Determination of the secondary structure of proteins in different environments by FTIR-ATR spectroscopy and PLS regression

ARTICLE *in* BIOPOLYMERS · NOVEMBER 2008

Impact Factor: 2.39 · DOI: 10.1002/bip.21022 · Source: PubMed

6 AUTHORS, INCLUDING:

Reinhard Ingemar Boysen
Monash University (Australia)
102 PUBLICATIONS **946** CITATIONS
SEE PROFILE

Bayden R Wood
Monash University (Australia)
146 PUBLICATIONS **2,827** CITATIONS
SEE PROFILE

Mustafa Kansiz
Agilent Technologies
12 PUBLICATIONS **406** CITATIONS
SEE PROFILE

Determination of the Secondary Structure of Proteins in Different Environments by FTIR-ATR Spectroscopy and PLS Regression

Yeqiu Wang,[1] Reinhard I. Boysen,[1] Bayden R. Wood,[2] Mustafa Kansiz,[2] Don McNaughton,[2] Milton T. W. Hearn[1]

[1] *Australian Research Council Special Research Centre for Green Chemistry, Monash University, VIC 3800, Australia*

[2] *Centre for Biospectroscopy and School of Chemistry, Monash University, VIC 3800, Australia*

Received 6 December 2007; revised 3 April 2008; accepted 2 May 2008
Published online 19 May 2008 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/bip.21022

ABSTRACT:

The secondary structures of proteins (α-helical, β-sheet, β-turn, and random coil) in the solid state and when bound to polymer beads, containing immobilized phenyl and butyl ligands such as those as commonly employed in hydrophobic interaction chromatography, have been investigated using FTIR-ATR spectroscopy and partial least squares (PLS) methods. Proteins with known structural features were used as models, including 12 proteins in the solid state and 7 proteins adsorbed onto the hydrophobic surfaces. A strong PLS correlation was achieved between predictions derived from the experimental data for 4 proteins adsorbed onto the phenyl-modified beads and reference data obtained from the X-ray crystallographic structures with r^2 values of 0.9974, 0.9864, 0.9924, and 0.9743 for α-helical, β-sheet, β-turn, and random coiled structures, respectively. On the other hand, proteins adsorbed onto the butyl sorbent underwent greater secondary structural changes compared to the phenyl sorbent as evidenced from the poorer PLS r^2 values (r^2 are 0.9658, 0.9106, 0.9571, and 0.9340). The results thus indicate that the secondary structures for these proteins were more affected by the butyl sorbent, whereas the secondary structure remains relatively unchanged for the proteins adsorbed onto the phenyl sorbent. This study has important ramifications for understanding the nature of protein secondary structural changes following adsorption onto hydrophobic sorbent surfaces. This knowledge could also enable the development of useful protocols for enhancing the chromatographic purification of proteins in their native bioactive states. © *2008 Wiley Periodicals, Inc. Biopolymers 89: 895–905, 2008.*

Keywords: ATR; FTIR; PLS; secondary structure; protein

Correspondence to: Milton T. W. Hearn; e-mail: milton.hearn@sci.monash.edu.au
Contract grant sponsors: Monash Synchrotron Fellowship Research Grant, Australian Synchrotron Program Fellowship Grant

This article was originally published online as an accepted preprint. The "Published Online" date corresponds to the preprint version. You can request a copy of the preprint by emailing the Biopolymers editorial office at biopolymers@wiley.com

INTRODUCTION

Fourier Transform InfraRed (FTIR) spectroscopy is a powerful technique that has been applied to elucidate the secondary structure of polypeptides and proteins.[1–3] In the 1950s, Elliot and Ambrose[4] demonstrated that an empirical correlation existed between the conformation of polypeptides and proteins and their respective amide modes in IR spectra. Later, this correlation was confirmed by theoretical calculations carried out by Krimm and coworkers.[5–8] The main advantages of FTIR spectroscopy over other techniques to determine conformation include the speed of spectral collection (typically from seconds to a few minutes), minimal sample preparation, and the applicability of rapid result processing methods using chemometric techniques.

The major bands obtained in the FTIR spectra of proteins and polypeptides include the amide I (mainly C═O stretching) and amide II (primarily C—N stretching) vibrations.[9–11] The sensitive infrared amide bands corresponding to the conformation of the peptide backbone of proteins (e.g., α-helices, β-sheet) have been established and documented.[9–11] Quantitative methods for the determination of the secondary structural content of proteins in aqueous solution were developed in the early 1990s. These approaches utilized chemometric methods including factor analysis, multiple linear regression,[12] second derivative analysis,[1] classical and partial least-squares methods,[2] Fourier self-deconvolution,[13,14] band curve-fitting procedures,[15] and neural network architectures.[3]

Second derivative analysis permits direct quantitative analysis of the secondary structural components of protein by revealing the bands contained within the original spectrum. Sloping baselines and broad spectral features are essentially removed by taking the second derivative of a spectrum and there is no need for baseline correction in further analysis. Dong[1] determined the relative amounts of the secondary structures of different protein directly from second derivative spectra relying solely on the amide I band by manually computing the areas under the bands assigned to a particular substructure. These results correlated strongly with the values determined for crystal structures using X-ray crystallography. However, the estimated value in terms of relative percentage for the major structure elements differed by ±5% from the values obtained from the X-ray crystal structure.

Dousseau and Pezolet improved the quantitative analysis of IR spectra by applying the partial least squares (PLS) method utilizing the amide I and II bands,[2] whereas Cai and Singh[16] applied this method using the amide I and III bands for several globular proteins of known structure and gained further improvement. These studies showed a strong correlation for α-helical and β-sheet structures with, however, a poor correlation for β-turn and random coiled structures. Multivariate spectral calibrations are now standard methods for performing quantitative spectral analyses in IR spectroscopy. PLS and principal components regression (PCR) have become the most common multivariate methods for quantitative spectral analyses.[2,16–19] In such studies, the PLS method is applied to analyze the correlation between the spectrum of the sample and the reference structural data of the sample and is a much more objective tool compared to the more subjective methods of deconvolution or curve-fitting, both of which can produce artifacts.

Moreover, it can be noted that the PLS method does not require any knowledge about the data, because it deals with errors in both *x*- and *y*-variables.[20] Consequently, it has flexibility to predict and fit to the reference content by adding more principal components (PCs or factors). However, only careful analysis of the regression coefficient plots can ensure that the results obtained are "real" and not erroneous.[21] Erroneous *R*-values can arise if the regression coefficients are not correlated with real secondary structural bands. Others have recognized this problem and developed a classical least squares (CLS)/PLS hybrid algorithm for spectral analysis,[22] which takes advantages of qualitative interpretation from CLS and quantitative analyses from PLS. In this study, we have examined the regression coefficient plots to determine qualitatively whether the spectral information is consistent with known secondary structural band and thus ensure that only the spectral features of the secondary structure being predicted are used in the calculation.

Although IR spectroscopy has been applied extensively to examine the structure of proteins in solution, only a few articles[23–26] have reported the use of FTIR spectroscopy in combination with Attenuated Total Reflectance (ATR) to obtain spectra of adsorbed proteins. In earlier studies, it was found that the IR spectra of proteins adsorbed to hydrophobic silica surfaces could not be interpreted.[27] In this article, we have applied the second derivative method to the qualitative analysis of secondary structures of globular proteins as monitored by FTIR-ATR spectroscopy and then employed a PLS method to quantitatively determine the secondary structure content of proteins in the solid state and when adsorbed onto polymer beads. The major aim of the study was to determine the effect of the surface chemistry of the polymeric beads on the secondary structure of the protein. To date, most published IR-PLS quantitative methods to determine protein secondary structure have been based on the analysis of the protein in aqueous solution.[2,16–19] When FTIR-ATR is used with proteins in aqueous protein solution, two limitations of this approach can arise.[25] First, there can be significant variations in the penetration depth with protein concentration, which results in a difference in the light absorption by water for concentrated and dilute solutions, and is especially evident in the amide I modes, and secondly a very thin layer of protein can adsorb onto the ATR crystal with changed secondary structure to that of the protein in bulk aqueous solution. To avoid the ubiquitous effects of water on the amide I mode, we have thus examined the spectral features of essentially dried proteins absorbed onto the surface of hydrophobic polymeric beads and compared these findings with the X-ray crystallographic structures of these same proteins. Although there have been previous reports that discuss the structural features of proteins adsorbed from the aqueous phase onto polymer substrates as assessed by various techniques, including deuterium exchange electrospray ionization mass spectrometry and Raman spectroscopy,[23,24,26–29]

to our knowledge none have combined IR spectroscopy and PLS to quantitatively analyze proteins adsorbed to hydrophobic sorbents.

Collectively, these findings have important ramifications for understanding the nature of the changes in the secondary structure of proteins following adsorption onto hydrophobic surfaces. The procedures described are rapid, robust, and readily adapted to different classes of proteins and different types of adsorbents. Moreover, the information provides direct insight into the conformational status of the polypeptide or protein when adsorbed onto a nonpolar stationary phase materials, such as those used in Hydrophobic Interaction Chromatography (HIC). This knowledge is anticipated to enable useful protocols to be established for the chromatographic purification of proteins in their native, bioactive states.

MATERIALS AND METHODS

Chemicals and Reagents

Ammonium sulfate (BDH chemicals, Australia Pty.), sodium phosphate and sodium dihydrogen orthophosphate (ICN Biomedicals, ACS reagent grade) were used as supplied. Water was distilled and deionized in a Milli-Q system (Millipore, Bedford, MA). The polymeric butyl and phenyl hydrophobic interaction sorbents were obtained from Tosoh Bioscience (Montgomeryville, PA).

FTIR Spectroscopy Instrumentation

Spectra were obtained at 8 cm^{-1} resolution with 50 coadded scans using a Golden Gate diamond ATR (SPECAC, P/N10500 series) coupled to a Bruker IFS-55 FTIR spectrometer using OPUS 3.0 spectroscopic software (Bruker, Optik, Ettlingen, Germany) and equipped with a liquid-nitrogen-cooled Mercury-Cadmium-Telluride (MCT) detector. The spectral backgrounds were recorded on a clean ATR crystal for measurements with the amorphous proteins, whereas the backgrounds for the solution phase measurements were recorded using water deposited onto the ATR crystal. To assess the spectral reproducibility and to obtain sufficient spectral information for PLS modeling, multiple (8–10) spectra were recorded for each sample. All spectra were preprocessed with the Multiplicative Scatter Correction (MSC) to compensate for baseline effects in the raw data and second derivative spectra were computed to reveal the "hidden" bands in the original spectra. The PLS regression model was calculated using *The Unscrambler* software (V7.5 CAMO, Sweden).

Protein Sample Preparation

The proteins included α-chymotrypsin (bovine pancreas); α-lactalbumin (bovine milk); *β*-lactoglobulin (bovine milk); concanavalin A (jack bean); cytochrome c (horse heart); hemoglobin (human); IgG (bovine); lysozyme (hen egg white); myoglobin (horse heart and sheep muscle); ribonuclease A (bovine pancreas); and trypsin (bovine pancreas). All protein samples were purchased from Sigma-Aldrich except trypsin, which was purchased from Worthington Biochemistry Corporation, and used without further purification.

A dry amorphous sample of each protein was allowed to warm to room temperature to eliminate water condensation onto the sample and 0.2–0.3 mg placed directly onto the surface of the diamond element of the ATR accessory. Optimal contact with the ATR was achieved by direct pressure using the screw thread device of the Golden Gate. Aqueous solutions of the proteins were prepared by dissolving 5 mg protein in 1 mL Milli-Q water and 10 μL was deposited directly onto the surface of the diamond element. Polymer beads were obtained by filtering the polymer gel slurry through No. 1 filter paper to remove the water from the gel. The protein-coated polymer beads were prepared by adding 20 μL of the aqueous solution of each protein to an aliquot of the semi-wet polymer beads (equivalent to 0.2 mL of polymer slurry). The amount of protein added to the polymer beads was chosen to be much less than the adsorption capacity of the polymeric sorbent, thus ensuring the protein-coated polymer bead sample, after centrifugation and washing, contained no free protein. Samples of the protein-coated polymer beads in ammonium sulfate buffer were prepared by immersing the beads in 200 μL of 2 M ammonium sulfate (pH 7). After incubation for about 30 min, the protein-coated beads were recovered by centrifugation (5 min at 8000 rpm). The partly desiccated, air dried protein-coated beads were then loaded onto the ATR surface and compressed using the screw thread pressure device to achieve optimal contact. This compression also had the effect of removing any residual buffer/water from the sample.

PLS (Partial Least Square) Method

Partial Least Squares (PLS) is a multivariate statistical method that can be applied to provide mathematical models that relate the IR spectra of proteins to its secondary structure. The mathematical model (regression calibration model) can then be used to predict the content of the secondary structure of unknown protein samples. The PLS regression, the infrared spectra, and the reference set of secondary structure measurements are modeled as linear combinations of a set of orthogonal components, which are linear combinations of the original IR spectra. These linear combinations are chosen in such a way that they have maximum correlation with reference secondary structure measurements. The PLS method and preprocessing were performed in *The Unscrambler 7.5* (CAMO, ASA, Norway).

A salient feature of *The Unscrambler* PLS is the graphical representation of results in the form of plots of "predicted versus measured," "regression coefficients versus variables," and "root mean square error (RMSE) variance versus factor." The plot of "predicted versus measured" is useful to check the quality of the regression model fitted to the data. Predicted values should be as close as possible to the measured values (reference content) and give rise to a correlation close to 1.0 with a low root mean square error of prediction (RMSEP).

The plot of "RMSE versus factor" is a plot of the average error versus principal components (PCs), for either the calibration or the validation error with a lower value indicating a better prediction. The RMSEC or RMSEP plotted against the number of PCs provides an estimate of the fit of the model to new data not present in the calibration, and is known as the validation. The ideal validation model should show a decreasing RMSE with increasing number of

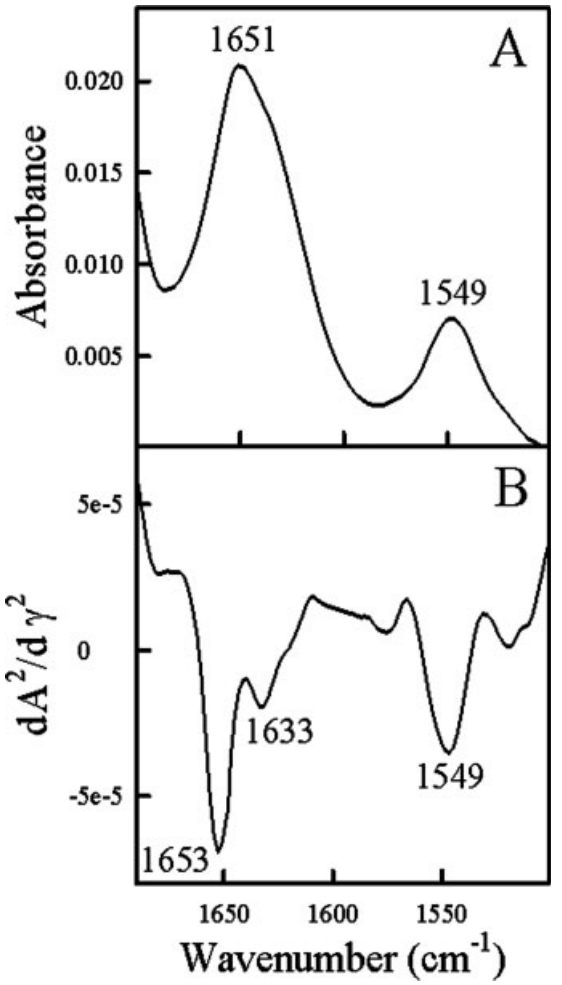

FIGURE 1 Myoglobin on butyl sorbent after subtraction of butyl sorbent spectrum. A, Absorbance spectrum; B, second derivative of absorbance.

PCs. If an increase in RMSE is observed for increasing PCs, then the model has over-fitted the measured data and the resulting prediction could be erroneous.

The plot of "regression coefficients versus wavenumber" helps to identify the important spectral features that are principally responsible for the linear correlation. In the case of proteins, these bands are usually associated with the amide modes of a particular secondary structure in the protein. Different regression coefficients are correlated to different PCs and give rise to different regression correlations. Consequently, it is important to analyze plots of regression coefficient versus PC numbers to ensure the correlation is based on spectral features of the secondary structure and not from nonspecific matrix absorbance effects. Thus an examination of the regression coefficient plots for each PC is essential in determining the number of PCs to be used for prediction, because ideally only the spectral features directly related to a particular secondary structure should be used in the prediction. The regression coefficient plot summarizes the relationship between all infrared spectra for a given reference content of the secondary structures derived mainly from X-ray crystallographic results.

Prior to PLS regression the spectrum was preprocessed by taking the second derivative spectra to minimize baseline effects, using the Savitsky-Golay algorithm with five smoothing points. Spectral derivatives can also be considered as a pseudo resolution enhancement technique, because they are able to highlight slight variations in the slope and contours of bands and hence increase the accessible spectral information. Alternatively, a multiplicative scatter correction (MSC) method was used to compensate for baseline effects in the data.

RESULTS AND DISCUSSION

The first aim of this study was to ascertain whether protein bands in the IR spectrum can be distinguished from underlying polymer sorbent bands and whether there are any significant changes in band structure between amorphous and adsorbed proteins. Figure 1 shows the absorbance and second derivative spectra of myoglobin (horse heart) on the butyl sorbent along with a spectrum of the protein after the spectrum of the butyl sorbent has been subtracted. The absorbance spectrum shows the amide I mode at 1651 cm^{-1} and the amide II mode at 1549 cm^{-1}. The second derivative spectrum shows a band at 1633 cm^{-1}, which is assigned to the β-pleated sheet secondary structural motif. The contribution of water to this band is minimal as the bead is essentially dried through the centrifugation and compression handling/air-drying steps required to obtain good contact of the sample with the ATR diamond.

To develop a calibration model it was necessary to first qualitatively analyze the amide bands of known structure proteins with their corresponding secondary structure band assignment. Second, it was necessary to quantitatively analyze the secondary structure of known proteins where the calibration model is set up using the reference content of X-ray determinations based on the qualitative analysis of the amide band assignments. It was then possible to carry out the pre-

Table I Content (%) of Protein Secondary Structures Determined by X-ray Crystallography

Proteins	α-Helix	β-Sheet	β-Turn	Random	Reference
α-Chymotrypsin	8	50	27	15	X-ray[30]
α-Lactalbumin	39	6	25	30	X-ray[31]
IgG	3	67	18	12	X-ray[30]
Concanavalin A	3	60	22	15	X-ray[30]
β-Lactoglobulin	15	50	20	15	X-ray[32]
Cytochrome c	48	10	17	25	X-ray[30]
Haemoglobin	87	0	7	6	X-ray[30]
Lysozyme	45	19	23	13	X-ray[30]
Myoglobin	85	0	8	7	X-ray[30]
Ribonuclease A	23	46	21	10	X-ray[30]
Trypsin	9	56	24	11	X-ray[30]

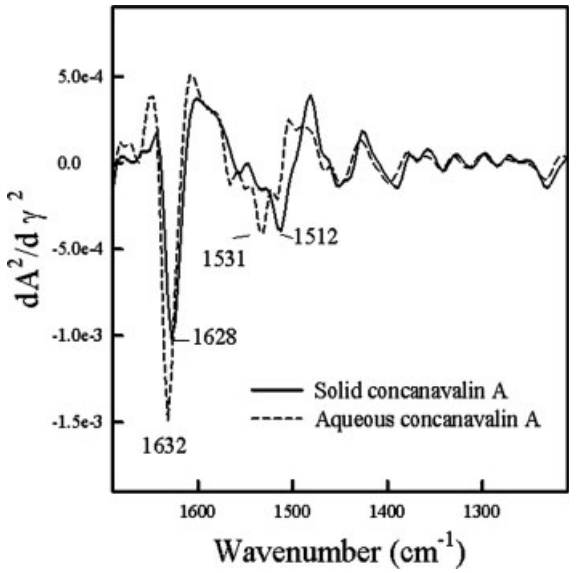

FIGURE 2 Second derivative spectra for concanavalin A in aqueous and solid states.

diction of the unknown structure of proteins by using the calibration model proteins of known structure. The secondary structure content of the proteins used in this study in the crystalline state as determined by X-ray crystallography is listed in Table I.

Qualitative Analysis of Protein Conformation from Second Derivative Spectra

Figure 2 shows the IR spectrum of concanavalin A in water and in the solid state. It can be seen that the β-sheet amide I band has a small shift while the β-sheet amide II band shifts dramatically from 1512 to 1531 for concanavalin A in the solid versus the aqueous state. Shifts to high wavenumber values were typically observed for most proteins in water relative to the solid and these are summarized in Table II. The shift to higher wavenumber value may result from the stronger interaction of hydrogen bonding between proteins and water molecules. The amide II band shifts are generally greater than those for amide I bands for most proteins examined. It should be noted that the OH-bending mode from water contributes significantly to the amide I mode, which causes radical changes in the band shape and amide I versus amide II ratio. Because the protein on the bead is essentially dried, we have developed our PLS models based on X-ray crystallographic data as opposed to NMR spectroscopic data which is often used to characterize the secondary structures

Table II Comparison of Distinctive Amide Bands of Proteins in Solid and in Aqueous State

	Solid State		Aqueous State	
Proteins	Amide I (cm^{-1})	Amide II (cm^{-1})	Amide I (cm^{-1})	Amide II (cm^{-1})
α-Lactalbumin	1643	1531	1647	1547
α-Chymotrypsin	1632	1512	1636	1543
β-Lactoglobulin	1624	1537	1628	1558
Concanavalin A	1628	1512	1632	1531
Cytochrome c	1651	1539	1655	1551
Haemoglobin	1651	1535	1655	1543
IgG	1632	1512	1635	1558, 1539
Lysozyme	1643	1535	1651	1543
Myoglobin (horse)	1647	1535	1653	1549
Myoglobin (sheep)	1651	1543	1655	1547
Ribonuclease A	1639	1547	1635	1543
Trypsin	1632	1516	1632	1554, 1520

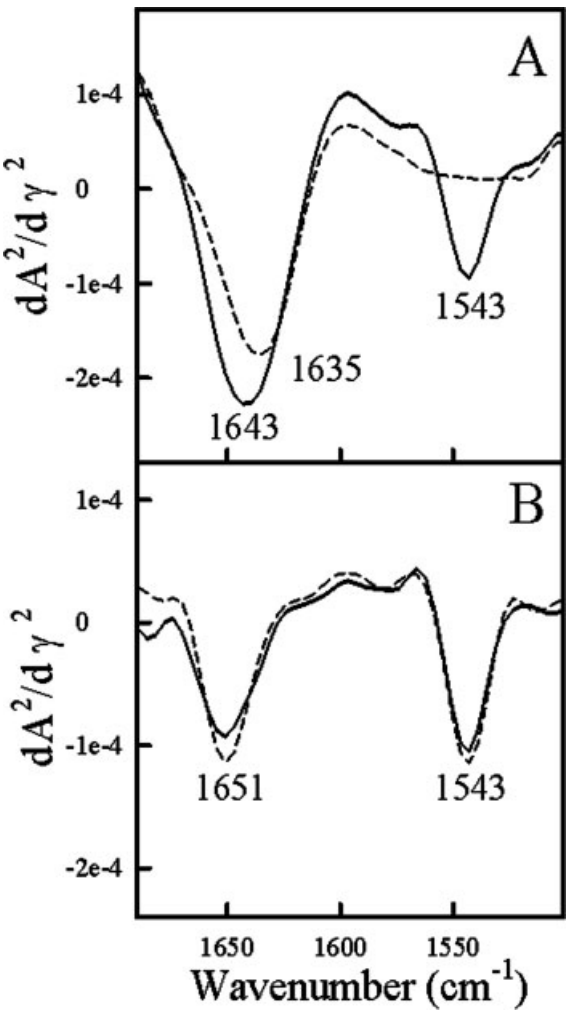

FIGURE 3 Second derivative spectra for lysozyme. A, solid line adsorbed on butyl sorbent, dashed line pure butyl sorbent; B, solid line aqueous, dashed line adsorbed on butyl sorbent with butyl sorbent subtracted.

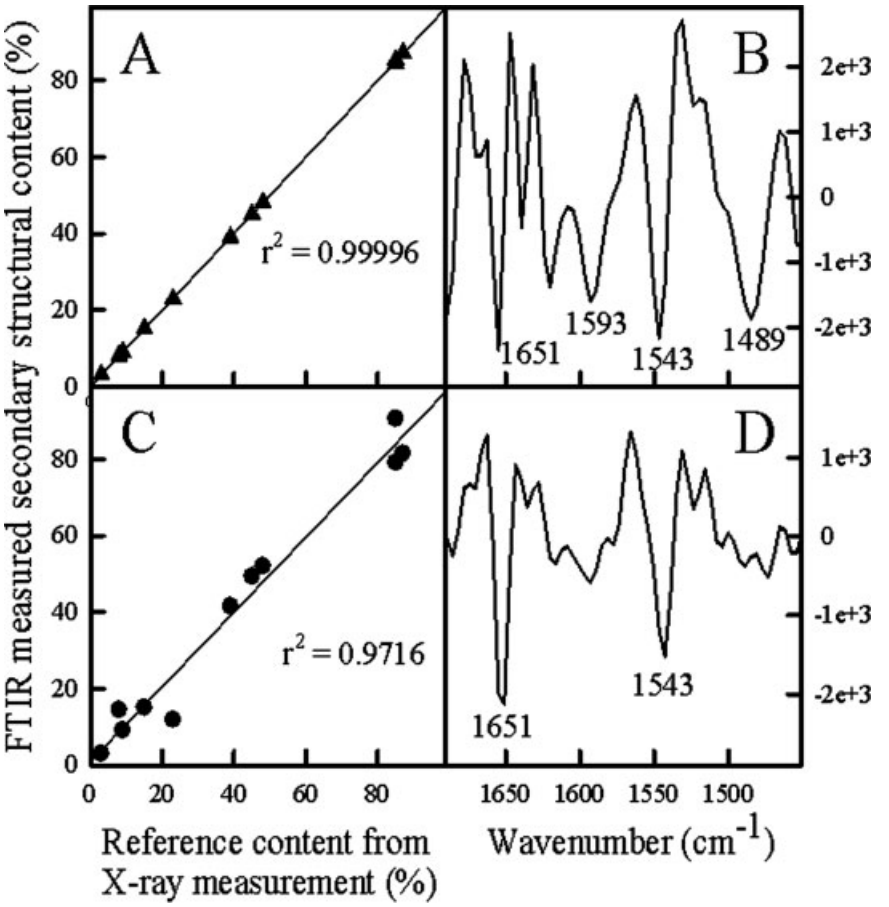

FIGURE 4 Calibration models for α-helical structure. A, correlation plot; B, regression coefficient plot for model with 10 PCs; C, correlation plot; and D, regression coefficient plot for model with 3 PCs.

of proteins in the solution phase. Other effects, such as the contributions from the variation in penetration depth as a function of protein concentration and differences in the secondary structure content of the very thin layer of bound protein, which is deposited onto the ATR crystal, compared to the protein in the bulk aqueous solution,[25] are also minimized, because any excess unbound protein was removed by the centrifugation/washing procedures employed, and the protein is then essentially dried in the bound state on the polymer bead.

All of the protein samples except lysozyme were amorphous. Previous IR work[33] has indicated that the amorphous forms have a greater tendency to aggregate compared to crystalline forms. Hence, they exhibit greater differences in terms of band position compared to the purely crystalline state. Pikal and Rigsbee[34] used insulin as a model large polypeptide, which was characterized by size exclusion chromatography, reversed-phase liquid chromatography (RP-HPLC), differential scanning calorimetry (DSC), and FTIR spectroscopy. Contrary to Shenoy's conclusion, these investigators found amorphous insulin to be far more stable than crystalline insulin under all conditions investigated. Both crystalline and amorphous insulin retain some higher order structure when dried, but the secondary structure is significantly perturbed from that characteristic of the native solution state.

Figure 3 depicts the second derivative spectra of lysozyme in the aqueous and adsorbed state (on the butyl sorbent). The amide I and II bands are at 1655 and 1543 cm^{-1} for lysozyme adsorbed onto the butyl sorbent in ammonium sulfate buffer (2 M, pH 7). The band intensity was dramatically decreased in the original spectra, with the second derivative indicating minor changes for lysozyme when adsorbed onto the butyl sorbent compared to aqueous lysozyme. This finding suggests that lysozyme undergoes some structural changes upon interaction with the butyl sorbent in the presence of 2M ammonium sulfate buffer. The combination of large band shifts observed for the amorphous form, the slight changes observed in the presence of ammonium sulfate

Table III PLS Prediction Results of the Secondary Structure of Amorphous Proteins

Form	Statistics Results	α-Helix	β-Sheet	β-Turn	Random
Amorphous	Correlation	0.9908	0.9721	0.8732	0.9000
	RMSEP	4.31	6.07	3.40	3.25
	PCs	4	4	4	4

buffer and the differences in backgrounds for different polymer beads (butyl versus phenyl) leads to the conclusion that it is necessary to build separate calibration models for each protein environment.

PLS Regression Calibration Models

To quantify the secondary structure of unknown proteins, a protein database of calibration spectra of proteins with known secondary structure, taken from Table I, was modeled. The inputs for the model consisted of four secondary structure variables (the percentage of α-helices, β-sheets, β-turns, and random coils) from 12 different proteins previously determined by X-ray crystallography.

The steps in the PLS regression calibration are delineated below:

1. The collection of representative spectra of samples spanning a range of the known secondary structural proteins, together with the secondary structural information from X-ray crystallography.
2. Calibration and validation, i.e., fitting the PLS model while performing a leave-out-one cross validation. Each sample is left out of the model, while the PLS model is formed on the remaining data. The excluded samples are then used as test samples.
3. Careful examination of the prediction error vs. factor and regression coefficient plots to determine the optimal number of factors for prediction.
4. Prediction of the unknown samples using the PLS model with optimal number of factors.

The method employed to select the number of factors in a PLS model was based on choosing those factors that gave rise to a minimum in the plot of prediction error versus number of factors. Figure 4 shows the correlation of prediction with reference for the α-helical structure of proteins for a model consisting of 12 protein samples.

Each point in the diagram of Figures 4A and 4C is derived from the average of 8 replicate spectra for each protein sample. The correlation and error are (r^2) 0.99996 and 0.577, respectively, using 10 PCs (Figure 4A). The corresponding plot of regression coefficients is shown in Figure 4B. The previously identified amide bands from the α-helical structure appear at 1651 and 1543 cm^{-1}. However, bands not attributed to the α-helical protein structure also appear at 1689, 1593, and 1489 cm^{-1}. Thus, although the correlation coefficient of the model is high, the correlation is erroneous, because other nonprotein bands contribute to the correlation. Figures 4C and 4D show the calibration results for selecting 3 PCs. Although the correlation is distinctly worse for 3 PCs (0.9716 compared to 0.99996), the regression coefficient diagram in Figure 4D indicates that the amide modes at 1651 and 1543 cm^{-1} are the most important factors in the model. The lack of spurious bands in the regression coefficient diagram indicates the correlation is based on contributions from the actual protein secondary structural bands, hence making the model and prediction more robust for the introduction of new protein examples.

The same phenomenon has been observed for other models predicting secondary structure of proteins in different environments. The use of models containing a high number of PCs is only valid when a large data set containing all possible combinations of secondary structure types is available. In this study, the criterion for choosing the optimal model to estimate the secondary structure of proteins is a combination of selecting PCs with a low Root Mean Square Error of

Table IV PLS Prediction Result of the Secondary Structure of Protein on Sorbent

Form	Statistics Results	α-Helix	β-Sheet	β-Turn	Random
Butyl sorbent	Correlation	0.9466	0.9833	0.9575	0.8873
	RMSEP	8.11	3.41	1.91	3.95
	PCs	5	7	7	7
Phenyl sorbent	Correlation	0.9797	0.9610	0.9640	0.8568
	RMSEP	5.157	5.406	1.720	4.323
	PCs	8	8	8	7
Butyl sorbent with ammonium sulphate	Correlation	0.9060	0.8426	0.9414	0.9256
	RMSEP	10.81	9.291	2.31	3.23
	PCs	7	8	10	10

Table V Comparison of the Secondary Structure Content of Insulin Determined by X-ray Crystallography, NMR-, and FTIR-Spectroscopy, and Neural Network (NN) Analysis

Methods	α-Helices	β-Sheets	Other	References
X-ray (insulin in	47.7	0		1ZEH (PDB)[35]
crystalline form)	61	15	23	Ref. 30
NMR (insulin	43.8	0		1MHI (PDB)[36]
in solution)				
FTIR and NN	50	19	16	Ref. 3
(insulin in H_2O)				

Prediction (RMSEP) and analyzing the regression coefficient diagram to ensure that only the structurally correlated bands are associated with the chosen PC. Using this criterion, the statistical results of the calibration model of amorphous proteins are displayed in Table III. An excellent correlation was achieved when the data for these bound proteins in amorphous phase was correlated with the X-ray crystallography data.

Table IV details the results of the prediction of secondary structure of proteins adsorbed on butyl sorbent, phenyl sorbent, and butyl sorbent with ammonium sulfate. The results indicate that the highest correlation for α-helical structure occurs when the protein adsorbs onto the phenyl sorbent ($r = 0.9797$), whereas the highest correlation for β-pleated sheet is obtained when the protein is adsorbed onto the butyl sorbent ($r = 0.9833$). In the presence of ammonium sulfate the prediction of both α-helical and β-pleated sheet is dramatically reduced, $r = 0.9060$ and $r = 0.8426$, respectively.

Quantitative Estimate of the Secondary Structure of Proteins by PLS in Differing Environments

To test the calibration model of proteins, insulin in the amorphous phase was selected as the test example, with the prediction being 45.5% α-helical, 15.1% β-sheet, 20.6% β-turn, and 18.7% random coil structure. For comparison Table V lists the secondary structural content of insulin determined by different methods and in different environments.

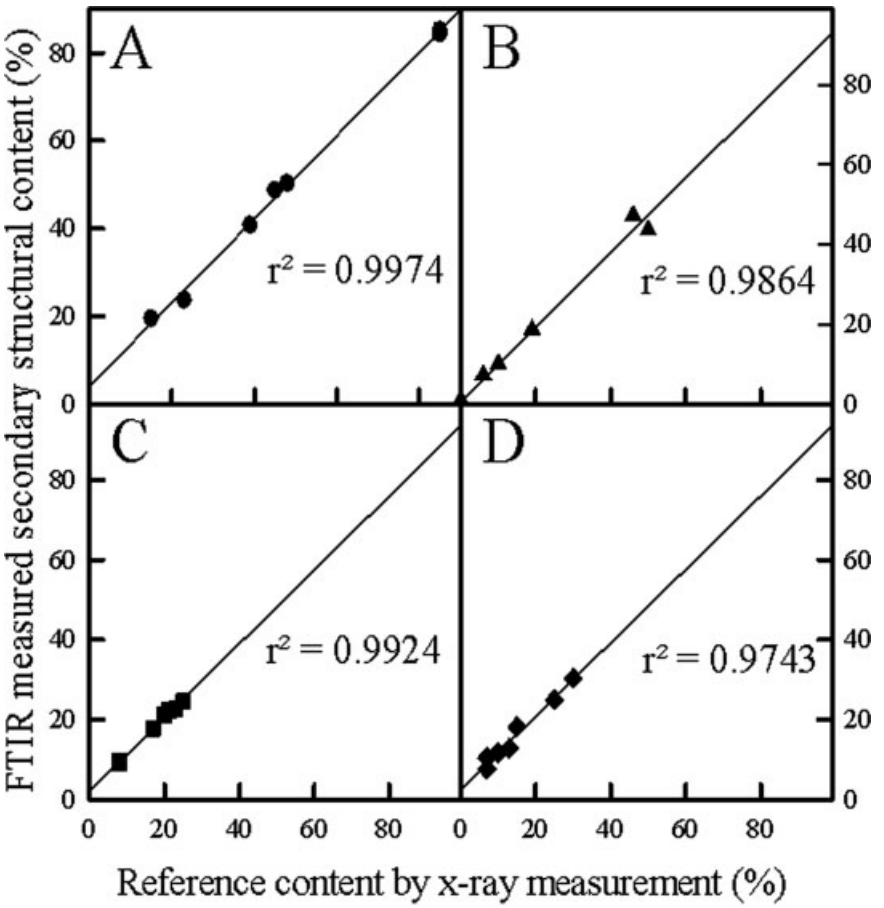

FIGURE 5 Prediction of the secondary structures of amorphous proteins on phenyl sorbent. A, α-helices; B, β-sheets; C, β-turns; and D, random coils.

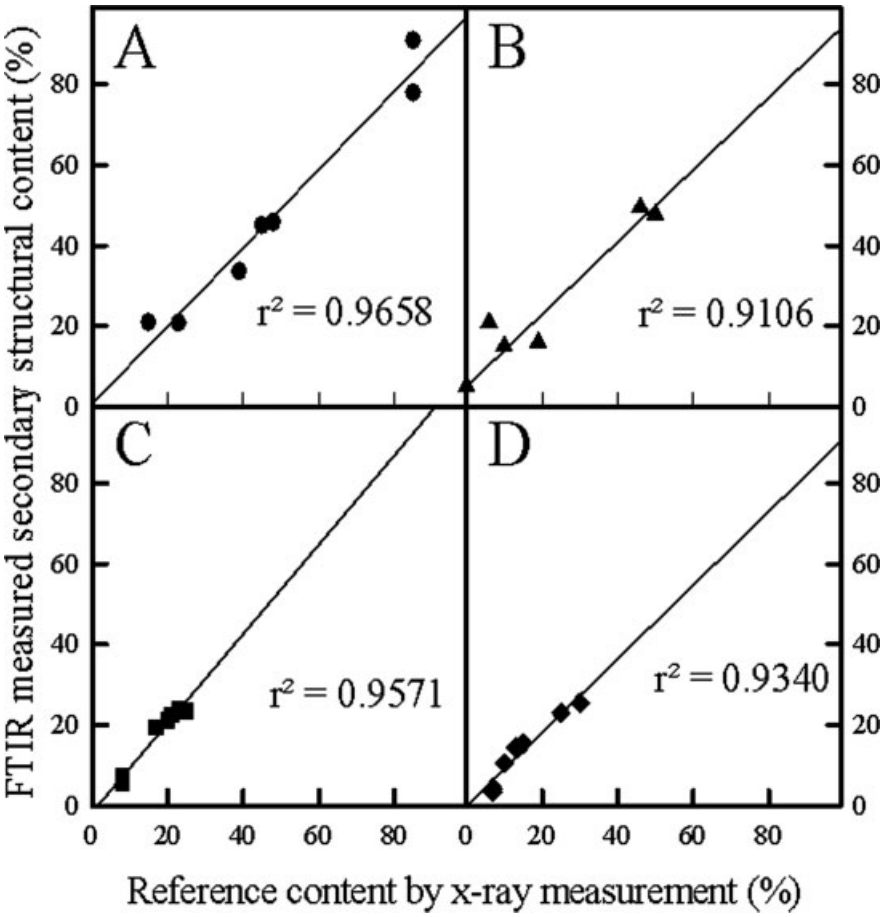

FIGURE 6 Prediction of the secondary structures of amorphous proteins on butyl sorbent. A, α-helices; B, β-sheets; C, β-turns; and D, random coils.

All of these methods give rise to different values for the secondary structural content of insulin in the different environments. Although the absolute values of insulin secondary structure are different for the different techniques, all of them are consistent in that they reveal that the secondary structure of insulin is predominantly in the α-helical conformation. The best results were achieved when the FTIR spectra of the amorphous protein was calibrated against the X-ray crystallographic data.

The individual models for different sample environments provide a good correlation for the secondary structure calibration model as delineated in Tables III and IV. As noted above, the prediction of unknown secondary structure of proteins requires the PLS model to be calibrated in the same environment. Consequently, the analysis was performed on two separate models, namely protein adsorbed onto phenyl sorbent and protein adsorbed onto butyl sorbent. The 7 proteins used in the analysis were cytochrome c, ribonuclease A, lysozyme, myoglobin (horse heart and sheep muscle), α-lactalbumin, and β-lactoglobulin. The individual predicted secondary structures of the proteins are shown in Figures 5 and 6 for these proteins with the phenyl and butyl sorbents, respectively.

A very good correlation between prediction from experimental analysis and reference data from X-ray crystallographic determination for the secondary structures of 4 proteins adsorbed onto the phenyl-substituted beads is shown in Figure 5 for which r^2 is 0.9974, 0.9864, 0.9924, and 0.9743 for α-helical, β-sheet, β-turn, and random coiled structures, respectively. A relatively poor correlation of the structural content for these proteins on the butyl sorbent was observed (see Figure 6), compared to the protein adsorbed onto the phenyl sorbent, (r^2 are 0.9658, 0.9106, 0.9571, and 0.9340 for α-helical, β-sheet, β-turn, and random coiled structures, respectively). The r^2 values for these predictions are different to the correlation results displayed in Table IV. The results indicate that the secondary structure is modified for some proteins bound to the butyl sorbent while the secondary structure remains relatively unchanged for the proteins adsorbed onto the phenyl sorbent. In independent

experiments, the hydrophobic interaction chromatographic and micro-calorimetric properties of these proteins in the presence of these butyl and phenyl sorbents as the stationary phases has been systematically investigated by our group (Ref. 37 and Hearn et al., unpublished data) and have indicated that the butyl ligand, when compared with the phenyl ligand, has higher hydrophobicity and influences protein structure to a greater extent. The conclusions reached from the FTIR-ATR measurements and the prediction models are supported by these observations.

CONCLUSIONS

The approach delineated in this article enables measurement of the secondary structural motifs of proteins in the amorphous or bound states with an excellent correlation to the reference content determined by X-ray crystallography, even for random coiled structures. This approach shows significant improvement relative to previously published methods.[1,2,16] The adaptation and improvement compared to the previous study include: (1) The Golden Gate diamond single reflection ATR accessory which provides an optimal condensed beam directly onto the samples and results in highly reproducible spectra with satisfactory water band elimination. Elimination of this water band is primarily the result of the sample handling methods, compression and intimate contact on the diamond crystal. (2) A larger variety of different proteins compared to previous studies. (3) The *Unscrambler* software, that enables direct analysis of the regression coefficients to facilitate better data interpretation and aid in the selection of the optimal number of PCs to produce the most accurate and robust models, free of any spurious correlations.

Although the major bands in FTIR spectra of proteins in the aqueous (neat water) state show that the secondary structures are quite similar to those in crystalline form determined by X-ray crystallography, there are, however, band shifts observed for proteins between the solid and aqueous states. When proteins are adsorbed to the polymeric sorbent, distinctive bands of the proteins can be distinguished by subtracting the polymer bands with the aid of analysis using the second derivative of the original spectra. To quantify the secondary structure of unknown proteins, PLS calibration models of proteins with known secondary structure were developed. For the PLS model, statistically significant samples are critical in both evaluating the different calibration models and obtaining good calibration results. The more data that are modeled in a calibration data set, the better the results of the calibration model and ultimately the prediction capability. It was found that although a good correlation was obtained for the PLS calibration models, different secondary structural motifs might contribute to the same major bands in amide I or II regions. To resolve this problem, it is necessary to qualitatively analyze the regression coefficient diagram with corresponding structural bands as well as the prediction error versus factor plot.

As documented in this article, FTIR-ATR spectroscopy in combination with PLS analysis provides an easy tool to achieve this outcome and measure the secondary structure of unknown proteins in a variety of states including the amorphous state or when adsorbed onto butyl- or phenyl-functionalized sorbents. For example, the FTIR spectra showed minor structural changes when hen egg white lysozyme was adsorbed onto the butyl sorbent in ammonium sulfate buffer. Ammonium sulfate is an anti-chaotropic salt that promotes the binding of the protein to the hydrophobic ligands. Under these conditions, changes in the solvational state of the protein in the presence of ammonium sulfate solutions of high molarity, when compared with neat water, leads to some secondary structure reorganizations of the protein on interaction with the nonpolar ligands. These reorganizations, reflected as the changes observed in the FTIR-ATR spectra, parallel similar "molten globule" transitions and changes in the secondary structure and folding processes previously found,[31,37] for these proteins in bulk solution as measured by micro-calorimetry. Moreover, the results of the quantitative analysis indicate that the secondary structural changes of proteins tend to be greater when adsorbed to the butyl sorbent than the phenyl sorbent. This finding is consistent with the effective hydrophobicity of the butyl sorbent being stronger than the phenyl sorbent, i.e., the influence of the butyl ligands on the secondary structure of the proteins, as evidence from the FTIR-ATR spectroscopy and PLS analysis, is greater than that of the phenyl sorbent.

These investigations were supported by the Australian Research Council. The donation of butyl and phenyl hydrophobic interaction sorbents by Tosoh Bioscience is gratefully acknowledged. We also thank Mr. Finlay Shanks for technical support.

REFERENCES

1. Dong, A. Biochemistry 1990, 29, 3303–3308.
2. Dousseau, F.; Pezolet, M. Biochemistry 1990, 29, 8771–8779.
3. Severcan, M.; Severcan, F.; Haris, P. I. J Mol Struct 2001, 565/566, 383–387.
4. Elliot, A.; Ambrose, E. J. Nature 1950, 165, 921–922.
5. Jakes, J.; Krimm, S. Spectrochim Acta Part A 1971, 27, 35–63.
6. Krimm, S.; Bandekar, J. Adv Protein Chem 1986, 38, 181–364.
7. Parker, F. S. Applications of Infrared Spectroscopy in Biochemistry, Biology and Medicine. Plenum: New York, 1971; pp 232–270.

8. Susi, H. Methods Enzymol 1972, 26, 455–472.
9. Calvert, J. F.; Hill, J. L.; Dong, A. Arch Biochem Biophys 1997, 346, 287–293.
10. Henkel, B.; Bayer, E. J Pept Sci 1998, 4, 461–470.
11. Thamann, T. J.; Chao, R. S. Spectochim Acta Part A 1999, 55, 2261–2270.
12. Lee, D. C.; Haris, P. I.; Chapman, D.; Mitchell, R. C. Biochemistry 1990, 29, 9185–9193.
13. Byler, D. M.; Susi, H. Biopolymers 1986, 25, 469–487.
14. Susi, H.; Byler, D. M. Biochem Biophys Res Commun 1983, 115, 391–397.
15. Yang, W. J.; Griffiths, D.; Byler, D. M.; Susi, H. Appl Spectrosc 1985, 39, 282–287.
16. Cai S.; Singh, B. R. 2000. In Infrared Analysis of Peptides and Proteins. pp 117–129.ACS Symposium Series 750.
17. Haaland, D. M.; Thomas, E. V. Anal Chem 1988, 60, 1193–1202.
18. Haaland, D. M.; Thomas, E. V. Anal Chem 1988, 60, 1202–1208.
19. Thomas, E. V.; Haaland, D. M. Anal Chem 1990, 62, 1091–1099.
20. Hoy, M.; Steen, K.; Martens, H. Chemom Intell Lab Syst 1998, 44, 123–133.
21. Haaland, D. M.; Melgaard, D. K. Appl Spectrosc 1999, 53, 390–395.
22. Haaland, D. M.; Melgaard, D. K. Appl Spectrosc 2001, 55, 1–8.
23. Ball, A. R. A. L. Jones. Langmuir 1995, 11, 3542–3548.
24. Buijs, J. W.; Norde, W.; Lichtenbelt, J. W. T. Langmuir 1996, 12, 1605–1613.
25. Goldberg, M. E.; Chaffotte, A. F. Protein Sci 2005, 14, 2781–2792.
26. Peppas, N. A.; Wright, S. L. Eur J Pharm Biopharm 1998, 46, 15–29.
27. Giacomelli, C. E.; Bremer, M. G. E. G.; Norde, W. J Colloid Interface Sci 1999, 220, 13–23.
28. Hearn, M. T. W.; Quirino, J. P.; Whisstock, J.; Terabe, S. Anal Chem 2002, 74, 2107–2119.
29. Xiao, Y.; Jones, T. T.; Laurent, A. H.; O'Connell, P. P.; Przybycien, T. M.; Fernandez, E. J. Biotechnol Bioeng 2007, 96, 80–93.
30. Levitt, M. J Mol Biol 1977, 114, 181–293.
31. Xie, D.; Bhakuni, V.; Freire, E. Biochemistry 1991, 30, 10673–10678.
32. Brownlow, S. J. H. M.; Cabral, R.; Cooper, D. R.; Flower, S. J.; Yewdall, I.; Polikarpov, A. C.; North, L.; Sawyer. Structure 1997, 5, 481–495.
33. Shenoy, B.; Wang, Y.; Shan, W.; Margolin, A. L. Biotechnol Bioeng 2001, 73, 358–369.
34. Pikal, M. J.; Rigsbee, D. R. Pharm Res 1997, 14, 1379–1387.
35. Whittingham, J. L.; Edwards, D. J.; Antson, A. A.; Clarkson, J. M.; Dodson, G. G. Biochemistry 1998, 37, 11516–11523.
36. Jorgensen, A. M.; Kristensen, S. M.; Led, J. J.; Balschmidt, P. J Mol Biol 1992, 227, 1146–1163.
37. Lin, F.-Y.; Chen, W.-Y.; Hearn, M. T. W. Anal Chem 2001, 73, 3875–3883.

Reviewing Editor: Laurence Nafie