Protein Structural Segments and Their Interconnections Derived from Optical Spectra. Thermal Unfolding of Ribonuclease T_1 as an Example[†]

Petr Pancoska,[‡,§] Heinz Fabian,[||] Gorm Yoder,[‡] Vladimir Baumruk,[‡] and Timothy A. Keiderling*[,‡]

Department of Chemistry, University of Illinois at Chicago, 845 West Taylor Street, Chicago, Illinois 60607−7061, and Institute for Biochemistry, Max Delbruck Center for Molecular Medicine, Robert-Rossle-Strasse 10, D-13125 Berlin, Germany

Received May 17, 1996; Revised Manuscript Received July 16, 1996[⊗]

ABSTRACT: A novel descriptor for protein structure is examined here that goes beyond predictions of the average fractional components (FC) of a few conformational types and represents the number and interconnection of segments of continuous, well-defined secondary structural elements such as α-helices and β-sheets. This matrix descriptor can be predicted from optical spectra using neural network methods. The new matrix plus traditional FC descriptors can be quickly and generally obtained to provide a level of detail not previously derived from optical spectra and a discrimination between proteins that might otherwise be viewed as being very similar using just the FC descriptor. As an example of its potential utilization, this matrix descriptor approach was applied to an analysis of both the native state and the reversible thermal denaturation of ribonuclease T_1 in H_2O. Analyses of the FTIR spectral data indicate initial loss of the major helical segment at 50−55 °C but with little accompanying change in the number of sheet segments or the sheet FC values. Circular dichroism (CD) and vibrational CD data are also used to support this interpretation based on FC changes with temperature. Parallel analysis of the corresponding data for this protein in D_2O demonstrates that the method is sensitive to the match between the degree of H−D exchange used to prepare samples for the unknown and the reference data set.

Optical spectra such as electronic and vibrational circular dichroism (ECD and VCD), Fourier transform infrared (FTIR), as well as Raman spectra have been extensively used in the past to obtain estimates of the average fractional component (FC) of secondary structure in a protein (Keiderling, 1996; Keiderling & Pancoska, 1993; Johnson, 1985, 1988, 1990; Sreerama & Woody, 1994; Sauer, 1994; Pancoska et al., 1995; Pribic et al., 1993; Sarver & Kruger, 1991; Baumruk et al., 1996; Jackson & Mantsch, 1995; Surewicz et al., 1993; Lee et al., 1990; Dosseau & Pezolet, 1990; Williams, 1986; Rahmelow et al., 1993; van Stokkum et al., 1990). Such optical methods, which sample structure on an inherently rapid time scale, are particularly appropriate for studying protein-folding processes, since the intermediate species one wishes to probe are often unstable. Such dynamic structures are poorly suited to more precise, but slower time scale, NMR structural techniques or to X-ray diffraction analyses of the crystal-stabilized distribution of structures. Optical spectra can rarely yield the structural detail of those techniques but remain vitally useful for qualitatively monitoring structure, particularly for relative changes in a single or related protein (Keiderling et al., 1994). In a series of previous papers, we have demonstrated that optical spectroscopic methods when used for quantitative determination of average secondary structure parameters are inherently limited in their accuracy. More importantly, the best predicting secondary structure algorithms we and others have developed utilize only part of the statistically significant spectral information available (Pancoska et al., 1991, 1994; Baumruk et al., 1996; Pribic et al., 1993; Rahmelow et al., 1993; van Stokkum et al. 1990). The root causes for both of these developments are the fundamental assumptions used to form what we will refer to as the protein structural descriptor, that mathematical entity used in the analysis to summarize the protein conformation, typically in terms of secondary structure.

The typical picture of protein secondary structure in terms of a sum of components is much too idealized and restrictive to explain the variations seen in optical spectra. Describing the protein only in terms of FC values oversimplifies the nature of secondary structural segments, ignoring their inhomogeneities. Since the average helix and sheet segments are relatively modest in length, the end residues, which can vary sharply in local conformation, form a significant component of the protein's structure. Their spectral contribution varies directly with the number, *N*, of segments but not simply with the corresponding FC of the respective segments. Since optical spectroscopic techniques sense these interconnecting end residues differently from the central residues, this dependence on *N* can have a profound impact on the accuracy of the resultant FC values obtained from optical spectra. Knowing the number, *N*, of segments of a type, ζ, and the corresponding FC_ζ value then gives one a means of determining average ζ segment length which can be correlated to those spectral effects sensitive to length. Since structural segments are interconnected via their ends, inclusion of such segment interfaces becomes a logical next step in the development of a more detailed structural descriptor to extract from optical spectra.

[†] The work at UIC was supported by a grant from the National Institutes of Health (GM30147 to T.A.K.). Initial development of the matrix approach to spectral analysis was facilitated by a Joint National Science Foundation grant between Charles University and UIC (NSF INT 91-07588 to T.A.K. and P.P.) and grants from the Grant Agency of Czech Republic (GACR 203-93-0714 to P.P.).

* To whom correspondence should be addressed at UIC.

[‡] University of Illinois at Chicago.

[§] Permanent address: Department of Chemical Physics, Faculty of Mathematics and Physics, Charles University, Ke Karlovu 3, 121 16 Prague 2, Czech Republic.

[||] Max Delbruck Center for Molecular Medicine.

[⊗] Abstract published in *Advance ACS Abstracts,* September 1, 1996.

S0006-2960(96)01178-6 CCC: $12.00

As we have separately described in mathematical detail and preliminarily communicated as a test example (using VCD data), a new descriptor of protein secondary structure has been developed along these lines (Pancoska et al., 1994, 1997). The most simple and convenient form for this descriptor has an integer matrix form consisting of the numbers of "uniform" segments for each structural type being considered along its diagonal and the numbers of their respective (oriented) interconnections as the corresponding off-diagonal elements. While, in principle, this matrix can be of any desired dimension, the simplest practical size is 3 × 3, representing helices, sheet segments, and the rest ("other"). Such a matrix is the first step in going beyond the restrictive "vector" type descriptor of FC values in terms of extending the level of structural detail.

This matrix descriptor has a form and level of detail qualitatively different from the FC approach yet one that is still appropriate for abstraction from analysis of optical spectra by implicitly including end effects via the interconnections and average segment lengths via their numbers. The matrix form maintains the directionality of the interconnection information. Vibrational spectra are sensitive to the local perturbations in structure that must occur at the ends of uniform segments (Dousseau & Pezolet, 1990; Venyaminov & Kalnin, 1990b; Venyaminov et al., 1996), while ECD spectra are sensitive to segment length distributions (Yang et al., 1986; Venyaminov & Yang, 1996).

The analysis of conventional FC values is still necessary to fully utilize the information gained from this new matrix descriptor. However, the matrix goes beyond the FC "vector descriptor" by offering new insight into the distribution of the separately determined fractional secondary structure over the protein sequence. This is a first step toward experimentally predicting the sequence-specific secondary fold of a protein. It is hoped that, when used in concert with theoretical structure prediction algorithms based on the sequence (Holly & Karplus, 1989; Garnier et al., 1978; Fasman, 1989; Srinivasan & Rose, 1995; Yue & Dill, 1996), such an approach might eventually lead to more reliable estimates of protein structure that are directly coupled to experimental data and that could additionally be obtained on a relatively short time scale.

In our preliminary studies, it became clear that to predict the matrix descriptor from optical spectra it would be necessary to use an approach different from the factor-analysis (FA)-based restricted multiple regression (RMR) methods we have previously used for FC prediction based on ECD, VCD, and FTIR spectra (Pancoska & Keiderling, 1991; Pancoska et al., 1991, 1994, 1995; Baumruk et al., 1996; Bitto, 1993). Those methods, though computationally optimized, were not conceptually different from a number of other protein structure analysis algorithms for ECD and FTIR spectra (Seigel et al., 1980; Hennessey & Johnson, 1981; Provencher & Glockner, 1981; Compton & Johnson, 1986; Johnson, 1988; Manavalan & Johnson, 1987; Dousseau et al., 1990; Lee et al., 1990; Perczel et al., 1991; Sarver & Kruger, 1991a,b; Toumadje et al., 1992; van Stokkum, 1990; Pribic et al., 1993; Sreerama & Woody, 1993; Andrade et al., 1993; Merelo et al., 1994). Here, we instead employed a novel "cascade" type back-propagation neural network trained on spectra of a basis set of proteins to predict their matrix descriptors. We have separately demonstrated that the matrix descriptor is predictable to a level of accuracy roughly comparable to that obtained for the corresponding FC values in preliminary tests done for a set of 23 globular proteins of known structure using just their amide I′ (dissolved in D_2O) VCD spectra (Pancoska et al., 1994). This preliminary test used a modified "one-out" strategy to test predictive ability for a selected protein structure that was not used in training the network. This procedure was repeated 23 times for each protein in the set. Despite the fact that the predicted matrices have the same level of precision, when they are combined with spectrally predicted FC values, the information content of the protein secondary structure parameters derivable from such a spectral analysis is increased. In this paper, we both test if this new descriptor can be predicted for and give useful information about a set of spectra for proteins of truly unknown structure and if its use can be extended beyond VCD data.

Ribonuclease A and Ribonuclease T_1 as a Model System. To test our methods on a real example of unknown protein structures, we have chosen to use the thermal unfolding of ribonuclease T_1 (RNase T1) as a source of unknown protein structures. In addition to being an interesting protein in its own right, RNase T1 has some useful similarities to ribonuclease A (RNaseA), a member of our training set of proteins used for both the FC and matrix prediction tests (Pancoska et al., 1994, 1995; Baumruk et al., 1996). These two proteins provide an example of two biomolecules which have similar spectra and nearly identical FC descriptors in their native states but quite different folds (Figure 1) which results in segment layouts (Figure 2) that lead to significantly different matrix descriptors. If our spectra−structure correlation methods can predict reasonable representations of both the similar FC values and the different matrix descriptors for these two proteins in their native states by use of the same algorithm, that will provide a good test of the algorithm's reliability. Beyond that, previous studies have shown that RNaseT1 at neutral pH and moderate concentration can be reversibly denatured with substantial unfolding of its secondary structure by raising its temperature to above ∼55 °C (Fabian et al., 1993, 1994).

Since the matrix descriptor is a novel concept, it is useful here to illustrate its construction by example. In RNaseA, a 124-residue protein, 21% of α-helix is distributed between three short helical segments (6−8 amino acids each, Figure 1). By contrast, in RNaseT1, 17−20 amino acid residues form just a single α-helical segment, resulting in approximately the same fraction (18%) out of 104 total residues. The matrix descriptor quantitatively discriminates between these two quite different helical distributions. The sheet fraction in these proteins is distributed similarly in each protein, being 35 and 30% spread over six and seven segments for RNaseA and RNaseT1, respectively. The topological arrangements of these segments for the two proteins are schematically expressed in Figure 2. By counting the segments and interconnections, one can develop a segment descriptor in its simplest form as a 3 × 3 matrix, [HEC], as shown below.

$$\mathbf{RNaseA} = \begin{bmatrix} 3 & 1 & 2 \\ 0 & 6 & 6 \\ 3 & 5 & 9 \end{bmatrix} \quad \mathbf{RNaseT1} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 7 & 7 \\ 1 & 7 & 9 \end{bmatrix}$$

It can be noted that this definition does not assign any residues to the segment junctions but rather just counts their

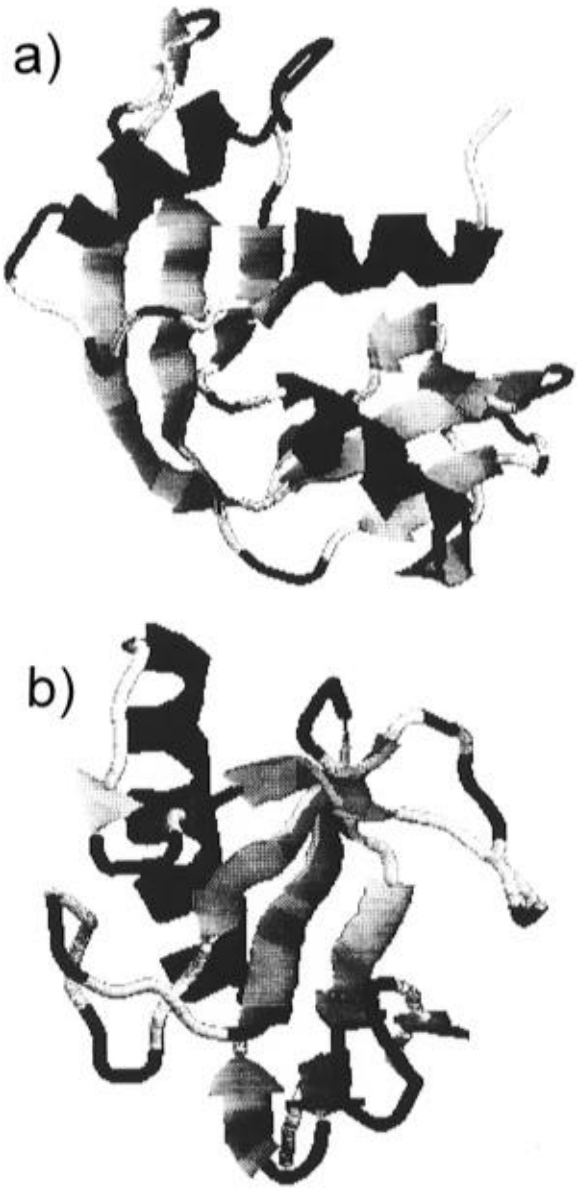

FIGURE 1: Schematic representations of the three-dimensional structure of (a) RNaseA (PDB file 3RN3) and (b) RNaseT1 (PDB file 9RNT) plotted using RasWin (v. 2.6, R. Sayle, Glaxo Research & Development, 1995).

existence. An alternative definition of the junctions as having a penetration of *j* residues into adjoining segments is equally compatible with this matrix descriptor form (Janota, 1993).

MATERIALS AND METHODS

All the spectral data used to develop the structural reference training sets were taken directly from our previous studies of FTIR of proteins in H_2O (Baumruk et al., 1996) and VCD of proteins in D_2O and ECD in H_2O (Pancoska et al., 1995) and used here without change. All of this data is available from the authors in machine readable format, and the latter (VCD in D_2O and ECD) set is additionally obtainable via the Internet.[1] The FTIR absorption spectra of these training set proteins in D_2O was obtained coincidentally with the VCD data collection. The same sampling procedures as described below for the VCD and measurement procedures similar to those for the RNaseT1 data collection (except for buffers) were used for the FTIR (D_2O) training set.

Materials

Recombinant ribonuclease T_1 was prepared and purified as described (Landt et al., 1992). To obtain defined buffer conditions for the FTIR studies, the lyophilized samples were dissolved, filtered at dilute buffer conditions through Sephadex G-25, and lyophilized again. Prior to the FTIR experiments, the samples were dissolved to yield buffer conditions of 10 mM cacodylate at pH 7.0 or pD 7.0 and protein concentrations of 30 and 10 mg/mL for measurements in H_2O and D_2O buffer, respectively. For ECD, samples were dissolved directly in H_2O-based 10 mM cacodylate buffer at pH 7.1 with a protein concentration of 0.8 mg per 2 mL or ~3 mM in peptide for $\Delta\epsilon$ determination.

Spectral Measurement

FTIR spectra were recorded as a function of temperature either on a Bruker IFS 66 (in D_2O at Berlin) or on a Digilab FTS-40A (in H_2O, at Winnipeg) FTIR spectrometer each equipped with a liquid nitrogen-cooled mercury cadmium telluride (MCT) detector and continuously purged with dry air. For each sample, 512 interferograms were coadded and Fourier transformed to generate a spectrum with a nominal 2 cm^{-1} resolution. The protein solutions were placed between a pair of CaF_2 windows separated by 6 μm for measurements in H_2O buffer or by 45 μm for samples in D_2O buffer. The sample temperature in both cases was controlled as previously described for the D_2O samples (Fabian et al., 1993). Pure solvent spectra were recorded under identical conditions and subtracted from the spectra of the proteins in the relevant solvent. Extreme care was taken to keep the same purge level during the measurements. Minor spectral contributions from residual water vapor were eliminated using a water vapor absorption spectrum measured under identical conditions as described earlier (Fabian et al., 1993).

VCD spectra over the amide I′ band of RNaseT1 in D_2O were obtained on the UIC dispersive instrument using techniques already described in detail in the literature (Keiderling, 1981, 1990; Pancoska et al., 1989, 1991). Briefly, a sample having a concentration of about 3 mg per 100 μL in D_2O was placed in a homemade variable-temperature cell (Wang, 1993) based on two CaF_2 windows separated by a 50 μm thick Teflon spacer contained in a small brass mount, which incorporated two o-ring seals in order to inhibit leakage. This cell was then tightly fit into a double-walled brass jacket through which water was pumped from a thermostatically controlled bath (Neslab). The bath temperature was regulated to achieve a constant setting as monitored by a thermocouple placed in the outer jacket of the cell. A total of eight VCD scans at each temperature interval were performed with each scan requiring approximately 30 min. Measurements were made at 10 °C intervals ranging from 20 to 70 °C, going beyond the transition temperature and then back to room temperature again over the course of 2 days. Because of the much longer time scales necessary for collection of the VCD data, somewhat different thermal behavior was encountered as compared to that seen in the FTIR experiments. FTIR

[1] As material appended to Pancoska et al. (1995).

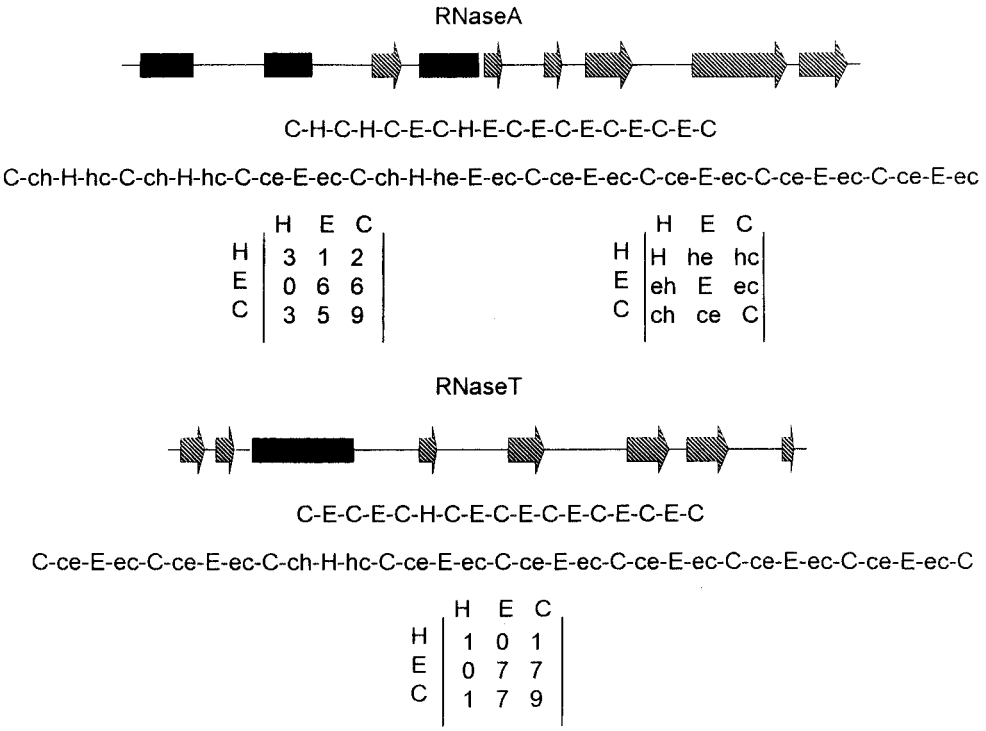

FIGURE 2: Schematic derivation of 3×3 [HEC] matrix descriptors for RNaseA (top) and RNaseT1 (bottom). Secondary structure segments (rectangle → helix, H; arrow → sheet, E; and line → other, C) are shown first to scale, then as a list of segments, and finally with interconnections inserted in lowercase symbols where, for example, hc and ch represent the ordered contacts of segments H and C (hc = H → C, ch = C → H). The [HEC] matrix is constructed by arranging the numbers of secondary and contact structures in a 3×3 scheme as shown.

spectra of these same samples were run on a BIO-RAD FTS-60 spectrometer as a control.

ECD spectra were measured using a JASCO J-600 spectropolarimeter with a cylindrical 0.05 cm path quartz cell (Precision Cells) held in a standard (JASCO) thermostated cell holder. The relationship of the bath and sample temperature and settling time was determined by precalibration of a remote sensing thermocouple. During the measurement, the temperature of the sample was monitored by the thermocouple which was in direct contact with the solution. Spectra in the range of 185−265 nm were measured as the average of five scans using 45 min for each temperature point. Upon the solution being returned to room temperature after the heating cycle and allowed an added delay of 70 min for refolding, the full initial spectrum was recovered.

Mathematical Methods

(1) Pretransformation of the Spectra. Ideally, one should determine the concentration and path length under which the spectra were run to create a basis set of spectra in terms of the molar extinction coefficient. For several technical reasons, it is not possible to do this with sufficient accuracy for our spectra, especially for those obtained on samples in H_2O. Consequently, each FTIR spectrum in the training set was normalized by the maximum of its amide I band to make a first-order correction for the concentration−path length variations in our training set samples. With this procedure, we approximate the actual extinction coefficient at the frequency of normalization with a common "effective value" which loses any absolute intensity variation with structural type but maintains relative intensity changes of the component bands. While this problem has been recognized before (Surewicz et al., 1993; Mantsch et al., 1986; Lee et al., 1990; Venyaminov & Kalnin, 1990a; Jackson & Mantsch, 1995), we feel this normalization is the most reasonable method for correction of the more serious, and spurious, intensity variations due to concentrations. To enhance the band shape variability of the FTIR spectra, the average training set spectrum[2] has been subtracted from the normalized spectrum of each protein to create a set of difference FTIR (DF) spectra that were used as the protein spectral input information for all subsequent calculations (Baumruk et al., 1996).

Native state FTIR spectra for RNaseT1 and RNaseA were treated in the same manner as the training set. However, for the temperature dependent RNaseT1 data, there are two contrasting points of view. The formally consistent method of treating all of the data is to normalize each spectrum, just as was done for the training set. We have done that to provide a totally consistent treatment of the DF spectral data and term it to be our *reference analysis*.

This formally consistent treatment may not be an optimal one for deriving the maximum amount of structural information from the data available in this specific case of a thermal denaturation. Since the temperature dependent RNaseT1 data were measured for a single sample in the same cell with no change in conditions other than temperature, the relative concentration−path length correction can in fact be made for these samples by referencing the high-temperature spectra to the native state spectrum. Thus, as a test of an alternative approach, the temperature dependent FTIR spectra of RNaseT1 were renormalized by using the amide I maximum of the *T*

[2] To obtain the average spectrum for creation of the DF set, the average of the normalized FTIR spectra for just the proteins in the training set was calculated (separately for measurements in H_2O and D_2O).

= 20 °C spectrum for all the higher-temperature spectra, thereby preserving the FTIR intensity changes. The DF spectra formed from these normalized RNaseT1 spectra were used as input for our second structural analysis, termed the *intensity-enhanced analysis*.

VCD spectra are treated a bit differently. For FC determination, following our usual methods (Pancoska et al., 1995), the amide I′ VCD intensity is normalized to the peak absorption intensity (giving a measure of $\Delta A/A$ at that frequency), and regression relationships are developed using these corrected intensities.

For the matrix descriptor determination using a neural network analysis of each type of spectral data, we scaled the input spectral intensities to fall in the interval from −0.8 to 0.8 (Pancoska et al., 1996). This was accomplished by dividing all the intensities in the training set plus those of the unknown samples by 1.2 times the absolute value of the largest normalized peak or through in the training set. Such an overall correction preserves the relative intensities for the analysis.

(2) FC Determination. For prediction of FC values from the RNaseT1 DF spectra, we used our optimized algorithm, based on the FA RMR method which we have described in detail separately (Pancoska et al., 1979, 1995; Baumruk et al., 1996). The RMR models were optimized for predictive accuracy on the appropriate reference sets of DF (FTIR) (Baumruk et al., 1996), ECD, and VCD (Pancoska et al., 1994, 1995) spectra for 19 proteins of known structure dissolved in H_2O and 23 known proteins in D_2O, as has been described. The RMR for DF spectra of proteins in D_2O was similarly determined using FA coefficients derived from FTIR spectra obtained at UIC (unpublished data) for a set of 22 known proteins.

(3) Generation of the Matrix Descriptors. The atomic coordinates for all the proteins in the training set as well as that for RNaseT1 were obtained from the Brookhaven Protein Data Bank.[3] The X-ray coordinates were processed by the Kabsch and Sander (1983) DSSP program to obtain assignments of amino acid residues to helix (H, includes α-helices and 3_{10}-helices), sheet (E, includes antiparallel and parallel strands), and other (C, the remainder, all residues not covered by the first two) secondary structure types. The segment descriptors based on these three classes of structure will be referred to here as [HEC] matrices to distinguish them from other more complex versions that are still fully consistent with the matrix descriptor concept (Pancoska et al., 1994). For all the proteins of known structure, [HEC] matrices were generated from the output of the DSSP program using our own Turbo Pascal (Borland, v. 7.0) program (Janota, 1993).

(4) Neural Network Calculations. Using the NeuralWorks Professional II/Plus package, v. 5.20 (NeuralWare Inc., Pittsburgh, PA), we composed a pseudo-cascade topology yielding an optimal back-propagation neural network for prediction of [HEC] matrices from DF and VCD spectra as shown schematically in Figure 3. In this representation, the double arrows represent the neuron layers with the actual number of neurons used indicated below the arrow. Two branches of hidden layer neurons (left and right) process the input spectral intensities independently using different schemes

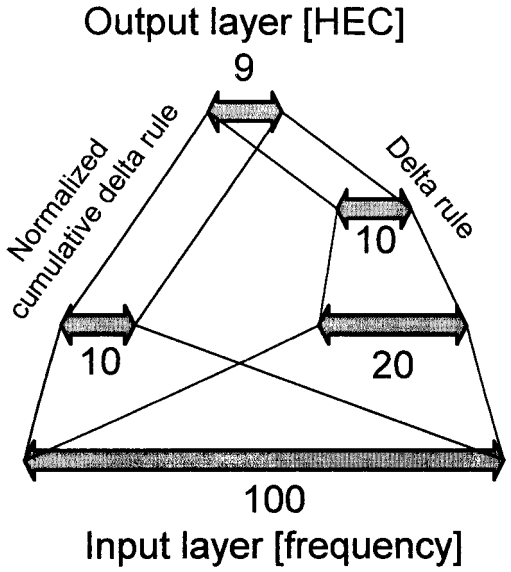

FIGURE 3: Topology of the back-propagation neural network used for calculation of the [HEC] matrix descriptors from spectra. Double arrows represent neuron layers with the number of neurons indicated by the numerals. Differences in the schemes used to update weights in the two branches are indicated.

for updating the weights as indicated on each side. The different weight update algorithms used in the final design prevent the neurons in the shorter (left) branch from saturating, which was found to occur in all tests with networks having the same weight update methods in both branches. The information from the left and right branches is then combined in the output layer, which consists of the nine matrix elements composing the [HEC] matrix. A hyperbolic tangent was used as the transfer function.

Several other network designs were tested having different numbers of hidden layers, different numbers of neurons in them, and different algorithms used in the respective branches for updating the weights. The above network topology was the best design in terms of learning rate and generalization ability for prediction of the matrix elements that resulted from these preliminary trials.

Just as it is necessary to scale the input spectra to the range of values in the interval of transfer function values, the output [HEC] matrix must be similarly normalized, which represents a much more important issue for the analysis. From extensive testing, we found that the best performance of the network in terms of prediction of [HEC] elements from the spectra of the training set proteins was obtained when the [HEC] matrix of each protein was individually normalized by dividing all the elements by its largest matrix element. Unfortunately, this method of output normalization is not general, since one cannot back-scale the results obtained for an "unknown" protein because its largest element is also unknown.

However, for the thermal unfolding of RNaseT1 presented here as a special but practical case, the problems of this optimal normalization scheme can be avoided to a large extent. Assuming that the RNaseA structural parameters would be closely enough related to those of RNaseT1, we used the following method. The network was trained to generate individually normalized [HEC] matrices using a training set that did not contain data for RNaseA. The quality of prediction of the normalized [HEC] matrix of RNaseA from its DF spectra was used as a criterion for optimization

[3] For detailed file names of the training set proteins, see Table 1 of Pancoska et al. (1995); for RNaseT1, the 9RNT structure file was used.

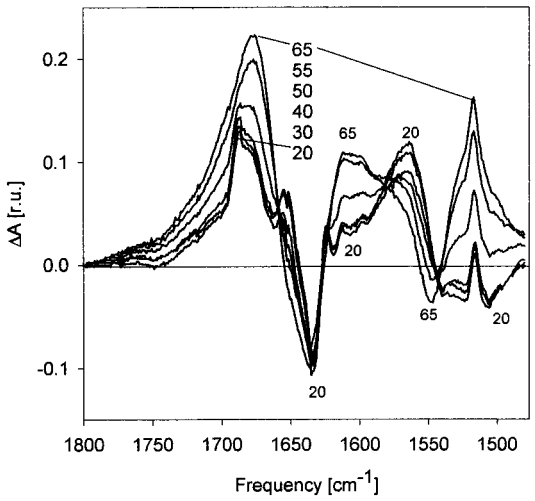

FIGURE 4: Temperature dependence of the amide I and II DF (difference FTIR) spectra of RNaseT1 in H_2O (intensity-enhanced normalization, temperature in degrees Celsius, DF intensity in relative units).

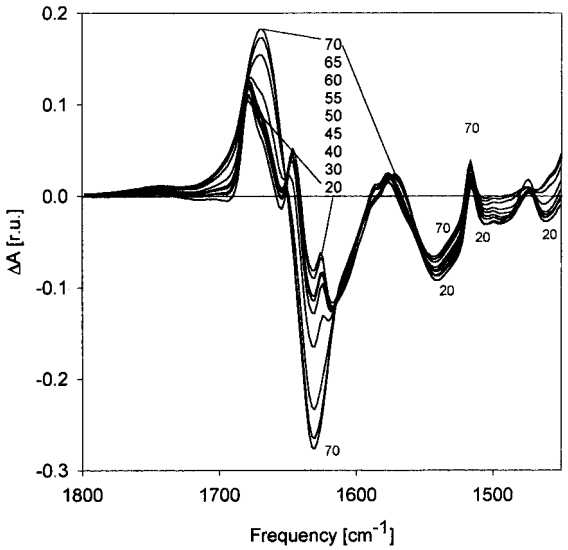

FIGURE 5: Temperature dependence of the amide I′ DF spectra of RNaseT1 in D_2O (as in Figure 4).

of the network using the "save-best" training variant[4] (Pancoska et al., 1996; NeuralWare, 1993). This approach is reasonable in that it ensures that the network is optimized to predict the structural descriptor of a protein that is structurally related to the unknown protein. The method also prevents overtraining of the network. The data for the unknown protein structures are then input to the optimized network, and normalized [HEC] matrices are generated. To recover interpretable [HEC] elements for RNaseT1 and its thermally unfolded states, the matrix element used to normalize the original [HEC] descriptor of RNaseA is applied to the NN output for RNaseT1. Comparison of this back-scaled predicted matrix for the native structure RNaseT1 sample with the expected [HEC] matrix of RNaseT1 (from the X-ray parameters) is a natural validity test for the approximations used in our method.

RESULTS

(1) Spectral Data. Experimental DF spectra (after subtraction of the average normalized absorption spectrum) for the temperature variation of the amide I and II (H_2O) and amide I′ (D_2O) FTIR bands of RNaseT1 are shown in Figures 4 and 5, respectively. These spectra are the ones used for the intensity-enhanced calculations. The reference analysis gives a data set of the same form except that the zero crossings are shifted and the intensity variations of the 1670, 1610, and 1520 cm^{-1} regions in the H_2O data set are more pronounced. The two sets of spectra (Figures 4 and 5) are characterized by reasonably well-defined "isosbestic" points, suggesting that they could be consistent with a two-state process. Up to 40 °C for the H_2O and 50 °C for the D_2O data sets, the spectra are fairly constant. Then in each case, the intensity on the high-frequency side of the main amide I or I′ feature grows. At higher temperatures in D_2O, the DF intensity at 1630 cm^{-1} also decreases, and in H_2O, the amide II intensity at ~1520 cm^{-1} sharply increases as does that at ~1600 cm^{-1}.

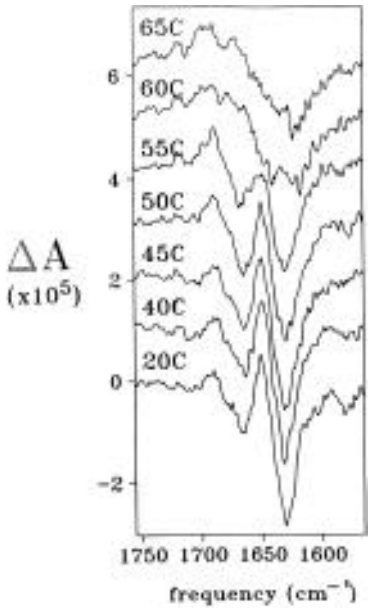

FIGURE 6: Temperature dependence of the amide I′ VCD spectra of RNaseT1 in D_2O (normalized by the peak amide I′ absorbance, temperature in degrees Celsius).

VCD spectra obtained for RNaseT1 in D_2O over the same temperature range are shown in Figure 6. The band shapes at the lower temperatures are characteristic of an $\alpha-\beta$ protein (W-shaped, Pancoska et al., 1991) and remain fairly conserved up to 50 °C. The 55 °C result is qualitatively similar but varies in the relative intensities of the negative VCD bands at 1630 and 1670 cm^{-1} and the positive features at 1650 and 1685 cm^{-1}. At higher temperatures, the high-frequency negative band (characteristic of helix) disappears and the VCD spectrum becomes less structured. The 65 °C VCD spectra are consistent with the spectra we have obtained for other proteins under denaturing and aggregating conditions (Yasui et al., 1990). At 70 °C (spectra not shown), precipitation of the protein was observed, which may have been exacerbated by the longer VCD time scale.

The ECD spectra for RNaseT1 from 20 to 65 °C are shown in Figure 7. The native state pattern is faithfully preserved

[4] After a predefined number of training iterations (e.g. 500 or 1000), the software interrupts the training and uses the RNaseA DF spectra to calculate its [HEC] prediction, which is then compared with its expected (correct) matrix descriptor. If the prediction has a smaller error than was found in the previous test, the network is stored. For the final prediction of matrices for the unknown structures, the stored network is used.

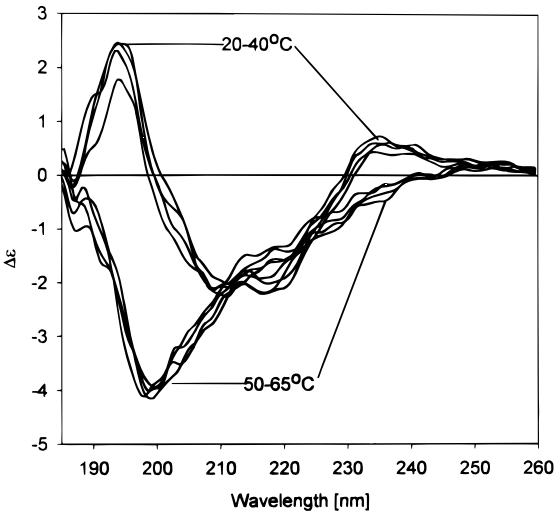

FIGURE 7: Temperature dependence of the far-UV ECD spectra of RNaseT1 in H_2O−buffer at pH 7 (temperature in degrees Celsius). Data are in molar CD on a per unit amide basis.

Table 1: Predictions of the Fractional Composition of (a) RNaseA and (b) RNaseT1 in Their Native State Using FA RMR Based on Various Spectroscopic methods[a]

spectral method	helix	sheet	turn	bend	other	Σ[d]
(a) RNaseA						
DF H_2O[b]	20	26	13	14	25	96.5
DF H_2O ref[c]	19	25	13	15	26	97.0
DF D_2O[b]	14	33	11	17	25	99.3
DF D_2O ref[c]	14	32	11	15	24	97.0
VCD D_2O	30	31	11	11	17	100.2
ECD H_2O	24	24	14	14	22	97.5
X-ray (KS)	21	35	11	14	19	100.0
(b) RNaseT1						
DF H_2O[b]	20	20	13	15	25	94.0
DF H_2O ref[b]	18	21	13	15	26	93.7
DF D_2O[b]	22	25	14	17	22	100.2
DF D_2O ref[c]	26	20	14	20	24	103.2
VCD D_2O	24	13	13	15	31	96.6
ECD H_2O	15	29	13	16	20	92.6
X-ray (KS)	18	30	24	9	19	100.0

[a] FC values predicted using a set of reference spectra without ribonuclease A. [b] Intensity-enhanced normalization of FTIR spectra was used. [c] Reference normalization of FTIR spectra was used (see text for explanation). [d] Control sum of predicted FC values (expected ΣFC = 100%).

to 40 °C and at 50 °C sharply changes to the high-temperature pattern which is maintained to 65 °C (whereupon the heating was terminated). The low-temperature spectra are characteristic of a mixed helix−sheet structure with enough coil component to diminish the 195 nm band and with the negative feature at 222 nm distorted by a weak positive band at 235 nm. The high-temperature spectrum indicates a large increase in the disordered (or "coil") component but still retains evidence of residual structure. The ECD intensity at 222 nm undergoes very little change in this process even though the 195 nm changes clearly indicate a substantial loss of helix for high-temperature RNaseT1.

(2) FA RMR Prediction of FC Values. To establish how well the predictions work for these proteins and for the sake of comparing them, the predicted FC values for the native forms of RNaseA and RNaseT1 are presented in parts a and b of Table 1, respectively, along with the X-ray-derived values for evaluation of the error of prediction. On the whole, the native state FC predictions for helix in RNaseT1 were quite good and those for sheet were uniformly low. These FC values are slightly sensitive to the normalization model used,[5] with the intensity-enhanced analysis giving higher helix and lower sheet values than the reference analysis for the H_2O data and vice versa for the D_2O data.

Table 2: Predictions of Fractions of Secondary Structures of RNaseT1 at Different Temperatures from FA RMR Results Based on DF (FTIR) Spectra in H_2O

temperature (°C)	helix	sheet	turn	bend	other	Σ
Reference Normalization						
20	18	21	13	15	26	93.7
30	17	21	14	15	26	93.1
40	18	21	14	15	25	93.7
50	14	22	14	16	25	91.9
55	10	23	15	17	25	90.4
65	7	23	15	18	26	88.6
Intensity-Enhanced Normalization						
20	20	20	13	15	25	94.0
30	20	20	13	15	25	93.6
40	20	20	13	15	25	93.5
50	18	19	14	15	25	92.0
55	15	19	14	16	26	90.3
65	13	18	15	16	26	88.2

Table 3: Predictions of Fractions of Secondary Structures of RNaseT1 at Different Temperatures from FA RMR Results Based on DF (FTIR) Spectra in D_2O

temperature (°C)	helix	sheet	turn	bend	other	Σ
Reference Normalization						
20	26	20	14	20	24	103.2
30	23	20	14	21	24	101.4
40	20	19	14	22	25	99.9
45	18	19	14	23	25	98.2
50	16	18	14	24	26	97.4
55	11	16	13	27	26	92.1
60	0	13	12	32	25	80.9
65	−6	11	11	34	24	75.0
70	−9	11	11	35	25	72.6
Intensity-Enhanced Normalization						
20	22	25	14	17	22	100.2
30	19	25	13	18	23	99.2
40	18	25	13	18	24	97.3
45	16	25	13	18	24	96.9
50	15	24	13	18	25	94.9
55	10	23	13	16	24	86.5
60	2	22	12	13	21	70.2
65	−2	21	12	12	21	63.4
70	−5	21	12	12	21	61.4

The predicted FC values for RNaseT1 at temperatures from 20 to 70 °C are listed in Tables 2−4 for values derived from the DF (H_2O), DF (D_2O), and VCD (D_2O) data sets, respectively. At the lower temperatures, the overall secondary structure composition of RNaseT1 is predicted to remain reasonably the same as the native structure as would be expected from the consistency of those spectra with that of the native state. In all cases, the main trend in the predicted FC values is the decrease of the content of helical secondary structure, beginning at 50−55 °C and continuing at higher temperatures. In general, all methods predict much smaller changes in the sheet content over this temperature range, but the VCD and DF (D_2O) predictions (Tables 3 and 4) do

[5] This is due to different spectral intensities being used with the two normalization schemes which, in turn, affects the subsequent regression calculation. Despite this, the variations are not significant when compared to the overall error level.

Table 4: Predictions of Fractions of Secondary Structures of RNaseT1 at Different Temperatures from FA RMR Results Based on VCD Spectra in D_2O

temperature (°C)	helix	sheet	turn	bend	other	Σ
20	24	13	13	15	31	96.6
40	25	14	13	14	29	94.3
45	26	14	13	14	30	97.9
50	26	8	15	16	31	95.6
55	22	10	15	17	30	93.6
60	13	17	16	21	31	98.2

Table 5: Neural Network Predictions of Matrix Descriptors for Native State RNaseT from VCD and DF FTIR Spectra[a]

method	calculated		corrected
FD (H_2O)	$\begin{vmatrix}1&0&1\\0&7&6\\1&7&9\end{vmatrix}$	→	$\begin{vmatrix}1&0&1\\0&7&7\\1&7&9\end{vmatrix}$
FD (D_2O)[b]	$\begin{vmatrix}1&0&1\\0&7&7\\1&6&9\end{vmatrix}$	→	$\begin{vmatrix}1&0&1\\0&7&7\\1&7&9\end{vmatrix}$
VCD (D_2O)[b]	$\begin{vmatrix}1&0&1\\0&6&6\\2&6&8\end{vmatrix}$	→	$\begin{vmatrix}1&0&1\\0&6&6\\1&6&7\end{vmatrix}$
FD (D_2O) (pre-exchanged)	$\begin{vmatrix}4&0&4\\0&5&4\\3&5&9\end{vmatrix}$	→	$\begin{vmatrix}4&0&4\\1&5&4\\3&5&9\end{vmatrix}$

[a] For FTIR spectra, the results for reference and intensity-enhanced normalization of the spectra were identical. [b] Measured without H−D pre-exchange.

decrease substantially at ∼55 °C. The highest-temperature (60 °C) VCD data (Table 4) significantly deviate from the trends which may be an indication of aggregation.

The overall tendency is for the turn and bend fractions to remain effectively constant. Perhaps more interesting is the overall insensitivity of the other fraction to the temperature changes, since one might expect loss of structure at high temperature to result in gain of disordered component which should be represented in this analysis by other. Another way to monitor this aspect of the quantitative FC results is to compare the sum of predicted FC values (last column, Tables 2−4). These sums are generally lower than 100% and, except for the VCD (D_2O) analysis, tend to decrease with increasing temperature. The low-temperature sums for the D_2O DF analysis are a bit better than for the H_2O DF analysis with little difference found between normalization methods. However, the sums for the DF (D_2O) analysis at or above 60 °C become so low (<70%) as to cause the quantitative results with this data set to lose credibility (see Discussion). While all the methods used are consistent with a transition temperature at ∼55 °C, the DF (D_2O) predictions for helix change much more gradually from 20 °C than do those from the DF (H_2O) or VCD (D_2O) data. The lower-concentration ECD spectra show a very sharp transition at 50−55 °C, but by contrast, quantitative analyses of the ECD spectra as temperature is varied yield ambiguous results which will be addressed further in the Discussion.

(3) Neural Network Predictions of the [HEC] Matrix Descriptor for RNaseT1. In all cases, the network correctly predicted the native state [HEC] matrix for ribonuclease A (see Figure 2) from its respective spectra (not shown). Quite pleasingly (Table 5), the native state RNaseT1 matrix (Figure 2) was also satisfactorily predicted by the DF (H_2O) analysis with both normalizations and fairly well by the VCD analysis

Table 6: Neural Network Predictions of Matrix Descriptors for RNaseT from DF FTIR Spectra in H_2O at Various Temperatures

	reference normalization		intensity-enhanced normalization	
	calculated	corrected	calculated	corrected
20−50 °C	$\begin{vmatrix}1&0&1\\0&7&7\\1&7&9\end{vmatrix}$	$\begin{vmatrix}1&0&1\\0&7&7\\1&7&9\end{vmatrix}$	$\begin{vmatrix}1&0&1\\0&7&7\\1&7&9\end{vmatrix}$	$\begin{vmatrix}1&0&1\\0&7&7\\1&7&9\end{vmatrix}$
55−60 °C	$\begin{vmatrix}0&0&0^a\\1&8&7\\0&8&9\end{vmatrix}$	$\begin{vmatrix}0&0&0\\0&8&8\\0&8&9\end{vmatrix}$	$\begin{vmatrix}0&0&0\\0&8&8\\0&8&9\end{vmatrix}$	$\begin{vmatrix}0&0&0\\0&8&8\\0&8&9\end{vmatrix}$

[a] Typical error; all calculated matrices can be transformed into a corrected result.

(with an effective error of only one sheet segment).[6] The DF (D_2O) matrix prediction using data for RNaseT1 which was H−D pre-exchanged at high temperature (bottom line, Table 5) was incorrect; however, that for our RNaseT1 sample dissolved directly in D_2O gave very good agreement, just as was found with the H_2O data.

The entries in the tables marked calculated are the matrix descriptors that resulted from back-scaling (with the RNaseA norm) the neural network predicted normalized matrices. This answer contains all the errors of our method that survive conversion to integer form, a process which does effect some smoothing of the prediction variations. Those entries marked corrected are the closest form of the [HEC] matrix whose elements obey the sum rules (Pancoska et al., 1994, 1997). As a constraint, the corrections were made to affect the minimal number of matrix elements predicted by the network from experimental data.

Results of predictions for the 3 × 3 helix−sheet−other [HEC] matrix descriptors of the connectivity of secondary structure segments in RNaseT1 from DF (H_2O) spectra at various temperatures are summarized in Table 6. For the RNaseT1 H_2O DF spectra up to 50 °C, the intensity-enhanced analysis yields the same [HEC] matrix as for the native state. At 55 °C and higher, the helical segment is predicted to disappear and the number of sheet segments to increase by one. As for the FC predictions, the number of other segments remains constant throughout this temperature range. The matrices predicted with the reference normalization are very similar, differing only for the 55 and 60 °C spectra and then only by one or two segments in the off-diagonals.

On the contrary, for the RNaseT1 D_2O DF spectra, the matrix descriptor analysis does not do so well for either normalization. This calculation predicts four helical segments for the native state (Table 5) which is then reduced by one as temperatures increase to 55 °C. However, this value returns back to four with a decrease in the number of sheet segments at even higher temperatures. While it is a very good experimental method for obtaining consistent data, high-temperature H−D pre-exchange creates spectra for which the network is not properly trained. This is confirmed by our success in using the same network to correctly predict the native form of the [HEC] matrix of RNaseT1 from the room-temperature, native state DF (D_2O) spectrum of RNaseT1, but without prior H−D exchange (Table 5).

[6] It might be noted that the N and C termini are both defined as C segments in our scheme which means that there is an extra C segment as compared to the sum rule based on segment interconnections (Pancoska et al., 1997). Our presentations here are corrected for this definitional restriction.

This consistency between methods (FTIR and VCD) for native state proteins is a striking strength of the new descriptor and our design of a network to predict it. However, as the sample is heated, the VCD-derived matrix deviates from the DF (H_2O) result and actually shows an increase in the number of helix segments. This deviation from the expected pattern is undoubtedly also affected by the change in deuteration as the sample is heated. The VCD sample was not pre-exchanged at elevated temperatures, as was the DF (D_2O) sample, but certainly becomes more thoroughly exchanged as the spectra are collected. This explains the parallel difficulties for matrix prediction at high temperatures with the VCD (D_2O) and DF (D_2O) data sets. Since these two D_2O-based data sets have a systematic error, their matrix descriptors are not presented.

DISCUSSION

Native State Structures. We have shown that a new descriptor for protein secondary structure is recoverable using neural network analyses of optical spectra. For native state RNaseT1, conventional FC values (Table 1) and the new matrix descriptors (Table 5) were predicted within reasonable error from both FTIR and VCD spectra. The matrix descriptor can be predicted with about the same accuracy as can the conventional FC descriptor, thus offering an enhancement in obtaining new added structural detail at a comparable level of accuracy.

One aspect of this matrix descriptor is that it only counts segments of secondary structure and their interconnections but does not restrict their order, other than to account for the interconnections. Thus, added information is needed to place the segments on the primary sequence. Coupling the matrix and FC predictors can provide a means of converting integer [HEC] matrix elements into information about the average lengths of those segments. Another source of information can come from sequence-based secondary structure prediction algorithms which can be used to locate these segments on the sequence, but now with direct experimental input about the protein structure in question.

Since RNaseT1 is predicted to be ~20% helical with only one segment, we know that this must be a segment of ~20 residues. On the other hand, the sheet segments must be relatively short for the 20−30% sheet fraction predicted to be distributed over seven segments. If one couples this insight into average segment length with a structure prediction algorithm based on primary sequence data, it is clear that locating the helix will require identification of a substantial stretch of relatively high helical propensity. This will give us new means of identifying and experimentally constraining predictions of secondary structure components in globular proteins.

To suggest possible placements for the helical and sheet segments, as shown in Figure 8, three different primary sequence structure prediction algorithms were utilized. They encompass three different strategies that these types of calculations currently use. The MSI-Biosym implementation of the original Chou−Fasman (1989) algorithm (C−F) represents a method which uses single amino acid residue propensities to form helix, sheet, and other secondary structure. The Holly−Karplus (1989) method (H-K) represents a neural network-based algorithm, in which the

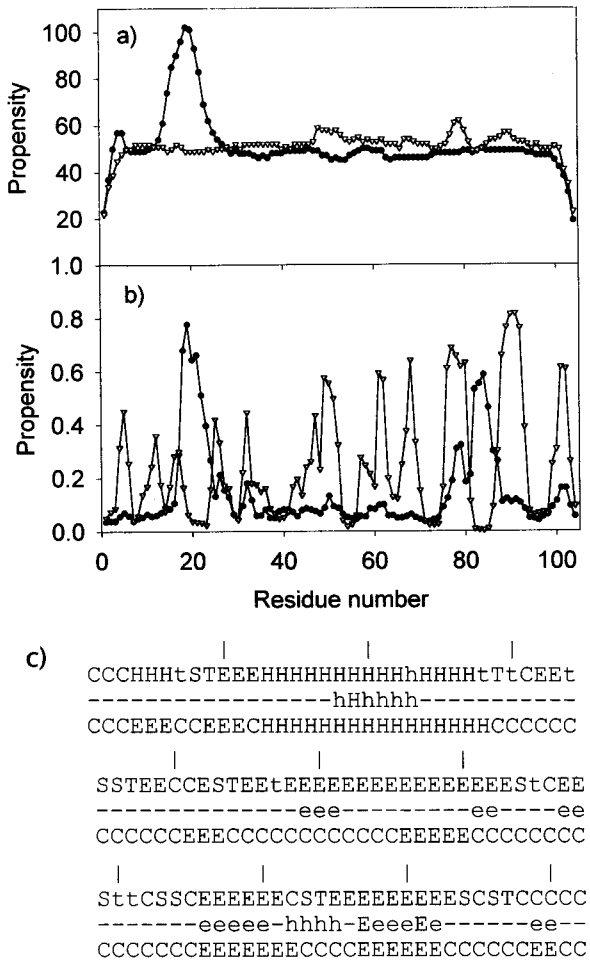

FIGURE 8: Propensities for helix (circles) and sheet (triangles) secondary structure derived from the primary structure of RNaseT1 using (a) the C-F method and (b) the H-K method. (c) Summary of the H-K and C-F predictions compared to the SOPMA (in this case the same as the X-ray) result [helix, H and h; sheet, E and e; other, C and −; and turn (H-K prediction only), T and *t*.

information for secondary structure assignment of any amino acid residue in the sequence is taken from a "window" of 16 neighboring residues. Finally, the SOPMA method of Geourjon and Delegae (1994, 1995) utilizes multiple alignments of protein sequences for secondary structure prediction. Both the C-F and H-K methods suggest a helical segment should occur from about residue 15 to 25. While this is short compared to our predicted length of ~20 residues, it is centered right over the helix seen in the RNaseT1 crystal structure (positions 13−29 according to the KS assignment) and furthermore is favored over any other possibilities due to its dominant propensity plus length. The SOPMA method predicts the longer helical segment (13−28), but that is to be expected since the method found a match with the RNaseT1 structure in the data base. The alternate helix position indicated in the H-K prediction is disfavored by our matrix prediction result indicating that just one long helical segment occurs. The discrepancy between the helical segment length predicted from the sequence with the first two algorithms and that indicated from optical spectra and the crystal structure may be due to added stabilization of long helical structures which is sometimes not included in the conventional propensity methods. Our experimentally determined matrix descriptor, indicating the number of segments, aids in sorting out such potential structures.

Further improvement of the structural interpretation of RNaseT1 from our data would require location of the sheet segments. Due to their higher number in RNaseT1 and to the more continuous variability of typical sheet segment lengths as found in globular proteins, their predicted placements are much less clear. Both sequence predictive methods would suggest the possibility of two short segments on the N terminal side of the helix if that segment started after residue 15; otherwise, only one would be predicted there since the matrix prediction indicates a coil segment should both lead and follow the helix. Experimentally, two sheet segments are found there, but the helix starts at residue 13 (Figure 2). The C-F predictions for sheet segments following the helix are less useful, but the H-K prediction does have a series of short segments of sheet structure predicted later in the sequence which in some sense parallels our spectral prediction of short average length for sheet segments. Three of these align well with the crystal structure sheet segments.

In summary, by determination of the matrix descriptor, our native state picture of the RNaseT1 secondary structure as obtained from optical spectra has substantially improved over that given by just FC values. The above predictions of segment layout qualitatively fit the patterns of a single helix stabilized by an "environment" of a series of short sheet segments as seen in Figure 2 for the RNaseT1 crystal structure.

Temperature Dependent Spectral Analysis. For interpretation of the RNaseT1 temperature dependent data, it is important that the quantitative analyses yield numbers that are sensible with regard to the qualitative interpretations of the various spectral types since these non-native states probably have some residual relationship to the native structure. It is this continuous change from the native state (known structure) to the high-temperature states (unknown structures) that give this study a reference point. We have found analyses that work as well as ones that do not. Their comparison yields information about the best way to approach such studies in the future.

Overall, our spectral data (particularly the ECD and VCD) qualitatively indicate a loss of helical character at higher temperatures as enumerated above (Results). The relative sharpness of the temperature dependence and consistency with a possible two-state transition is dependent on concentration. Qualitative evaluation of the DF (FTIR, H_2O) plots (Figure 4) is more difficult, but the positive increase in intensity both above and below the main α-helical contribution is consistent with a loss of helix (well-defined, narrow absorbance) and gain of disordered (hence, broad absorbance) structure. None of these techniques results in final high-temperature spectra consistent with a random coil polypeptide, but all imply that some level of structure remains, though qualitative analysis of our spectra does not uniquely indicate its nature and has difficulty in determining zero contribution from a given structural type. Previous IR analyses (Fabian et al., 1993) have attributed this residual to turns with a full loss of sheet contribution, but those were based on frequency assignments which are tenuous at best for such variable structures (Pancoska et al., 1993; Surewicz et al., 1993; Jackson & Mantsch, 1995). We and others (Dukor & Keiderling, 1991; Tiffany & Krimm, 1968) have shown that left-handed helical segments, which can be as short as a turn, do contribute to the ECD and VCD in the same manner as do coils. Furthermore, short, twisted extended (sheetlike) segments could contribute to at least the VCD and IR spectra in a similar manner as that ascribed to coils (Pancoska et al., 1989; Baumruk & Keiderling, 1993). Thus, these ambiguities of interpretation imply the qualitative analyses of the RNaseT1 high-temperature spectra are consistent with there being residual, short extended segments as well as turns.

Turning to the quantitative analyses, in both the DF (H_2O and D_2O) and VCD (D_2O) analyses of the conventional FC values, we see the same loss of helical structure trends with temperature. By contrast, the sheet component in the various analyses stayed remarkably constant for the H_2O study but did change more for the D_2O-based ones which are less well referenced. The bend, turn, and other fractions were also relatively constant.

Surprisingly, quantitative analyses of the ECD data did not show the large change in helix as might have been expected. It seems that the dependence of these ECD-based predictions on the coefficient of just one subspectrum that has a shape much like that of a typical α-helix ECD (Pancoska et al., 1995) leads the quantitative scheme to exhibit an insensitivity to structures of low helical content. This could be exacerbated by the rather unusual positive ellipticity at 233 nm found for the native state RNaseT1 which may arise from aromatic contributions. The near constant ellipticity at 222 nm resulting with increased temperature may be the focus of this insensitivity and the source of problems with earlier ECD-based analyses of the RNaseT1 unfolding (Oobatake et al., 1979). This RNaseT1 temperature study provides another example of the benefits of using multiple techniques to monitor structural change, as analysis of any one could have artifacts.

Coupling ECD and VCD in a combined quantitative analysis, which we have previously shown to enhance accuracy (Pancoska et al., 1995), leads to prediction of a more gradual change of the helix contribution and an increase in the other component but also underestimates the sheet and overestimates the helix fractions for the native state (results not shown). These coupled results for the higher temperatures should not be taken too literally since our qualitative observations indicate that the thermal processes are affected by the protein concentration to the point where the ECD and VCD data show different transition temperatures.

The above addresses the prediction of conventional (FC) secondary structure descriptors from the temperature dependent spectra. The point of this paper is that we now can also compute the new matrix descriptor and, for the DF (H_2O) temperature dependence data, obtain a result that is consistent with the qualitative spectral interpretation as well as the quantitative FC value determinations. The helical segment is predicted to disappear at 55 °C, and the matrices predicted at higher temperatures suggest an added sheet segment appears. Though an integer change, this is significant in terms of the form of the predicted matrices, the main structural change being the loss of the main helical fragment in RNaseT1.

Because we consider this work to be an "introductory" study of the application of the matrix descriptor, we will here go beyond this primary conclusion of the NN analysis and also comment on some other, less certain details, notably the apparent increase in the number of sheet segments at high temperatures. This may actually be a constraint of the matrix model which, when there are no helical segments,

becomes a 2 × 2 matrix which has only one independent variable; i.e. by definition of the matrix descriptor form, all elements must be equal. The training of the NN naturally recognizes the fundamental sum rule character of all the matrix descriptors used. In other words, by loss of a helical segment, the sum rules would force loss of another segment or the gain of a sheet segment and appropriate interconnection elements. During its training, the NN encoded information into its weights that make them valid for discrimination between helix, sheet, and other segments as they are represented in the training set of native protein structures. Thermal denaturation might produce structural segments which are not typical of those in the reference data set. Nevertheless, to generate its output, the network will use the similarity of all the spectral features to the spectra in the reference set. We can independently quantify the similarity between the spectra by the elements of the correlation matrices, calculated as part of the FA RMR analysis. Alternatively, cluster analysis can be used (Pancoska et al., 1991) for this purpose. By these independent methods, the high-temperature spectra of RNaseT1 were found to be most similar to spectra of those reference set proteins which have significant fractions (19−32%) of sheet. The network therefore quantifies what is seen qualitatively; i.e. at higher temperatures, the spectra of RNaseT1 lose much of their helical feature but remain similar to spectra of proteins which contain a distribution of sheet, turns, bends, and other conformations.

Comparison with Previous Studies. Previous studies of the thermal denaturation using frequency assignment of Fourier self-deconvolved (FSD) FTIR absorption spectra (Fabian et al., 1993, 1994) were interpreted in terms of substantial loss of helix *and* sheet fraction. While our quantitative and qualitative analyses support the loss of helix, our quantitative analyses of the data all agree that the loss of sheet is not so significant as previously suggested. This contrast in the previous FSD−FTIR analysis and the present band shape-based FC and matrix descriptor analyses of the same data set suggests two interpretations. One might assume that, despite all of their successes with native state structural prediction, our methods could have some hidden difficulty. One potential difficulty of a training set-based prediction method is the correlation between α-helix and β-sheet FC values that we have previously found in a neural network analysis of the proteins in the data bank (Pancoska et al., 1992). However, in our recent parallel analysis, the matrix descriptor does not show a correlation between the number of helical and sheet segments in this representation of protein structure (Pancoska et al., 1997). Since FC and matrix analyses are yielding systematic and consistent results for the temperature denaturation of RNaseT1, this source of error must not be dominant.

A second inherent difficulty of all such empirical studies is that they are not referenced to a training set containing disordered structures. In practice, it is impossible to have such a training set. There are no unstructured proteins for which we know the structure (i.e. this would be self-contradictory) and upon which one could base a regresision analysis. Even denatured proteins (e.g. in high concentrations of urea) are usually inappropriate since the denaturants affect the spectra. Thermal denaturation seems attractive, but from the shape of the spectra presented here, it is clear that substantial "structure" can remain. We feel that one must be resigned to the limitations of secondary structure prediction methods and must be aware of those limitations in applying such techniques to folding studies. Therefore, one must resort to the alternate logical approach of using these perhaps limited techniques and evaluating the information they provide. The systematic, continuous evolution of spectra and structure with temperature in the RNaseT1 case suggests this data set to be an ideal medium for a first test of probing the residual structural elements in a denatured protein.

The contrast between our band shape-based FTIR analyses and the previous frequency assignment-based FSD−FTIR analyses can be in fact meaningfully interpreted. Both approaches have established themselves as capable of characterizing protein secondary structure. Yet, due to the differences in their physical bases, these interpretive methods sense the structure differently and lead to different conclusions. The key to a possible interpretation is that the number of segments (sheet or other) did not drop after the transition. Thus, the high-temperature structure is implied to consist of eight sheet segments which on average must be quite short (three or four residues). These segments have lost one of their dominant organizational components, interaction with the long helical segment. Thus, these residual sheet segments in the high-temperature structure should be more distorted and probably of shorter length than in the native state.

If these sheet segments exist, why do they not contribute to the FSD−FTIR? If they are short and distorted, their frequency dispersion becomes much greater. Since the FSD method is most sensitive to sharp spectral components (Surewicz et al., 1993; Jackson & Mantsch, 1995), it is possible that their FSD−FTIR contribution is greatly diminished. As an example of the sensitivity limits of FSD−FTIR, in a recent study of calmodulin, the 8% β-sheet component in two β-sheets proved hard to detect (Fabian et al., 1996). The band shape methods applied here to FTIR data are also subject to frequency shifts, but this is generally a higher-order effect. The VCD and ECD methods are even less sensitive to these effects since their band shapes arise from the relative geometries of interacting dipoles, in this case those on the amide functions, rather than the frequency dispersion of dipole strength.

A natural extension of our matrix descriptor would be to represent the interconnections of segments by assigning one or two "head-and-tail" residues in each segment as the off-diagonal terms. If applied to the RNaseT1 case, due to the short overall length of the sheet segments, many of the sheet segments would actually disappear from the diagonal after eliminating "end residues", and consequently, the quantitative band shape analyses and frequency assignment analyses would come even closer together. This "zero-penetration" [HEC] matrix form was adopted in this paper for simplicity as an initial test of the new method. Future study will be directed at attempts to optimize the matrix form (Janota, 1996).

Deuteration Effects. Comparison of H_2O- and D_2O-based FTIR spectral analyses for RNaseT1 at higher temperatures clearly indicates that the degree of H−D exchange is the culprit in the failed DF (D_2O) analyses. The NN parameters must be obtained from a training set of data appropriate to the unknown being studied. For D_2O-based studies, all the proteins in our training set were exchanged to an equilibrium condition for the native state (Pancoska et al., 1989, 1991);

however, the temperature variation DF (D_2O) analysis was done with fully exchanged proteins (Fabian et al., 1993), and the level of exchange in the VCD study changed as the protein unfolds, resulting in poorly predicted [HEC] matrices. The higher-temperature RNaseT1 states are unfolded to a greater extent than are any of the training set proteins which could also encompass structural elements not represented in the training set. If aggregation occurs for the higher-temperature VCD studies, due to the longer scan times, these structures would also not be well-represented in the training set. Thus, the variances between our matrix descriptor analyses can be understood, if not overcome.

This sensitivity of the matrix descriptor prediction algorithm to the degree of H−D exchange (even for native states) is a new development found in this study. Its source may arise from the NN using the entire spectrum, while the FC determination uses just the simplified representation made available by FA of the input data. Alternatively, it may be an intrinsic property of the higher-level structural sensitivity of the matrix descriptor. Certainly, deuterated segments are known to have different vibrational spectroscopic responses than do protonated ones. We would expect these to impact the analysis at some level. While this H−D exchange sensitivity can be viewed as a limitation of the method, at the same time, it suggests further development of the technique. Since different segments are likely to be protected or exposed to H−D exchange, matrix descriptor analyses of spectra from variably deuterated samples may eventually be able to take advantage of this sensitivity to provide more site-specific interpretations of secondary structure from optical spectral data. Such H−D exchange studies for some small globular proteins have been shown to yield useful structural information using FTIR-based analyses (Backmann et al., 1996).

Summary. Our qualitative and quantitative analyses of all the optical spectra (FTIR, ECD, and VCD) for the thermal denaturation of RNaseT1 lead to the conclusion that the 60 °C temperature state is one with significant remaining elements of secondary structure, but with most of its helical component lost. In particular, the new matrix descriptor analysis no longer recognizes a helical segment or any interconnections of helices to other segments. By contrast, the sheet component is predicted to continue contributing to the structure in terms of FC value as well as segment number. The result is consistent with ECD and VCD spectra interpretations. This band shape-based analysis result conflicts with previous interpretations of the FSD−FTIR spectra which could have been reduced in sensitivity due to the segment length effects. While the methods used may formally lack a basis for application to denatured proteins, the consistency of information derived from various techniques suggests a possible interpretation that the high-temperature state still maintains a number of very short, presumably dynamically distorted segments having dihedral angles in the β or extended region of conformational space. Our multiple-technique analyses suggest that what is analyzed as sheet with the FC methods is most probably conformationally quite different from idealized sheet segments, due to end effect perturbations. Clearly, additional studies on other protein systems will be required to establish the generality of these interpretations and the reliability of applying such analytical techniques to denatured proteins.

Without the added information we have derived from analysis of this novel matrix descriptor, it could not be possible to suggest an explanation for the apparent discrepancy between the previous FSD analysis and the present band shape-based results. The consistency of all of our analyses suggests that, while the residual high-temperature structure may certainly involve turns, to explain the spectra it should also have short segments whose ϕ,ψ angles are similar to those characteristic of extended strands. Since this high-temperature state has been found to have IR spectral properties similar to those obtained for this protein and that of RNaseA under denaturing conditions (Fabian & Mantsch, 1995), it is important to study this system further to determine what aspects of the protein structure are actually preserved with each type of denaturation and what types of structure can be detected with each technique. The differentiation of denatured states obtained under alternate conditions is becoming more broadly recognized. It is however possible that this thermally induced intermediate is only kinetically stable and that with time would unfold more and aggregate to an irreverisbly denatured form. That would certainly be consistent with the results of our VCD analyses. Finally, it is important to realize that while the bulk of this matrix descriptor interpretation of the structure is dependent on the DF (H_2O) data, which are systematically referenced to the training set, the interpretations of all the spectra are qualitatively consistent.

ACKNOWLEDGMENT

We are grateful to Ulrich Hahn, Universitat Leipzig, for the gift of the RNaseT1 protein. One of us (H.F.) thanks Henry H. Mantsch, Institute of Biodiagnostics, Winnipeg, for the opportunity to record FTIR spectra of RNaseT1 there.

SUPPORTING INFORMATION AVAILABLE

FTIR spectra of 23 proteins, in digital form, are available. Access information is given on any current masthead page.

REFERENCES

Andrade, M. A., Chacon, P., Merelo, J. J., & Moran, F. (1993) *Protein Eng. 6*, 383−390.
Backmann, J., Schultz, C., Fabian, H., Hahn, U., Saenger, W., & Nauman, D. (1996) *Proteins: Struct., Funct., Genet. 24*, 379−387.
Baumruk, V., & Keiderling, T. A. (1993) *J. Am. Chem. Soc. 115*, 6939−6942.
Baumruk, V., Pancoska, P., & Keiderling, T. A. (1996) *J. Mol. Biol. 259*, 774−790.
Bitto, E. (1993) Diploma Thesis, Charles University, Prague.
Chang, C. T., Wu, C. S. C., & Yang, J. T. (1978) *Anal. Biochem. 91*, 13−31.
Chou, P. Y., & Fasman, G. D. (1978) *Methods Enzymol. 47*, 45−148.
Compton, L. A., & Johnson, W. C., Jr. (1986) *Anal. Biochem. 155*, 155−167.
Dousseau, F., & Pezolet, M. (1990) *Biochemistry 29*, 8771−8779.
Dukor, R. K., & Keiderling, T. A. (1991) *Biopolymers 31*, 1747−1761.
Dukor, R. K., Pancoska, P., Prestrelski, S., Arakawa, T., & Keiderling, T. A. (1992) *Arch. Biochem. Biophys. 298*, 678−681.
Fabian, H., & Mantsch, H. H. (1995) *Biochemistry 34*, 13651−13655.
Fabian, H., Schultz, C., Naumann, D., Landt, O., Hahn, U., & Saenger, W. (1993) *J. Mol. Biol. 232*, 967−981.

Fabian, H., Schultz, C., Backmann, J., Hahn, U., Saenger, W., Mantsch, H. H., & Naumann, D. (1994) *Biochemisty 33*, 10725−10730.
Fabian, H., Yuan, T., Vogel, H. J., & Mantsch, H. H. (1996) *Eur. Biophys. J. 24*, 195−201.
Fasman, G. D. (1989) in *Prediction of Protein Structure and the Principles of Protein Conformation*, Plenum, New York.
Geourjon, C., & Delegae, G. (1994) *Protein Eng. 7*, 157−164.
Geourjon, C., & Delegae, G. (1995) *Comput. Appl. Biosci. 11*, 681−684.
Hennessey, J. P., Jr., & Johnson, W. C., Jr. (1981) *Biochemistry 20*, 1085−1094.
Holly, L. H., & Karplus, M. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 152−156.
Jackson, M., & Mantsch, H. H. (1995) *Crit. Rev. Biochem. Mol. Biol. 30*, 95−120.
Janota, V. (1993) Diploma Thesis, Charles University, Prague.
Janota, V. (1996) Ph.D. Thesis, Charles University, Prague.
Johnson, W. C., Jr. (1985) *Methods Biochem. Anal. 31*, 61−163.
Johnson, W. C., Jr. (1988) *Annu. Rev. Biophys. Chem. 17*, 145−166.
Johnson, W. C., Jr. (1990) *Proteins: Struct., Funct., Genet. 7*, 205−214.
Kabsch, W., & Sander, C. (1983) *Biopolymers 22*, 2577−2637.
Keiderling, T. A. (1981) *Appl. Spectrosc. Rev. 17*, 189−226.
Keiderling, T. A. (1990) in *Practical Fourier Transform Infrared Spectroscopy. Industrial and Laboratory Chemical Analyses* (Ferraro, J. R., & Krishnan, K., Eds.) pp 203−284, Academic, San Diego.
Keiderling, T. A. (1996) in *Circular Dichroism and the Conformation of Biomolecules* (Fasman, G. D., Ed.) pp 555−598, Plenum, New York.
Keiderling, T. A., & Pancoska, P. (1993) in *Biomolecular Spectroscopy Part B* (Clark, R. J. H., & Hester, R. E., Eds.) pp 267−315, Wiley, Chichester.
Keiderling, T. A., Wang, B., Urbanova, M., Pancoska, P., & Dukor, R. K. (1994) *Faraday Discuss. 99*, 263−285.
Landt, O., Zirpel-Giesebrecht, M., Milde, A., & Hahn, U. (1992) *J. Biotechnol. 24*, 189−194.
Lee, D. C., Haris, P. I., Chapman, D., & Mitchell, R. C. (1990) *Biochemistry 29*, 9185−9193.
Manavalan, P., & Johnson, W. C., Jr. (1987) *Anal. Biochem. 167*, 76−85.
Mantsch, H. H., Casal, H. L., & Jones, R. N. (1986) in *Spectroscopy* (Clark, R. J. H., & Hester, R. E., Eds.) Vol. 13, pp 1−46, Wiley & Sons, London.
Merelo, J. J., Andrade, M. A., Prieto, A., & Moran, F. (1994) *Neurocomputing 6*, 443−454.
NeuralWare (1993) in *Neural Computing. A technology Handbook for Professional II/PLUS and NeuralWorks Explorer*, pp 63−84, NeuralWare, Inc., Technical Publications Group, Pittsburgh, PA.
Oobatake, M., Takahashi, S., & Ooi, T. (1979) *J. Biochem. (Tokyo) 86*, 89−98.
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.
Pancoska, P., Blazek, M., & Keiderling, T. A. (1992) *Biochemistry 31*, 10250−10257.
Pancoska, P., Wang, L., & Keiderling, T. A. (1993) *Protein Sci. 2*, 411−419.
Pancoska, P., Bitto, E., Janota, V., & Keiderling, T. A. (1994) *Faraday Discuss. 99*, 287−310.
Pancoska, P., Bitto, E., Janota, V., Urbanova, M., Gupta, V. P., & Keiderling, T. A. (1995) *Protein Sci. 4*, 1384−1401.
Pancoska, P., Janota, V., & Keiderling, T. A. (1966) *Appl. Spectrosc. 50*, 658−668.
Pancoska, P., Janota, V., Nesetril, J., & Keiderling, T. A. (1997) *Protein Sci.* (submitted for publication).
Perczel, A., Hollosi, M., Tusnady, G., & Fasman, G. D. (1991) *Protein Eng. 4*, 669−679.
Pribic, R., van Stokkum, I. H. M., Chapman, D., Haris, P. T., & Bloemendal, M. (1993) *Anal. Biochem. 214*, 366−378.
Provencher, S. W., & Glöckner, J. (1981) *Biochemistry 20*, 33−37.
Rahmelow, K., Hubner, W., & Ackermann, T. (1993) in *Fifth International Conference on the Spectroscopy of Biological Molecules* (Theophanides, T., Ed.) pp 139−140, Kluwer Academic Publishers, Dordrecht.
Sarver, R. W., & Krueger, W. C. (1991a) *Anal. Biochem. 194*, 89−100.
Sarver, R. W., & Krueger, W. C. (1991b) *Anal. Biochem. 199*, 61−67.
Sauer, K., Ed. (1994) Biochemical Spectroscopy, in *Methods in Enzymolgoy*, Vol. 246, Academic Press, San Diego.
Seigel, J. B., Steinmetz, W. E., & Long, G. L. (1980) *Anal. Biochem. 104*, 160−167.
Sreerama, N., & Woody, R. W. (1993) *Anal. Biochem. 209*, 32−44.
Sreerama, N., & Woody, R. W. (1994) *J. Mol. Biol. 242*, 497−507.
Srinivasan, R., & Rose, G. D. (1995) *Proteins: Struct., Funct., Genet. 22*, 81−99.
Surewicz, W., Mantsch, H. H., & Chapman, D. (1993) *Biochemistry 32*, 389−394.
Tiffany, L., & Krimm, S. (1968) *Biopolymers 6*, 1767−1770.
Toumadje, A., Alcorn, S. W., & Johnson, W. C., Jr. (1992) *Anal. Biochem. 200*, 321−331.
Urbanova, M., Pancoska, P., & Keiderling, T. A. (1993) *Biochim. Biophys. Acta 1203*, 290−294.
van Stokkum, I. H. M., Spoelder, H. J. W., Bloemendal, M., van Grundelle, R., & Groen, F. C. A. (1990) *Anal. Biochem. 191*, 110−118.
Venyaminov, S. Yu., & Kalnin, N. N. (1990a) *Biopolymers 30*, 1259−1271.
Venyaminov, S. Yu., & Kalnin, N. N. (1990b) *Biopolymers 30*, 1273−1280.
Venyaminov, S. Yu., & Yang, J. T. (1996) in *Circular Dichroism and the Conformation of Biomolecules* (Fasman, G. D., Ed.) Plenum, New York.
Venyaminov, S. Yu., Braddock, W. D., & Prendergast, F. G. (1996) *Biophys. J. 70*, A65.
Wang, L. (1993) Ph.D. Thesis, University of Illinois at Chicago, Chicago.
Yang, J. T., Wu, C.-S. C., & Martinez, H. M. (1986) *Methods Enzymol. 130*, 208−269.
Yasui, S. C., Pancoska, P., Dukor, R. K., Keiderling, T. A., Renugopalakrishnan, V., Glimcher, M. J., & Clark, R. C. (1990) *J. Biol. Chem. 265*, 3780−3793.
Yue, K., & Dill, K. A. (1996) *Protein Sci. 5*, 254−261.

BI961178U