Matheson, I. B. C., Lee, J., & Muller, F. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 948-952.
Mitchell, G., & Hastings, J. W. (1969) *J. Biol. Chem. 244*, 2572-2576.
O'Connor, D. V., & Phillips, D. (1984) *Time Correlated Single Photon Counting*, Academic Press, London.
O'Kane, D. J., & Lee, J. (1985a) *Biochemistry 24*, 1467-1475.
O'Kane, D. J., & Lee, J. (1985b) *Biochemistry 24*, 1484-1488.
O'Kane, D. J., Ahmad, M., Matheson, I. B. C., & Lee, J. (1986) *Methods Enzymol. 133*, 109-128.
O'Kane, D. J., & Lee, J. (1986) *Methods Enzymol 133*, 149-172.
O'Kane, D. J., Karle, V. A., & Lee, J. (1985) *Biochemistry 24*, 1461-1467.
Small, E. D., Koka, P., & Lee, J. (1980) *J. Biol. Chem. 255*, 8804-8810.
Spencer, R. D., & Weber, G. (1970) *J. Chem. Phys. 52*, 1654-1663.
Tu, S.-C. (1979) *Biochemistry 18*, 5940-5945.
Tu, S.-C. (1982) *J. Biol. Chem. 257*, 3719.
Tu, S.-C. (1986) *Methods Enzymol. 133*, 128-139.
Ward, W. W. (1979) *Photochem. Photobiol. Rev. 4*, 1-57.
Vervoort, J., Muller, F., Lee, J., van der Berg, W. A. M., & Moonen, C. T. W. (1986) *Biochemistry 25*, 8062-8067.
Visser, A. J. W. G., & Lee, J. (1980) *Biochemistry 19*, 4366-4372.
Visser, A. J. W. G., & Lee, J. (1982) *Biochemistry 21*, 2218-2226.
Vos, K., Van Hoek, A., & Visser, A. J. W. G. (1987) *Eur. J. Biochem. 165*, 55-63.
Ziegler, M. M., & Baldwin, T. O. (1981) *Curr. Top. Bioenerg. 12*, 65-113.

How To Determine Protein Secondary Structure in Solution by Raman Spectroscopy: Practical Guide and Test Case DNase I

Bernd M. Bussian[‡] and Chris Sander[*,§]

Biocomputing Programme, EMBL, Meyerhofstrasse 1, D-6900 Heidelberg, West Germany, and Institut für Physikalische Chemie, Technische Hochschule, Petersenstrasse 20, D-6100 Darmstadt, West Germany

Received August 9, 1988; Revised Manuscript Received December 1, 1988

ABSTRACT: For many proteins available in large (milligram) quantities, a three-dimensional structure determination by X-ray or NMR methods is very difficult, impossible, or too costly. In these cases, spectroscopic determination of secondary structure content can be a valuable source of partial information about protein structure in solution. In particular, Raman spectroscopy can be used to determine to fair accuracy the helix and sheet content of a globular protein. However, technical difficulties have hampered the routine application of the method: (1) The large background signal of aqueous solvent in the amide I region is difficult to subtract accurately. (2) The reference data set of Raman spectra of proteins with known crystal structure is incomplete, and the assignment of secondary structure in a known crystal structure is not unambiguous. (3) The mathematical problem of extracting structure information from the spectra is ill determined; i.e., there are many apparently satisfactory solutions for a given spectrum. We have now partly solved and partly sidestepped these problems by improving and simplifying existing methods. Here, we give a step-by-step outline of a procedure intended for routine determination of the percentage of α-helix and β-sheet from the amide I Raman spectra of proteins in solution. Its main features are (a) an uncomplicated procedure for solvent subtraction, aided by use of a divided spinning cell technique, (b) a numerically stable data handling algorithm, and (c) a clear statement of expected accuracy. In our hands, using the reference spectra of Williams (1983), helix content can be determined to an accuracy of 6 percentage points (largest error 12%) and β-sheet content to an accuracy of 5 percentage points (largest error 7%). However, the experimental distinction between parallel and antiparallel β-sheet does not appear possible without a significant expansion of the set of reference proteins. As a test we have measured the Raman spectrum of DNase I, a known structure treated as unknown, and derive 14% α-helix and 22% β-sheet content, compared to X-ray derived values of 20% helix and 25% sheet (hydrogen bonds per 100 residues). The error, −6% for helix and −3% for sheet content, is typical. The method can be a tool for checking the structural purity of genetically engineered proteins, detecting major structural alterations of mutant proteins, and providing a priori information as input to predictions of protein structure.

In Raman spectroscopy, incident photons scatter inelastically off the sample. In the amide I band of proteins the energy difference between incident and scattered beam corresponds to vibrational modes around 1660 cm^{-1} involving a few atoms primarily in peptide units (Carey, 1982), with a major component of C–O bond stretching. When peptide units are involved in secondary structure H-bonds, the normal modes are perturbed and the spectral lines shifted by different amounts for the different types of secondary structure. This fact is exploited empirically to derive information about the amount of secondary structure of proteins in solution. The method was developed by Garfinkel and Edsall (1958), Miyazawa (1960), Krimm and Bandekar (1986 and earlier papers), and Peticolas et al. (1979). An implementation by Thomas and

[*] Author to whom correspondence should be addressed.
[‡] Institut für Physikalische Chemie.
[§] Biocomputing Programme.

0006-2960/89/0428-4271$01.50/0

Agard (1984), and similarly that by Byler and Susi (1988), extracts information from the spectrum by iterative Fourier deconvolution. Williams (1983, 1986), on the other hand, simply expresses any new spectrum as the superposition of a set of reference spectra. Williams recently applied his method to proteins of unknown X-ray structure, e.g., human leukocyte interferon (Williams, 1985) or sarcoplasmic reticulum calcium pump protein (in lipids) (Williams et al., 1986), and extended it to include the amide III region (Williams, 1986).

At present, further improvement of Williams' very useful method and its routine application are hindered by the following problems: (1) subtraction of solvent background is difficult because of the proximity of the *H_2O* bending mode at about 1645 cm^{-1} to the amide I region; (2) criteria for the localization of secondary structure states in crystal structures vary depending on the authors involved; (3) reference spectra vary in quality with some structural types underrepresented, such as proteins with mixed α/β topology, leading to difficulties in the interpretation of some spectra.

We present here improvement or clarification on each of these three aspects: (1) a new simplified subtraction procedure; (2) use of an automatic procedure for the definition of secondary structure; (3) use of a novel numerical procedure by Provencher and Glöckner (1981) and Provencher (1982a,b) and restriction of the analysis to only three types of secondary structure.

Here, our goal is to provide a practical guide to Raman amide I determination of protein secondary structure sufficiently useful for a nonspecialist with milligrams of soluble protein in hand and access to an appropriate spectrometer.

METHOD: HOW TO TAKE A RAMAN SPECTRUM OF A PROTEIN

Preparing the Protein Sample

One needs two samples, the protein sample containing protein and buffer and the background sample containing only the buffer. It is crucial to achieve identical conditions in the two samples so that the solvent contribution to the spectrum is identical in both.

One typically needs several milligrams of protein in aqueous solution. In a capillary of 10–50 μL or in a cuvette of about 100 μL this amounts to a protein concentration of 20–100 mg/mL. If necessary, increase concentration by removing excess solvent by complete lyophilization and redissolving in buffer. For further purification, equilibrate against the buffer by gel filtration, e.g., with 1 mL of Sephadex G-25/100 μL of protein solution, soaked overnight in buffer.

Whatever the precise details of protein purification, which may vary from case to case, it is essential to end up with a protein concentration of about 2–5% by weight dissolved in a buffer identical with the reference buffer solution. This can usually be achieved by dialyzing overnight.

Measuring the Spectrum

We restrict ourselves to proteins in solution (as distinct from protein in crystals). There are four aspects of the experimental procedure: (a) laser source; (b) sample and sample holder; (c) data recording equipment; (d) data collection protocol.

(*a*) *Laser Source and Monochromator*. The scattering geometry (Figure 1a) employs 90° viewing optics where the inelastically scattered light is collected along an axis perpendicular to the incident beam. The incident beam has the green (514.53 nm) or red line (413.13 nm) of an argon ion or krypton ion laser. The beam diameter is about 2 mm, focused down to about 15 μm by a glass lens of 50- or 60-mm focal length.

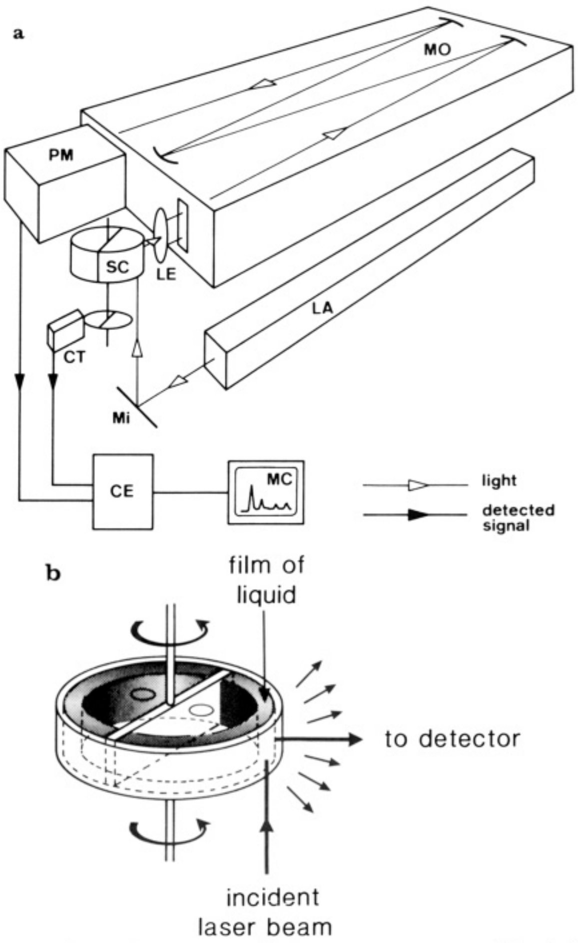

FIGURE 1: Scattering geometry of a Raman spectrometer with spinning divided cell. (a) Experimental setup of a Raman spectrometer. Abbreviations: LA, laser tube; Mi, mirror; SC, spinning divided cell; LE, focusing lens; MO, monochromator; PM, photomultiplier; CE, counting electronics; CT, channel trigger; MC, minicomputer. (b) Schematic of a spinning divided cell.

The laser power at the sample ranges from 100 mW up to 1 W; i.e., power density at the sample is of the order of 10^5 W/cm^2. As the spectral resolution required is about 2–5 cm^{-1}, an ordinary commercial monochromator, such as the SPEX 1403 double monochromator used by us, is adequate.

(*b*) *Sample and Sample Holder*. You may use a stationary cell, a simple spinning cell, or a partitioned spinning cell. The stationary cell is the simplest to handle but does not allow as high a laser intenity as a spinning cell. A spinning cell (Kiefer & Bernstein, 1971) continuously mixes the sample, rapidly dissipating heat and reducing the effect of optical perturbations by statistical averaging. A partitioned spinning cell (Figure 1b) has two identical sectors allowing simultaneous measurement of two solutions under identical conditions (optical path, temperature, degree of mixing, drift). This greatly simplifies the subtraction of the buffer spectrum from the protein plus buffer spectrum.

(*i*) *Stationary Cell*. Typically, a temperature-controlled brass block serves as sample holder in which either a capillary is mounted (at least 5 μL of solution) or a 1-cm path-length glass cuvette with planar windows is mounted (about 0.1–0.2 mL of solution). If desired, the temperature of the sample can be monitored with a small thermistor inserted directly into the capillary and positioned a few millimeters from the laser beam. Protein concentration is 5–10% by weight, so one needs on the order of 1 mg or more of protein. The protein may or may not be reusable after the experiment depending on the amount of thermal damage.

(*ii*) *Spinning Cell.* A spinning cell is a hollow glass cylinder about 60 mm or less in diameter and 5–10 mm tall (available from Hellma, Müllheim, Baden, West Germany). Fill only about 5% of the volume of the cell by adding about 0.5–0.8 mL of solution to each of the two chambers. As the cell spins at about 10 revolutions per second, the solution coats the wall while the laser beam passes through the cylinder tangentially. As a spinning cell allows a higher incident laser intensity, the protein concentration needed is only about 2–5% by weight. However, thermostating is not as easily done. Total amount of protein needed is about 10 mg, which is generally reusable after the experiment (shown by activity measurements before and after data collection, e.g., for adenylate kinase and DNase I; unpublished data).

The main advantage of the stationary cell technique appears to be that less protein material is required; that of the spinning cell technique, reduced damage and lower protein concentration, increasing the likelihood for being able to take the protein out of the preparation pathway and feeding it back after the spectroscopic experiment without alteration.

(*c*) *Data Recording Equipment.* The recording part comprises a commercial double monochromator, a photomultiplier, and real-time two channel counting electronics (figure 1a). Here we have used a SPEX 1403 double monochromator and an RCA 34031 photomultiplier which fed two ELSCINT real-time single channel photon counters. The monochromator and the two photon counters were controlled by a personal computer. For details see Eysel and Bussian (1985).

(*d*) *Data Collection Protocol.* The quantitative analysis of the amide I region requires data from 1500 to 1800 cm^{-1} at a resolution better than 5 cm^{-1}. The digitizing increment should be somewhat smaller. For example, a spacing of one wavenumber per data point allows subsequent five- to nine-point smoothing of the spectrum (Bussian & Härdle, 1984).

As the protein might be irreversibly denatured by the intense laser radiation, the total exposure time should be kept as low as possible, especially for a stationary cell. However, an initial exposure of, say, 10 min up to 1 h with gradually increasing laser intensity may be necessary in order to achieve photobleaching of flourescence, before data collection begins.

To minimize the effects of slow systematic drift or sudden discontinuities in the spectrometer or sample when a stationary cell is being used, take repeated scans sweeping rapidly over the 300 steps in the wavenumber interval, and then remove faulty scans and sum the remaining scans at the end to increase signal to noise ratio. A total of 10–20 scans with a scanning speed of 1 cm^{-1}/s gives good results (total data collection time on one data point of 10–20 s). For a spinning cell, sudden discontinuities are less likely, so that fewer scans may be used, with more time spent on each data point.

Finally, the data are passed to a digital computer for further analysis.

Reducing the Raw Spectral Data

(*a*) *Subtraction* of *Solvent Background.* Ideally, the buffer spectrum (*b*) is subtracted from the protein plus buffer spectrum (*pb*) such that the difference (*p*) represents that of the protein only. In practice, this is done by an empirically best choice of a scale factor α in

$$p_i = (pb)_i - \alpha b_i \qquad (1)$$

where *i* runs over all data points.

(*i*) *Single Cell.* α is chosen according to the following criteria. The difference spectrum should have a flat base line between 1740 and 1800 cm^{-1}, and the extrapolation of this base line must be consistent with the base line near 1500 and 1580 cm^{-1}. Alternatively, particular molecular groups with distinct and strong Raman lines can be used as an internal standard, if they are present in both solutions in identical solvent conditions. In that case, α is chosen such that these lines are eliminated fully in the difference spectrum (example: peak in HEPES buffer near 1050 cm^{-1}, Figure 4).

(*ii*) *Divided Cell.* Subtract directly the buffer spectrum from the protein spectrum; i.e., use $\alpha = 1.0$. For consistency, check that the resulting spectrum satisfies the same criteria (base line and/or internal standard) as for a single cell (see previous paragraph). If the criteria are not satisfied, adjust α until they are, taking into consideration the difference in water concentration between the two solutions and the different scattering intensity of protein and buffer.

(*b*) *Removal of Noise and Reduction of Data Points.* Smooth the spectrum to a resolution of 5 cm^{-1}, e.g., by five-point or seven-point polynomial smoothing on one-per-wavenumber data points. It may be necessary to remove bad data points (false very large or very small values). Here, a method of weighted means (Bussian & Härdle, 1984) with a "tuning parameter" of $\chi = 0.9$ for intensity values normalized to the interval 0.0–10.0 worked well. Report the smoothed protein spectrum at discrete points spaced 5 cm^{-1} apart and normalize.

Method: How To Extract Structural Information from the Spectrum

Simple Analysis of the Spectrum by a Visual Method

The vibrational amide I spectrum of a peptide unit at 1640–1780 cm^{-1} is dominated by the carbonyl stretching band which is influenced by the local hydrogen-bonding pattern. Prevalent hydrogen-bonding conformations in proteins are classified as α-helix and β-sheet, with the remainder called loop or coil. Here, the basic assumption is that proteins with similar percentage of secondary structure, regardless of the size of the protein, have similar amide I bands (Lord, 1977).

The simplest way to assess the secondary structure amount of one protein is to compare visually the profile of the amide I region with spectra of other proteins of known structure (Figure 2). An example—extracting structural information from the Raman spectrum of lysozyme—illustrates two ways of doing this.

(a) Compare the lysozyme spectrum with spectra of proteins of *pure* structure type, e.g., the all-α-helical proteins like myoglobin, hemoglobin, carp parvalbumin, and hemerythrin and the all-β-sheet proteins like concanavalin A, trypsin, chymotrypsin, and elastase (see Figure 2). Estimate the percentage contribution to the lysozyme spectrum from the α and β class, e.g., spectrum of lysozyme = 1/2(spectrum of myoglobin) + 1/2(spectrum of concanavalin A). This implies that lysozyme secondary structure is roughly half α-helix and β-sheet.

(b) Compare lysozyme with proteins from all structure classes and pick the most similar spectrum (Figure 2). Say, the spectrum of lysozyme is approximately equal to that of carboxypeptidase. Thus the estimate of percentage of α-helix and of β-sheet in lysozyme is almost identical with the percentage of α-helix and of β-sheet in carboxypeptidase.

The visual comparison requires an intuitive notion of similarity of two curves. This intuition can be quantified by using the sum of squared deviations over all data points as a measure of dissimilarity.

Mathematical Analysis of the Spectrum

There is a basic assumption in the following mathematical analysis of the spectrum. If the spectrum of a protein can be

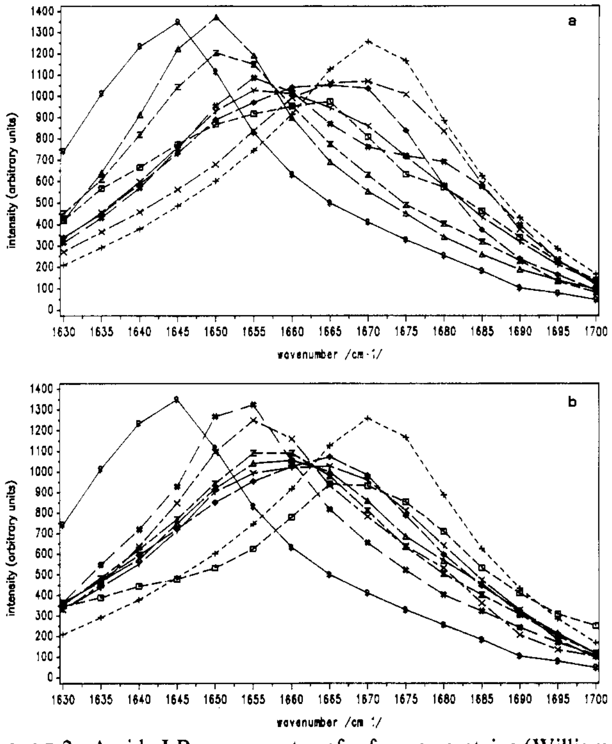

FIGURE 2: Amide I Raman spectra of reference proteins (Williams, 1983) with varying secondary structure content. The notation is according to the Brookhaven Protein Data Bank (if deposited): PLL, polylysine; CNA, concanavalin A; INS, insulin; FD, fd-phage; MLT, mellitin; TLN, thermolysin; TPI, triosephosphate isomerase; CRN, crambin; PTI, pancreatic trypsin inhibitor (bovine); LYZ, lysozyme (hen egg); ICB, calcium binding protein; MHR, hemerythrin; LDH, lactate dehydrogenase; CPA, carboxypeptidase; ADH, alcohol dehydrogenase; AVN, avidin. Spectra are digitized at 1-cm^{-1} intervals. The spectra of pure α-helical proteins peak at around 1645 cm^{-1}; those of pure β-type proteins, at around 1670 cm^{-1}. Proteins of mixed α–β-type have broader spectra with an intermediate peak. Proteins are divided into two groups for visual clarity only. The amide I spectrum of a protein of unknown structure can be visually compared with the reference spectra in (a) or (b) for a rough estimate of secondary structure type. (a) First group of nine reference proteins: PLL (♀); CNA (+); CRN (□); TIM (◇); FD (△); INS (#); MLT (Z); PTI (X); TLN (Y). (b) Second group of an additional seven reference proteins plus polylysine (PLL) and concanavalin A (CNA) from the first group, for orientation: PLL (♀); CNA (+); AVN (□); ADH (◇); LDH (△); MHR (#); LYZ (Z); ICB (X); CPA (Y).

fit by linear superposition of the spectra of a set of reference proteins (Figure 2), then the derived property "secondary structure content" of that protein can be calculated as an analogous linear superposition of secondary structure contents of the reference proteins. In mathematical terms

$$\sum_j A_{ij}\gamma_j = S_i \tag{2}$$

where S_i is the ith spectral data point of the spectrum being analyzed and j refers to the jth reference protein. The jth reference spectrum (the jth column of the matrix A) contributes to the fit with weight γ_j.

The secondary structure content f_k is then calculated by

$$f_k = \sum_j \gamma_j g_{kj} \tag{3}$$

where k denotes the secondary structure class (α-helix, β-sheet, etc.). The matrix elements g_{kj} are the kth secondary structure amount of the jth protein. Each protein contributes secondary structure proportional to the previously determined weight γ_j.

There are a number of algorithms to calculate the linear superposition weights γ for a best fit. Generally, these methods minimize the sum of squared deviations between the observed and calculated spectrum.

In our hands, the method of Provencher and Glöckner (1981) and Provencher (1982a,b) yields reasonable results when modified to focus on a good fit in regions of large intensity of the reference proteins. This is done in the constrained least-squares fit of CONTIN[1] by simply weighting the deviations between observed and calculated spectrum (to be minimized) by the sum of intensities of the reference spectra:

$$B_i = \sum_j A_{ij} \tag{4a}$$

$$w_i = \frac{B_i - \min B_i}{\max B_i - \min B_i} \tag{4b}$$

The accuracy of the spectroscopic secondary structure estimates calculated in this fashion is apparent from the comparison of calculated values with crystallographically derived values (Figure 3) for the set of reference proteins. While there is good correlation between the two for α-helical and total β-sheet content, the distinction between β-parallel and β-antiparallel cannot be made with reasonable accuracy. The overall accuracy is about the same for practical purposes as that reported by Williams (1983, 1986).

APPLICATION: TEST CASE DNASE I

The spectrum of DNase I serves as an example to show how structural information can be extracted from the amide I region (Figures 4 and 5). DNase I solution was prepared as in Suck (1982). The DNase I concentration was 2.3% by weight. The buffer solution contained 5 nM HEPES, 1 nM $CaCl_2$, 0.5 mM NaN_3, and 0.1 mM PMSF at pH 7.5. Pure buffer spectra and protein plus buffer spectra were then recorded by the divided spinning cell technique (see Measuring the Spectrum).

The raw spectrum (Figure 4a,b) contains experimental noise, and the amide I region is dominated by the broad water band around 1645 cm^{-1}. The water-subtracted (α = 1.35, determined by matching the internal standard peak near 1050 cm^{-1}) and smoothed (Bussian & Härdle, 1984) difference spectrum represents the contribution of DNase I only (Figure 4c). The absence of the major HEPES-related band near 1050 cm^{-1} in the difference spectrum is a good measure of the success of the subtraction procedure: after subtraction a line of medium intensity is revealed in this region at 1032 cm^{-1}, most likely due to phenylalanine.

Outside of the amide I region, the quality of the experimental spectrum is apparent from the presence of a number of marker bands, in particular for the aromatic residues tryptophan (at 759, 877, around 1010, and 1554 cm^{-1}), phenylalanine (at 1002 and 1032 cm^{-1}), and tyrosine (at 643 and around 1195 cm^{-1}). Also, lines of aromatic ring deformations at 1585 and 1616 cm^{-1} on the shoulder of the broad water band are clearly visible after subtraction (Figure 4c). The maximum intensity in the amide I region is at 1662 cm^{-1}. The secondary structure analysis is performed with the difference spectrum in the range 1630–1700 cm^{-1} (Figure 5). Interference of the nearest aromatic side-chain peak at 1616 cm^{-1} in this range is small (about 10% of amide I peak intensity at 1630 cm^{-1} and less than 2% at 1640 cm^{-1} [Figure 1 in Williams (1986)]). This interference is further suppressed by the bell-shaped choice of weights (eq 4b).

By visual comparison with reference proteins (cf. Simple Analysis of the Spectrum by a Visual Method), DNase I is estimated to be most similar in secondary structure content

[1] The program CONTIN can be obtained from S. Provencher, MPI Biophysikalische Chemie, D-3400 Göttingen, FRG, and the modified routine USRWT ("user weights") from the authors.

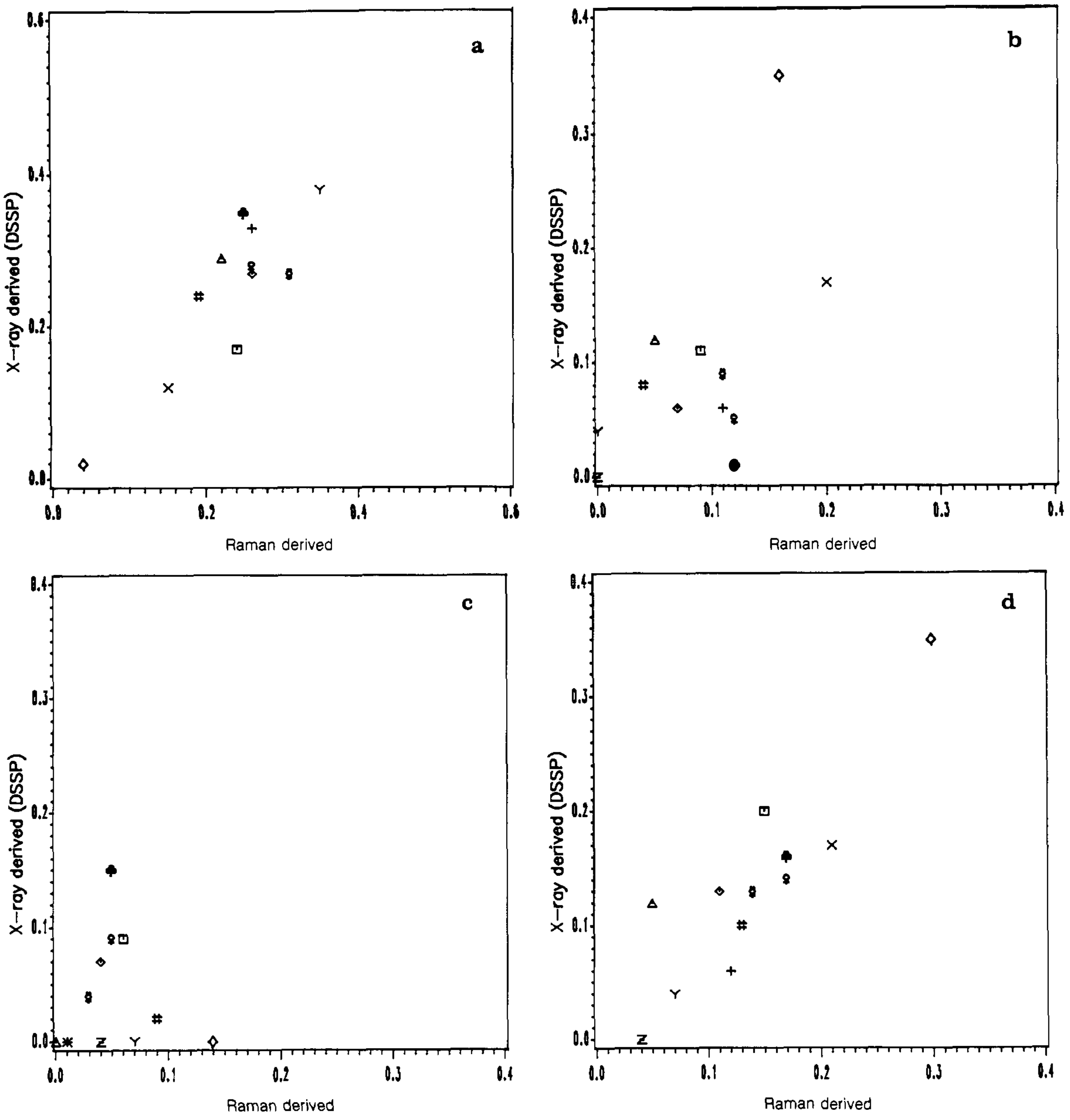

FIGURE 3: Correlation of crystallographically and spectroscopically determined secondary structure content (evaluation of accuracy). X-ray-derived structure content is calculated from the three-dimensional crystallographic coordinates by the program DSSP (Kabsch & Sander, 1983). Raman-derived secondary structure content comes from the fit (Provencher & Glöckner, 1981) of the amide I spectrum of one protein at a time in terms of the remaining 12 reference spectra. The more scattered the values, the less accurate is the expected accuracy of the spectroscopic estimates. Clearly, the distinction between antiparallel (b) and parallel (c) β-sheet is much less accurate than the determination of total α-helical (a) and total β-sheet (d) content. Proteins are CPA (♀), CRN (+), ADH (□), LDH (diamond with tick), INS (▲), LYZ (#), MLT (Z), PTI (X), ICB (Y), CNA (◇), TLN (female symbol with double quote), and TIM (cloverleaf). Abbreviations: avdv, average deviation; stdv, standard deviation; corl, correlation coefficient. (a) α-helix: avdv = 2.6%; stdv = 5.9%; corl = 0.95. (b) Antiparallel β-sheet: avdv = 0.6%, stdv = 7.7%; corl = 0.57. (c) Parallel β-sheet: avdv = −1.1%; stdv = 6.3%; corl = 0.03. (d) Total β-sheet: avdv = −0.5%; stdv = 4.2%; corl = 0.88.

to bovine pancreatic trypsin inhibitor (PTI). The more detailed analysis by the nonlinear fitting procedure of Provencher and Glöckner provides a fit of the amide I spectrum (Figure 5) as a linear superposition of 13 reference spectra not including that of DNase (Table I) with coefficients γ_j quantifying the contribution of protein j. Equation 3 then yields the secondary structure content. There is approximate agreement between the visual and the mathematical procedure (Table II). In fact, the mathematical fit also identifies PTI as a structurally close relative, along with concanavalin A (Table I). All other reference proteins contribute considerably less. To achieve a proper balance of α-helical contributions, some predominantly α-helical proteins appear with a negative coefficient. The deviation of the Raman-derived secondary structure content of DNase I from the X-ray derived values is −6% for helix content and −3% for sheet content, within the expected range of error.

DISCUSSION

We have improved the use of Raman spectroscopy for secondary structure analysis of proteins by application of the divided spinning cell technique and simplification of solvent subtraction, use of a stable method of data analysis, use of an objective definition of secondary structure, and careful calibration of expected accuracy.

For proteins, the spinning cell technique is a clear improvement over conventional Raman spectroscopy. Due to the inherently weak signal of the Raman effect, one needs high laser beam intensity, long exposure times, and fairly high protein concentrations in order to obtain a good signal-to-noise

Table I: Contribution of Reference Proteins to the Calculated Spectrum of DNase I in the Amide I Region

protein	linear coeff (10^{-4})	error (10^{-4})	% α[a]	% β[a]	% undefined[a]	% α[b]	% β[b]	% undefined[b]
CPA	1129	55	40	30	30	28	14	58
LDH	1058	139	42	24	34	27	13	60
MHR	-207	236	80	0	20			
LYZ	852	193	46	19	35	24	10	66
MLT	-794	290	85	0	15	71	0	29
PTI	2056	161	26	45	29	12	17	71
ICB	-1348	365	71	8	21	38	4	58
CNA	2416	247	20	65	33	2	35	63
TIM	802	193	52	24	24	35	16	49
INS	1247	268	55	21	24	29	12	59
ADH	1099	139	29	37	34	17	20	63
TLN	845	58	40	30	30	27	13	60
CRN	838	236	46	22	32	33	6	61

[a] Percent secondary structure from Williams (1983), according to Levitt and Greer (1977). [b] Percent secondary structure from DSSP defined as number of hydrogen bonds of type α (or β) per 100 residues (Kabsch & Sander, 1983), except for MHR.

Table II: Secondary Structure Content of DNase I

method	% α	% β	% undefined
Raman derived, visual	26	45	29
Raman derived, mathematical, Levitt and Greer sec struct[a]	22	44	33
Raman derived, mathematical, DSSP sec struct[a]	14	22	64
X-ray derived, DSSP sec struct[a]	20	25	55

[a] sec struct = secondary structure definition.

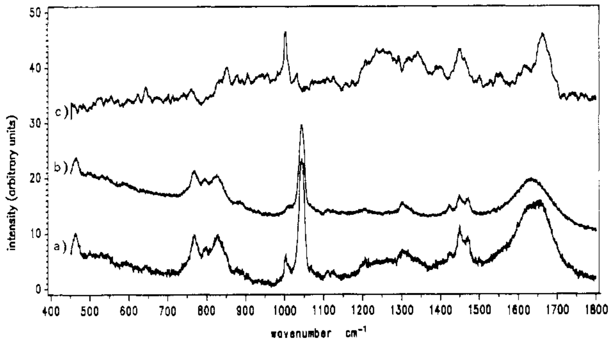

FIGURE 4: Raman spectra of DNase I and a HEPES buffer solution. Data from 450 to 1800 cm^{-1}, at 1-cm^{-1} intervals. (a) Raw data, protein plus buffer; (b) raw data, buffer only; (c) smoothed and normalized difference (a − b), representing the contribution of protein only. Experimental details: HEPES buffer at pH 7.5, protein concentration 2.3% (weight), laser line 514 nm, monochromator slit width 5-7.5-7.5-5 cm^{-1}, data accumulated for 20 s at each cm^{-1} value.

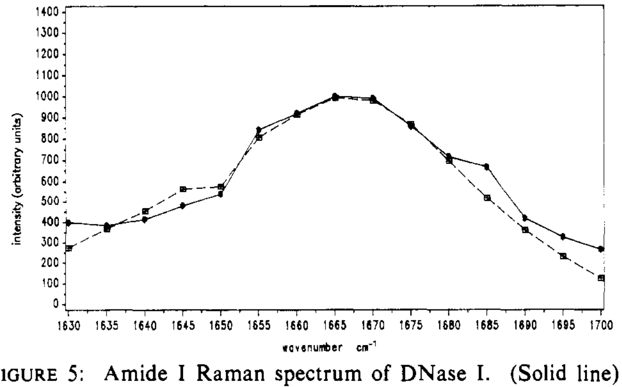

FIGURE 5: Amide I Raman spectrum of DNase I. (Solid line): Experimental spectrum, buffer subtracted and smoothed (Bussian & Härdle, 1984) and digitized at 5-cm^{-1} intervals. (Dashed line): Best fit as a linear superposition of reference spectra with linear coefficients as in Table I; in the peak region the fit is very close to the experimental data which is achieved by use of eq 4b. The aromatic ring deformation contributes some intensity in the 1630-cm^{-1} region and may therefore perturb the result for α-helix structure. This fit is the basis for estimating the secondary structure content of DNase I.

ratio. These conditions tend to damage proteins by denaturation or chemical modification. A spinning cell, however, has improved heat dissipation and keeps the protein in gentle conditions even at high laser intensities and long exposure times. Often, the protein remains usable (retains activity) after it leaves the spectrometer. The sophistication of dividing the spinning cell permits simultaneous recording and direct subtraction of two spectra taken in identical conditions.

Solvent subtraction is particularly simple in that spectra of both protein and protein plus buffer are measured in identical conditions. Subtraction of the spectral contribution of aromatic rings near the amide I region (nearest peak at 1616 cm^{-1}), a fairly complicated procedure in the work of Williams (1986) (p 316) is rendered unnecessary by choosing 1630 cm^{-1} as the lower limit of the range used in the secondary structure analysis and by using weights (eq 4b) which concentrate on the center of the amide I region and are small at the edges.

The ambiguities inherent in fitting the observed spectrum as a superposition of linearly dependent reference spectra is solved stably by use of the constrained least-squares method of Provencher and Glöckner (1982) and Provencher (1982a,b). Their procedure gives an unambiguous decomposition of experimental spectra, with a best linear combination of reference spectra as reflected by a best choice of linear coefficients γ. The better of the two methods used by Williams (1983, 1986), based on singular value decomposition, appears to be capable of yielding spectral fits of similar quality, but the criteria for choosing the best solution among those offered by the method have not yet been computerized. Neither the deconvolution method of Thomas and Agard (1984) nor that of Byler and Susi (1988) has to our knowledge been applied to the complete set of reference proteins.

For the analysis of the amide I spectrum in terms of secondary structure content, the simple visual method is adequate for a rough estimate. For better reproducibility and accuracy the constrained least-squares procedure should be used.

As the secondary structure estimate echoes the definition of secondary structure in the reference set of proteins, it is important to use a physical meaningful definition of secondary structure. As the definition of Kabsch and Sander (1983) used here is based on the detection of H-bonding patterns, it more precisely reflects the abundance of the molecular groups contributing to the Raman amide I band than that of Levitt and Greer (1977). In addition, the corresponding computer program (DSSP) is publicly available[2] and can thus be con-

[2] The program DSSP (Kabsch & Sander, 1983) can be requested from the electronic mail fileserver NETSERV@EMBL.BITNET (send the mail message "help").

sistently applied as new structures are included in the reference data set.

By independently checking the claim [Williams, 1983; retracted in Williams (1986)] that the present set of reference spectra allows distinction between parallel and antiparallel β-structure (it does not), we gain a clear statement of the accuracy one can expect in routine application of Raman spectroscopy to globular proteins.

Most likely the accuracy of the Raman-derived secondary structure estimate could be significantly improved by extending the set of reference spectra. It seems particularly important to include proteins of all topological classes, especially of proteins containing domains with both α-helices and β-sheets, such as parallel β-barrels and nucleotide binding folds. It is our hope that several of the groups involved in Raman spectroscopy of proteins will help to improve the method by contributing new reference spectra of proteins of known structure.

Acknowledgments

We thank Dietrich Suck for supplying DNase I and Hans H. Eysel for beam time on his Raman spectrometer and for helpful discussions. We thank the referees for detailed suggestions.

Registry No. DNase, 9003-98-9.

References

Bussian, B. M., & Härdle, W. (1984) *Appl. Spectrosc. 38*, 309.

Byler, D. M., & Susi, H. (1988) *J. Ind. Microbiol. 3*, 73–88.

Carey, P. R. (1982) in *Biochemical Applications of Raman and Resonance Raman Spectroscopies*, Chapter 4, Academic Press, New York.

Eysel, H. H., & Bussian, B. M. (1985) *Spectrochim. Acta 41A*, 1149.

Garfinkel, D., & Edsall, J. T. (1958) *J. Am. Chem. Soc. 80*, 3818.

Kabsch, W., & Sander, C. (1983) *Biopolymers 2*, 2577.

Kiefer, W. (1973) *Appl. Spectrosc. 27*, 253.

Kiefer, W. (1977) *Adv. Infrared Raman Spectrosc. 3*, 1.

Kiefer, W., & Bernstein, H. J. (1971) *Appl. Spectrosc. 25*, 500.

Krimm, S., & Bandekar, J. (1986) *Adv. Protein Chem. 38*, 181–364.

Levitt, M., & Greer, J. (1977) *J. Mol. Biol. 114*, 181–239.

Lord, R. C. (1977) *Appl. Spectrosc. 31*, 187.

Miyazawa, T. (1960) *J. Chem. Phys. 32*, 1647.

Peticolas, W. L. (1979) *Methods Enzymol. 61*, 425.

Provencher, S. (1982a) *Comput. Phys. Commun. 27*, 213.

Provencher, S. (1982b) *Comput Phys. Commun. 27*, 229.

Provencher, S., & Glöckner, W. (1981) *Biochemistry 20*, 33.

Sargent, D., Benevides, J. M., Yu, M. H., King, J., & Thomas, G. J., Jr. (1988) *J. Mol. Biol. 199*, 491–502.

Suck, D. (1982) *J. Mol. Biol. 162*, 511.

Thomas, G. J., Jr., & Agard, D. A. (1984) *Biophys. J. 46*, 763–768.

Williams, R. W. (1983) *J. Mol. Biol. 166*, 581.

Williams, R. W. (1985) *J. Biol. Chem. 260*, 3937–3940.

Williams, R. W. (1986) *Methods Enzymol. 130*, 311–331.

Williams, R. W., McIntyre, J. O., Gaber, B. P., & Fleischer, S. (1986) *J. Biol. Chem. 261*, 14520–14524.

Sequence-Dependent Termination of Bacteriophage T7 Transcription in Vitro by DNA-Binding Drugs[†]

Robin J. White and Don R. Phillips*

Biochemistry Department, La Trobe University, Bundoora, Victoria 3083, Australia

Received September 13, 1988; Revised Manuscript Received December 7, 1988

ABSTRACT: An in vitro T7 bacteriophage transcription system has been utilized in which the RNA was initiated to a specific length (defined by the absence of the appropriate nucleoside triphosphate). When the DNA–RNA–RNA polymerase ternary complex was exposed to nonsaturating levels of DNA-binding ligands (i.e., a small fractional occupancy at each site), and the RNA transcript then allowed to elongate in the presence of all four nucleoside triphosphates, there was a synchronous increase of RNA lengths up to sites occupied by ligands. A unique characteristic is that bacteriophage transcription was completely terminated at every ligand site, in contrast to bacterial RNA polymerases where "read-through" past drug sites occurs and results merely in a delay of transcription at each site due primarily to dissociation of drug from the DNA. Similar termination of transcription at each drug site was observed with T3 and SP6 RNA polymerases. The termination at drug sites in the bacteriophage system results in RNA of specific lengths which define the location of ligand sites, and the RNA concentration provides a measure of relative ligand occupancy at that site. Termination of transcription was observed with four drugs with relatively long DNA residence times (half-life ⩾300 s at 20 °C for nogalamycin, actinomycin, mithramycin, and echinomycin) but to a lesser extent with drugs of intermediate residence times [a bis(thiadaunomycin) and an acridine–tripyrrole, with half-lives of 230 and 7 s, respectively, at 20 °C].

The technique of footprinting has been widely employed in recent years to define the sequence specificity of DNA-binding drugs and proteins. Several different DNA cleaving agents and procedures have been employed [e.g., DNase I, MPE-Fe(II), and EDTA-Fe(II)], and these have recently been reviewed (Dabrowiak, 1983; Dervan, 1986; Tullius, 1987). This technique has been especially successful in delineating the boundaries of large ligands with long residence times, such as specific DNA-binding proteins. For example, the size of

[†] This work was supported by a grant from the Australian Research Grants Scheme.