Protein Secondary Structure from Circular Dichroism Spectroscopy

Combining Variable Selection Principle and Cluster Analysis with Neural Network, Ridge Regression and Self-consistent Methods

Narasimha Sreerama and Robert W. Woody

Department of Biochemistry and Molecular Biology
Colorado State University, Fort Collins, CO 80523, U.S.A.

Different approaches to improve the analysis of protein secondary structure from circular dichroism spectra are compared. Grouping proteins based on the similarity of their circular dichroism spectra, using cluster analysis methods, was utilized as a new way of implementing variable selection. The performance of three basic methods (neural networks, ridge regression and singular value decomposition) was evaluated in combination with three approaches to improve the predictions; namely, variable selection, cluster analysis and the self-consistent method. Cluster analysis performed on the basis set proteins resulted in three clusters, subanalyses of which provide a new way of performing variable selection. The neural network with two hidden layers performed better than that with one hidden layer and was combined with variable selection. Inclusion of the variable selection principle improved the performance of all three basic methods. While the neural network method performed slightly better than the other two methods at the basic level, the inclusion of variable selection led to similar performance indices for all three methods.

Keywords: circular dichroism; protein secondary structure; neural networks; cluster analysis; variable selection

1. Introduction

Circular dichroism (CD) spectroscopy is a valuable tool to study the conformation of biomolecules, particularly polypeptides and proteins. Different secondary structures of proteins (α-helix, β-sheet, β-turns and unordered) give characteristic CD spectra (Brahms & Brahms, 1980; Woody, 1985; Yang et al., 1986). A variety of methods has been proposed to estimate the fractions of protein secondary structure from the CD spectrum (Greenfield & Fasman, 1969; Chen & Yang, 1971; Chen et al., 1974; Chang et al., 1978; Brahms & Brahms, 1980; Bolotina et al., 1980; Hennessey & Johnson, 1981; Provencher & Glückner, 1981; Manavalan & Johnson, 1987; van Stokkum et al., 1990; Panecoska & Keiderling, 1991; Perczel et al., 1991; Böhm et al., 1992; Sreerama & Woody, 1993).

The CD spectrum of a protein ($C_i$) can be represented as a linear combination of secondary structural component spectra ($S_{jk}$) as:

$$C_i = \sum_{j} f_j S_{jk},$$

where $f_j$ is the fraction of the $j$th secondary structure (Woody, 1985; Yang et al., 1986; Johnson, 1988, 1990). It should be recognized that this is an approximation, because not all residues of a given secondary structural type make identical contributions to the protein CD, and because features other than secondary structure contribute. These factors are discussed in the penultimate section of the paper. The analytical methods developed for deconvoluting the protein CD spectrum and estimating the fractions of secondary structures are based on the linear relation between the spectrum and the secondary structure. The earlier methods made use of the CD spectra of model polypeptides in specific secondary structures.
as the pure component spectra, $S_{ij}$, and estimated secondary structural fractions by obtaining a least-squares fit for $C_2$. (Greenfield & Fasman, 1969; Rosenkranz & Scholten, 1971; Brahms & Brahms, 1980). The model polypeptide spectra were later replaced by the pure component spectra derived from the CD spectra of a set of proteins with known secondary structures (Saxena & Wetlauffer, 1971; Chen et al., 1974; Chang et al., 1978; Bolotina et al., 1980). Subsequently, methods based on least-squares procedures were developed to provide a stable solution to the set of linear equations relating the CD spectra of proteins in the basis set to their structures, and to estimate the secondary structure of proteins outside the basis set (Provencher & Glöckner, 1981; Hennessey & Johnson, 1981; Pancoska & Keiderling, 1991). More recently, neural network methods have been used to relate the CD spectra to the secondary structure and solve for the unknown structures (Bohm et al., 1992). While the methods cited above make use of the structural information from X-ray diffraction, the convex constraint analysis described by Perzel et al. (1991) does not require the input of structural information for deconvoluting the CD spectra.

There are at least six different methods for deconvoluting CD spectra; namely, ridge regression (Provencher & Glöckner, 1981), singular value decomposition (Hennessey & Johnson, 1981; Compton & Johnson, 1986; Manavalan & Johnson, 1987; van Stokkum et al., 1990; Tounadje et al., 1992; Srerama & Woody, 1993), neural network method (Bohm et al., 1992), principal component factor analysis (Pancoska & Keiderling, 1991), convex constraint analysis (Perzel et al., 1991) and the simple least-squares procedure (Chen & Yang, 1971; Chen et al., 1974; Bolotina et al., 1980). Of these methods, principal component factor analysis is mathematically similar to singular value decomposition (Pancoska & Keiderling, 1991), ridge regression uses a regularizer with the least-squares method (Provencher & Glöckner, 1981), and convex constraint analysis does not use the X-ray information for deconvoluting the CD spectra (Perzel et al., 1991). The original papers used different assignments of secondary structures for the X-ray structures (Chang et al., 1978; Hennessey & Johnson, 1981; Levitt & Greer, 1977; Kabach & Sander, 1983) and different sets of reference spectra.

In this work, we have considered three of these methods (ridge regression, singular value decomposition and neural network) for a detailed comparison with a common basis set of proteins and the Kabach & Sander (1983) assignment of secondary structures. We have examined the inclusion of a variable selection principle into these methods for improvements in secondary structure estimation. We present the results of combining the variable selection principle, as implemented in the locally linearized model, and the cluster analysis method, with the ridge regression, self-consistent and neural network methods.

2. Materials and Methods

(a) Basis set

The basis set consisted of an all-$\alpha$ polypeptide, poly(L-glutamic acid), and the following 16 proteins: Bence-Jones protein, prealbumin, rubredoxin, chymotrypsin, elastase, papain, thermolysin, lysozyme (egg white), subtilisin BPN, glyceraldehyde-3-phosphate dehydrogenase, flavodoxin, lactate dehydrogenase, triosephosphate isomerase, cytochrome c, hemoglobin and myoglobin. The first five are $\beta\beta$ proteins, the next three are $\alpha + \beta$ proteins, the next five are $\alpha/\beta$ proteins and the last three are $\alpha$ proteins (Levitt & Chothia, 1976). The CD data used were in the range 175 to 260 nm at a resolution of 1 nm, unless otherwise noted. The CD spectra of these proteins were kindly provided to us by Dr. W.C. Johnson, Jr. The Kabach & Sander (1983) method was used to assign the secondary structure elements. Seven secondary structure elements identified by the Kabach & Sander (1983) method (in their notation: H, $\alpha$-helix; G, 3$\alpha$-helix; I, $\pi$-helix; E, $\beta$-sheet; B, $\beta$-bridge; T, $\beta$-turns; S, bends) were used to calculate the secondary structure fractions as follows: H + G, $\alpha$-helix ($\alpha$); E + B, $\beta$-sheet ($\beta$); T + S, turns (T); and I + unassigned, unordered (U).

(b) Computer programs

The computer programs for performing the locally linearized method, the self-consistent method and the cluster analysis were developed in our laboratory. A modified version of CONTIN (Provencher, 1982), the computer program for performing ridge regression, was provided by Dr S. Yas. Venyaminov. The computer program VARSELIC for performing the variable selection method was provided by Dr W. C. Johnson, Jr. Neural network calculations were performed using NeuralWorks Professional II/PLUS software, NeuralWorks, Inc., Pittsburgh.

3. Results and Discussion

The analysis was carried out by removing one protein at a time from the basis set and predicting its secondary structure using the remaining proteins in the basis set. The results from different methods were compared using the correlation coefficient ($r$) and the r.m.s. differences ($\delta$) between the X-ray and the predicted structures of all proteins in the basis set. (The definitions of $r$ and $\delta$ are given in Table 1.) Two sets of $r$ and $\delta$ were calculated. The first set, which contained $r$ and $\delta$ for each secondary structure separately, gives the performance of a given method with respect to each secondary structure prediction (denoted by $r_s$, $\delta_s$; $r_f$, $\delta_f$; $r_f$, $\delta_f$; and $r_U$, $\delta_U$ for the four secondary structures, respectively). The second set, calculated by considering all secondary structure fractions collectively, indicates the overall performance of a given method (denoted by $r$, $\delta$).

(a) Methods based on singular value decomposition

The original papers describing the variable selection method (Manavalan & Johnson, 1987), the locally linearized model (van Stokkum et al., 1990) and the self-consistent method (Srerama & Woody,
used the singular value decomposition algorithm to solve the set of linear equations relating the CD spectra of basis set proteins to their secondary structures. The differences between these methods are in the selection of the eigenvectors and in performing the variable selection.

The principle of variable selection is quite simple. The proteins in the basis set that contribute adversely to the solution, because they do not have factors that are found in the protein analyzed or have factors that are not found in the protein being analyzed, need to be removed from the basis set for a better solution. Since one does not know initially which of the proteins should be removed, proteins are deleted in a systematic manner and the calculation is performed on all possible combinations of the basis set proteins until a satisfactory solution is obtained (Manavalan & Johnson, 1987). A more efficient way of performing the variable selection is the locally linearized model (LL) of van Stokkum et al. (1990), wherein the proteins in the basis set are arranged in the order of increasing r.m.s. distance of the CD spectra from that of the protein analyzed, and the more distant proteins are deleted in a systematic manner.

Prediction improves when the protein analyzed is included in the basis set (Provencher & Glöckner, 1981; Hennessey & Johnson, 1981; Woody, 1985), since the solution is biased towards the test protein structure. This feature is utilized in the self-consistent method (SC) of Srerarama & Woody (1993), which includes the spectrum of the protein analyzed in the basis set and makes an initial guess for the unknown secondary structure. The solution replaces the initial guess and the process is repeated until self-consistency is attained. The self-consistent method also incorporates the variable selection principle in the LL framework.

Hennessey & Johnson (1981), in the first application of singular value decomposition, used the first five singular values that were found to reproduce the CD spectra to within the experimental error. We followed a similar approach with the basis set considered in this work and the results are reported, for comparison, under HJ. The variable selection method (Manavalan & Johnson, 1987) also retained the first five significant eigenvalues. The results reported under VS (Table 1), which were kindly provided by W. C. Johnson, Jr., were obtained with a basis set of 22 proteins (17 proteins of the basis set used in this paper and five additional proteins: ribonuclease A, subtilisin, coxboxinpeptidase, concanavalin A and hemerythrin). Results for the 17 proteins of our basis set were used in obtaining the performance indices given in Table 1 for VS. (According to the nomenclature followed here, Manavalan and Johnson’s variable selection method can be considered as a combined approach and can be termed as HJ/Vs.) The number of significant eigenvalues was varied in the locally linearized approach from one to seven, depending on the number of proteins in the reduced basis set, and the allowed solutions were selected and averaged using the constraints, $f_j > -0.05; \sum f_i - 1.0 \leq 0.1$ (van Stokkum et al., 1990; Srerarama & Woody, 1993). The locally linearized method is equivalent to HJ/LL and the results are given under LL, following Srerarama & Woody (1993). The self-consistent method incorporates the LL and the self-consistent approaches in the HJ method. The results are tabulated under SC, keeping the original abbreviation (Srerarama & Woody, 1993), though the method is equivalent to HJ/LL/SC. The results from HJ, VS, LL and SC methods are given in Table 1.

The HJ method gave better performance indices for the a-helix fraction; the overall performance, however, was worse. Improvement over the HJ method was obtained from VS, LL and SC methods for almost all of the proteins, except cytochrome c and prealbumin. The results from VS, LL and SC were comparable (Table 1). This is because, as pointed out earlier, LL is actually an efficient way of performing VS and is equivalent to VS (Srerarama & Woody, 1993; W. C. Johnson, Jr., personal communications). The variable selection, in principle, removes the dependence of the prediction on the basis set as long as the important proteins for a given protein are included.

(b) Performance of different neural network topologies

In an artificial neural network, neurons are grouped in several interconnected layers; a neuron in a given layer is connected to all neurons in the next layer through numerically weighted connections; information is passed through these connections and processed in the neurons (Rummelhart & McClelland, 1986; Simpson, 1990). The network learns by adjusting the numerical weights iteratively until all patterns presented to the input layer are correctly projected on the output layer in the training phase. The weights thus determined can be used to obtain outputs corresponding to new inputs in the test phase. We used the back-propagation algorithm (Rummelhart et al., 1986; Böhm et al., 1992), in which random weights are assigned to the neural connections, and for each input pattern the output is calculated as either a linear or a sigmoidal function of the input; the sigmoidal transfer function is known to be superior (Rummelhart & McClelland, 1986; Eberhart & Dobbins, 1990). The difference in the actual output and the calculated output is the residual error, which is now used to change the weights of the neural connections. The process is repeated until the residual error in the output reaches a minimum, when the training is complete.

In the application to CD analysis, CD data at wavelengths from 178 to 260 nm form the 83 input neurons, the secondary structure data form the four output neurons, and there are one or more hidden layers with predetermined numbers of neurons. The patterns corresponding to the basis set proteins, excluding the test protein, are presented in the training phase and the output for the test protein gives its secondary structure. Böhm et al. (1992) used 13 proteins in the first application of neural network methods for the analysis of CD spectra, with the back-propagation algorithm and a linear transfer.
function. They used one hidden layer of 45 neurons connecting 83 input neurons, corresponding to 83 wavelengths from 178 to 260 nm, and five output neurons, corresponding to five secondary structural elements. They varied the basis set proteins by considering various sets of nine proteins in the learning cycle and obtaining predictions for the remaining four, so this represents an implementation of the variable selection principle.

We performed analyses with different topologies of neural networks (different numbers of neurons, and different numbers of hidden levels), before attempting the combination with variable selection. The different topologies that we considered were: (1) NN with no hidden layer and sigmoidal transfer function (NN0), that, in principle, is equivalent to the least-squares procedure; (2) NN with one hidden layer and sigmoidal transfer function (NN1); (3) NN with two hidden layers and sigmoidal transfer function (NN2). More extensive calculations were carried out using NN1 to determine the number of neurons necessary for the hidden layer to get better results. The results obtained with 5, 10 and 15 neurons in the hidden layer were inferior, and those obtained with 20 or more neurons were comparable. We chose 45 neurons for the hidden layer to be consistent with the selection of Böhm et al. (1992). For NN2, we chose the 40–20 topology. We also considered linear transfer function in our networks, but the results were inferior to those obtained with sigmoidal transfer function.

The networks converged to residual r.m.s. error in the output (0.01 to 0.08) with less than 200,000 cycles. To avoid overtraining the network, output was obtained during intermediate stages of learning (approximately every 25,000 cycles) and these intermediate outputs converged. The network with NN2 topology had the least residual error in the output, and the residual error increased as the number of hidden layers increased. Results obtained with NN0, NN1 and NN2 networks are summarized.

Table 1

<table>
<thead>
<tr>
<th>Method</th>
<th>Alpha-helix</th>
<th>Betasheet</th>
<th>Beta-turns</th>
<th>Unordered</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\delta_1$</td>
<td>$\tau_1$</td>
<td>$\delta_2$</td>
<td>$\tau_2$</td>
</tr>
<tr>
<td>PG</td>
<td>0.11</td>
<td>0.535</td>
<td>0.15</td>
<td>0.558</td>
</tr>
<tr>
<td>PG/LL</td>
<td>0.09</td>
<td>0.638</td>
<td>0.09</td>
<td>0.845</td>
</tr>
<tr>
<td>PG/CA</td>
<td>0.08</td>
<td>0.963</td>
<td>0.10</td>
<td>0.816</td>
</tr>
<tr>
<td>PG/SC</td>
<td>0.23</td>
<td>0.615</td>
<td>0.23</td>
<td>0.184</td>
</tr>
<tr>
<td>PG</td>
<td>0.07</td>
<td>0.969</td>
<td>0.11</td>
<td>0.701</td>
</tr>
<tr>
<td>PG1/LL</td>
<td>0.07</td>
<td>0.962</td>
<td>0.16</td>
<td>0.896</td>
</tr>
<tr>
<td>PG1/CA</td>
<td>0.06</td>
<td>0.977</td>
<td>0.08</td>
<td>0.871</td>
</tr>
<tr>
<td>PG1/SC</td>
<td>0.08</td>
<td>0.964</td>
<td>0.11</td>
<td>0.732</td>
</tr>
<tr>
<td>HJ</td>
<td>0.05</td>
<td>0.970</td>
<td>0.13</td>
<td>0.656</td>
</tr>
<tr>
<td>VS</td>
<td>0.07</td>
<td>0.970</td>
<td>0.16</td>
<td>0.810</td>
</tr>
<tr>
<td>LL</td>
<td>0.06</td>
<td>0.963</td>
<td>0.08</td>
<td>0.883</td>
</tr>
<tr>
<td>SC</td>
<td>0.08</td>
<td>0.962</td>
<td>0.07</td>
<td>0.887</td>
</tr>
<tr>
<td>SC/CA</td>
<td>0.07</td>
<td>0.980</td>
<td>0.09</td>
<td>0.805</td>
</tr>
<tr>
<td>NX0</td>
<td>0.08</td>
<td>0.952</td>
<td>0.14</td>
<td>0.571</td>
</tr>
<tr>
<td>NN1</td>
<td>0.10</td>
<td>0.922</td>
<td>0.13</td>
<td>0.828</td>
</tr>
<tr>
<td>NN2</td>
<td>0.10</td>
<td>0.922</td>
<td>0.11</td>
<td>0.730</td>
</tr>
<tr>
<td>NN/LL</td>
<td>0.09</td>
<td>0.946</td>
<td>0.11</td>
<td>0.760</td>
</tr>
<tr>
<td>NX/CA</td>
<td>0.08</td>
<td>0.956</td>
<td>0.08</td>
<td>0.806</td>
</tr>
</tbody>
</table>

The r.m.s. deviation $\delta$ and correlation coefficient $\tau$ were calculated using equations:

$$\delta = \sqrt{\frac{1}{N} \sum (f_i - \bar{f})^2}$$

and

$$\tau = \frac{\sum (f_i - \bar{f}) (f_i' - \bar{f}')}{\sqrt{\sum (f_i - \bar{f})^2 \sum (f_i' - \bar{f}')^2}}$$

where $f_i$ and $f_i'$ are the X-ray and CD estimates of secondary structure fractions of X samples.

1 PG, Provencher & Gleckner (1981) (ridge regression) method with solution suggested by the program CONTIN; PG/LL, PG combined with locally linearized approach (varying the number of basis set proteins); PG/CA, PG combined with cluster analysis; PG/SC, PG combined with self-consistent method; PG1, PG method with solution having the least standard error as given by the program CONTIN; PG1/XX, PG1 combined with XX (LL, SC and CA); HJ, Henriksen & Johnson (1981) (SV regression with a significant eigenvector); VS, Vera & Johnson (1981) (variable selection method); LL, van Stokkum et al. (1990) (varying both the number of proteins and the significant eigenvalues); SC, Sturman & Woody (1993) (self-consistent method); SC/CA, cluster analysis combined with self-consistent method; NN0, neural network method with no hidden layers and sigmoidal transfer function; NN1, neural network with 1 hidden layer of 45 neurons and sigmoidal transfer function; NN2, neural network with 2 hidden layers (40 and 20 neurons) and sigmoidal transfer function; NN/LL, neural network with 2 hidden layers (40 and 20 neurons) and sigmoidal transfer function combined with locally linearized approach, NN/CA, neural network with 2 hidden layers (40 and 20 neurons) and sigmoidal transfer function combined with cluster analysis.
in Table 1. The solutions from all NN topologies gave $\Sigma f_i$ close to 1-0 and $f_i \geq 0-0$, except for poly(Glu), myoglobin and hemoglobin. Hemoglobin was predicted to have $-0-01\%$ $\beta$-sheet by NN2 topology, while all topologies predicted negative fractions ($-0-05$ to $-0-01$) for the $\beta$-sheet in poly(Glu) and myoglobin.

Comparison of NN0, NN1 and NN2 indicates the better performance of NN2 topology (Table 1). The NN0 topology, which has hidden layer, gave a better correlation coefficient and R.M.S. error for the helix and was comparable with NN1 in its performance. The performance indices for NN2 topology were slightly inferior to the NN0 topology for the $\alpha$-helix, while those for the other three secondary structures were superior. The CD spectra have experimental noise and contributions from aromatic residues, disulfide bonds, etc., superimposed on the linear combination of component spectra, which influence the learning phase. These additional contributions, which are non-linear, are probably handled better with the inclusion of hidden layer neurons, thus improving the performance. However, this is achieved at the expense of $\alpha$-helix prediction. The reduction in the performance indices of $\alpha$-helix prediction is, however, small. The better performance of NN0 in predicting the $\alpha$-helix fraction illustrates the dominance of the $\alpha$-helix spectrum in the protein CD spectrum, which is predicted well with simple linear methods. Flexibility in the operational basis set is introduced by hidden layers in NN1 and NN2 topologies, which improves the overall analysis. A similar reduction in the performance for $\alpha$-helix was observed with introduction of basis set flexibility (variable selection) in the HJ method (Manavalan & Johnson, 1987; Table 1), but the overall performance improved.

\section{Combining neural networks with variable selection}

The results for NN0, NN1 and NN2 topologies indicated the better performance of NN2 topology. We combined the NN2 topology with the LL approach, and the method is denoted as NN/LL. The proteins in the basis set, with the test protein removed, were arranged in the order of increasing R.M.S. distance from the test protein spectrum. The analysis was performed with this set as the basis set, and the results correspond to the standard NN2 analysis. Now the protein with the largest R.M.S. distance in the basis set was removed and the analysis was performed with the remaining proteins. The process was repeated until six proteins remained in the basis set. Each protein had ten solutions corresponding to ten deletions of proteins from the basis set. Most of the solutions were acceptable, with $|\Sigma f_i - 1| \leq 0-02$ and $f_i \geq 0-0$. The average of a set of solutions where the R.M.S. difference of each solution in the set with any other solution in the set was less than 0-1 was taken as the final solution. The $\beta$-sheet fraction for myoglobin and hemoglobin, which were negative for the NN2 method, became positive as the more distant proteins were deleted. Only poly(Glu) was predicted to have a negative fraction for the $\beta$-sheet. The results from the NN/LL method showed improvement over the NN2 method with better performance indices (Table 1). The predictions for individual proteins showed mixed success for NN/LL (data not presented); substantial improvements were obtained for myoglobin, flavodoxin, prealbumin and Bence-Jones protein; worse predictions for elastase and subtilin BPN; and comparable results for the rest. Cytochrome $c$, lysozyme and rubredoxin did not show any improvement.

\section{Combining ridge regression with variable selection and self-consistent methods}

The ridge regression procedure used in the method of Provencher & Glockner (1981), abbreviated PG, fits the CD spectrum of the test protein ($C_i^{\text{test}}$) as a linear combination of the CD spectra of $N$ reference proteins by minimizing the function:

$$\sum_{i=1}^{N} (C_i^{\text{obs}} - C_i^{\text{calc}})^2 = \alpha \sum_{j=1}^{N} (v_j - v_j)^2$$

where $C_i$ is the spectrum at $n$ wavelengths, $\alpha$ is the regularizer and the $v_j$ determine contributions from the reference spectra. The relations $f_i \geq 0-0$ and $\Sigma f_i = 1-0$ are used as constraints. The method gives several solutions for a range of values of $\alpha$. Smaller values of $\alpha$ give solutions similar to the normal least-squares and larger values tend to give solutions biased towards certain proteins, limiting the number of degrees of freedom. The program CONTIN selects a solution based on several criteria, but with a warning against using it blindly (Provencher, 1982b). In our application we found that this solution was not always the best; there were solutions closer to the X-ray structure of the test protein that were rejected by the program. Manavalan & Johnson (1987) used two other criteria in comparing their results with the PG method, selecting the solution with five degrees of freedom and the solution that gives $\Delta$ less than 0-22, and obtained improved correlation. We have performed analyses by considering solutions based on two criteria. The first criterion considers the solution chosen by the program with $\alpha$ close to 0-5, and the results are given under PG. The second criterion selects the solution with the least standard error (the number of degrees of freedom for this solution was less than that for the chosen solution) from among the set of solutions given by the program CONTIN, and the results are under PG1.

The ridge regression procedure optimizes the contributions from different proteins for the final solution. If one of the proteins is not contributing or contributing negatively then the corresponding

\footnote{Supplementary data giving detailed results for individual proteins are available \textit{via} anonymous ftp at internet address: 129.82.125.151; login name: ftp; directory: pub/CD spectroscopy; files: table1.jnb to table3.jnb.}
Table 2  
Overall performance of different methods

<table>
<thead>
<tr>
<th>Method</th>
<th>δ</th>
<th>r</th>
<th>δ</th>
<th>r</th>
<th>δ</th>
<th>r</th>
<th>δ</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC</td>
<td>0.11</td>
<td>0.783</td>
<td>0.08</td>
<td>0.990</td>
<td>0.08</td>
<td>0.894</td>
<td>0.19</td>
<td>0.406</td>
</tr>
<tr>
<td>PG1</td>
<td>0.07</td>
<td>0.908</td>
<td>0.07</td>
<td>0.916</td>
<td>0.07</td>
<td>0.926</td>
<td>0.07</td>
<td>0.906</td>
</tr>
<tr>
<td>NN‡</td>
<td>0.08</td>
<td>0.882</td>
<td>0.08</td>
<td>0.900</td>
<td>0.07</td>
<td>0.922</td>
<td>0.07</td>
<td>0.930</td>
</tr>
<tr>
<td>HJ</td>
<td>0.11</td>
<td>0.792</td>
<td>0.07</td>
<td>0.932</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The corresponding values for the SC/CA combination were 0.08 and 0.902. The results from the HJ/CA combination were unacceptable. The NN/SC combination was not performed.

‡ The entry under HJ/SC combination corresponds to HJ/SC/LL due to inclusion of LL approach in the SC method (Sreerama & Woody, 1993).

The performance of the PG/LL and PG1/LL methods was similar to the NN/LL method, by arranging the basis set proteins with hemoglobin as the test protein and removing the more distant proteins systematically. The selection of the final solution from among the solutions under PG/LL and PG1/LL was not as straightforward as in the case of the SV-based methods, where the set of allowed solutions satisfying the criteria \( |\sum_j f_j - 1.0| \leq 0.10 \) and \( f_j \geq 0.05 \) were averaged. In the ridge regression method, those criteria are used as constraints and all solutions satisfy this selection criteria. We obtained the final solution by selecting solutions with less than 10% r.m.s. differences and averaging them, as was done in the NN/LL method. Results are summarized in Table 1.

We have also combined the ridge regression method with the self-consistent method. Here, the protein analyzed was included in the basis set and an initial guess was made for its structure; we considered the structure of the protein with the CD spectrum most similar to that of the test protein as the initial guess (Sreerama & Woody, 1993). The solution obtained replaced the initial guess and the process was repeated until self-consistency was attained. The two criteria for selecting a solution from ridge regression led to two separate analyses which are reported under PG/SC and PG1/SC.

The PG/LL method showed substantial improvements over PG. The performance indices were markedly better, particularly for β sheet and turns for which the r.m.s. error decreased by about 50% and the correlation coefficient increased by about 50%. This difference in the performances of PG and PG/LL was due to the selection of the correct solution, one closest to the X-ray structure, in the PG/LL method. As pointed out above for the PG method, the program did not always select the correct solution. As the more distant proteins were removed, the solution selected by the program improved and became closer to the X-ray structure. The predictions by the PG/LL method for lysozyme, prealbumin, glycer aldehyde 3-phosphate dehydrogenase, flavodoxin and rubredoxin were worse than those from PG but most of the other proteins showed improvements. The main advantage of PG/LL over PG was in the removal of the ambiguity in selecting a solution based on the value of the regularizer.

The PG/SC combination performed very poorly. The r.m.s. deviations and the correlation coefficients were poor. This is mainly due to the selection of the solution suggested by the program. If the second guess, the solution obtained at the first iteration, deviates further from the X-ray structure then the final solution will be very different from the X-ray structure. The basic premise of the self-consistent method, that the successive iterations will be closer to the true structure, is violated. In fact, this happened with most proteins, which resulted in the worst performance indices for the PG/SC combination. This was not the case with PG1/SC combination, where the solution with the least standard error from the set of solutions given by CONTIN was selected. This normally selected the solution with the least degrees of freedom, generally close to five, and was qualitatively similar to the PG3 of Manavalan & Johnson (1987). The improved performances of PG1 (and its combinations with LL and SC) over PG are mainly due to this difference, and Manavalan & Johnson (1987) obtained similar improvements.

The performances of PG1, PG1/LL and PG1/SC were similar. The PG1 method gave slightly better performance indices for the unordered fraction and slightly poorer indices for β sheet and turns than the PG1/LL method. Overall performance indices of these three methods were comparable (Table 2). The PG1 method performed slightly better than the PG/LL method (δ and r: 0.07 and 0.905 for PG1; 0.08 and 0.900 for PG/LL). The similarity in the performances of PG1 and PG1/LL methods confirm the implicit inclusion of the variable selection principle in the ridge regression method, and that the
PG and VS methods are equivalent (W. C. Johnson, Jr. private communications). The biasing effect due to inclusion of the test protein CD spectrum in the PG1/SC method did not improve the results.

(e) Cluster analysis

The variable selection method, more particularly the locally linearized approach, modifies the basis set so that proteins with spectral characteristics similar to that of the test protein are always included in the analysis. Pattern recognition techniques, such as cluster analysis (Jardine & Sibson, 1968; Shara et al., 1986) (CA), which determine groups or clusters of samples in a given set of samples based on the similarity between the samples, can be used to classify proteins based on the similarities in their CD spectra. Proteins with similar CD spectra, and hence similar structural characteristics, will be grouped together and form a cluster. Using these subgroups of proteins for the CD analysis of any member has been suggested as an improvement over the variable selection method (Pancoska & Keiderling, 1991). The similarity index between two samples can be calculated using a defined distance metric, such as the r.m.s. difference, the absolute difference, the principle components, singular value decomposition coefficients, etc. We used the r.m.s. distance between the CD data (ΔC) of samples as the distance metric and followed the centroid method (Shara et al., 1986) for determining clusters. We have used the absolute distance between the CD data and the coefficients from singular value decomposition as the distance metric, and the results were similar.

The results of cluster analysis are presented in the form of a connection dendrogram (Shara et al., 1986; Jardine & Sibson, 1968), with the relative dissimilarity indices and the tertiary structure class of the protein shown in Figure 1. Poly(Glu) formed a separate cluster. It was not included in Figure 1 so that the relative dissimilarity index for the other members of the basis set could range from 0 to 1. (The formation of a separate cluster by poly(Glu) raises the question of whether it is appropriate to retain it in the basis set. However, poly(Glu) does not influence the analysis of proteins with low-α content as it is always excluded from the analysis due to variable selection. Inclusion of poly(Glu) improves the analysis of high-α proteins and hence it was retained in the basis set.) The proteins form three distinct clusters, with five or six members in each, depending on the secondary structural content, which may be called high-α, high-β and mixed α-β clusters. Clustering could not be explained based on the different tertiary structure classes, although all high-β proteins (Bence-Jones protein, prealbumin, elastase, chymotrypsin and rubredoxin) belong to the β class and formed a separate cluster from the other three classes. The cluster analysis was not able to separate the α/β and α + β classes, and cytochrome c, an αα protein, was clustered with the mixed α-β cluster, which contained mostly α/β and α + β proteins. The proteins thermolysin, lactate dehydrogenase and triosephosphate isomerase were clustered with hemoglobin and myoglobin, forming the high-α cluster. These clusterings are, however, consistent with the discriminating power of CD with respect to the secondary structure of proteins rather than the tertiary structure. Efforts to classify proteins into different tertiary structure classes based on CD spectra have been reasonably successful (S. Yu, Venyaminov, personal communication).

Pancoska & Keiderling (1991) have used principal component factor analysis coefficients in their application of cluster analysis to vibrational circular dichroism and CD spectra, and compared the clusters with those obtained with X-ray secondary structural fractions. The general results from their cluster analysis of CD spectra are similar to ours, despite the differences in the basis sets. In both analyses, the high-α proteins, myoglobin and hemoglobin, are placed in a cluster at one end, and the high-β proteins, elastase and chymotrypsin, are placed at the other end. Lysozyme and cytochrome c are grouped with an intermediate cluster in both, but are grouped with papain and added to the mixed α-β/high-β cluster in our analysis, and they are grouped with triosephosphate isomerase and added to the high-α cluster in the analysis described by Pancoska & Keiderling (cluster Aα + Bα in their paper). Our analysis places papain closer to lysozyme, while it is added to the high-β proteins in the analysis described by Pancoska & Keiderling (cluster Cβ + Cβ). Pancoska & Keiderling place triosephosphate isomerase closer to lysozyme and the cluster was added to the high-α cluster (cluster Aα + Bα), while our analysis places triosephosphate isomerase with the high-α cluster. Papain has about 25% α and triosephosphate isomerase has about 50% α, and our results are consistent with the information in the CD spectra. Increasing the number of proteins in the basis set might lead to slightly different clustering, but the differences are likely to be minimal.

Figure 1. Cluster analysis of basis set proteins; a connection dendrogram representation.
Combining cluster analysis with PG, NN and SC methods

Analyzing the CD spectrum of a protein using the other members of the cluster as the basis set could provide an improvement in the results since these proteins are similar to the protein analyzed. The locally linearized approach incorporates this feature only partially, by arranging proteins in the order of increasing dissimilarity indices with the test protein. To test the performance of secondary structure analysis in conjunction with the cluster analysis we have performed subanalyses of the proteins belonging to the three clusters shown in Figure 1. We added poly(Glu) to the high-α cluster. The subanalyses were performed by considering proteins in each cluster as a separate basis set and each protein was analyzed with the other members of the cluster to which it belongs. The subanalyses were performed using LL, SC, NN, PG and PG1 methods, thus effectively combining CA with these methods, and the results are listed under LL/CA, SC/CA, NN/CA, PG/CA and PG1/CA, respectively, in Table 1. We did not obtain acceptable solutions for some of the proteins from the LL/CA method, indicating the failure of the analysis, which is mainly caused by the paucity of information content resulting from the small number of proteins in the basis set (five or six). We did not get acceptable solutions for some proteins (Bence-Jones protein, prealbumin, rubredoxin and papain) from the SC/CA method beyond the first few iterations. One of the reasons for this might be the rigid constraints, $|\Delta f| - 1.0 < 0.05$ and $f_1 \geq 0.025$, used in the SC method (Sreearam & Woody, 1993). However, relaxing these limits to $0.1$ and $-0.05$, respectively, did not solve this problem. We considered the solution that had the least r.m.s. difference with the previous solution. The solutions from PG/CA and PG1/CA were always acceptable, as negative fractions are not allowed and the sum of fractions is always $1.0$. All solutions from NN/CA were acceptable; poly(Glu) was predicted to have a negative $f_{\beta} (-0.02)$, and for all proteins the sum of fractions was close to $1.0$.

The results obtained by combining cluster analysis with a parent method (PG, PG1, NN or SC) were comparable with those obtained with that same method in combination with LL, except for a few proteins. The predictions for the α-helix were better and those for the unordered were worse with a parent method/CA combination than the parent method/LL combination (Table 1). For turns, the PG method showed no difference in the performance with regard to combinations with CA or LL, while the PG1 method performed better in combination with CA. For β-sheet, both PG and NN performed better when combined with CA. The PG1/CA combination performed distinctly better for α-helix, β-sheet and turns than the PG1/LL combination, but worse for the unordered.

The NN method seems to perform better with CA as the framework for introducing the variable selection principle in the analysis. Improved results were obtained for all proteins, except hemoglobin and poly(Glu), with NN/CA as compared with NN/LL, and the performance indices were better for NN/CA (Table 2). The NN/CA method had a lower r.m.s. difference and a higher correlation coefficient than NN/LL (Table 2). With the choice of CA, the number of proteins used in the learning phase and the total r.m.s. error are small. The combination of NN with CA is doubly advantageous as compared with the combination with LL. Firstly, since the r.m.s. error is small, the non-linear errors in the network are also small, making learning more efficient. The smaller number of proteins requires less computer time for obtaining the solution (combination with LL requires 20 to 25 times more computer time than that with CA). Secondly, the results are better than those from the combination with LL.

The performance of the PG/CA method was comparable with that of the PG/LL method. The predictions for the α-helix were better when PG was combined with CA than with LL, and those for the β-sheet and the unordered were worse. There were some differences in the predictions for some proteins but the r.m.s. difference and the correlation coefficient were almost identical. The PG1/CA method performed better than PG1 and PG1/LL methods for α-helix, β-sheet and turns, while PG1 performed better for the unordered. The overall performance indices for the PG1/CA were slightly better than that for PG1 (P = 0.926 for PG1/CA and 0.908 for PG1). As pointed out earlier, the ridge regression method is, in principle, equivalent to variable selection, and the performances of PG1 and PG1/LL were similar. The inclusion of cluster analysis in the ridge regression method, however, provides a way of explicitly including the variable selection principle by considering only the important proteins in the basis set, and this improves the prediction.

The performance of SC/CA was inferior to that of SC (the SC method incorporates the LL approach). Slight improvements in the performance indices for the α-helix and the turns for SC/CA were offset by the performances for β-sheet and unordered (Table 1). The correlation coefficient decreased from 0.930 for SC to 0.897 for SC/CA (Table 2).

Comparison of different methods

A comparison of the combined methods (Table 1), excluding the PG/SC method for reasons discussed earlier, indicates that the α-helix fraction was predicted the best; $\delta_1$ varied from 0.06 to 0.09 and $r_1$ from 0.977 to 0.946. The performance indices for other secondary structural fractions were less consistent. The performance indices for the β-sheet, $\delta_1$, varied from 0.07 to 0.11 and $r_1$ from 0.887 to 0.732. The ranges of correlation coefficients for the turns and unordered were similar to that of $r_1$ ($r_2$ from 0.973 to 0.778; $r_{\Delta}$ 0.796 to 0.486), but the r.m.s. deviations were smaller ($\delta_{\Delta}$ 0.04 to 0.06; $\delta_{\Delta}$ 0.05 to 0.08).

The α-helix spectrum dominates the protein CD (Johnson, 1990; Sreearam & Woody, 1993) and was expected to be predicted well. Though high
correlation coefficients were obtained for the α-helix (\(r_s > 0.946\)), the r.m.s. deviations were larger than those for turns and unordered. This discrepancy deserves an explanation. The average fractions of the four secondary structures in the basis set were different: α-helix, 36.5%; β-sheet, 20%; turns, 22%; unordered, 21.5%. The dynamic ranges of the secondary structure fractions, \(f_{\text{max}} - f_{\text{min}}\), were 0.07, 0.50, 0.34 and 0.36 for α-helix, β-sheet, turns and unordered, respectively. The r.m.s. deviations should be divided by either the amount of secondary structure present in the basis set (Sreerama & Woody, 1983) or by the dynamic range of fractions (Pancoska et al., 1992) to obtain a more reasonable comparison.

The \(\delta_s\) normalized in this way is comparable with or better than the normalized \(\delta_\varepsilon\). Among the individual secondary structural fractions, α-helix and turns were predicted better than the other two, and the predictions for the β-sheet were relatively poorer.

Neural network-based methods perform the best among the three methods. The basic NN (NN2 in Table 1) gives solutions that satisfy the criteria, \(|\Sigma f_s - 1.0| < 0.10\) and \(f_s > 0.05\), quite well and yields the best performance indices when compared with the basic PG and HJ methods, and performs comparably with the methods incorporating the variable selection principle. Among the various methods listed in Table 1, the NN/CA combination has the best performance indices for individual secondary structural fractions: r.m.s. \(< 0.08\) and correlation coefficient \(> 0.712\). The PG and SVD-based methods gave similar performance indices when combined with the variable selection principle. The PG-based methods have higher r.m.s. and lower correlation coefficients for the β-sheet than the SVD-based methods, while the SVD-based methods showed lower correlation coefficients for turns and the unordered.

A slightly different picture emerges from Table 2, which gives the overall performance of these methods. The \(\delta_s\) and \(r\) values indicate that SC (equivalent to HJ/SC(LL)) and LL (equivalent to HJ(LL)), the SVD-based methods, perform the best (\(\delta_s < 0.07\); \(r\), 0.932, 0.950), with NN/CA and PG/CA close seconds (\(\delta_s < 0.07\); \(r\), 0.926, 0.922). The methods PG1/LL, PG1/SC and NN/LL follow them. Based on these values from Table 2, we can order the performance of the methods as follows:

\[
\text{SC} \approx \text{LL} \approx \text{NN/CA} \approx \text{PG1/CA} > \text{PG1/LL} \approx \text{PG1/SC} \approx \text{NN/LL} \approx \text{PG/LL} > \text{NN} > \text{PG} \approx \text{HJ}
\]

It is very clear that the inclusion of the variable selection principle improves the results in the PG, PG1 and NN methods. Of the two approaches for the implementation of variable selection, CA is a better choice for NN and PG1, whereas PG performed equally well with both, which is also likely to be the case for PG1. One may consider the basic method, be it PG, NN or SVD (HJ), as the first level method and the methods that incorporate the variable selection principle as the second level method. The performances of the first level methods are varied. However, the differences between the performance indices of the second level methods are minimal. This leads to the conclusion that the analysis performed with any of the second level method should be equally reliable.

The basic assumption of these methods of analysis is the simple linear relation between the secondary structure and the CD spectrum. There are, however, other contributions to the CD of proteins that are not explicitly treated in these analyses. They are: contributions from the tertiary structure, contributions from the aromatic residues and disulfide bonds in the far UV, and effects of geometric variability, such as chain/strand-length dependence, distortions, etc. These have been thoroughly reviewed (Manning, 1989). Efforts have been made to include the chain-length dependence of the α-helix in the least-squares method (Chang et al., 1978). Theoretical calculations have predicted that an α-helix must contain at least three turns to exhibit the α-helix spectrum, and smaller helices give smaller amplitudes (Manning & Woody, 1991). Such geometric variability leads to non-linear relations between the structure and the spectra.

Aromatic contributions from tryptophan, tyrosine and phenylalanine could be significant, as has been demonstrated theoretically (Manning & Woody, 1989; Manning et al., 1992; our unpublished results), and may be a major source of the error in secondary structure estimations. Anomalies in the CD spectra of concanavalin A, DNase I, gene 5 protein, etc. have been assigned to aromatic contributions. The quantitative estimation of such contributions has thus far eluded researchers. Bolotina & Lugauskas (1985) have tried to calculate aromatic contributions by subtracting the secondary structure contributions from the protein spectra with some success. Perelz et al. (1992) have also tried to correlate one of the "pure component" spectra generated by the convex constraint analysis with aromatic contributions and found a modest correlation with the percentage of aromatic residues. The major problem with all these approaches is that the aromatic contributions vary widely in magnitude and sign from one protein to another, even among homologous proteins.

Although none of the methods discussed in this paper explicitly considers aromatic contributions, end effects, or distortions from regular secondary structure, the protein CD spectra of the basis set proteins include such contributions. Earlier methods of CD analysis that used a fixed basis set, i.e., a single set of reference spectra for all proteins, were unable to take advantage of this implicit information content. With a flexible basis set, first introduced by Provencher & Glöckner (1981), basis set proteins with CD spectra more closely resembling that of the protein under examination will be given greater weight in the operational basis set, while those with very different CD spectra will be given little or no weight. The CD spectra are determined by the secondary structure and by the additional factors
previously discussed, such as aromatic contributions, end effects, and distortions from regular geometry. Therefore, these additional factors are included in the analysis, albeit in an indirect way. This is the basic reason behind the improved performance of the methods using the variable selection, cluster analysis, and neural network methods with one or more hidden layers. Of course, if the test protein has additional contributions that are very different from any of the proteins in the basis set, the analysis will be subject to larger error. However, such problems will frequently be revealed by failure to meet the conditions for a satisfactory analysis. Further expansion of the basis set to include a wider range of proteins should make such cases increasingly rare.

4. Summary and Conclusions

We have examined new strategies for improving the prediction of protein secondary structure fractions from circular dichroism spectra and applied them in three basic methods of analysis. The methods (singular value decomposition, ridge regression, and neural networks) were compared with a common basis set of proteins. Three approaches to improve the analyses (variable selection as implemented in the locally linearized model, cluster analysis and self-consistent method) were combined with the basic methods. Cluster analysis, wherein the proteins with similar CD spectral characteristics are grouped into clusters, provided a new way of performing the variable selection, retaining only the important proteins for the analysis. The self-consistent method provided the means for using the test CD spectrum to influence the solution.

Slight improvements were obtained by incorporating cluster analysis into ridge regression and variable selection into neural network, and marked improvements were obtained by incorporating variable selection and self-consistent methods into singular value decomposition methods. The performance of variable selection as implemented by the locally linearized model and the self-consistent method with singular value decomposition, ridge regression with cluster analysis and neural network with variable selection was similar. The performance of other combinations was slightly inferior. The neural network method performed better at the basic level. All these methods were able to handle the random experimental noise quite well. The predictions for individual proteins were similar for most of the combined methods, implying that an analysis performed with any of the methods incorporating the variable selection is equally reliable within the basic assumption of the linear relation between the secondary structure and secondary structure spectra. It appears that we have reached a limit with this kind of analysis of CD spectra. Further improvements may be possible, should one be able to incorporate the non-linear contributions to the protein CD spectra, which might arise from aromatic side-chains, length dependence of helix/strand and their geometric variations, and tertiary interactions, into the analysis.

Thanks are due to Dr W.C. Johnson, Jr. for providing the CD spectra of the proteins and the VARSELEC program, and to Dr S. Yu. Venyaminov for providing the modified CONTIN program. Thanks are due to Dr A. V. M. Woody for a critical reading of the manuscript. This work was supported by NIH grant GM22994.

References


*Edited by F. Cohen*

(Received 25 January 1994; accepted 22 June 1994)