THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

Extracting Infrared Spectra of Protein Secondary Structures Using a Library of Protein Spectra and the Ramachandran Plot

James V. Coe,*[†] Steven V. Nystrom,[†] Zhaomin Chen,[†] Ran Li,[†] Dominique Verreault,[†] Charles L. Hitchcock,[‡] Edward W. Martin, Jr.,[§] and Heather C. Allen[†,‡]

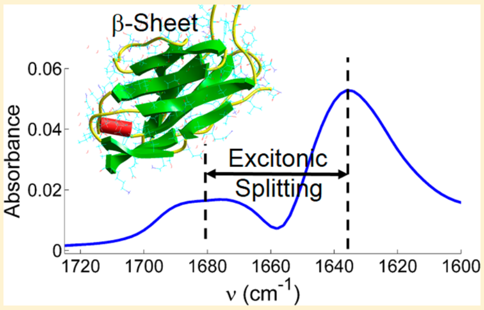
[†]Department of Chemistry and Biochemistry, The Ohio State University, 100 West 18th Avenue, Columbus, Ohio 43210-1173, United States

[‡]Department of Pathology, The Ohio State University, 4132 Graves Hall, 333 West 10th Avenue, Columbus, Ohio 43210, United States

[§]Department of Surgery, Division of Surgical Oncology, The Ohio State University, 410 West 10th Avenue, Columbus, Ohio 43210, United States

S *Supporting Information*

ABSTRACT: Infrared (IR) spectra from 1200 to 1800 cm^{-1} of the pure α-helix and β-sheet secondary structures have been extracted using a covariant least-squares procedure which relates a library of 40 infrared (IR) solution protein spectra from the work of Dong, Carpenter, and Caughey and amino acid fractions of the proteins based on assignments by STRIDE (secondary *str*ucture *ide*ntification) of Eisenhaber and Argos. The excitonic splitting of the β-sheet structures is determined for this library of solution proteins. The method is extended to find a set of spectral basis functions that analyze IR spectra of protein samples for α-helix and β-sheet content. A rigorous error analysis including covariance, the correlations between the input library spectra, was used to justify the results and avoid less meaningful results. The utility of the results on α-helix and β-sheet regions is demonstrated by detecting protein changes due to cancer in imaging Fourier transform IR (FTIR) spectra of liver tissue slices. This work ends with a method to extract IR spectra of less prominent torsional angle distributions.



■ INTRODUCTION

The infrared (IR) spectra of proteins are well-known to be sensitive to protein secondary structures.[1,2] For example, the shape of the protein amide I band (~1600−1700 cm^{-1}) is very different in a protein dominated by α-helix than one dominated by β-sheet.[3,4] The amide I band is the strongest IR band in most tissue slices[5−7] and is important in diagnosing biological and medical samples. Theoretical treatment of the vibrational spectra of α-helix and β-sheet structures[8−11] remains a challenging problem because the structures are large and have many subunits with variable chain length, a variety of interstrand interactions, twisting of strands, solvent and pH interactions. A very good review[12] of the interactions associated with amide I band spectra describes the importance of through-space transition moment coupling between neighboring amino acids with lesser contributions from H-bonding, through-bond coupling, and solvation interactions. The amide I vibration (~1665 cm^{-1}) is dominated by the atoms in the backbone of the protein (C=O stretching with contributions from out-of-phase CN stretching vibration, CCN deformation, and the NH in-plane bend). Since the backbone torsional angles (defined in Figure 1a) involve the same atoms as the amide I band vibration, these properties are intrinsically linked. Figure 1b shows a Ramachandran plot of the occurrence of amino acid torsional angles for the library proteins of this work. Each amino acid in a protein makes a contribution to the amide I band shape that is primarily mediated by transition moment coupling to neighbors whose orientations and distances vary with protein secondary structure. These interactions are very different in the α helix and β sheet structures producing very different IR spectra.

Since tissue samples generally have varieties of proteins, the amide I band is typically a broad unresolved band with many subtle, but telling inflections. While there has been much work[12] using second derivatives[13] and deconvolution[14,15] to identify the prominent features of the IR spectra of the most common secondary protein structures,[1,2,12] the spectra are highly overlapped. The most desirable feature of this work and its multiple regression predecessors is the extraction of the whole IR spectrum of a protein secondary structure. This is associated with some difficulty since the IR spectra of various protein secondary structures are highly correlated—some of the secondary structure groups are overlapped on the Ramachandran plot and others correspond to multiple structure groups

Received: August 18, 2015
Revised: September 22, 2015
Published: September 23, 2015

DOI: 10.1021/acs.jpcb.5b08052
J. Phys. Chem. B 2015, 119, 13079−13092

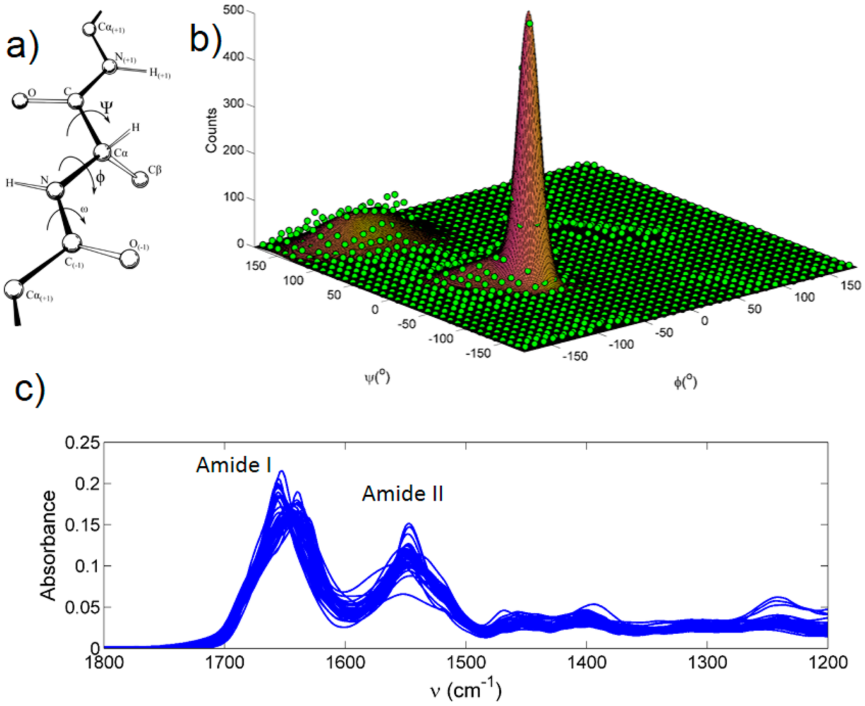


Figure 1. (a) Protein backbone definitions of the ω, φ, and ψ torsional angles which also involve atoms important in the amide I and II vibrations. (b) Ramachandran plot with counts of the number of amino acid residues occurring out of 9313 in the protein library that fall within 10° square bins of the φ, ψ torsional angles (green dots). These distributions were fit to rotating two-dimensional Gaussians. (c) Normalized IR spectra of 40 proteins from the database of Dong, Carpenter, and Caughey are overlapped to reveal variations in the amide I and II band regions.

on the Ramachandran plot. The spectra of protein secondary structures are extracted in this work by relating the fractions of secondary structures in library proteins to the IR spectra of the library proteins (normalized protein spectra are shown in Figure 1c). This linear least-squares relation has been considered[4,16−23] by many investigators and justified with a Beer's law interpretation.[18,20] Statistical methods of factor analysis[17] or single value decomposition,[21−23] partial least-squares,[16,18,24,25] and multivariate curve resolution-alternating least-squares[20] attest to the difficulty of the endeavor. The current approach is direct least-squares, but it is called generalized least-squares as it uniquely allows for the correlation between the input library of protein IR spectra. Note that the input IR spectra are strongly correlated and ignoring this fact can lead to underestimated uncertainties. The current effort distinguishes itself by considering covariance. Fits were performed without weights (as most previous work), with weights, and with the covariance between input spectra. Rigorous covariant error analysis and correspondence to the Ramachandran plot are critical features distinguishing the current method from previous multivariate approaches.

The following sections describe the basic methods and results including: the input data for the least-squares analysis, an analysis of the torsional angle distributions of common protein secondary structures, a general least-squares procedure relating secondary structure amino acid fractions and library IR spectra producing isolated spectra of α-helix and β-sheet, a set of spectral basis functions for analysis of α-helix and β-sheet content from the protein library spectra, an application of the spectra basis functions for detecting protein changes due to cancer in imaging FTIR spectra of liver tissue slices, and a technique for analysis of spectra from overlapped regions on the Ramachandran plot, i.e., the IR Ramachandran ellipse method. All of this includes a rigorous error analysis which justifies the results and approach.

Input Data—Protein Spectra and Secondary Structure Groups. The input library of protein IR spectra come from the database of Dong, Carpenter, and Caughey (www.unco.edu/nhs/chemistry/faculty/dong/irdata.htm)[26−29] and the normalized IR spectra are overlaid in Figure 1c revealing notable variations in the amide I and II spectral regions. The selected library proteins consist of 40, water-soluble (H_2O buffer), short-chain proteins with amino acid chain lengths varying from 55 to 757 as listed in Table 1. The IR spectra come from solutions of the proteins (typically ~5 mg protein/mL in a 10 mM phosphate buffer solution at pH 7.3). They were recorded with a 6 μm path length over the range 1200−2000 cm^{-1} at 4 cm^{-1} resolution along with H_2O buffer spectra in the same cell and a subtraction protocol removed the water contributions to the protein spectrum.[30] Since the spectra do not all share the same wavenumbers, all of the spectra were interpolated to a common set from 1200 to 2000 in 2 cm^{-1} steps. The region in these spectra from 1840 to 1920 cm^{-1} is baseline and the standard deviation of the baseline noise was used to determine weights for each library spectrum, while spectra were only used over the reduced range from 1200 to 1800 cm^{-1} in this work. Since the spectra have variable concentrations and chain lengths, each was normalized so that the dot product of each spectrum with itself was one over the interval from 1200 to 1800 cm^{-1} (Figure 1c). This was a critical aspect missing from our early efforts. All 40 of the protein library spectra were stacked as row vectors into one matrix, the **Y** matrix, which is $m_s \times n$ in dimension where m_s is the number of protein library spectra and n is the number of steps in the IR spectra from 1200 to 1800 cm^{-1}. The **Y** matrix was 40 × 301 in this work. Each column of **Y** corresponds to the spectral

Table 1. Protein Library List[a]

no.	name	protein	PDB	no. of AA	fraction α	fraction β
1	a1pi	α-1-proteinase inhibitor (human)	1KCT	375	0.0880	0.0800
2	bsa	albumin (bovine serum, A-0281 Sigma)	4F5S	583	0.7204	0.0000
3	albumnhu	albumin (human serum)	1E7I	582	0.7096	0.0000
4	alcdehho	alcohol dehydrogenase (equine liver)	6ADH	374	0.1738	0.2219
5	alcdehye	alcohol dehydrogenase (baker's yeast)	2HCY	347	0.2767	0.2911
6	bfgf	basic fibroblast growth fac. (recom., human)	1BFG	126	0.0000	0.4127
7	carbanhy	carbonic anhydrase (bovine erythrocytes)	1V9E	259	0.0734	0.3050
8	concanv	concanavalin A (jack bean)	3CNA	237	0.0000	0.4304
9	chymbov	α-chymotrypsin (bovine pancreas)	1YPH	131	0.0000	0.3511
10	cytreho4	cytochrome *c* (reduced; equine heart)	2GIW	104	0.4038	0.0000
11	cytoxho4	cytochrome *c* (oxidized; equine heart)	1AKK	104	0.3942	0.0385
12	cytoxtun	cytochrome *c* (oxidized; tuna heart)	3CYT	103	0.4563	0.0388
13	cytoxiso	cytochrome *c* (oxidized; baker's yeast)	2LIR	108	0.3611	0.0000
14	dnase1	deoxyribonuclease I (bovine pancreas)	1DNK	250	0.2720	0.2960
15	elastspo	elastase (porcine pancreas)	2V35	240	0.0667	0.3458
16	enolase	enolase (baker's yeast)	3ENL	436	0.3899	0.1743
17	rfxiii	factor XIII (recombinant; homodimer; human)	1F13	722	0.1260	0.4072
18	apoferit	ferritin (apo, horse spleen)	4DE6	168	0.7738	0.0000
19	fibrgnhu	fibrinogen (human plasma)	3GHG	401	0.3541	0.1970
20	hbcohu	hemoglobin (carboxy; human)	1K0Y	141	0.7163	0.0000
21	hbmethor	hemoglobin (aquomet; equine)	1NS6	141	0.7589	0.0000
22	iggbov	immunoglobulin G (bovine)	1GB1	56	0.2500	0.4107
23	interfhu	interferon-γ (recombinant; human)	1EKU	252	0.7024	0.0000
24	lalbnca	α-lactalbumin (Ca-bound; bovine milk)	1F6S	122	0.3443	0.0820
25	ldhrab	lactic dehydrogenase (rabbit muscle)	3H3F	331	0.4109	0.2145
26	blgabov	β-lactoglobulin A (bovine milk)	1CJ5	162	0.0556	0.3457
27	blgbbov	β-lactoglobulin B (bovine milk)	4IBA	157	0.1146	0.4140
28	len	light-chain LEN (human recombinant)	2LVE	113	0.0000	0.5133
29	lysozyme	lysozyme (chicken egg white)	1AZF	129	0.3333	0.0620
30	ovalbum	ovalbumin (chicken egg)	2FRF	152	0.7632	0.0000
31	papain	papain (papaya latex)	9PAP	211	0.2322	0.1801
32	rnasea	RNase A (bovine pancreas)	2QCA	124	0.1935	0.3306
33	subtilis	subtilisin Carlsberg (*Bacillus licheniformis*)	1SBC	274	0.3139	0.1642
34	sodoxbov	Cu,Zn-superoxide dismutase (ox.; bov. liver)	1CB4	151	0.0397	0.4040
35	sodrebov	Cu,Zn-superoxide dismutase (red.; bov. liver)	1SXN	151	0.0331	0.4172
36	staphnuc	staphylococcal nuclease (recombinant)	1NUC	135	0.2741	0.3111
37	tim	triosephosphate isomerase (rabbit muscle)	1R2S	247	0.4372	0.1579
38	trypsnb	trypsin (bovine pancreas)	4I8L	223	0.0807	0.3363
39	trypgenb	trypsinogen (bovine pancreas)	1TGN	222	0.0811	0.3468
40	sti	trypsin inhibitor (soybean)	1BA7	169	0.0000	0.4260

[a] Each protein has an IR spectrum from the library of Dong, Carpenter, and Caughey (2nd column) and a corresponding file from the RCSB Protein Data Bank (4th column). The fractions of α-helix and β-sheet are from STRIDE H and E groups.

absorption value of each library protein at one specific wavenumber and this is the dependent variable in the least-squares process.

A Ramachandran plot,[31] a plot of occurrence of amino acid residues as a function of the φ and ψ backbone dihedral angles, is shown for the library proteins in Figure 1b (with torsional angles defined in Figure 1a). The bin size for the two-dimensional histogram in Figure 1b was 10° for both torsional angles and a home-written Matlab routine counted and plotted the number of amino acids falling within each bin area. The total number of amino acids in the library (9313) enables counts to be converted to fractions for later work.

The torsional angle information in Ramachandran plots and its direct connection to IR spectra can be profitably explored in terms of well-defined secondary structures which complement torsional angle data with additional knowledge such as hydrogen bonding. STRIDE, a knowledge-based method for assigning protein secondary structure improving on the "old" DSSP method, was used to assign secondary structures.[32] A Web site (http://webclu.bio.wzw.tum.de/cgi-bin/stride/stridecgi.py) accepts a protein data bank file (*.pdb) as input, and outputs text files with STRIDE secondary structure assignments. Pertinent sections of the outputs (i.e., the sequential list of the protein's amino acids with columns of other data including secondary group assignments, torsional angles, and protein identity) were gathered into one Excel file for the whole protein library. It was read as input for our Matlab programs and has been added as Supporting Information. STRIDE identified six secondary structures groups for this library of proteins including "H" for α-helix (0.3163), "G" for 3_{10} helix (0.0321), "C" for coil (0.1730), "T" for turns (0.2540), "E" for extended conformation or β-sheet (0.2095), and "B" for bridge (0.0148), where the fractions of the total amino acid count are given in parentheses out of a total of 9313 amino acids. A Ramachandran dot plot (one dot for each amino acid) is shown in Figure 2 for each of these groups. The α-helix is defined by hydrogen bonding between the i and $i+4$ amino acids along the helix. The STRIDE α helix group also contains the turns at the beginning and end of the alpha chain. As can be seen in Figure 2 (top left), the proper α helix distribution is tightly clustered about the torsional angles φ_c, $\psi_c = -63°, -42°$, which compares well to the crystallographic determinations of α-helix by Hovmöller and co-workers[33] of φ_c, $\psi_c = -63.8°$, $-41.1°$, so this is the right-handed α helix structure. The α-helix turns have a smaller fraction with more spread than the helix proper. This is the most concentrated secondary structure group in terms of the torsional angle distributions and therefore is the easiest for extraction of an IR spectrum. Another well-defined group includes the 3_{10}-helix which has hydrogen bonding between i and $i + 3$ amino acids [Figure 2 (top right)]. This group is ~10 times less abundant than α-helix, but highly overlapped with the α-helix distribution. The second most concentrated distribution is that of the extended β strands in Figure 2 (bottom left). This region encompasses the crystallographic determinations for parallel and antiparallel β-sheets with average values[33] of φ_c, ψ_c = −116°, +128° and −122°, +136°. Again, STRIDE includes the turns which spread to many other regions of the Ramachandran plot. While the distribution of torsional angles is more spread than that of α-helix, it is still highly concentrated with a set of torsional angles that are very different than α-helix and that includes parallel and antiparallel β-sheet structures. In contrast, the groups designated as coils and turns, which together include ~43% of the amino acids, are spread all over the Ramachandran plot [Figure 2 (middle left and right)]. Unlike α-helix and β-sheet, coils and turns have a variety of torsional angles and therefore a variety of contributing IR spectra associated with each group. More analytical work is required to characterize the IR spectra of these groups.

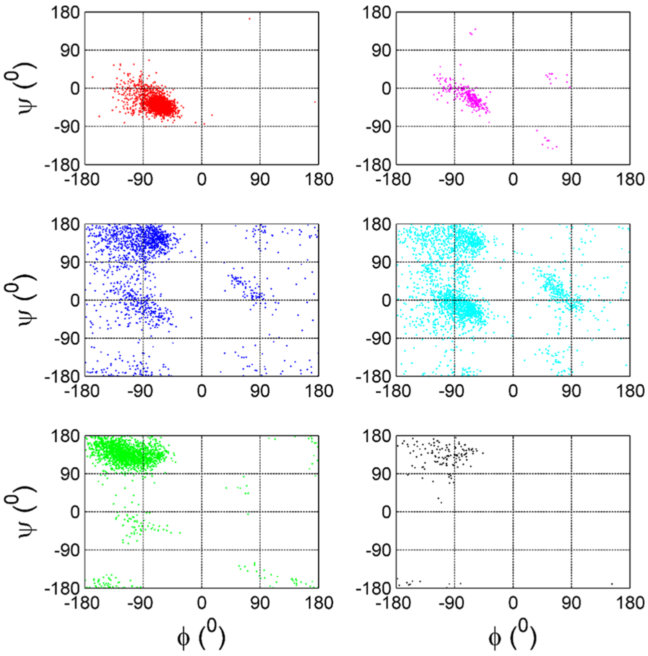


Figure 2. Dot plots of the torsional angles of amino acids of library proteins that have been classified as groups by the STRIDE analysis including α-helix (top left, STRIDE "H"), 3_{10}-helix (top right, STRIDE "G"), coils (middle left, STRIDE "C"), turns (middle right, STRIDE "T"), β-sheet or extended strands (bottom left, STRIDE "E″), and bridges (bottom right, STRIDE "B"). The α-helix, 3_{10}-helix, and β-sheet groups are dominated by single distributions, while the coils and turns groups are much more spread out over a variety of structures.

The fractions of α helix (STRIDE group "H") and β sheet (STRIDE group "E") for each library protein are given in Table 1. It can be seen in Table 1 that the fraction of α-helix in individual library proteins varies from 0.000 to 0.759 while the fraction of β-sheet varies from 0.000 to 0.513. These library protein fraction values are incorporated into an **X** matrix which is the independent variable in the least-squares fit. The **X** matrix is $m_s \times n_g$ where m_s is the number of protein library spectra and n_g is the number of secondary structure groups. The **X** matrix is 40 × 2 or 40 × 3 in this work depending on whether two or three groups were used. Each column contains the fraction of amino acids in a specific secondary structure for each library protein. The two data sets, represented by the matrices **X** and **Y**, comprise the independent and dependent variables of the least-squares procedure, respectively. The MATLAB programming language and computational package from MathWorks was used extensively to deal with these matrices.

Fitting Torsional Angle Distributions. The torsional angle distributions of the STRIDE secondary structure groups have been fit to rotating, two-dimensional Gaussian functions in order to better explore the relationship between multiple structure groups and IR spectra. The most prevalent distributions within and between groups were fit with a nonlinear least-squares routine (using the "fminsearch" function in Matlab) to the following form of a rotating two-dimensional Gaussian:

$$f(\phi, \psi) = A\mathrm{e}^{-[(\cos^2\theta/2\sigma_a{}^2+\sin^2\theta/2\sigma_b{}^2)(\phi-\phi_c)^2+2(-\sin(2\theta)/4\sigma_a{}^2+\sin(2\theta)/4\sigma_b{}^2)(\phi-\phi_c)(\psi-\psi_c)+(\sin^2\theta/2\sigma_a{}^2+\cos^2\theta/2\sigma_b{}^2)(\psi-\psi_c)^2]} \tag{1}$$

Table 2. Fit Parameters for Rotating 2D Gaussian Functions of the Torsional Angle Distributions Shown in Figure 3 Following Eq 1

name	A	φ_c (deg)	σ_a (deg)	ψ_c (deg)	σ_b (deg)	θ (deg)	int. count
α-helix	554.915(18)	−63.3794(23)	9.829(19)	−41.926(4)	6.935(19)	59.38(3)	2377
β-sheet	61.297(29)	−115.310(7)	29.549(29)	135.461(4)	16.628(29)	23.74(8)	1892
3_{10}-helix	34.64(3)	−61.720(5)	15.58(3)	−25.233(17)	6.28(3)	57.400(21)	213
T1	41.15(3)	−79.622(10)	30.60(3)	−15.94(5)	13.53(3)	45.49(3)	1070
T2	46.63(4)	−74.413(11)	22.76(4)	143.099(6)	19.48(4)	42.99(24)	1299

DOI: 10.1021/acs.jpcb.5b08052

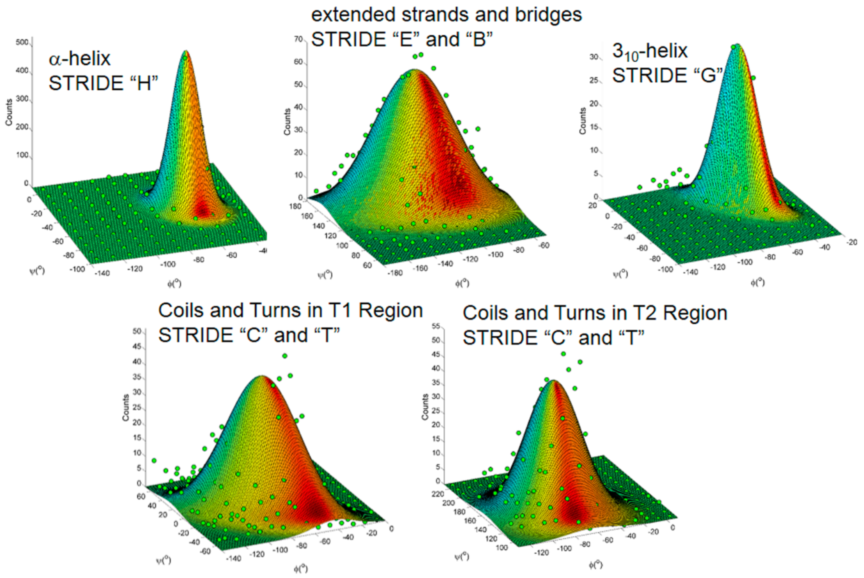


Figure 3. Nonlinear least-squares fits of STRIDE group torsional angle distributions to rotating, two-dimensional Gaussian functions. Fit parameters are given in Table 2.

where A is the distribution intensity in amino acid counts, φ_c, ψ_c are the center torsional angles, σ_a and σ_b are the standard deviations of the distributions along the major and minor axes, and θ is the rotation from the φ axis. The numerical fitted results are given in Table 2, pictured as a surface in Figure 1b, and given as individual fits in Figure 3. All of these fits were accomplished with ($\Delta\varphi$ = 10°) × ($\Delta\psi$ = 10°) bins so they could be readily compared. Only data within the fitting regions for each distribution, as shown in Figure 3, were used for fits to minimize interactions with other distributions. The bridge group (STRIDE "B") was merged with the extended strands (STRIDE "E") because of the extensive overlap of their most dense regions as can be seen in Figure 2. The number of amino acid counts within a rotating, 2D Gaussian distribution is $2\pi A\sigma_a\sigma_b/(\Delta\varphi\Delta\psi)$ and this quantity is included in Table 2. It allows one to estimate the fraction of turns included in the STRIDE groups. For instance, the STRIDE group "H" is α-helix and contains 2377 amino acids in the α-helix proper. This can be compared to 2946 for the whole group suggesting that 19.3% are turns. Furthermore, this corresponds to α-helix proper average lengths of about 8.4 amino acids plus two turns.

The five regions of torsional angle distributions are not always simply related to secondary structures, so it is important to identify the ones that are simply related in order to get all positive least-squares spectral solutions as demanded by the physics of IR spectra. The sharpest distribution from the solution protein library (standard deviations of 9.8° and 6.9°) is for α-helix and was centered at φ_c, ψ_c = −63.38°, −41.93° in agreement with crystallographic determinations.[33] The 3$_{10}$-helix distribution is displaced to φ_c, ψ_c = −61.72°, −25.23°, but it is much smaller and overwhelmed by the α-helix. Both the α-helix and 3$_{10}$-helix distributions are extensively overlapped by a nearby distribution that we have called T1. This distribution has prominent contributions from both the Coil and Turn STRIDE groups ("C" and "T"). These groups were merged and fit in this region revealing a common center for T1 of φ_c, ψ_c = −79.6°, −15.9°, with widths of 30.6° and 13.5°, a slightly different rotation than α-helix, and encompassing 11.5% of the library amino acids. The β-sheet region is the second most intense region with a fitted center at φ_c,ψ_c = −115.31°, −135.46° encompassing the crystallographic determinations for parallel and antiparallel β-sheets with average values[33] of φ_c,ψ_c = −116°, +128° and −122°, +136° considering the fitted widths of 29.6° and 16.6°. Given great overlap, the STRIDE bridge group ("B") was merged in the fit and later analysis, so this region of β-sheets also includes twisted strands. Finally, the β-sheet distribution is extensively overlapped by a distribution centered at φ_c, ψ_c = −74.41°, − 143.10° which we have labeled T2. This group has prominent contributions from both the Coil and Turn STRIDE groups ("C" and "T"). One must consider the overlap of T2 with the β-sheet region, as well as T1 with the α helix region in order to extract IR spectra of protein secondary structures. With this background, the paper proceeds to the basic linear least-squares relation between IR library spectra and amino acid fractions of secondary structure groups.

Linear Least Squares Relation Between Spectra and Fractions. The current method obtains an ordinary linear least-squares solution, however it is performed at each and every wavenumber of the IR spectrum; i.e., the full process is a multivariate linear regression. To better understand this, consider three secondary structure groups [α helix (α), the β region (β), and other (O)] at one wavenumber, for example at 1200 cm^{-1}. Then the linear least-squares relation is

$$\begin{bmatrix} y_{1,1200\text{cm}^{-1}} \\ y_{2,1200\text{cm}^{-1}} \\ \vdots \\ y_{m_s,1200\text{cm}^{-1}} \end{bmatrix} = \begin{bmatrix} x_{\alpha,1} & x_{\beta,1} & x_{O,1} \\ x_{\alpha,2} & x_{\beta,2} & x_{O,2} \\ \vdots & \vdots & \vdots \\ x_{\alpha,m_s} & x_{\beta,m_s} & x_{O,m_s} \end{bmatrix} \begin{bmatrix} b_{\alpha,1200\text{cm}^{-1}} \\ b_{\beta,1200\text{cm}^{-1}} \\ b_{O,1200\text{cm}^{-1}} \end{bmatrix} \tag{2}$$

where the left-hand column of y values contains the absorbance of each library protein at the selected wavelength, the x values are the fractions of amino acids in each secondary structure group for each protein, and the b values are the IR spectra of the secondary structure groups at the selected wavelength. The method models each library protein's spectrum with a linear combination of the fractions and the secondary group spectra, $y_i = x_{\alpha,i}b_\alpha + x_{\beta,i}b_\beta + x_{O,i}b_O$. Upon extending the ordinary least-squares procedure to all wavenumbers, then the relation becomes a multivariate regression which in matrix form is

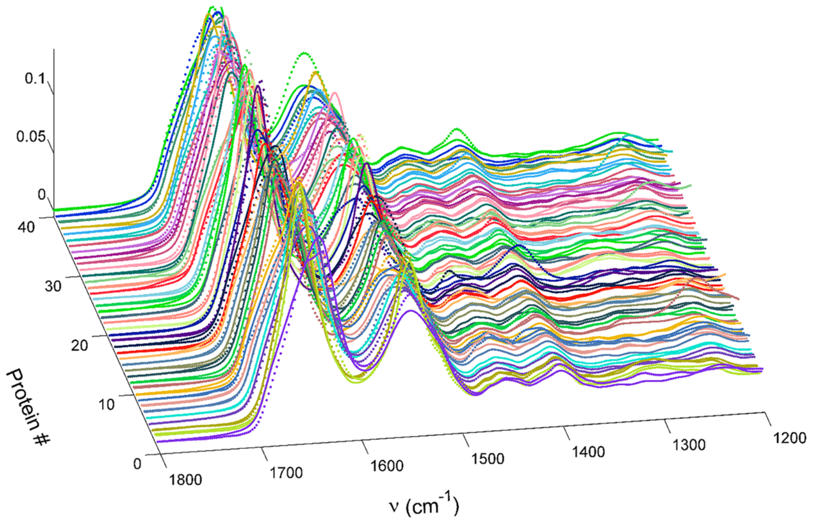


Figure 4. Fits of the normalized input library spectra with the covariant fitting option. The input **Y** (dots) and calculated **Ŷ** (lines) protein library spectra. The average error at 1654 cm^{-1} was 11.0%. The quality of the fit varied by less than 0.5% with the three weighting schemes.

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{B} \tag{3}$$

where the matrix **B** contains the IR spectra of the three groups of protein secondary structures as rows, just as the **Y** matrix (defined earlier) contains the library protein IR spectra as rows. The number of rows in both **X** and **Y** is the number of library proteins ($m_s = 40$), while the number of columns in **Y** and **B** is the number of steps in the IR spectra ($n = 301$) in this work. There exist a variety of multivariate statistical analyses[4,16−23,34] for extracting information from IR spectra, however the strength of this work arises from its connection to the Ramachandran plot, not the mathematics. Its validity follows from three stages of error analysis, including calculations without weights, using weights from the baselines of the library input spectra, and using covariance between the input library spectra which are highly correlated. The general least-squares solution to eq 3 in matrix form is

$$\hat{B} = (\mathbf{X}^{\mathrm{T}} \cdot \mathbf{W} \cdot \mathbf{X})^{-1} \cdot \mathbf{X}^{\mathrm{T}} \cdot \mathbf{Y} \tag{4}$$

where the ’hat’ indicates a fitted result and **W** is the weighting matrix which is a square matrix of dimension m_s x m_s. The matrix **W** equals the identity matrix for unweighted least-squares (**W** = **I**) and it has the reciprocal of each library spectrum’s variance for weighted least-squares

$$\mathbf{W} = \begin{bmatrix} \frac{1}{{\sigma_1}^2} & 0 & \ldots & 0 \\ 0 & \frac{1}{{\sigma_1}^2} & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \ldots & \frac{1}{{\sigma_{m_s}}^2} \end{bmatrix} \tag{5}$$

These weights were chosen for the library spectra by calculating the standard deviation of the baseline noise, σ_i, of each library spectrum in a baseline region from 1840 to 1920 cm^{-1} of the normalized spectrum. The values of σ_i range from 0.00003 to 0.00029 in absorbance units of the normalized library spectra. These can be compared to the normalized average absorbance at 1654 cm^{-1} of 0.16 normalized absorbance units (see Figure 2a) giving errors of ∼0.09% for the amide I band of the input library spectra. The most general least-squares approach takes account of the significant correlation between the input library spectra. The correlation between different pairs of input library spectra varies from 0.802 to 0.998. This case is called a general least-squares problem or a least-squares fit with covariance. In such a case, **W** is a nondiagonal matrix with correlation coefficients between each pair of library protein spectra in the off-diagonal positions. The general least-squares problem is solved formally by decomposing the **W** matrix into two matrices by QR factorization, which in turn are used to reweight the **X** and **Y** matrices in such a way that the whole problem can be rewritten as a simple least-squares[35,36] (MATLAB’s “lscov.m” routine). Once the results of eq 4 are calculated for any of the three options with **W**, then the library protein spectra are calculated with **Ŷ** = **X**·**B̂**, where the “hats” in general indicate fitted or calculated values. For example, **Y** contains the input protein spectra, while **Ŷ** indicates the fitted spectra as calculated using **B̂**. Since both **X** and **Y** are normalized quantities, it can be presumed that the output group spectra **B̂** are also normalized. In fact, the use of a group with small fractions does produce a raw solution with high absorbance. The raw solutions and their errors have been multiplied by the amino acid weighted fractions of the corresponding secondary structure groups to compensate for this effect.

There are error statistics to consider for the fitting of the spectra of both the library proteins and the protein secondary structure groups. The error statistics for the library spectra involve the *rows* of **Y** and **Ŷ** and the variances for each library spectrum are

$${\sigma_{Y,i}}^2 = \frac{[\mathbf{Y}(i,:) - \hat{Y}(i,:)] \cdot [\mathbf{Y}(i,:) - \hat{Y}(i,:)]^T}{m_s - n_g} \tag{6}$$

where $i = 1, 2, ..., m_s$ is an index over the library spectra. The notation $(i,:)$ means all of the elements across row i, so this amounts to a sum of the errors squared across the IR spectrum for each library protein. The error statistics for the fitted group spectra **B̂** of protein secondary structures involve the *columns* of **Y** and **Ŷ** and are given as a mean square of errors at each wavenumber in the spectrum as

DOI: 10.1021/acs.jpcb.5b08052
J. Phys. Chem. B 2015, 119, 13079−13092

$$mse_j = \frac{\mathbf{Y}(:,j)^T \cdot \mathbf{W} \cdot \mathbf{Y}(:,j) - \mathbf{Y}(:,j)^T \cdot \mathbf{W} \cdot \hat{Y}(:,j)}{n - n_g} \quad (7)$$

where $j = 1, 2, ..., n$ is an index for the wavenumbers in the spectrum. The notation $(:,j)$ means all of the elements down the column j, so this is an assessment across the library proteins at each wavenumber. The variance-covariance matrix for the $\hat{\mathbf{B}}$ parameters is calculated at each wavenumber step of the spectrum as

$$\hat{V}_j = (\mathbf{X}^T \cdot \mathbf{W} \cdot \mathbf{X})^{-1} mse_j \quad (8)$$

where again $j = 1, 2, ..., n$ which steps through the wavenumbers. The estimated standard deviations of the fitted spectra of protein secondary structures are obtained from the square root of the diagonal elements of $\hat{\mathbf{V}}_j$ at each wavenumber (index j steps through wavenumbers). To summarize, the inputs are **X**, **Y**, **W** and the outputs are $\hat{\mathbf{B}}$ and $\hat{\mathbf{Y}}$ and their errors. The MATLAB programs and input data to perform the above calculations have been provided as Supporting Information to this paper.

IR Spectra of STRIDE Groups. Calculations were performed by programming eqs 3-8 directly into MATLAB and then checked[35,36] with the MATLAB "lscov" function which uses QR factorization (without the need to calculate inverses) to get $\hat{\mathbf{B}}$, the errors in $\hat{\mathbf{B}}$, and mse_j. In this work, both methods got exactly the same answers. Calculations were first pursued with two groups (STRIDE α-helix and "other") using all three weighting schemes (unweighted, weighted, and covariant). Errors in both the fitted library spectra eq 6 and the extracted group spectra (eqs 7 and 8) warrant consideration. The fitted library spectra for the covariant case are shown in Figure 4. The symbols represent the normalized input spectra (**Y**) and the lines are the fits ($\hat{\mathbf{Y}}$). While spectral errors were calculated at all wavelengths, the largest errors averaged to 11.0% at the maximum of the amide I band for the covariant fit, 10.9% for the weighted fit, and 10.5% for the unweighted fit, i.e. the library fit errors varied little with the weighting scheme. The fitted group spectra are shown in Figure 5 for the three weighting cases. The error at 1654 cm^{-1} is 2.9% for unweighted, 3.8% for weighted, and 0.9% for covariant. Weighting matters more for the group spectra results ($\hat{\mathbf{B}}$) than the fitted spectra ($\hat{\mathbf{Y}}$) and the spectrum of STRIDE α-helix group ("H") is very well determined. This might be expected since the STRIDE α-helix distribution of torsional angles is intense and narrow. The α-helix group spectra have prominent peak maxima at 1656 and 1548 cm^{-1} for the amide I and II bands which agree with many tabulations in the literature.[1,2,13,16,37,38]

Two group calculations were repeated for the combination of the STRIDE "E" and "B" groups which we have called β-sheet, but it represents both β-sheets and twisted β-strand structures. The results are presented in Figure 6. The error at 1638 cm^{-1} is 3.0% for unweighted, 3.3% for weighted, and 0.8% for covariant. Again, the spectra are very well determined, the covariant result has the lowest error, and the STRIDE merged group ("E" and "B") is dominated by a single torsional angle distribution. The β-region spectra have a prominent amide I peak maxima at 1638 cm^{-1} agreeing with many tabulations in the literature[1,2,13,16,37,38] as well as another band extending from 1670 to 1688 cm^{-1} which is dramatically different than the IR spectrum of α-helix. The IR spectrum of β-sheet exhibits an excitonic splitting in the amide I band. This splitting has been denoted $2D_{01}$ (due to through space coupling of transition moments between H-bonded strands[12]) and has an average value of 40 cm^{-1} which may vary from 34 to 52 cm^{-1} over this library of proteins, i.e. $D_{01} \approx 20$ cm^{-1}, varying from 17 to 26 cm^{-1}. The amide II band also shows a splitting of about 40 cm^{-1}.

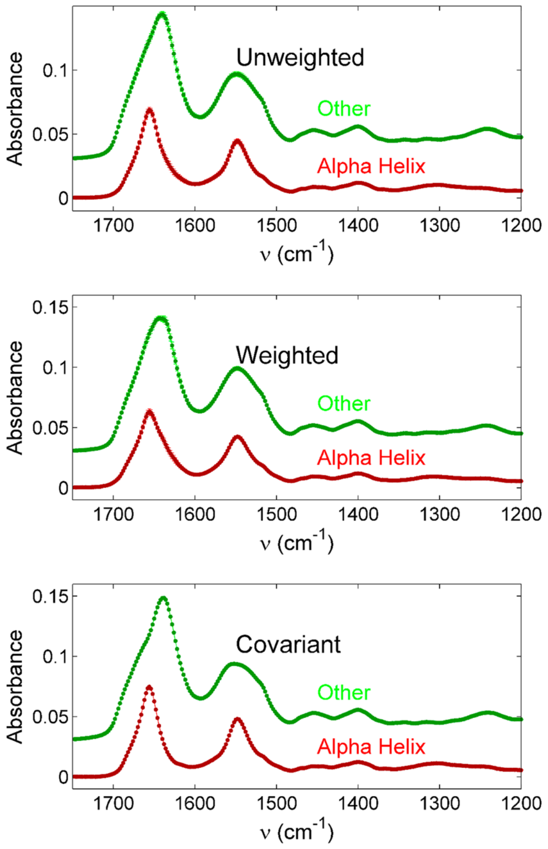


Figure 5. Two group fits of the STRIDE α-helix and all other groups as "other". The top results are unweighted, middle is weighted, and the bottom is covariant. The errors are plotted as error bars, but they are small in all three weighting cases. Recall that this result includes the contributions of the turns in the α-helix spectrum.

Unfortunately, none of the other STRIDE groups work very well with two group fits. This result arises from the existence of multiple distributions of torsional angles (and therefore different contributing IR spectra to the group spectra) as is evident from Figure 2. One might expect that a three group analysis would work easily, since the α-helix and β-sheet spectra are so well determined, however the errors and results change fairly dramatically with more groups as will be shown in the next section.

Three Group Analysis for Spectral Basis Functions. Our primary motivation in undertaking this project was that these results might enable distinction of α-helix and β-sheet content of IR protein spectra in general. Lee et al.[17] in 1990 on an IR library of 18 proteins, Dousseau and Pézolet[4] in 1990 with a library of 13 proteins, Sarver and Krueger[21] in 1991 with 10 proteins (combined with circular dichroism data), Rahmelow and Hüber[19] in 1996 on a library of 39 proteins with cross-validation analysis, Cai and Singh[18] in 2004 with a library of 18 proteins, and Navera, Tauler, and de Juan[39] in 2005 with 24 proteins all have demonstrated the ability to predict α-helix and β-sheet content using multivariate methods on IR protein libraries. However, it would be difficult for a

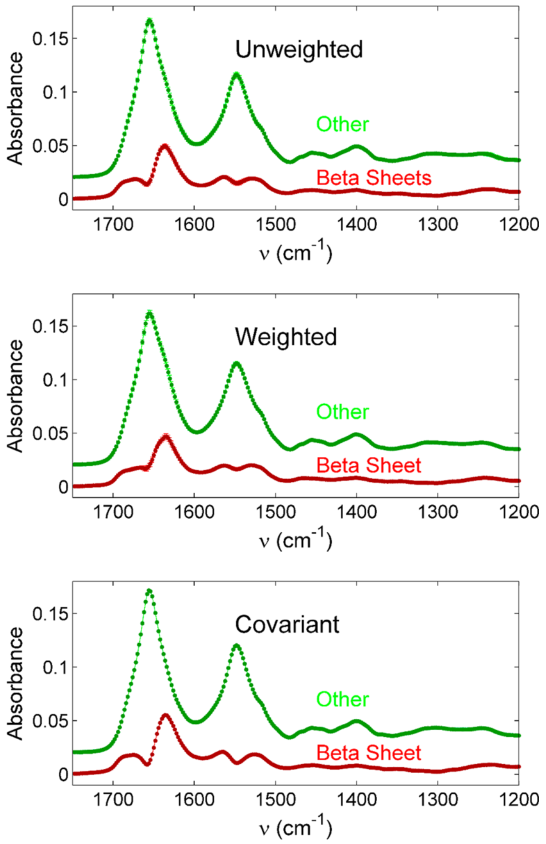


Figure 6. Two group fits of the combined STRIDE "E" and "B" groups vs all other groups as "other". The top results are unweighted, middle is weighted, and the bottom is covariant. The errors are plotted as error bars, but they are small in all three weighting cases. Recall that this result includes the contributions of the turns at the ends of the β-sheets.

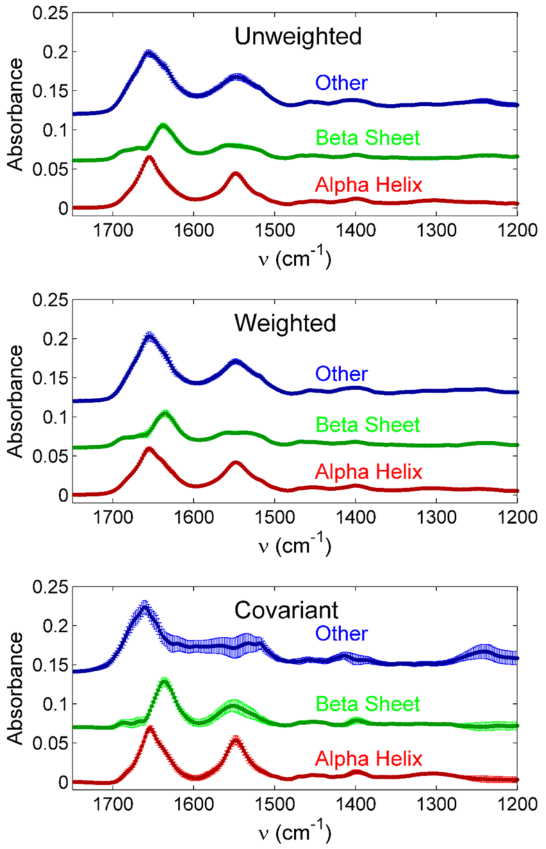


Figure 7. Three group fits with unweighted (at top), weighted (middle), and covariant (bottom) results including error bars. The error is now bigger with the covariant method and the best result is the unweighted result at top.

reader to employ the above-mentioned results to their own work as they would likely need digital copies of the secondary group spectral determinations. So, three group analyses were performed in order to obtain a set of spectral functions that others could use to analyze for α-helix and β-sheet content of their own IR protein spectra.

A three group analysis with the STRIDE α-helix as one group, the combined STRIDE "E" and "B" groups as a second (β-region), and with all of the rest as the third is shown in Figure 7. The errors at 1654 cm^{-1} for the α-helix group are 3.4%, 4.7%, and 5.7% for unweighted, weighted, and covariant, respectively. With two group analysis, the error was 0.9% for covariant weights. The errors at 1638 cm^{-1} for the β-sheet group are 7.2%, 9.3%, and 9.3% for unweighted, weighted, and covariant, respectively. Previously with two groups, the error was 0.8% for covariant weighting. The three group covariant errors are considerably bigger for the "other" group, the group spectra change more with weighting, and all of the errors are bigger. In fact, the errors are unacceptably large for all attempts (to date) with more than three groups.

The unweighted results in Figure 7 (top) have potential utility in characterizing changes in protein secondary structure, so digital files of the unweighted group spectra of Figure 7 and their errors are provided in the Supporting Information for use as spectral basis functions for protein characterization. Considering that the results in Figure 7 were obtained by modeling the library protein spectra, then the group spectra should serve as a set of spectral basis functions for least-squares analysis of α-helix, β-sheet, and "other" content of general protein spectra. This idea of analyzing for α-helix and β-sheet structure was tested by performing a nonnegative least-squares fit (using MATLAB's lsqnonneg function) of each of the library protein IR spectra using the spectral group basis functions (of Figure 7 top), $\mathbf{y}_i = f_\alpha \mathbf{b}_\alpha + f_\beta \mathbf{b}_\beta + f_O \mathbf{b}_O$, where $\mathbf{b}_\alpha$,$\mathbf{b}_\beta$, and $\mathbf{b}_O$ are the spectral basis functions and f_α, f_β, and f_O are the fit parameters. The constraint of nonnegativity was important to avoid negative fractions and the fitted coefficients were normalized so that they summed to one for comparison to the fractions given in Table 1. Plots of the fitted fractions vs the known fractions (Table 1) for α-helix, β-sheet, and "other" are given in Figure 8 which uses a number to identify each library protein. Each of these sets was fit to a line (blue curves) as a guide for comparison to the ideal result of a diagonal line extending from (0,0) to (1,1). The slopes and intercepts (given in the Figure 8 caption) are in fact close to one and zero, respectively. There is some spread in the results (the standard deviation of the difference between the Table 1 fractions and the nonnegative and normalized fitted fractions was 0.17 for α-helix and 0.13 for the β-sheet and 0.18 for the "other" group). These spreads are perhaps a little larger than literature reports, but they correspond to a larger library. In spite of some spread, these spectral functions are clearly capable of discerning α-helix, β-sheet, and "other" content.

Application for Liver Cancer Detection. Given the success illustrated in Figure 8, the unweighted group spectra of Figure 7 (top) were used as calibrant spectra or spectral basis

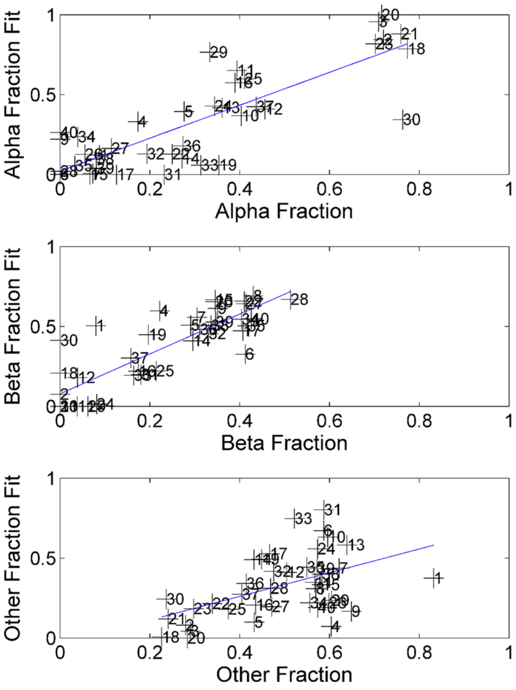


Figure 8. Unweighted, three group results from Figure 7 were used as basis functions for a nonnegative least-squares fit of each library IR spectrum. The fitted fractions are plotted against the known fractions from Table 1 with numbers from Table 1 to identify the proteins. The data were fit to a line (blue curves) for comparison to the ideal correlation trends which would extend from (0,0) to (1,1). The fitted α fraction line is 1.031 (α fraction) + 0.021, the β region line is 1.247 (β fraction) + 0.078, and the "other" line is 0.741 (other fraction) − 0.036, where the true slope would be 1 and the true intercept zero.

functions in analyzing imaging FTIR spectra of a human liver tissue slice containing a tumor. Since the liver interfaces between the digestive and circulatory systems, it is a frequent site for metastases from cancers in other organs. Considering that IR spectroscopy is sensitive to molecular level biochemical changes, nondestructive, and involves no labeling or staining, it has potential to someday be used as a real-time intraoperative diagnostic tool.[40] The subject tissue contained a liver metastasis of colorectal origin which was surgically removed from a consenting patient (IRB no. 2011C0085) at the time of a planned liver resection at the University Hospital (Ohio State University, Columbus, OH). The tissue was snap frozen in liquid nitrogen without formalin fixation or dehydration procedures. A cryostat section of ∼2−3 μm thickness was obtained at −20 °C. The acquisition of data has been previously described[41,42] and briefly involves collection of an FTIR spectrum at every 6.25 μm by 6.25 μm pixel of liver tissue in an area of 2200 μm by 1200 μm, i.e., spectra at each of 67 584 image pixels. The spectra were recorded with a PerkinElmer Spotlight 300 imaging FTIR with a 16 element MCT array detector, 4 cm^{-1} resolution, 750−4000 cm^{-1} range, and 16 scans per pixel, which required about 12 h of scanning.

A hematoxylin and eosin (H&E) stained image from an optical microscope of the tissue was obtained after the IR imaging and is shown in the top left of Figure 9 (top left). The dark purple and light pink glandular structure reveals a tumor in the bottom half of the image, the less textured region of orange/pink color indicates the nontumor region in the top half, and the purple regions in the top half are rich in lymphocytes indicating strong immunological response toward the cancer. Both the IR spectra at each image pixel (all 67 584 pixels) and the α helix and β region IR spectral basis functions from Figure 9 (top left) were normalized over the region from 1200 to 1800 cm^{-1} so that the inner product of any spectrum with itself was one. Scores at each pixel were obtained from the inner product of each normalized liver tissue pixel spectrum with the normalized spectra of α-helix and then the β-sheet region [i.e., normalized versions of the spectra in Figure 7 (top)]. The contour plot of at the bottom of Figure 9 reveals the occurrence of scores that fall within two-dimensional bins of the α-helix and β-sheet region scores. The contour plot is highly structured and separates tumor and nontumor regions. Important regions in the contour plot were encircled with colored ellipses. Image pixels with scores falling within the ellipse were plotted with the corresponding color in the top

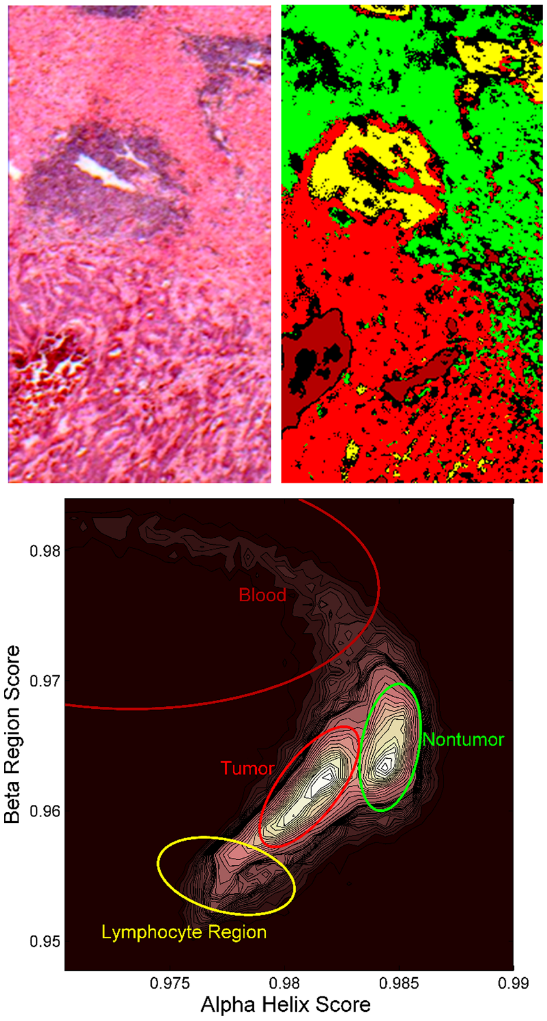


Figure 9. Optical microscope image (top left) of an H&E stain of human liver tissue shows a tumor in the bottom half of the image. Before staining, imaging FTIR spectra were recorded at each pixel in this 2.2 × 1.2 mm region. At bottom is a contour plot of the occurrence of α helix and β region scores for all pixels in the image, where the scores are normalized dot products of the IR pixel spectra and the basis spectra of α-helix and the β-region from Figure 4. Colored ellipses on the contour plot indicate regions of interest and images pixels having scores that fall within the colored ellipses have been plotted with the same color in the top right image. There is excellent separation of the tumor and nontumor regions using the 3 group, basis spectra of α-helix and the β-region from Figure 4.

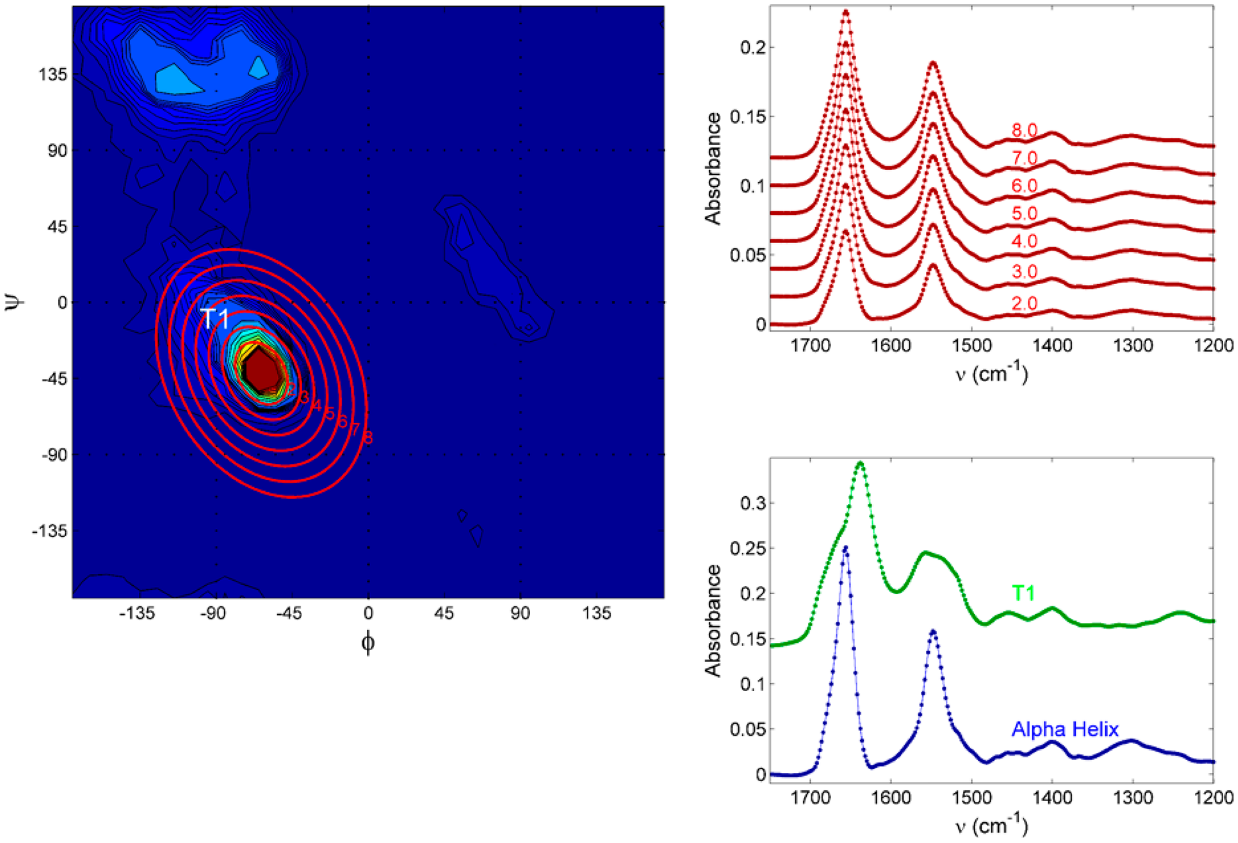


Figure 10. Ramachandran plot on the left has ellipses (red) centered on the α-helix distribution at 2, 3, 4, 5, 6, 7, and 8 standard deviations in width. An IR spectrum was extracted for each of these ellipses (top right) which have subtle changes as they include more of the T1 distribution with increasing ellipse size. A least-squares decomposition was performed (bottom right) using torsional angle distributions to divide the inside ellipse fractions into α-helix and T1 which reveals the nature of the IR spectrum of the T1 distribution, while also getting an α-helix spectrum in good accord with earlier work. The T1 spectrum has peaks 1638, 1556, 1452, 1400, and 1240 cm^{-1}, but it is distinguished by being broader with many unresolved peaks.

right image of Figure 9 where red is tumor, green is the nontumor region, yellow indicates the lymphocyte rich regions, and dark red indicates a blood rich region.

The α-helix and β-sheet region basis functions (from Figure 7) provide excellent separation of tumor and nontumor regions. They also reveal regions of strong immunological response, as well as blood rich regions. Considering that normal liver protein is dominated by albumin (roughly 80%) which is an α-helix dominated protein, it should not be surprising that the protein changes associated with a liver tumor involve a decrease in α-helix protein. This is definitively indicated by the tumor region occurring to lower values of the normalized α-helix score. While there are many other interesting effects to be explored, it is clear that the results of the STRIDE groups for α-helix and β-sheet region have utility in diagnosing protein changes associated with cancer.

IR Ramachandran Ellipse Method. The overlapped and multiple torsional angle distributions of STRIDE groups make it much more difficult to extract further information from the protein library. A method, called the Ramachandran ellipse method, has been developed to begin to deal with these complexities. The user specifies the constants of a parametrized ellipse on the Ramachandran plot in order to define new group fractions for amino acids falling within the ellipse. Note that fractions now come from the ellipse rather than STRIDE definitions. The ellipse is defined with the parameter t which varies from 0 *to* 2π in radians:

$$\phi = \phi_c + a\cdot\cos(t)\,\cos(-\theta) - b\cdot\sin(t)\,\sin(-\theta)$$

$$\psi = \psi_c + a\cdot\cos(t)\,\sin(-\theta) + b\cdot\sin(t)\,\cos(-\theta) \qquad (9)$$

where φ_c, ψ_c is the ellipse center, $2a$ and $2b$ are the major and minor axis widths respectively, and θ is the tilt axis of the ellipse's major axis to the φ axis of the Ramachandran plot. A Matlab routine loops through the library protein data and counts the amino acids that fall within the ellipse for each protein. An amino acid is inside an ellipse if the following criterion is true

$$\left[\frac{(\phi - \phi_c)\cos(\theta) + (\psi - \psi_c)\sin(\theta)}{a}\right]^2 + \left[\frac{(\phi - \phi_c)\sin(\theta) - (\psi - \psi_c)\cos(\theta)}{b}\right]^2 \leq 1 \qquad (10)$$

The trick is to choose a set of ellipses which yield all positive IR group spectra while the variation of fractional contributions of STRIDE groups and/or T1 and T2 groups is knowable. This is accomplished by centering ellipses on the two biggest distributions. Figure 10 shows ellipses (red) centered on the STRIDE α-helix group at multiples of 2, 3, 4, 5, 6, 7, and 8 of the distribution standard deviations which are used to define fractions of amino acid for each library protein for inside and

Table 3. Fractions of α-Helix, 3_{10}-Helix, and T1 That Fall Within the Ramachandran Plot Ellipses of Figure 10 As Determined by Numerical Integration[a]

	factor						
	2	3	4	5	6	7	8
α-helix	0.8916	0.8232	0.7635	0.7219	0.6935	0.6750	0.6494
3_{10}-helix	0.0319	0.0528	0.0623	0.0635	0.0620	0.0605	0.0582
T1	0.0765	0.1240	0.1738	0.2146	0.2445	0.2645	0.2925

[a] The ellipses are centered on the α-helix distribution and have widths that are multiple factors of the α-helix standard deviations (9.829° and 6.935°) along the major and minor axes, respectively.

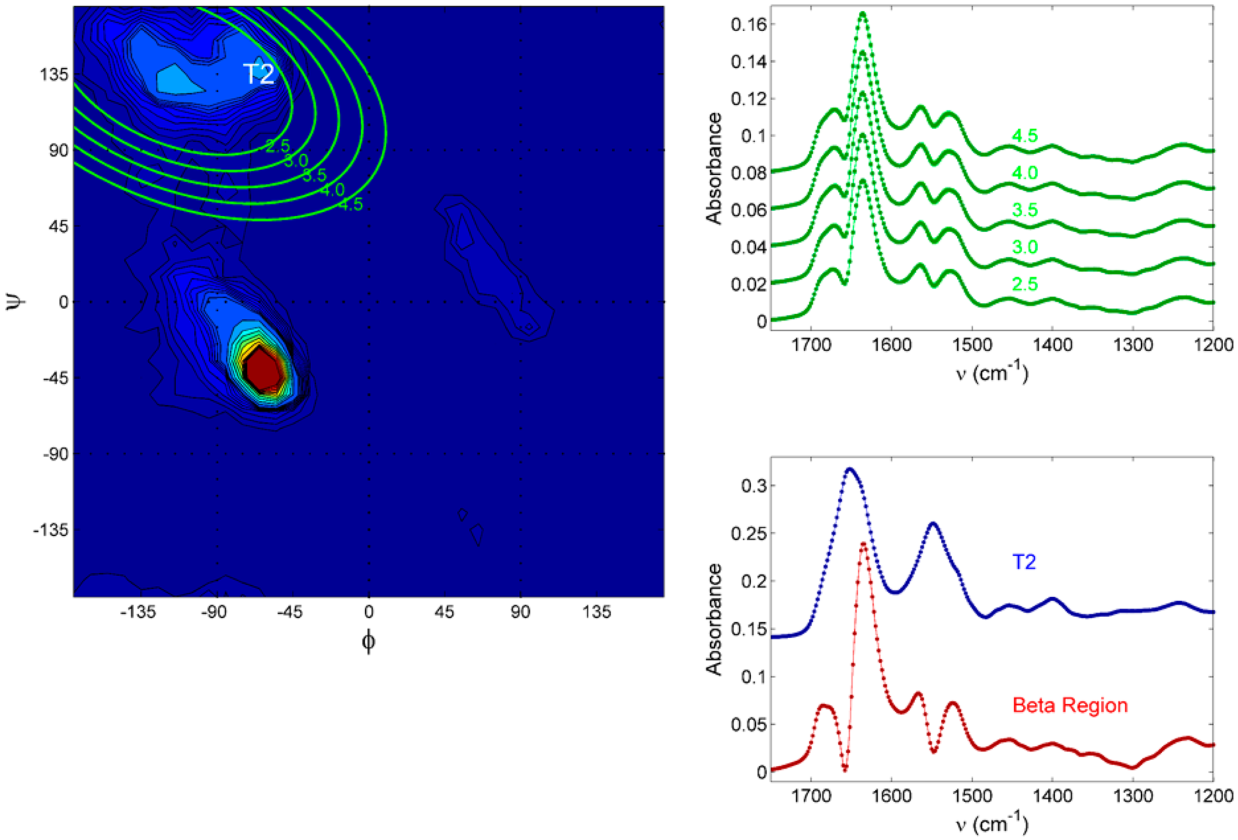


Figure 11. Ramachandran plot on the left has ellipses (green) centered on the β-sheet region distribution at 2.5, 3.0, 3.5, 4.0, and 4.5 standard deviations in width. An IR spectrum was extracted for each of these ellipses (top right) which have subtle changes as they include more of the T2 distribution with increasing ellipse size. A least-squares decomposition was performed (bottom right) using fitted torsional angle distributions to divide the inside ellipse fractions into β-sheet region and T2 revealing the nature of the IR spectrum of the T2 distribution, while also getting an β-sheet region spectrum in good accord with earlier work. The T2 spectrum has peaks 1652, 1548, 1454, 1400, and 1244 cm^{-1}, but it is also distinguished by being broader with many unresolved peaks.

outside of the ellipse. An IR spectrum was extracted for each ellipse using these fractions and a two group method as described earlier, i.e. by solving eq 3, but with fractions from ellipses on the Ramachandran plot. The resulting "inside" of the ellipse spectra are shown at the top right of Figure 10. Note that the spectra generally resemble the α-helix spectra of Figure 5, but they get a bit broader as the ellipse gets bigger due primarly to contributions from the T1 group. Finally, we use the fits of Figure 3 and Table 2 and numerical integration to define new fractions of each group distribution that are within the ellipse, i.e. torsional angle fits determine a breakdown of the ellipse fractions into two groups. For example, the smallest ellipse in Figure 10 contains 0.891 α-helix, 0.032 3_{10}-helix, and 0.077 T1, while the largest ellipse has 0.675 α-helix, 0.060 3_{10}-helix, and 0.265 T1 as shown in Table 3. The range of meaningful change for these factors starts when all-positive IR spectra are obtained, due to domination by the prominent group, and ends when the less prominent group is largely encompassed. The approach uses the same eqs 3-8 to extract group spectra, however the input data are the spectra on the top right of Figure 10 and the fractions are given in Table 3. We were not able to extract the spectrum of 3_{10}-helix due to its small fractional representation, so its fraction was merged with the α helix to obtain the results shown in the bottom right of Figure 10. This shows a group spectrum dominated by α-helix (blue), and a broader spectrum (green) attributed to the T1 distribution. Recall that both the STRIDE coil and turn ("C" and "T") groups have prominent contributions from T1. This spectrum is no doubt broader than the α-helix and β-sheet spectra from earlier sections because it corresponds to a more heterogeneous distribution of torsional angles. The same approach was applied to the β-sheet and T2 regions in Figure

11 where again the T2 region has major contributions from both the STRIDE coil and turn ("C" and "T") groups, but with different proportions. The ellipses are centered on the β-sheet distribution and the standard deviations along the major and minor axes are multiplied by factors of 2.5, 3.0, 3.5, 4.0, and 4.5. The smallest ellipse is 73% β-sheet and 27% percent T2 which varies to 60% β-sheet and 40% percent T2 with the largest ellipse as shown by the fractions in Table 4. The method produces group spectra of the β-sheet region with an excitonic splitting similar to Figure 6 and isolates the IR spectrum of the T2 region. The T2 spectrum is broad and more similar to the T1 spectra than the spectra of α-helix or β-sheet.

Table 4. Fractions of β-Sheet Region and T2 Distributions That Fall Within the Ramachandran Plot Ellipses of Figure 11 As Determined by Numerical Integration[a]

	factor					
	2.0	2.5	3.0	3.5	4.0	4.5
β-sheet	0.7301	0.6722	0.6308	0.6094	0.6009	0.5984
T2	0.2699	0.3278	0.3692	0.3906	0.3991	0.4016

[a] The ellipses are centered on the β-sheet region distribution and have widths that are multiple factors of the β-sheet region standard deviations (29.549° and 16.628°) along the major and minor axes, respectively.

■ CONCLUSIONS

The interplay between torsional angle distributions (which are strongly correlated with IR spectra) and traditional secondary structure determinations (torsional angles and H-bonding) are addressed by use of the Ramachandran plot which is the most distinguishing feature of this work. Unweighted, weighted, and covariant least-squares approaches have been compared which relate the fraction of protein secondary structures in a library of proteins to the IR spectra of each protein. The method obtains IR spectra of the secondary structure groups as defined by a structural method, STRIDE, which considers both hydrogen bonding and torsional angles—not just the torsional angles as displayed by a Ramachandran plot. The approach yields the whole spectrum of prominent secondary structure regions including regions where different secondary structures share similar spectral regions—if the torsional angles are dominated by one distribution, such as for α-helix and β-sheet.

While the relationship between fractions and IR spectra is straightforward, $\mathbf{Y} = \mathbf{X}\cdot\mathbf{B}$, an extensive error analysis including both the fitting of the library IR spectra $\hat{\mathbf{Y}}$ and the errors in the predicted spectra of secondary groups $\hat{\mathbf{B}}$ was required to inspire confidence in the results. Basically, the covariant method works best for two groups, can be extended with less confidence to three groups, and becomes unwieldy for more than three groups. At three groups, we found that the unweighted results had the lowest error and these results were used for a set of spectral basis functions to be used in protein analysis. The three group spectral basis set of α-helix, the β-sheet region, and "other" (as given in Figure 7 top) has the ability to detect α-helix, β-sheet region, and "other" contributions to the IR spectra of proteins as shown in Figure 8. This ability was applied to the detection of liver cancer in Figure 9. Scores were obtained by taking normalized dot products of the basis set spectra of α-helix, the β-sheet region, and "other" with measured IR spectra from the pixels of a human liver sample as recorded with an imaging FTIR microscope. There is great potential to characterize and interpret protein changes associated with liver tumors[41] and thin tissue slices in general.

An extension of the technique to ellipses on a Ramanchandran plot enables the analysis to extend to groups with multiple and overlapped structural contributions. Torsional angle distributions have been identified and called T1 and T2 in this work, but they do not correspond to single STRIDE secondary structure groups. Basically, T1 and T2 are the most prominent torsional angle distributions in the STRIDE secondary structure groups called coils and turns ("C" and "T"). The STRIDE coil or turn groups get different weighted spectral contributions from T1 and T2 as well as contributions from a variety of other smaller torsional angle distributions (see Figure 2 middle). The most important covariant extractions in this work are given in Figure 12 which includes the two group extraction of the STRIDE α-helix spectrum, the two group extraction of the STRIDE β-sheet region, as well as the IR Ramachandran ellipse method extractions of the T1 and T2 regions which make the biggest spectral contributions to both the STRIDE coil and turn groups ("C" and "T"), but are not themselves STRIDE groups, i.e. the IR spectra of the STRIDE Coil and Turn groups will have major contributions from both T1 and T2, but with different weights.

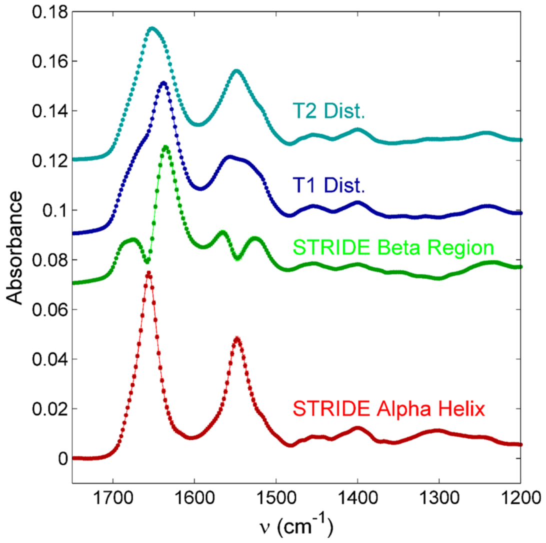


Figure 12. Summary of IR spectra extracted in this work. The α-helix and β-sheet regions were two-group, covariant extractions using STRIDE definitions, and the T1 and T2 groups are torsional distributions from the Ramachandran plot that are the major components of the STRIDE coil and turn groups which consist of a variety of torsional angle distributions.

All of this work was accomplished with a library of only 40 IR protein spectra, so larger and more varied protein libraries would likely help to improve the statistics of the error analysis. Having 10 times as many library spectra might enable extraction of IR spectra of groups at the 5% level, like the 3_{10}-helix. One might also choose library proteins that are better suited to a specific task, for instance cancer work might profitably include collagens and mucins in the protein library. Considering that the current work averages over, or averages away, the effect of amino acid side chain groups, future work with bigger protein libraries might be able to isolate these contributions provided that one includes sufficient examples of proteins enriched in amino acids with a particular side chain in

DOI: 10.1021/acs.jpcb.5b08052

the training library. The approach might also be profitably combined with other spectroscopies like circular dichroism,[21,23,24] Raman,[43] or attenuated total reflection IR data.[44] In general, future work will include the development of strategies for extracting IR spectra for less common structures in order to obtain new IR metrics for tissue protein diagnostics.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpcb.5b08052.

> MATLAB programs and corresponding input data files, illustrating the covariant least-squares fitting of protein secondary groups for the two group case (ZIP)

AUTHOR INFORMATION

Corresponding Author

*(J.V.C.) E-mail: coe.1@osu.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the National Cancer Institute and the National Institutes for Health for Grant NIH R21 CA167403.

REFERENCES

(1) Barth, A. Infrared Spectroscopy of Proteins. *Biochim. Biophys. Acta, Bioenerg.* **2007**, *1767*, 1073−1101.

(2) Tatulian, S. A. Structural Characterization of Membrane Proteins and Peptides by FTIR and ATR-FTIR Spectroscopy. In *Lipid-Protein Interactions: Methods and Protocol*, Kleinschmidt, J. H., Ed.; Springer Science+Business Media: New York, 2013; Vol. *974*.

(3) Jackson, M.; Mantsch, H. H. Protein Secondary Structure from FT-IR spectroscopy: Correlation with Dihedral Angles from Three-dimensional Ramachandran Plots. *Can. J. Chem.* **1991**, *69*, 1639−42.

(4) Dousseau, F.; Pezolet, M. Determination of the Secondary Structure Content of Proteins in Aqueous Solutions from Their Amide I and Amide II Infrared Bands. Comparison Between Classical and Partial Least-Squares Methods. *Biochemistry* **1990**, *29*, 8771−9.

(5) Chiriboga, L.; Yee, H.; Diem, M. Infrared Spectroscopy of Human Cells and Tissue. Part VII: FT-IR Microspectroscopy of DNase- and RNase-treated Normal, Cirrhotic, and Neoplastic Liver Tissue. *Appl. Spectrosc.* **2000**, *54* (4), 480−485.

(6) Chiriboga, L.; Yee, M.; Diem, M. Infrared Spectroscopy of Human Cells and Tissue. Part VI: A Comparative Study of Histopathology and Infrared Microspectroscopy of Normal, Cirrhotic, and Cancerous Liver Tissue. *Appl. Spectrosc.* **2000**, *54* (1), 1−8.

(7) Diem, M.; Chiriboga, L.; Yee, H. Infrared Spectroscopy of Human Cells and Tissue. VIII. Strategies for Analysis of Infrared Tissue Mapping Data and Applications to Liver Tissue. *Biopolymers* **2000**, *57* (5), 282−290.

(8) Wu, H.; Canfield, A.; Adhikari, J.; Huo, S. Quantum Mechanical Studies on Model α-pleated Sheets. *J. Comput. Chem.* **2009**, *31* (6), 1216−1223.

(9) Welch, W. R. W.; Keiderling, T. A.; Kubelka, J. Structural Analyses of Experimental 13C Edited Amide I′ IR and VCD for Peptide β-Sheet Aggregates and Fibrils Using DFT-Based Spectral Simulations. *J. Phys. Chem. B* **2013**, *117* (36), 10359−10369.

(10) Kubelka, J.; Keiderling, T. A. Differentiation of β-Sheet-Forming Structures: Ab Initio-Based Simulations of IR Absorption and Vibrational CD for Model Peptide and Protein β-Sheets. *J. Am. Chem. Soc.* **2001**, *123* (48), 12048−12058.

(11) Zanetti Polzi, L. Z.; Daidone, I.; Amadei, A. A Theoretical Reappraisal of Polylysine in the Investigation of Secondary Structure Sensitivity of Infrared Spectra. *J. Phys. Chem. B* **2012**, *116* (10), 3353−3360.

(12) Barth, A.; Zscherp, C. What Vibrations Tell about Proteins. *Q. Rev. Biophys.* **2002**, *35* (04), 369−430.

(13) Dong, A.; Huang, P.; Caughey, W. S. Protein Secondary Structures in Water from Second-Derivative Amide I Infrared Spectra. *Biochemistry* **1990**, *29*, 3303−3308.

(14) Byler, D. M.; Susi, H. Examination of the Secondary Structure of Proteins by Deconvolved FTIR Spectra. *Biopolymers* **1986**, *25*, 469−487.

(15) Rahmelow, K.; Huebner, W. Fourier Self-deconvolution: Parameter Determination and Analytical Band Shapes. *Appl. Spectrosc.* **1996**, *50*, 795−804.

(16) Navea, S.; Tauler, R.; de Juan, A. Application of the Local Regression Method Interval Partial Least-squares to the Elucidation of Protein Secondary Structure. *Anal. Biochem.* **2005**, *336*, 231−242.

(17) Lee, D. C.; Haris, P. I.; Chapman, D.; Mitchell, R. C. Determination of Protein Secondary Structure using Factor Analysis of Infrared Spectra. *Biochemistry* **1990**, *29*, 9185−93.

(18) Cai, S.; Singh, B. R. A Distinct Utility of the Amide III Infrared Band for Secondary Structure Estimation of Aqueous Protein Solutions using Partial Least Squares Methods. *Biochemistry* **2004**, *43*, 2541−2549.

(19) Rahmelow, K.; Hubner, W. Secondary Structure Determination of Proteins in Aqueous Solution by Infrared Spectroscopy: A Comparison of Multivariate Data Analysis Methods. *Anal. Biochem.* **1996**, *241*, 5−13.

(20) Shariati-Rad, M.; Hasani, M. Application of Multivariate Curve Resolution-alternating Least Squares (MCR-ALS) for Secondary Structure Resolving of Proteins. *Biochimie* **2009**, *91*, 850−856.

(21) Sarver, R. W., Jr.; Krueger, W. C. An Infrared and Circular Dichroism Combined Approach to the Analysis of Protein Secondary Structure. *Anal. Biochem.* **1991**, *199* (1), 61−7.

(22) Sarver, R. W., Jr.; Krueger, W. C. Protein Secondary Structure from Fourier Transform Infrared Spectroscopy: A Data Base Analysis. *Anal. Biochem.* **1991**, *194* (1), 89−100.

(23) Compton, L. A.; Johnson, W. C., Jr. Analysis of Protein Circular Dichroism Spectra for Secondary Structure using a Simple Matrix Multiplication. *Anal. Biochem.* **1986**, *155* (1), 155−67.

(24) Navea, S.; Tauler, R.; Goormaghtigh, E.; de Juan, A. Chemometric Tools for Classification and Elucidation of Protein Secondary Structure from Infrared and Circular Dichroism Spectroscopic Measurements. *Proteins: Struct., Funct., Genet.* **2006**, *63*, 527−541.

(25) Haaland, D. M.; Thomas, E. V. Partial Least-squares Methods for Spectral Analyses. 1. Relation to Other Quantitative Calibration Methods and the Extraction of Qualitative Information. *Anal. Chem.* **1988**, *60* (11), 1193−202.

(26) Dong, A.; Huang, P.; Caughey, B.; Caughey, W. S. Infrared Analysis of Ligand- and Oxidation-induced Conformational Changes in Hemoglobins and Myoglobins. *Arch. Biochem. Biophys.* **1995**, *316*, 893−8.

(27) Dong, A.; Huang, P.; Caughey, W. S. Redox-dependent Changes in Beta-sheet and Loop Structures of Cu,Zn Superoxide Dismutase in Solution Observed by Infrared Spectroscopy. *Arch. Biochem. Biophys.* **1995**, *320*, 59−64.

(28) Huang, P.; Dong, A.; Caughey, W. S. Effects of Dimethyl Sulfoxide, Glycerol, and Ethylene Glycol on Secondary Structures of Cytochrome c and Lysozyme as Observed by Infrared Spectroscopy. *J. Pharm. Sci.* **1995**, *84*, 387−92.

(29) Dong, A.; Kendrick, B.; Kreilgard, L.; Matsuura, J.; Manning, M. C.; Carpenter, J. F. Spectroscopic Study of Secondary Structure and Thermal Denaturation of Recombinant Human Factor XIII in Aqueous Solution. *Arch. Biochem. Biophys.* **1997**, *347*, 213−20.

(30) Dong, A.; Caughey, W. S. Infrared Methods for Study of Hemoglobin Reactions and Structures. *Methods Enzymol.* **1994**, *232*, 139−75.

(31) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of Polypeptide Chain Configurations. *J. Mol. Biol.* **1963**, *7*, 95−99.

(32) Frishman, D.; Argos, P. Knowledge-based Protein Secondary Structure Assignment. *Proteins: Struct., Funct., Genet.* **1995**, *23* (4), 566−79.

(33) Hovmoller, S.; Zhou, T.; Ohlson, T. Conformations of Amino Acids in Proteins. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2002**, *58*, 768−76.

(34) Dousseau, F.; Pezolet, M. Determination of the Secondary Structure of Proteins in Aqueous Solution using IR Spectroscopy. *Biochemistry* **1990**, *29*, 8771.

(35) Graybill, F. A. *Theory and Application of the Linear Model*; Duxbury Press: North Scituate, MA, 1976.

(36) Strang, G. *Introduction to Applied Mathematics*; Wellesley-Cambridge Press: Wellesley, MA, 1986.

(37) Arrondo, J. L. R.; Goni, F. M. Infrared Studies of Protein-Induced Perturbation of Lipids in Lipoproteins and Membranes. *Chem. Phys. Lipids* **1998**, *96* (1−2), 53−68.

(38) Kong, J.; Yu, S. Fourier Transform Infrared Spectroscopic Analysis of Protein Secondary Structures. *Acta Biochim. Biophys. Sin.* **2007**, *39*, 549−559.

(39) Navea, S.; Tauler, R.; De Juan, A. Application of the Local Regression Method Interval Partial Least-squares to the Elucidation of Protein Secondary Structure. *Anal. Biochem.* **2005**, *336* (2), 231−242.

(40) Diem, M.; Miljkovic, M.; Bird, B.; Chernenko, T.; Schubert, J.; Marcsisin, E.; Mazur, A.; Kingston, E.; Zuser, E.; Papamarkakis, K.; Laver, N. Applications of Infrared and Raman Microspectroscopy of Cells and Tissue in Medical Diagnostics. *Spectroscopy* **2012**, *27* (5−6), 463−496.

(41) Chen, Z.; Butke, R.; Miller, B.; Hitchcock, C. L.; Allen, H. C.; Povoski, S. P.; Martin, E. W.; Coe, J. V. Infrared Metrics for Fixation-Free Liver Tumor Detection. *J. Phys. Chem. B* **2013**, *117*, 12442−12450.

(42) Coe, J. V.; Chen, Z. M.; Li, R.; Butke, R.; Miller, B.; Hitchcock, C. L.; Allen, H. C.; Povoski, S. P.; Martin, E. W. Imaging Infrared Spectroscopy for Fixation-Free Liver Tumor Detection. *Proc. SPIE* **2014**, *8947*, 89470B.

(43) Tuma, R. Raman Spectroscopy of Proteins: From Peptides to Large Assemblies. *J. Raman Spectrosc.* **2005**, *36* (4), 307−319.

(44) Glassford, S. E.; Byrne, B.; Kazarian, S. G. Recent Applications of ATR FTIR Spectroscopy and Imaging to Proteins. *Biochim. Biophys. Acta, Proteins Proteomics* **2013**, *1834* (12), 2849−2858.