for structure determination in that it shows that a novel pattern of factors needed to reconstruct the spectrum of an unknown protein is not a reliable indication of its structure.

Fig. 2. Test of potential predictors for the structure prediction accuracy.

(A) Relationship between the FCdet error (FC determination error for α-helix) and the spectral residual from reconstructing unknowns with the factors from a PLS-1 model. The residual is characterized by the sum of the absolute values of the difference between the actual and reconstructed spectra at each point. These data were obtained from a cross validation of hybrid RaSP50 IR + CD spectra. Spectral preprocessing parameters were optimal. (B) Mahalanobis distances for the factor scores (significant factors only) in the FC determination of FCE. These data were obtained from a cross validation of hybrid RaSP50 IR + CD spectra. Spectral preprocessing parameters were optimal for FC det determination. C. Comparison of the sum of all determined structures for an unknown (residual for SFCdet) with FC det determination error. ○, FCdet errors of individual proteins; compared with ·, FCdet. These data were obtained from a cross validation of hybrid RaSP50 IR + CD spectra.
of a failed analysis. Conversely, the results are not necessarily reliable for unknowns whose score vectors fall within the same region of f-space as the reference-spectra score vectors.

**Do the structure fractional contents total 100%?**

Another potential measure of statistical analysis error for an unknown is the results themselves. Because the FC values should account for all the residues in a protein, they should total 100%. If the total is not close to 100%, then it is reasonable to question the analysis results. The variable selection method [11], as well as others [35,41] use this as a criterion for evaluating the quality of analysis results. In particular, this was found to be very useful to build the SelCon method [5]. As for the residuals and the scores, the determination errors for individual proteins are compared with this accuracy measure in Fig. 2C. Again, there is no apparent relationship between these quantities. Therefore the sum of FC\(_{\text{det}}\) values cannot be used to diagnose analysis accuracy or failure. This finding is important: it indicates that the determination accuracy for each secondary structure type is independent of the other structures that have been analyzed. Therefore, it is not appropriate to disregard the analysis results in their entirety if a single determined FC is questionable.

**IR/CD comparison**

We propose that a more reliable method of evaluating analysis results is the consistency between analysis results from different structure-sensitive techniques, such as IR and CD. Because infrared and CD spectra depend on different phenomena, particular structural distortions are likely to have a different effect on each of these spectral types, and that these differences can be used to evaluate analysis accuracy. In the simplest case, it is necessarily true that when the FCs obtained from separate analyses of IR and CD spectra of a protein are very different, then at least one of the determined structures must be `incorrect’. For convenience, we will refer to this difference, specifically FC\(_{\text{IR}}\) − FC\(_{\text{CD}}\), as ΔIRCD\(_{\text{det}}\).

To illustrate the type of information that is available from ΔIRCD\(_{\text{det}}\), the FC\(_{\text{H}}\) determination errors for individual proteins from cross validation of IR-only and CD-only data is plotted against ΔIRCD\(_{\text{det}}\) in Fig. 3. An intuitive relationship is revealed by this figure: when the FC\(_{\text{H}}\) determined with CD alone is lower than the FCH determined with CD alone is lower than the FC\(_{\text{H}}\) from IR only (ΔIRCD\(_{\text{det}}\) is positive), then the CD analysis result tends to strongly underestimate the actual FC\(_{\text{H}}\). A similar relationship holds when the FC\(_{\text{H}}\) from IR is the lower value. The linear regression correlation coefficients (R\(^2\)) for the data plotted in Fig. 2A,B are 0.345 and 0.434, respectively, which indicate a definite relationship. It appears that ΔIRCD\(_{\text{det}}\) is the only quantity examined in this study which has any significant correlation with analysis error.

In an attempt to evaluate the potentiality of the test provided by the ΔIRCD\(_{\text{det}}\) measure, the ΔIRCD\(_{\text{det}}\) value for each protein was used to divide the RaSP50 members into two subsets with different analysis characteristics. The first subset was defined as the proteins with ΔIRCD\(_{\text{det}}\) (absolute value of ΔIRCD\(_{\text{det}}\)) smaller than 6%. The rms FC\(_{\text{H}}\) determination errors calculated for this subset of proteins were rmsE\(_{\text{H,IR}}\) = ± 4.82% H and rmsE\(_{\text{H,CD}}\) = ± 4.46% H. Combining these results with the \(\alpha\)-helix \(\sigma_{\text{CS,subset}}\) for these 27 proteins in the \(\zeta\) score equation gives \(\zeta_{\text{subset}}\) values of 4.24 and 4.58, respectively (compare with data in Table 1). If the hybrid IR + CD spectra analysis results for these same proteins is considered, the rms error is ± 4.46% H.

These results show that the margin of error for FC\(_{\text{H}}\) determination is reduced when the results from separate IR and CD analyses are similar (|ΔIRCD\(_{\text{det}}\| < 6\%\)). However, this accounts for just over half of the proteins in the RaSP50 set. If we consider the remaining proteins, overall the larger determined FC\(_{\text{H}}\) was more accurate for 74% of the proteins in this second RaSP50 subset. For \(\beta\)-sheet, a similar observation was made for the ΔIRCD\(_{\text{det}}\) > 6% E proteins, but IR was also more accurate for eight out of 11 proteins. Therefore, the more accurate result is likely to come from IR analysis. In conclusion, ΔIRCD\(_{\text{det}}\) can be used to identify proteins with anomalous spectra (|ΔIRCD\(_{\text{det}}\| > 6\%\)), and therefore assist in the identification of failed analyses.
Comparison of different statistical analysis algorithms

Thus far, the discussion has focused on the optimization of input data for protein structure analysis methods. We will now briefly address the role that the algorithms themselves play in analysis accuracy. Recently, Sreerama and Woody [6] demonstrated on a large set of CD spectra that the algorithm used (CONTIN, SELCON or CDSSTR) has little effect on the rms. We have used the RaSP50 set to compare different methods on the IR, CD and combined IR/CD on the broad range of structures represented in RaSP50. It was found (Table 3) that the choice of analysis method has only a small effect on analysis accuracy.

By examining selected literature data, it can be observed that there is a relationship between the number of protein folds represented in a basis set and the rms. Contrary to what would be expected, the general trend is for the error to increase with the number of proteins used (e.g [41]). For the CD analyses the relationships for FC_H and FC_E are well represented by straight lines (not shown). Combining this observation with the frequency of anomalous spectra just described suggests that those authors who have introduced more spectra into their reference sets have increased the number of proteins with anomalous spectra. Through this, they have degraded the quality of the spectra-structure relationship in their statistical models. We suggest that for small protein basis sets we obtain primarily a measure of the internal consistency (lack of anomalous spectra) rather than their expected performance in general. In order to test this hypothesis, a RaSP50 subset was assembled using 16 of the most common proteins in the IR studies with attention given to maintaining a broad FC distribution. This set, RaSP16, was tested both in cross validation and on the spectra of the full RaSP50 set. We found that the rms values for the RaSP50 set are generally lower than for the RaSP16 set. However, when the RaSP16 statistical model was used to determine the structures of all the proteins in the RaSP50 set, its accuracy was ≈ 28% (relatively) worse than when predictions were made with the RaSP50 model. This shows that there is information in the RaSP50 statistical model that is lacking from the RaSP16 model. In a second step, we randomly generated hundreds of different other 16-protein databases. Even though the accuracy of the secondary structure prediction evaluated in cross validation yielded generally RMSs better than the RaSP50, none of them was able to satisfactorily describe the RaSP50 proteins left out when building the 16-protein subset. It is impossible that the proteins in the RaSP50 set are representative of all possible structural distortions, so one can ask how much information may be lacking from the RaSP50 model. Of course, a definitive answer to this question cannot be given, but it is possible to estimate the amount of information describing anomalous signal is contained in the RaSP50 statistical model. This can be done by comparing the standard errors from cross validation and self-validation of the set. Consider that when an anomalous spectrum is removed from the set during cross validation, if the information needed to model that spectrum is not contained in the remaining spectra then the FC determination error for that protein will be high. This in turn will increase the rms. However, in self validation, all the information contained in the full basis set can be used to model each spectrum. Therefore, the difference between the standard errors of self and cross validations gives an indication of the completeness of the information contained in the basis spectra. For example, the standard errors of self validation for the RaSP16 set were 3.45% H and 2.58% E which are essentially half as large as the RaSP50 rms values.

The latter analyses demonstrate that when the proteins are carefully selected for the diversity in their structure, no small subset of the database contains the necessary information to describe the entire set. One conclusion of the paper is therefore that large protein databases, observing stringent selection criteria, are necessary for the prediction

Table 3. Performance comparison of different analysis algorithms with the RaSP50 set. PCA-MR, principal component analysis followed by multiple regression, constrained to a 100% total; PLS, simultaneous partial least-squares analyses of all structure classes, constrained to a 100% total; PLS-1, separate partial least-squares analyses of each structure type with the use of weighting during the spectral decomposition step. SelCon has been described in detail in [5].

<table>
<thead>
<tr>
<th>Data</th>
<th>Algorithm</th>
<th>⊗-Helix (H)</th>
<th>β-Sheet (E)</th>
<th>Turn (T)</th>
<th>Σ Other (C + G + B + S)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RMS</td>
<td>ζ</td>
<td>R²</td>
<td>RMS</td>
</tr>
<tr>
<td>IR (Amide)</td>
<td>SelCon3</td>
<td>8.5</td>
<td>2.6</td>
<td>0.92</td>
<td>7.73</td>
</tr>
<tr>
<td>I + II</td>
<td>PCA-MR</td>
<td>6.91</td>
<td>3.2</td>
<td>0.95</td>
<td>7.64</td>
</tr>
<tr>
<td></td>
<td>PLS</td>
<td>7.29</td>
<td>3.03</td>
<td>0.94</td>
<td>7.58</td>
</tr>
<tr>
<td></td>
<td>PLS-1</td>
<td>7.16</td>
<td>3.09</td>
<td>0.95</td>
<td>7.36</td>
</tr>
<tr>
<td>IR + CD</td>
<td>SelCon3</td>
<td>7.57</td>
<td>2.91</td>
<td>0.939</td>
<td>7.97</td>
</tr>
<tr>
<td></td>
<td>PCA-MR</td>
<td>6.83</td>
<td>3.24</td>
<td>0.95</td>
<td>7.23</td>
</tr>
<tr>
<td></td>
<td>PLS</td>
<td>6.8</td>
<td>3.25</td>
<td>0.95</td>
<td>6.97</td>
</tr>
<tr>
<td></td>
<td>PLS-1</td>
<td>6.73</td>
<td>3.28</td>
<td>0.95</td>
<td>6.68</td>
</tr>
<tr>
<td>CD</td>
<td>SelCon3</td>
<td>8.15</td>
<td>2.71</td>
<td>0.91</td>
<td>10.43</td>
</tr>
<tr>
<td></td>
<td>PCA-MR</td>
<td>7.97</td>
<td>2.77</td>
<td>0.93</td>
<td>9.37</td>
</tr>
<tr>
<td></td>
<td>PLS</td>
<td>7.72</td>
<td>2.86</td>
<td>0.94</td>
<td>9.47</td>
</tr>
<tr>
<td></td>
<td>PLS-1</td>
<td>7.7</td>
<td>2.87</td>
<td>0.94</td>
<td>9.22</td>
</tr>
</tbody>
</table>

a The correlation coefficient (R) between the determined and actual FCs for the full RaSP50 set.
of unknown proteins. A second important conclusion of the paper is that only the comparison of analyses run on CD and IR spectra independently is able to identify failed solutions in the absence of known structure.

As far as the specific case of membrane protein is concerned, the issue has been raised a number of times but it is now definitively settled. Fasman claimed that the transmembrane and peripheral helices could be distinguished on the basis of their deconvolved CD spectrum [46,47]. Wallace investigated the performance of soluble protein sets of CD spectra in analyzing membrane protein CD spectra. The conclusion was that the soluble protein reference set of CD spectra yields inaccurate results for membrane protein CD spectra [48,49]. Conversely, Sreerama and Woody showed recently for 13 membrane proteins that there is no systematic difference in the CD spectra of soluble and membrane proteins. Yet, they reported that increasing the number of proteins in the CD spectrum database from soluble proteins is an important factor to improve the prediction. Similarly, the additional inclusion of the CD spectra of membrane proteins brings a small but significant additional improvement. In the field of infrared spectroscopy, anomalous frequencies have been described for the transmembrane helices of proteins such as bacteriorhodopsin [50] but most of the membrane proteins have been shown to have a typical helix contribution originating from the transmembrane helices. In conclusion, more work is needed to assess the quality of the prediction from membrane proteins.

Acknowledgements

This work was funded by an ARC grant (Action de Recherche Concertée, Communauté française de Belgique, Belgium). Dr Goormaghtigh is Research Director at the Belgian National Fund For Scientific Research, Belgium.

References


