Relationships between Secondary Structure Fractions for Globular Proteins. Neural Network Analyses of Crystallographic Data Sets†

Petr Pancoska,*,‡ Martin Blazek,§ and Timothy A. Keiderling*,‖

Department of Chemical Physics, Faculty of Mathematics and Physics, Charles University, 121 16 Prague 2, Czechoslovakia, Computational Center, Czechoslovak Academy of Sciences, 170 00 Prague 7, Czechoslovakia, and Department of Chemistry, University of Illinois at Chicago, Box 4348, Chicago, Illinois 60680

Received April 22, 1992; Revised Manuscript Received August 4, 1992

ABSTRACT: The relationship between the fractions of protein secondary structural components as determined from X-ray crystallographic data by the procedures of Kabsch and Sander (KS) and of Levitt and Greer (LG) is analyzed by neural network analysis of these two tabulations of literature data. A linear relationship between the KS and LG reductions of X-ray data to secondary structure descriptors is demonstrated by a regression analysis of the relationships between these sets of structural parameters. Back-propagation neural network analysis was then used to derive equations for determination of the most probable fractions of β-sheet, bend, turn, and "other" conformations given the fraction of α-helix in a globular protein. The deviation of the X-ray values for β-sheet from that determined with these equations was shown to have a variance that exponentially decreased with increasing fraction of α-helix. A second neural network analysis showed that knowledge of both the α-helical and β-sheet fractions in a protein significantly reduces the uncertainty in prediction of the other components of the secondary structure. These analyses provide insight into the nature of the data sets derived from crystal structures. Since these compilations of crystal structure data are commonly used as reference information for quantitative evaluation of spectra (for example, FTIR, Raman, and electronic or vibrational circular dichroism) in terms of secondary structure, such internal correlations in the reference sets may have significant effects on the stability of spectroscopic analyses derived from them.

Understanding of a protein structure is normally considered to be a first step in elucidating its function. Historically, spectroscopic techniques (such as circular dichroism, infrared, and Raman spectroscopies) have been used as a first step in determining average structural characteristics for any new, poorly characterized protein. These techniques are particularly useful with regard to determination of the average secondary structure characteristic of a protein.

Many calculational schemes used for quantitative analysis of spectroscopic data of proteins use an empirical approach. Their general scheme is in all cases similar, relying on the availability of a reference set of samples. Until now the reference information has been derived almost exclusively from interpretation of the results of X-ray crystallography.

Underlying the structural application of spectroscopic methods is the implicit assumption that the protein structure as found in the crystalline state is conserved in aqueous solution. There is substantial evidence that this assumption is valid for many proteins (Tu, 1986). But there are also a few examples of the opposite behavior (Urbanova et al., 1991). Given this form of structural reference, the above empirical methods can be viewed as trying to project from the solution spectra what the conformation of a protein would be in its crystalline phase. This extrapolation would be easily justified in the cases where there is a well-defined relationship between the protein crystal and solution structure. But this is a molecular property over which one has no influence. In this paper we focus on the other, more formal, aspects of these methods which, if properly understood, could lead to improvement in their execution or, at least, a better appreciation of their limitations.

The first step of most empirical analyses represents a mathematical transformation of the experimental spectra into a more condensed form. An example of such a "condensation" of the experimental information is Fourier self-deconvolution of Fourier transform (FT)IR spectra, followed by curve fitting to determine the component band areas (Mantsch et al., 1986). Another approach is use of the singular value decomposition (Manavalan & Johnson, 1987, 1985), principal component (Pancoska et al., 1979, 1991), or convex constraint decomposition (Perczel et al., 1989, 1991) algorithms to reexpress the experimental CD or FTIR spectra of a protein reference set as a linear combination of a small number of (common) basis spectra. Other approaches have been proposed over the years (Fasman, 1989; Manning, 1989).

Parameters resulting from these various transformations are then related by some variant of the least-squares methodology to appropriate descriptors of protein crystal structure. Typically, only fractional secondary structure parameters have been considered. The spectroscopic descriptors of protein secondary structure are thus usually in the form of fractions or percentage of α-helix, β-sheet, and "other" conformations (which can encompass or further subdivide out contributions of bends, turns, etc.). These percentages of individual secondary structures are averaged properties of the crystal structure calculated as the fraction of residues in the protein which have been classified as belonging to a given conformational type by some algorithm.

There are several objective algorithms available for the classification of protein secondary structures using X-ray diffraction determined atomic coordinate data (Kabsch &

† This research was funded in part by a grant from the National Institutes of Health (GM 30147 to T.A.K.) and by an International Cooperation Grant from the National Science Foundation (INT91-07588 to P.P. and T.A.K.).

* To whom correspondence should be addressed.
‡ Charles University.
§ Czechoslovak Academy of Sciences.
‖ University of Illinois at Chicago.

0006-2960/92/0431-10250$03.00/0 © 1992 American Chemical Society

Sander, 1983; Levitt & Greer, 1977; Unger et al., 1989). The fractional composition (FC) of secondary structure types determined from these algorithms differ due to the fact that different criteria are used in each method for inclusion of a residue in a particular type. Each method has its own adherents, which stems from the different bases used for classification in each scheme. For example, the KS scheme uses hydrogen-bonding patterns to delineate structural types and also partitions the protein into a large number of specific components. On the other hand, the LG method has fewer types and a more relaxed criterion for a given residue being classified into the main ones. Thus the LG method tends to identify larger fractions of α-helix and β-sheet for a given crystal structure than does the KS result. This is often evidenced in terms of differences in the lengths of series of residues that are assigned to a uniform secondary structure along the protein chain. These intrinsic ambiguities in the X-ray information reduction must also affect the precision of any spectral analysis that subsequently uses that reduction to formulate a reference data set.[1]

Implicit in previously proposed spectral analyses is an assumption that the X-ray based secondary structure descriptors are independent. To our knowledge, no attempt has been published to look for interrelations among fractional secondary structure descriptors for various sets of proteins. However, a referee has noted that two previous papers have noted in passing that the secondary structure components must not be independent (Siegel et al., 1980; Hennessey & Johnson, 1981). Due to the statistical character of the methods relating the spectral parameters to the X-ray descriptors, such an interdependence, if any, can have important consequences. From the spectroscopist's point of view, it is certainly interesting to confirm that a set of spectrally derived structural parameters reflect any such relationship that might exist in the X-ray set upon which it is based. In other words, one must establish that the "information" topologies of both the X-ray and spectroscopic data are comparable to support the use of spectroscopic analyses to infer protein secondary structure.

In this paper we present the results of our efforts to understand relationships between protein secondary structure parameters by using two of the most commonly used X-ray reduction tabulations, those of Kabsch and Sander (KS) and Levitt and Greer (LG), which frequently appear as reference bases for protein spectral analysis. First, in the next section it will be shown that these two methods for determination of fractional secondary structure are, to a level of accuracy that is practical for spectroscopic studies, interconvertible despite differences in the authors' algorithms.

Next, we address the question of interrelations between the fractions of various secondary structure types for individual proteins forming the KS or the LG reference sets. In one respect, this provides another view of the similarities and dissimilarities of the reference information. More importantly, we are able to show that approximate relations of this type do exist between structural types. The results of two different neural network strategies which were found to be instrumental in this task are discussed in the second part of this paper. Finally, a suggested protocol for proper use of these relations which should be considered when analyzing results of spectroscopic methods for structural determination as applied to proteins is presented.

[1] In this paper much of the emphasis is on the KS data set because its tabulated values are available for a large number of proteins and also because it is provided standardly with the Brookhaven Protein Data Bank.

Table I: Precision of Linear Regression between LG and KS Secondary Structure Descriptors

	α-helix	β-sheet	bend	other
average error (%)[a]	4.2	3.1	4.0	4.9
error as percent of dynamic range[b]	5.4	6.5	21.5	13.6

[a] Average deviation of $FC_\zeta^{KS}(LG)$, calculated from LG data using eqs 1–4, from X-ray values, evaluated as $(1/22)\Sigma|FC_\zeta^{KS}(\text{X-ray}) - FC_\zeta^{KS}(LG)|$. [b] Dynamic range = $FC_\zeta^{max} - FC_\zeta^{min}$ for given type, ζ.

RELATIONSHIPS BETWEEN REFERENCE SETS

The KS and LG analyses of protein crystal structures, as published, overlap for 23 proteins. The following relationships were derived by linear regression analysis and allow one to recalculate the fractional component (FC_ζ, in percent) of various secondary structure types, ζ, from one algorithm to the other:

$$FC_\alpha^{KS} = -1.970 + 0.856FC_\alpha^{LG} \quad (r = 0.97, a = 0.99) \tag{1}$$

$$FC_\beta^{KS} = -1.789 + 0.660FC_\beta^{LG} \quad (r = 0.96, a = 0.99) \tag{2}$$

$$FC_o^{KS} = 12.377 + 1.402FC_o^{LG} \quad (r = 0.83, a = 0.99) \tag{3}$$

$$FC_b^{KS} = -1.268 + 1.029FC_{rt}^{LG} \quad (r = 0.71, a = 0.99) \tag{4}$$

For convenience we also present the inverse equations:

$$FC_\alpha^{LG} = 4.746 + 1.131FC_\alpha^{KS} \quad (r = 0.97, a = 0.99) \tag{5}$$

$$FC_\beta^{LG} = 4.875 + 1.401FC_\beta^{KS} \quad (r = 0.96, a = 0.99) \tag{6}$$

$$FC_{lt}^{LG} = 3.859 + 0.692FC_t^{KS} \quad (r = 0.58, a = 0.99) \tag{7}$$

$$FC_{rt}^{LG} = 2.559 + 0.375FC_b^{KS} \quad (r = 0.63, a = 0.99) \tag{8}$$

$$FC_o^{LG} = 1.375 + 0.358FC_o^{KS} \quad (r = 0.65, a = 0.99) \tag{9}$$

Here α is α-helix, β is β-sheet, b is bend, and t is turn as defined by KS, whereas lt is left turn and rt is right turn as defined by LG, o covers "other" conformations not included in the definitions of previous secondary structures, r is the correlation coefficient, and a is the significance level of r.

It can be seen from these regression equations that the LG algorithm yields systematically higher FC values for α-helix and β-sheet (slope >1 in eqs 5 and 6) at the expense of turn, bend, and "other" conformations. This quantitative result parallels the qualitative trend mentioned in the previous section. We have chosen the LG "right turn" structure as a counterpart to the "bend" conformation of Kabsch and Sander. This choice was based on the goodness of fit for various possible correlations, but other correlations are possible. Somewhat as a result, the bend fraction is clearly the most poorly determined of the four descriptors, even though the significance level of the regression is still ~0.99. As a measure of error, the average deviations for the FC values derived from these "system conversion" equations as compared to their original LG and KS values are listed in Table I.

In some sense, one can understand the regression lines as "averages" of the FC_f information in KS and LG algorithms in that they describe the commonality in the secondary structures determined for the 23 proteins considered. The degree of difference between the KS and LG algorithms should be related to the errors listed in Table I. In practice, if a spectroscopic technique is expected to yield secondary structure information to a precision much higher than indicated in Table I, the analysis will be sensitive to the reference set used. But these error limits would be, in fact, very tight for typical spectroscopic studies (Byler & Susi, 1986; Manavalan & Johnson, 1987; Pancoska et al., 1991; Pancoska & Keiderling, 1991; Dousseau & Pezolet, 1990; Server & Krueger; 1991). Thus, for normal uses of spectroscopy to determine structure, either reference set should suffice. This is not to say that the KS and LG values are the same; they are not, but they are systematically related in an average sense.[2]

The KS model assigns some residues to the "other" category (in our reduction of the KS categories) when alternate algorithms might assign them to helix or sheet components. This is not a "missed" structure but is instead an alternate categorization. Our analyses show that, from a spectroscopic error point of view, either type of categorization of the X-ray data can be used to fit spectroscopic data. The result for any given protein will be the classification of its secondary structure according to the method chosen within the error range of the spectroscopic analysis used. If classification according to another method is described, eqs 1–9 should allow an adequate conversion.

RELATIONS BETWEEN DIFFERENT FC VALUES

The concept that interrelations might exist between the fractions of various secondary structures in the set of protein X-ray data was stimulated by the results of our analyses of electronic and vibrational circular dichroism (CD) spectra of globular proteins (Pancoska et al., 1989, 1991; Pancoska & Keiderling, 1991). Using the principal component decomposition of these sets of CD spectra into linear combinations of orthogonal basis subspectra, two independent selective multiple regression schemes were developed to relate the resulting spectral coefficients to the X-ray derived secondary structure FC values. In our VCD studies in particular, we found that the simplest single-variable regressions for the α and β fractions had a dependence on the same subspectrum. This observation implied that there might be an interrelationship between these two regular secondary structures which was independent of the spectral analysis. Our studies had some parallel in the earlier observations by Hennessey and Johnson (1981) that five independent spectral parameters derived from electronic CD spectra could be used to predict eight secondary structure parameters.

Neural Network Tests of Data Set Structure. Our search for relationships within the KS and LG data sets was facilitated by extensive application of the neural network back-propagation algorithm (Simpson, 1990; Horejs & Kufudaki, 1990) as implemented on an ANZA Plus (HNC Inc., San Diego, CA) dedicated numerical coprocessor installed on a 386-level personal computer. Neural networks are effectively pattern recognition schemes (Simpson, 1990; Maggiora et al., 1991) which do not require a priori knowledge of classification or recognition rules. Instead, such rules are generated through the network architecture during the training session, when the mapping of the input to a known output is distributed over the network nodes ("neurons") by adaptive optimization of their interconnections ("synaptical weights"). Variation in the network topology, defined by number, layout, and connectivities of the nodes allows one to "tune" the network performance for specific tasks. The analysis of neuron "potentials" and synaptical weights of the trained network can be then used to formulate the relations between the input and output information.

Table II: Comparison of X-ray Values with FC_f^{KS} Values Output from 5–x–5 Neural Networks

	α-helix	β-sheet	bend	turn	other
5–1–5 network					
av deviation[a]	2.0	7.0	4.0	3.0	6.0
max deviation	10.0	28.0	14.0	13.0	6.0
av % of dynamic range	2.4	14.3	14.0	13.0	12.1
5–2–5 network					
av deviation[a]	1.0	1.0	3.5	3.0	2.5
max deviation	8.0	5.0	16.0	13.0	11.0
av % of dynamic range	1.2	2.0	12.2	13.0	5.0
5–3–5 network					
av deviation[a]	1.0	1.0	1.0	3.0	1.5
max deviation	7.0	3.0	3.0	14.0	6.0
av % dynamic range	1.2	2.0	3.5	13.0	3.0

[a] Calculated as $(1/62)\sum|FC_f^{KS}(\text{output}) - FC_f^{KS}(\text{input})|$.

We have applied several network topologies to test the internal structure of the KS and LG data sets. The first ones tried consisted of 5–x–5 networks, i.e., a neural network system with 5 neurons in the input layer, 5 neurons in the output layer and $x = 1, 2, 3$, or 4 neurons in the hidden layer. This network topology was used as an information filter to sort out the features with the most important variance in the set of X-ray FC data. During the training process, the 62 five-component vectors [FC_α, FC_β, FC_b, FC_t, FC_o] for the KS set and 29 [FC_α, FC_β, FC_{lt}, FC_{rt}, FC_o] for the LG data set were loaded into the both the input and output layer. In this manner, the network was trained to project the input FC vector onto itself through the hidden layer neurons. For $x < 4$, this represents a reduction of the dimensionality of the input matrix, necessarily forcing some of the input variables to become interrelated. Through successive calculations, it was found that for a network with $x = 1$ only the α-helical content is transferred from the input to the output layer without substantial distortion. The other FC values, as regenerated in the output layer, were found to be strongly nonlinearly related to the α-helical content. For $x > 1$, the other secondary structure fractions in the order β-sheet ($x = 2$), other ($x = 3$), and turn and bend (both at $x = 4$) were projected successfully onto the output layer. Table II summarizes the above observations in quantitative form using the average and maximal deviations of output of the 5–x–5 network from the X-ray determined FC_f^{KS} values.

It is important to note that, for all secondary structures, a substantial information flow between the corresponding input-and-output neurons has been directed through the "α-helical channel". This was made clear to us by the large relative magnitudes (weights) of the synaptical connections topologically related to the FC_α neurons in the network. In other words, the network used the variance in α-helical content to project the variance of other secondary structure types as well. Since the output values had some validity, even for small x, this implies that the $FC_{f\neq\alpha}$ values have some dependence on FC_α.

[2] Since the FC values from the KS and LG data sets describe the same molecular property (the X-ray determined conformation) one can assume that relationships similar to those in eqs 1–9 would be derivable for data sets resulting from other algorithms. If not, those algorithms would not describe the same secondary structure.

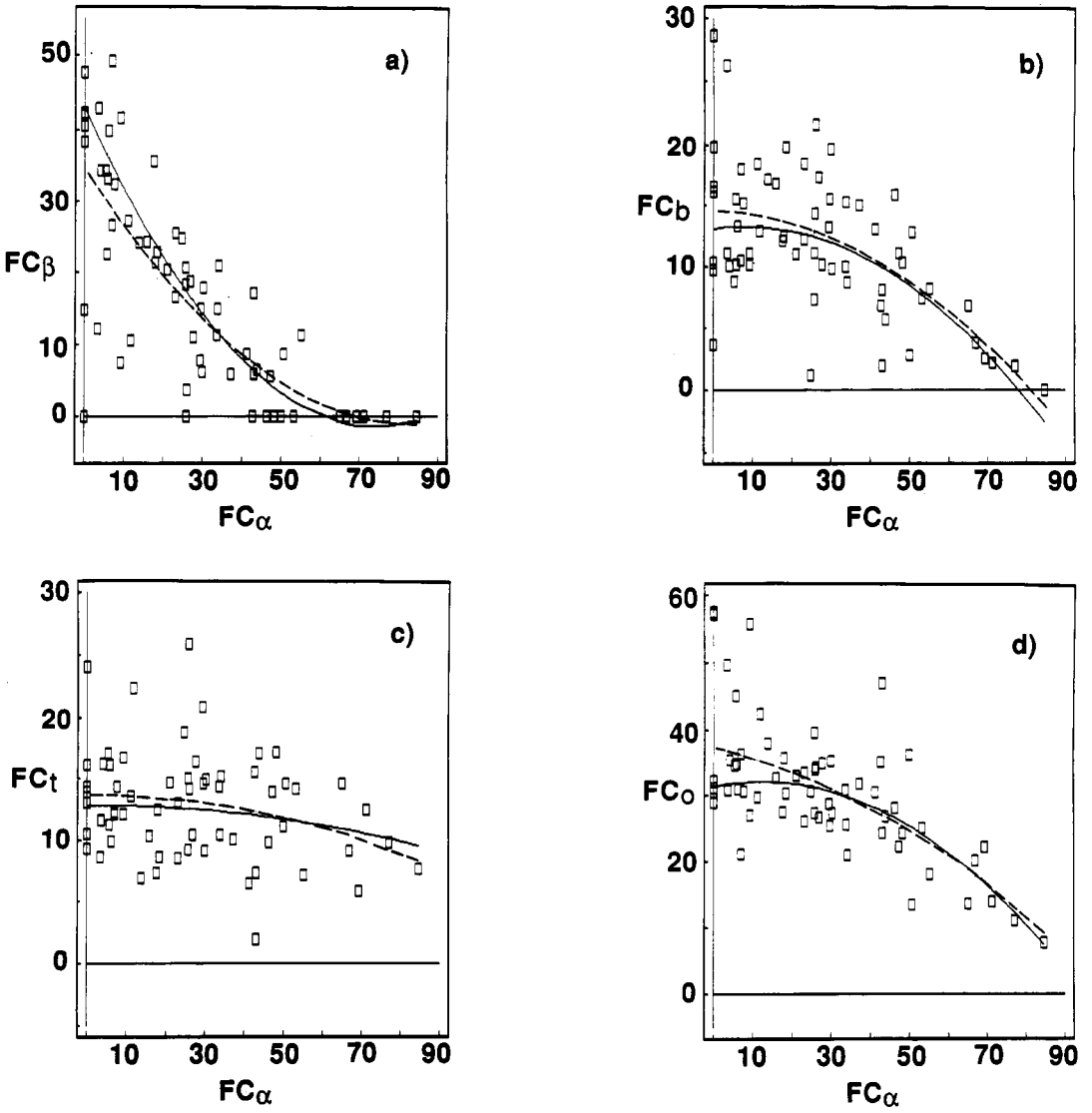

FIGURE 1: Relationship between FC_α^{KS} and (a) FC_β^{KS}, (b) FC_b^{KS}, (c) FC_t^{KS}, and (d) FC_o^{KS} for 62 proteins from the Kabsch and Sander (1983) analysis of crystal structures. The solid line indicates the neural network derived relationship (eqs 14–17 for a–d) and the dashed line the least-squares fit by a second-order polynomial (eqs 22–25 for a–d).

1–3–4 Neural Network Analysis. To analyze the interdependence of the FC values more directly, a 1–3–4 network architecture was employed. For training this second network, the input neuron was loaded with FC_α^{KS} or FC_α^{LG} and the output neurons with FC_β^{KS}, FC_b^{KS}, FC_t^{KS}, and FC_o^{KS} or with FC_β^{LG}, FC_{lt}^{LG}, FC_{rt}^{LG}, and FC_o^{LG}, respectively. We first tested the 1–3–4 network topology for FC interdependence prediction by using a set of artificial FC_ζ values ($\zeta = \beta$, b, t, o), calculated from FC_α^{KS} with hypothetical linear relationships chosen to cover approximately the range of FC_ζ values found in the KS data set. These test results confirmed that the initial relationships between FC_ζ values can be recovered in the following way: For a trained network, the hidden layer potentials were recorded and related separately by (linear) regression equations to the training set FC_α^{KS} values used as the input. The potentials of the output neurons were then related to the potentials of the hidden layer. Algebraic manipulation of these two sets of equations (hidden neuron potentials were excluded) exactly yielded the original relations used to generate the artifical data set.

The same procedure was then used to search for the unknown relationships between the FC_α^{KS} or FC_α^{LG} values with the FC values for the other structural types in the respective reference sets. Using the back propagation algorithm, 62, for KS, and 29, for LG, proteins have been used in training the respective networks. The input values and neuron potentials of the resulting networks were found to be satisfactorily related by approximation as second-order polynomials. The correlation coefficients obtained were higher than 0.94 (typically 0.99 and better) and the residual standard deviation of the fitted potential values (which are between 0 and 1) from the real ones ranged from 2.0% to 0.02%. For the worst fits in the set, a third-order polynomial was also tested, but the residual standard deviations dropped by only a factor of 1.8. The

Table III: Comparison of Error for Linear, Quadratic, and Neural Network Relations of FC_ζ with FC_α[a]

function	β-sheet	bend	turn	other
linear				
av error	7.3	3.8	3.4	5.8
% range	14.8	13.3	14.1	11.6
quad. LSQ[b]				
av error	6.4	3.8	3.3	5.5
% range	13.0	13.4	14.0	11.2
quad. NN[c]				
av error	6.4	3.7	3.2	5.2
% range	13.0	12.8	13.9	10.6

[a] Errors are as defined in Table I, except for 62 KS proteins. [b] From eqs 22–25. [c] From eqs 14–17.

effect of the third-order component on the final regression equation was minor due to the small coefficient obtained. The second-order polynomial approximation for the input/neuron potentials in our network can thus be considered to be the optimal one.

The above described algebraic manipulation of these regression equations for the KS protein fractional FC values yielded the following second-order relationships:

$$FC_\beta^{KS} = 0.81(FC_\alpha^{KS})^2 - 1.20FC_\alpha^{KS} + 0.430 \quad (14)$$

$$FC_b^{KS} = -0.27(FC_\alpha^{KS})^2 + 0.043FC_\alpha^{KS} + 0.131 \quad (15)$$

$$FC_t^{KS} = -0.05(FC_\alpha^{KS})^2 + 0.003FC_\alpha^{KS} + 0.128 \quad (16)$$

$$FC_o^{KS} = -0.47(FC_\alpha^{KS})^2 + 0.116FC_\alpha^{KS} + 0.313 \quad (17)$$

For the respective LG values, the following set of equations resulted:

$$FC_\beta^{LG} = 0.056(FC_\alpha^{LG})^2 - 0.745FC_\alpha^{LG} + 0.596 \quad (18)$$

$$FC_b^{LG} = 0.030(FC_\alpha^{LG})^2 - 0.150FC_\alpha^{LG} + 0.180 \quad (19)$$

$$FC_{lt}^{LG} = 0.027(FC_\alpha^{LG})^2 - 0.116FC_\alpha^{LG} + 0.104 \quad (20)$$

$$FC_o^{LG} = -0.029(FC_\alpha^{LG})^2 - 0.061FC_\alpha^{LG} + 0.140 \quad (21)$$

The KS relationships are decidedly curved, showing a fast drop to a low value for FC_β as FC_α increases and, by contrast, near constant values for the others with a fast fall off at high FC_α values. These relationships are shown as solid lines and compared to the KS data in Figure 1. By contrast, the LG relationships are much closer to linear, as evidenced by the small quadratic coefficients in eqs 18–21.

We also tried a more standard least-squares approach to determine the dependence of $FC_{\zeta\neq\alpha}^{KS}$ on FC_α^{KS} with the assumption that the second-order polynomial function found with the neural network method is a sufficient approximation for the regression function. The following equations resulted for the KS set:

$$FC_\beta^{KS} = 0.51(FC_\alpha^{KS})^2 - 0.849FC_\alpha^{KS} + 0.350 \quad (22)$$

$$FC_b^{KS} = -0.20(FC_\alpha^{KS})^2 + 0.015FC_\alpha^{KS} + 0.146 \quad (23)$$

$$FC_t^{KS} = -0.075(FC_\alpha^{KS})^2 + 0.014FC_\alpha^{KS} + 0.137 \quad (24)$$

$$FC_o^{KS} = -0.22(FC_\alpha^{KS})^2 - 0.140FC_\alpha^{KS} + 0.372 \quad (25)$$

These curves deviate from the neural network results, being lower for FC_β and higher for the others at low FC_α values as shown by the dashed lines in Figure 1, but both have surprisingly similar average errors.

Given the scatter in the data plotted in Figure 1, it might be thought that a linear fit would do just as well. In fact the average error obtained with a linear fit is somewhat worse (see Table III), but, more importantly, a linear fit misses some of the qualitative nature of the data. In particular, the linear equations would generate significantly negative β-sheet fractions for proteins having large α-helical components. Thus the quadratic form is needed to preserve the qualitative form of the interrelationship for the KS data set. Furthermore the quadratic form follows from the neural network analysis. A linear form cannot be used to express the real neuron potentials derived. This objective determination of functional form is a virtue of the neural network approach.

2–3–3 Neural network analysis. The next logical step in this search for interrelationships was to include information about another secondary structure into the neural network input. We chose to add FC_β to complement FC_α and developed a new predicting neural network with a 2–3–3 topology. The choice of FC_β follows from its representing the second most dominant secondary structure form and its having the second largest dynamic range. In addition, from a formal point of view, FC_β was the second variable that successfully came through the 5–x–5 compressing networks with $x \geq 2$.

As a consequence, FC_b, FC_t, and FC_o were related to FC_α and FC_β with a converged 2–3–3 neural network, using the same procedure as above. The input data have been related to the potentials of hidden layer neurons and these, in turn, to the output neuron potentials. The hidden layer variables were then eliminated algebraically as described above to get the following approximate analytical equations for the KS set:

$$FC_b^{KS} = -0.331FC_\alpha^{KS} - 0.305FC_\beta^{KS} + 0.267 \quad (26)$$

$$FC_t^{KS} = -0.104FC_\alpha^{KS} - 0.110FC_\beta^{KS} + 0.173 \quad (27)$$

$$FC_o^{KS} = -0.565FC_\alpha^{KS} - 0.584FC_\beta^{KS} + 0.560 \quad (28)$$

and for the LG set

$$FC_{rt}^{LG} = -0.461FC_\alpha^{LG} - 0.485FC_\beta^{LG} + 0.453 \quad (29)$$

$$FC_{lt}^{LG} = -0.134FC_\alpha^{LG} - 0.056FC_\beta^{LG} + 0.135 \quad (30)$$

$$FC_o^{LG} = -0.398FC_\alpha^{LG} - 0.406FC_\beta^{LG} + 0.406 \quad (31)$$

DISCUSSION

From a mathematical point of view, one can ask why we use the seemingly more sophisticated, complicated, and time-consuming neural network algorithm instead of an ordinary, nonlinear least-squares regression analysis to determine these relations. The value of the neural network lies in its ability to "generalize" without imposing any a priori functional relationship on the data. The algorithm seeks an overall trend, without being significantly affected by small numbers of outlier proteins that do not follow that trend. Additionally, it effects a "smoothing" of the scatter in the data. In Figure 1 we present a comparison of the neural network result to that of a general second-order polynomial least-squares fit of scatter plots of the FC^{KS} variables vs the FC_α^{KS} parameter for the 62 proteins in the KS set (eqs 22–25). The maximal deviation of the two types of curves generated by these two methods occurs for the lowest FC_α^{KS} values where the most highly scattered data points lie. Those points deviating substantially from the neural network line have a weight equivalent to those lying close to the line in the least-squares fit. On the other hand, the neural network algorithm automatically implements a meaningful weighting of the data to give what we term as a "most probable" prediction (see below). This leads to a deviation at low FC_α^{KS} values of the least-squares line from

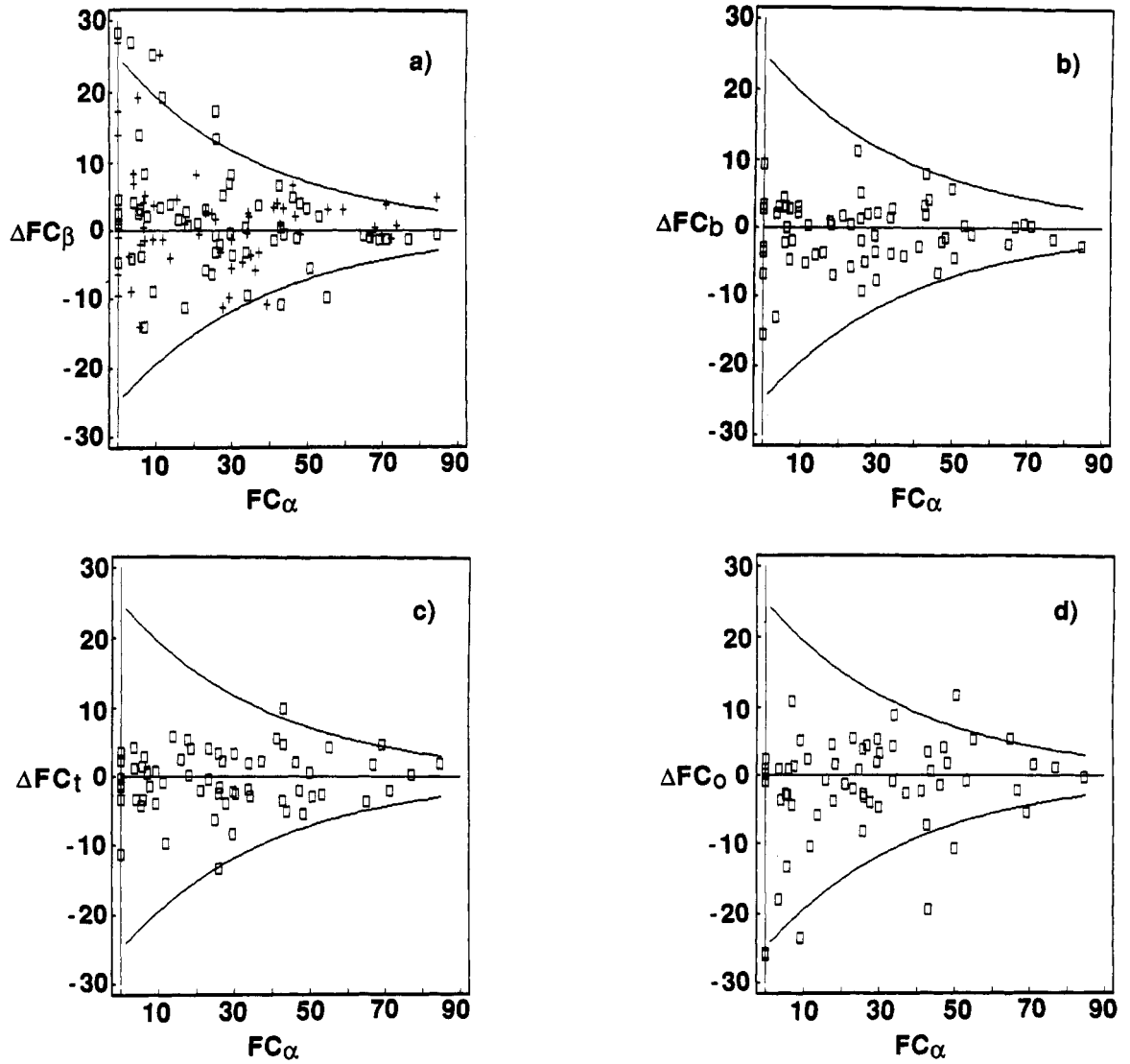

FIGURE 2: Deviations of (a) FC_β^{KS}, (b) FC_b^{KS}, (c) FC_t^{KS}, and (d) FC_o^{KS} from neural network most probable estimations based on eqs 14–17 for a–d, respectively. Squares indicate the KS data, and in panel a crosses indicate data from Qijan and Sejnowski (1988). The curved lines indicate the extent of the dispersion of ΔFC values using eq 32 with a preexponential of 0.25.

the neural network line toward, in one case, lower FC_β^{KS} and, for the other ζ cases, higher FC_ζ^{KS} values.

Another formal advantage of the neural network approach should be mentioned. The selection of the order of polynomials to use for development of approximate relations between the neuron potentials was more straightforward and less ambiguous than it would have been to choose a nonlinear form for a direct least-squares analysis of the original X-ray data. In the above comparisons of the two algorithms, we were able to choose a second-order polynomial fit directly since it had already been shown from the neural network to yield a satisfactory level of approximation.

The "transferability" of these results is an important question. We have tested this on the protein secondary structure descriptors calculated by Quijan and Sejnowski (1988) with the KS algorithm for a larger set of protein X-ray structural data. Their work lists FC_α^{KS} and FC_β^{KS} descriptors for 106 proteins, including all entries from the original KS set. Errors in the estimation of FC_β^{KS} values for the additional 48 proteins using eq 14 were found to be well within the error limits established using the original KS training set. This is illustrated in Figure 2a, which is a plot of the difference of the KS calculated FC_β values from the values predicted with eq 14 vs FC_α for all the 106 proteins. In terms of differences from the neural network curve, the added protein data points, indicated by crosses, are distributed well within the dispersion seen for the original 62 proteins, indicated with squares.

Physical Interpretation. To develop some understanding of the physical meaning of the above derived relationships, we analyzed the dispersion of the *differences* between the KS $FC_{\zeta\neq\alpha}$ values and those calculated from eqs 14–17 ($\Delta FC_{\zeta\neq\alpha}$). These are plotted as a function of FC_α^{KS} in Figure 2. The approximate character of relationships in eqs 14–17 is obvious; for proteins in the analyzed set there is no single $FC_{\zeta\neq\alpha}$ value for a given FC_α but rather a dispersion of observed FC_ζ values about an FC_ζ predicted by the equation. Aside from outliers, which are given reduced weight in the neural network analysis, deviations are randomly distributed around the calculated values and exhibit a systematic trend. They are largest for low FC_α values and diminish nonlinearly with increasing α-helix content as is most clear in Figure 2a for FC_β. Different explanations of this trend can be proposed:

(1) For a given FC_α value, the FC_ζ values for other secondary structures can be assumed to be completely independent. Within such a limiting model, the fractions of sheet, bend, turn, or other conformations could be as large as $(1 - FC_\alpha)$. This range should be reflected in the deviations from our

approximate relationships. Such large deviations are not found.

(2) As another extreme, it can be assumed that the remaining four types of secondary structure can occur with equal probability and achieve a uniform distribution. Their $FC_{\zeta\neq\alpha}$ values and their derivations from calculated ones should be then restricted by lines at $\pm 1/4(1 - FC_\alpha)$. This is approximately the case for the FC_β deviations seen at $FC_\alpha \approx 0$ (see Figure 2a), but such a simple model would not explain the variance seen in the other FC_ζ values, not would it encompass the FC_ζ values generated by eqs 14–17 at low FC_α values. Clearly, choosing five types of secondary structure is an arbitrary decision. The approximate behavior of the deviations in following this trend encouraged us to further consider such a model in an attempt to understand some overall characteristics of the KS data set. However, for higher FC_α, the available conformational space in the proteins is apparently more restricted as can be visualized by much lower dispersion of FC_ζ values at high FC_α than would be predicted by a simple linear variance.

(3) A suitable functional relationship for the restriction of the dispersion of $FC_{\zeta\neq\alpha}$ values at $FC_\alpha > 0$ can be derived as follows: Let us denote by x the total number of possible conformational states of amino acid residues not assignable to α-helical segments in a given protein: $x = k_1(1.0 - FC_\alpha)$. Assume that all residues may adopt β-sheet, bend, turn, or other conformations with equal probability. The number of possible states available (Cantor & Schimmel, 1980) for each nonhelical secondary structure type is then proportional to $\Omega = x!(x/4)!(x - x/4)!$. Standard analysis (McQuairre, 1973) of the dependence of Ω on $(x/4)!$ yields a Gaussian distribution function having a width (variance) $\sigma = (\Omega/2)^{1/2}$. Assuming x to be large and using the Stirling formula, the following equation for Ω can be derived: $\ln \Omega = K_2 x$, which would imply that $\sigma \sim K_1 \exp(-K_2 FC_\alpha)$, where K_1 and K_2 are combinations of constant terms and are not evaluated explicitly here since we are interested only in the functional form for the width of the distribution.

Using this functional relationship between σ and FC_α, the deviations (ΔFC_β) of the X-ray derived FC_β values from neural network values can be enclosed by the functions

$$\delta_\pm{}^\beta = \pm 0.25 \exp(-2.5 FC_\alpha{}^{KS}) \qquad (32)$$

Here the constant parameters were determined manually by trial and error to encompass the ΔFC range for all nonhelical secondary structures. For bend and turn, it is possible to reduce the preexponential term to 0.15 and encompass the ΔFC_ζ values, but, for the fraction of "other", we feel that the larger value (0.25) is still reasonable.

Our interpretation of the FC^{KS} values as calculated from eqs 14–17 can be understood in the following way: They are estimates of the most probable values of β-sheet, bend, turn, and other secondary structure fractions in a protein with a given fraction of α-helix. Simultaneously, using eq 32 and variants, one can determine the approximate uncertainty of such an estimation for a given $FC_\alpha{}^{KS}$ value.

The deviations of FC_ζ (X-ray) data from those calculated by eqs 22–24 as based on FC_α and FC_β are about the same for the bend and turn fraction as was seen in Figure 2b,c. However, as shown in Figure 3, the inclusion of the β-sheet fraction significantly reduced the scatter (σ) in the ΔFC_o plots, which relate the dispersion of the "other" component component to FC_α. The ΔFC_o values in Figure 3 are here more uniformly distributed around zero than in Figure 2d, particularly at low FC_α values.

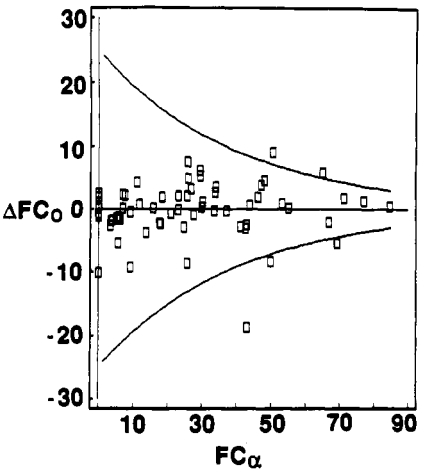

FIGURE 3: Deviations of $FC_o{}^{KS}$ from estimations based on the 2–3–3 neural network analysis (eq 28) showing the reduction in dispersion of values as compared to Figure 2d.

A simple understanding of the relationship of eqs 15–17 to 22–24 is to substitute eq 14 into eqs 22–24. For the KS-based analysis, new relationships functionally equivalent to those in eqs 15–17 are found. For some reason, the parallel conversion with the LG data does not work as well. Presumably this is a result of the reduced extent of the LG set and of the near linearity of the $FC_\alpha{}^{LG}$-dependent equations.

Implications and Possible Practical Applications. One must be careful not to overinterpret the interrelationships found here. Equations 14–31 have been derived using data for a set of similar globular, water-soluble, and crystallizable proteins. Thus they are most likely to be applicable to such proteins. If these equations were to be used for estimation of the secondary structure of a protein of unknown structure from knowledge of its α-helix content, some independent confirmation of the similarity (compatibility) of the protein in question with those in the "training set" should be established. Spectroscopic methods such as circular dichroism, infrared spectroscopy, and especially vibrational circular dichroism have the potential of providing such an experimental confirmation by analysis of the commonality of the spectral bandshape measured for the protein with those of the training set proteins. Additionally, these techniques can provide reasonable estimates of the α-helical content as an input to the derived equations, particularly when several methods are used to determine FC_α and their results are in agreement.

We have found the following procedure for establishing a relationship to the training set to be effective: First, the spectra of the unknown proteins are coprojected together with the spectra of the training set onto a set of orthogonal basis spectra using a modified factor analysis (Pancoska et al., 1979, 1991) or singular value decomposition computational scheme (Manavalian & Johnson, 1987). The resulting coefficients of these linear combinations are used as "coordinates" that define the spectral properties of the proteins in the set. These are input into a series of cluster analysis (CA) calculations (Sharaf et al., 1988) using a variety of CA algorithms (centroid type). The results of these calculations are evaluated in terms of the relative dissimilarity of the spectra of the proteins in the set. If these spectra reflect protein structure, the CA results will reflect any significant structural dissimilarity that would preclude use of our equations for estimating the most probable values of the other components of secondary structure from FC_α. If the unknown protein spectra cluster well with the training set spectra, then continued analysis is justified.

CONCLUSIONS

We have shown that two popular protein crystal structure analysis algorithms used to generate secondary structure data for reference sets for spectroscopic studies are indeed interconvertible at a practical level. The fractional coefficients that result from these analyses are not statistically independent. Through neural network analysis, we have studied the functional dependence of the various $FC_{\zeta\neq\alpha}$ values on FC_α and on FC_α and FC_β. By study of the variance of these $FC_{\zeta\neq\alpha}$ values with respect to the derived functional dependences, we have come to an understanding of the meaning of these interdependences. Our model is proposed to aid thinking about protein structure and to find a possible explanation for the oberved statistical behavior of these protein structural variables. In fact, its exact form is not the main point here.

In terms of spectroscopic utilization of these data sets, an appreciation for this variance in the protein structures used to calibrate one's technique is vital in reliably using the spectroscopically derived parameters. If a particular spectroscopic technique cannot achieve a less disperse variance than obtained by use of the equations presented here, that technique is not providing a description of the structure that is more reliable than a determination of FC_α or $FC_\alpha + FC_\beta$ alone. Simple measurement of average error alone will not yield this value because the training set used may be biased toward a region of low average variance.

From another point of view, these results make it clear why a technique such as electronic CD (in the UV), which is dominated by the α-helical contribution for protein samples, gives reasonable estimations of other $FC_{\zeta\neq\alpha}$ values. Our results clearly imply that a reliable α-helix estimation will lead to $FC_{\zeta\neq\alpha}$ estimations that have seemingly reasonable error bars when evaluated against a training set composed of proteins from a crystallographic compilation. But these results also show that such $FC_{\zeta\neq\alpha}$ values have not been actually determined by the spectra but have, in fact, derived from the structure of the training set.

Recognizing the internal structure of these data sets has let us propose methods for testing training set appropriateness to the questions at hand. From the complementary point of view, this study makes clear the need to carefully design one's training set to be appropriate. To improve on this situation, one should consider undertaking such spectral analyses for proteins by defined class rather than from a global point of view as is common now. We are now developing such focused spectral training sets using a variety of techniques in our laboratories.

ACKNOWLEDGMENT

We thank the IBM Corporation for establishment of a computational and electronic communication facility in Prague (IBM–Czechoslovak Academic Initiative), which was instrumental in this cooperation, and Charles University in Prague for facilitating the cooperation.

REFERENCES

Byler, D. M., & Susi, H. (1983) *Biopolymers 25*, 469–487.
Cantor, C. R., & Schimmel, P. R. (1980) in *Biophysical Chemistry*, Chapter 15, W. H. Freeman & Co., San Francisco, CA.
Dousseau, F., & Pezolet, M. (1990) *Biochemistry 29*, 8771–8779.
Fasman, G. D. (1989) *Prediction of Protein Structure and the Principles of Protein Conformation*, Plenum Press, New York.
Hennessey, J. P., & Johnson, W. C. (1981) *Biochemistry 20*, 1085–1094.
Horejs, J., & Kufudaki, O. (1990) in *Theoretical Aspects of Neurocomputing. Selected Papers from the Symposium on Neural Networks and Neurocomputing (Neuronet 1990)* (Novak, M., & Pelikan, E., Eds.) pp 44–52, World Scientific, Teaneck, NJ.
Kabsch, W., & Sander, C. (1983) *Biopolymers 22*, 2577–2637.
Levitt, M., & Greer, J. (1977) *J. Mol. Biol. 114*, 181–239.
Maggiora, G. M., Mao, B., Chou, K. C., & Narshiman, S. L. (1991) *Methods Biochem. Anal. 35*, 1–86.
Manavalan, P., & Johnson, W. C., Jr. (1985) *J. Biosci. (Suppl.) 8*, 141–149.
Manavalan, P., & Johnson, W. C., Jr. (1987) *Anal. Biochem. 167*, 76–85.
Manning, M. (1989) *J. Pharm. Biomed. Anal. 7*, 1103–1119.
Mantsch, H. H., Casal, H. L., & Jones, N. R. (1986) in *Spectroscopy of Biological Systems* (Clark, R. J. H., & Hester, R. E., Eds.) pp 1–46, John Wiley, Chichester.
McQuarrie, D. A. (1973) in *Statistical Thermodynamics*, Chapter 1, University Science Books, Mill Valley, CA.
Pancoska, P., & Keiderling, T. A. (1991) *Biochemistry 30*, 6885–6895.
Pancoska, P., Fric, I., & Blaha, K. (1979) *Collect. Czech. Chem. Commun. 44*, 1296–1312.
Pancoska, P., Yasui, S. C., & Keiderling, T. A. (1989) *Biochemistry 28*, 5917–5923.
Pancoska, P., Yasui, S. C., & Keiderling, T. A. (1991) *Biochemistry 30*, 5089–5103.
Perczel, A., Hollosi, M., Tusnady, G., & Fasman, G. D. (1989) *Croat. Chem. Acta 62*, 189–200.
Perczel, A., Hollosi, M., Tusnady, G., & Fasman, G. D. (1991) *Protein Eng. 4*, 669–679.
Qijan, N., & Sejnowski, T. J. (1988) *J. Mol. Biol. 202*, 865–888.
Sarver, R. W., Jr., & Krueger, W. C. (1991) *Anal. Biochem. 194*, 89–100.
Sharaf, M. A., Illman, D. L., & Kowalski, B. R. (1986) *Chemometrics*, John Wiley, New York.
Siegel, J. B., Steinmetz, W. E., & Long, G. L. (1980) *Anal. Biochem. 104*, 160–167.
Simpson, P. (1990) *Artificial Neural Systems: Foundations, Paradigms, Applications and Implementations*, Pergamon Press, New York.
Tu, A. T. (1986) in *Spectroscopy of Biological Systems* (Clark, R. J. H., & Hester, R. E., Eds.) pp 47–112, John Wiley, Chichester.
Unger, R., Hazel, D., & Sussman, J. L. (1989) *Proteins 5*, 355–373.
Urbanova, M., Dukor, R. K., Pancoska, P., Gupta, V. P., & Keiderling, T. A. (1991) *Biochemistry 30*, 10479–10485.