Advanced Statistical and Numerical Methods for Spectroscopic Characterization of Protein Structural Evolution

Victor A. Shashilov[†] and Igor K. Lednev*

Aegis Analytical Corporation, 1380 Forest Park Circle, Suite 200, Lafayette, Colorado 80026, and Department of Chemistry, University at Albany, State University of New York, 1400 Washington Avenue, Albany, New York 12222

Received April 14, 2009

Contents

1. Introduction

Significant advancement in laser technology, microfluidics, and the development of novel light detectors has dramatically improved spectroscopic methods for molecular characterization over the past decade. In many cases, nanoliter samples are sufficient for high-quality spectroscopic characterization, which opens the possibility of working with precious biological and chemical systems. Although the accumulation time is significantly reduced in a typical measurement, the amount of information hidden in digital data sets composed of hundreds and thousands points is increased dramatically. The golden rule of the previous generation of spectroscopists—if you do not see a change in the spectrum by the naked eye, then you are chasing a ghost—is no longer applicable. Various statistical methods, chemometrics in particular, have been developed for processing these data and extracting essential information about the composition and evolution of biological systems.[1] The structure and dynamics of proteins have been the focus of numerous spectroscopic

* Corresponding author. E-mail: lednev@albany.edu.
[†] Aegis Analytical Corporation.
[‡] University at Albany, State University of New York.

10.1021/cr900152h
Published on Web 07/02/2010

Victor Shashilov was born in Vitebsk (Belarus) in 1978. He received his M.S. in Physics from Belarusian State University in 2000. In 2000−2004, Victor worked in the group of Prof. Tolstorozhev in Minsk (Belarus) and studied 8-azasteroid compounds using submicrosecond time-resolved fluorescence. In 2004, he was admitted to the State University of New York, where as part of his Ph.D. studies he conducted research in the field of quantitative deep UV resonance Raman spectroscopy under the supervision of Professor Lednev. His areas of interest are quantitative modeling of resonance Raman scattering, signal separation, chemometrics, and protein folding. Victor received his Ph.D. in 2007 and is currently employed with Aegis Analytical Corporation, based in Colorado.

Igor K. Lednev is an associate professor at the University at Albany, State University of New York. He graduated from the Moscow Institute of Physics and Technology, Russian Federation, receiving his Ph.D. degree in 1983. Then Dr. Lednev worked at the Institute of Chemical Physics, Russian Academy of Sciences, as a group leader. Dr. Lednev joined the University at Albany in 2002. As an academic visitor, he worked in several leading laboratories around the world, including the United Kingdom, Japan, Canada, and Germany. Dr. Lednev's research is focused on the development and application of novel laser spectroscopy for biomedical and chemical studies. Most recently, a novel method combining deep ultraviolet Raman spectroscopy with post-mortem hydrogen−deuterium exchange and advanced statistical analysis has been proven to be uniquely suitable for structural characterization of proteins at all stages of amyloid fibrillation, the process associated with many neurodegenerative diseases. His research group is also working on the application of advanced spectroscopy for forensic purposes. Dr. Lednev is a recipient of the Research Innovation Award. He has coauthored over 100 publications in peer-reviewed journals.

studies.[2] However, the utilization of chemometrics as a routine tool for characterizing protein structure and dynamics still needs to be established. Furthermore, the improper use of statistical methods or misinterpretation of their results may often lead to an erroneous conclusion and to the wrong mechanistic knowledge of a biochemical process.

The aim of this review is to introduce a broad audience of chemists and biochemists to modern statistical and numerical methods used in the analysis of spectral data and, in particular, for the quantitative characterization of protein structural evolution. This paper is focused on the value, use cases, and practical side of the methods and is addressed to chemists and biologists using spectroscopy in their studies. Methods that have no records of application for the analysis of biospectroscopy data and have been used exclusively by statisticians are omitted from this review. Theoretical sections with algorithm descriptions are written for nonmathematicians and nonstatisticians and should be easily understood by chemists and spectroscopists with a strong quantitative background, e.g. those trained in physical or analytical chemistry. We hope that this review will be a useful reference for chemists and spectroscopists who are already using advanced statistical methods, and will also be a good introductory guide for those experimentalists who would like to begin using these powerful methods in their studies.

We provide a classification of methods followed by a short and a mainly qualitative description of each of the algorithms. For each class of methods, we review the recent literature reports and discuss future trends and new areas of application.

2. Classification of Methods

Each of the algorithms described here requires a properly constructed set of spectral data that normally consists of one or more spectra and is called the data set. Whenever possible, the experiment is designed such that the sufficient number of spectra is recorded for the analysis.

2.1. Multivariate Curve Resolution

Multivariate curve resolution (MCR) is a broad class of methods that extract the spectra of individual components and the contributions of the components to each spectrum of a data set. The contribution of each component to the spectrum is proportional to its concentration in the experimental sample. MCR methods are used when the spectra of one or more individual components are not known and/or the concentrations of one or more components in one or more experimental samples are unknown. An example of a MCR problem is a temperature-induced melting of the protein α-helix. The total number of spectroscopically distinguishable species including the reactant, product, and intermediates are not known; their spectra and melting curves are unknown as well. The goal of a MCR method in this case would be to find the number of species observed during the transition and calculate their spectra and melting curves. These four multivariate curve resolution methods have been used for studying protein folding problems:

(i) Alternating least-squares (ALS) combined with principal component analysis (PCA)
(ii) Independent component analysis (ICA)
(iii) Pure variable/spectrum methods
(iv) Bayesian source separation and maximum entropy method (MEM).

2.2. Calibration Algorithms

Calibration algorithms are used to estimate the concentration of one or more individual components using the spectrum of a sample. The spectra of individual components are not computed by calibration methods. The calibration method requires a training or calibration data set that consists of a number of spectra with known concentrations of individual components. An example of a calibration problem is to determine the secondary structure composition of an unknown protein given the IR spectra of 15 proteins with a known percentage of secondary structures, e.g. from X-ray data. This problem can also be solved by using MCR methods. We will discuss the pros and cons of either approach in section 5. The calibration algorithms discussed here include the following:

(i) Univariate calibration methods
(ii) Ordinary linear or classical least-squares
(iii) Partial least-squares (PLS)
(iv) Least squares support vector machines (LS-SVM)
(v) Artificial neural networks.

2.3. Classification Algorithms

Cluster analysis and classification methods are used in order to assign an unknown protein to one of the classes. Cluster analysis, or unsupervised pattern recognition, attempts to identify groups or clusters based on similarities in their spectra without any prior information about proteins. Methods that use known class membership are called classification or supervised pattern recognition. An example of a classification problem is to determine the blood type of a crime scene sample using its Raman spectrum. Training data sets with a few spectra for each type of blood would be needed in this case. We will review the following types of classification algorithms:

(i) Cluster analysis
(ii) Principal component analysis
(iii) Linear discriminant analysis
(iv) SIMCA
(v) Partial least-squares discriminant analyses.

2.4. Database Search and Feature Extraction Algorithms

Finally we will consider search and feature extraction algorithms for the exploration of large spectral databases. The primary application of database search methods is identification of biomolecules. Spectral library search is routinely used in MS/MS studies for protein identification. The database search is normally preceded by the feature extraction step, which reduces the dimensionality of raw instrumental output. This review covers the following types:

(i) Feature selection and extraction methods
(ii) Classical library search methods
(iii) Spectral search in Fourier, wavelet, and principal component domains.

3. Need for New Numerical and Statistical Methods

Numerical and statistical methods have been in the arsenal of biospectroscopists for years. Over the past decade, most effort has been focused on the development of applied multivariate methods and solving ill-posed problems of biospectroscopy. The examples of ill-posed problems are the fit of the curve by a sum of exponentials or the fit of the spectral bands by a sum of Gaussian or Lorentzian shapes. The former is common in florescence spectroscopy, where, for example, a multiexponential decay of tryptophan fluorescence can be evidence of different microenvironments of tryptophan chromophores.[3] Routinely, the multiexponential decay is subsequently fitted with one, two, three, etc. exponentials until the residuals of fitting are approximately normally distributed. Calculated exponential lifetimes are then assigned to different chromophors or groups of chromophors. Unfortunately, such an approach does not provide a unique solution. As shown by Lakowicz,[4] there is no unique way to recover fluorescence lifetime even in the case of two-exponential decay, because of the correlation between the amplitudes and exponential lifetimes. In other words, change in the lifetimes of one exponential can be compensated by change in the amplitudes and the lifetime of the other exponentials, and thereby, an infinite number of solutions arise. This means that the estimation of lifetimes based on the simple least-squares fit can easily lead to an erroneous conclusion. No solution to this problem has been proposed so far; the most robust approach in such cases is perhaps the Bayesian method.[5] Similarly, fitting a spectral band with different combinations of Gaussian shapes can yield the same quality of fit.[6,7] This makes the analysis of secondary structure content based on the fit of an amide band extremely difficult, if not impossible. In this case, again, the Bayesian approach may be a good alternative to the classical least-squares fit offered by most spectral processing programs. Despite the progress made in solving ill-posed spectroscopic problems, most of the researchers continue using conventional methods and remain unaware of their potential pitfalls.

4. Multivariate Curve Resolution (MCR)

MCR methods extract evolution profiles or concentrations and pure component spectra of individual protein structures from the spectral data of multicomponent systems. The pure component spectra can represent the pure secondary structure (α-helix, unordered structure, or β-sheet) or the reaction-specific species, such as an intermediate on the protein unfolding pathway. This class of methods is termed source separation or multivariate curve resolution (MCR). Application of MCR requires a number of spectra (called the data set below) recorded at different stages of protein structural evolution. A typical example of a data set is a series of IR spectra measured over the course of protein unfolding, where the spectral contribution of native protein decreases and that of unfolded protein increases across the data set.

The MCR problem is as follows

$$\mathbf{D} = \mathbf{C} \cdot \mathbf{S} + E \tag{1}$$

where **C** is the concentration matrix, **S** is the matrix of pure component spectra, and E is error (random or systematic). The matrix **D** representing the data set is assumed to be known while the matrices **C** and **S** are to be estimated. Representation 1 is not unique, as there is an infinite number of matrices **C** and **S** that satisfy eq 1. In fact, given any invertible matrix **T**, eq 1 can be rewritten as follows:

$$\mathbf{D} = \mathbf{C}^0(\mathbf{T}^{-1} \cdot \mathbf{T}) \cdot \mathbf{S}^0 = \mathbf{C}_1^0 \cdot \mathbf{S}_1^0 + E \tag{2}$$

where $\mathbf{S}_1^0$ and $\mathbf{C}_1^0$ are new pure component and concentration matrices. Equation 2 illustrates the *rotational ambiguity*[8] of MCR, meaning that the unique solution cannot be found unless the pure component spectrum and/or concentration matrices are constrained using available information about the pure individual components and their concentrations. Normally, pure component spectra and evolution profiles are sought to be non-negative. In addition, the concentration profiles can be required to be monotonic or have only one peak (unimodality constraint), to sum up to a constant at each experimental point to ensure the conservation of mass (closure constraint), to follow the specific kinetics or equilibrium scheme (hard constraint),[8] etc. Similarly, some of the pure component spectra or parts of the spectra can be constrained to known or anticipated spectral shapes. In the absence of any specific prior information, the pure component spectra can be constrained to be statistically independent,[9] to be sparse,[10] or to minimize the entropy of pure components,[11] etc.

Based on the type of constraints and the way to incorporate them into the algorithm, all MCR methods used in protein folding studies can be assigned to one of the following categories: (i) alternating least-squares (ALS), normally combined with principal component analysis (PCA), (ii) independent component analysis (ICA), (iii) pure variable/spectrum methods, and (iv) the Bayesian source separation and maximum entropy method (MEM).

4.1. Preliminary Data Analysis

4.1.1. Deducing the Number of Individual Components in the Data Set

Most MCR algorithms rely on the knowledge of the number of individual spectroscopically distinguishable components in the data set. Several methods have been proposed to help in choosing the correct number of factors.[12] Nevertheless, an unambiguous conclusion about the number of principal components in a data set is often difficult, if not impossible, because of random and systematic experimental errors.[12] The determination of the true number of significant factors for data with uncertainty is not a trivial task. No criterion for determining the number of factors is completely satisfactory when used alone.[12] Therefore, both empirical methods and methods requiring the knowledge of experimental error are normally utilized to draw a reliable conclusion. The empirical methods include the eigenvalue analysis, the target factor analysis, the evolving factor analysis, the cross-validation approach, the Malinowski indicator factor function, and others.[13] Perhaps, the most popular approach used to decide on the number of principal components nowadays is a cross-validation. To perform cross-validation, segments of the data are omitted during the PCA. Using one, two, three, etc. principal components, the omitted data is predicted and compared to the actual values. This procedure is repeated until every data element has been kept out at least once. The principal component model that yields the minimum prediction error for the omitted data is retained.

Retaining a small number of principal components can result in poor quality of fit and loss of relevant information. On the other hand, models with too many components account for a large fraction of noise in the raw data and have low predictive power. It is important to emphasize here that prior information about the biochemical process under study, if available, can be most helpful in deducing the number of individual components. Furthermore, there are cases where the number of principal components in the data set can be a criterion for discriminating between possible mechanisms of protein structural rearrangement. For example, we applied various chemometric methods to verify the presence of two principal components in electrospray ionization mass spectrometry (ESI-MS), fluorescence, and deep UV resonance Raman spectra of lysozyme recorded over the course of its irreversible heat-induced unfolding[14] and thus proved that the irreversible partial unfolding of lysozyme proceeded via a two-state transition. At the initial stage of fibrillation, partial denaturation of lysozyme, the presence of several protein conformations was evident. Abstract factor analysis (AFA),[12,15−19] cross-validation methods,[19,20] and evolving factor analysis (EFA)[21−24] have been utilized for analysis of lysozyme DUVRR and fluorescence spectra measured at various stages of denaturation.[25] Both methods unambiguously suggested the presence of two significant components in the data set. Multivariate curve resolution methods[24,26−30] resulted in a perfect fitting of the experimental spectra with two basis spectra, one of which was close to the spectrum of native lysozyme, and the other one was the spectrum of a partially unfolded intermediate.

However, a definite conclusion on the number of conformers cannot be made based solely on the above spectroscopic data even if the chemometric analysis strongly suggested the existence of two principal components. Therefore, electrospray ionization mass spectrometry (ESI-MS) was also utilized to address the hypothesis. The protein ion charge state distribution (CSD) envelopes of ESI-MS provide information on protein conformational changes. By using chemometric analysis, the CSD envelopes of the incubated lysozyme were well fitted with two principal components. Based on the spectroscopic/spectrometric data along with chemometric analysis, the partial unfolding of lysozyme during *in vitro* fibrillation was proved to be a two-state transition.

4.1.2. Reducing the Dimension of the Data Set

The majority of MCR methods exploit the covariance in the data to extract decorrelated spectral components as a first step of the analysis. After the number of individual components *n* has been found, *n* orthogonal or decorrelated components are extracted from the data set using eigenvector matrix decomposition to produce abstract individual component spectra and concentrations. This procedure is called dimension reduction or whitening.[31] Using these *n* individual components, the data set matrix can be reconstructed as follows:

$$\mathbf{D} = \mathbf{C}^0 \cdot \mathbf{S}^0 + E \tag{3}$$

where $\mathbf{S}^0$ and $\mathbf{C}^0$ are the matrices of abstract individual components and their contributions to experimental spectra, respectively. Matrices $\mathbf{S}^0$ and $\mathbf{C}^0$, however, normally do not have any physical meaning. The ultimate goal of the MCR algorithm is to find unique and physically meaningful pure component and concentration matrices by using transformation 2.

4.2. Alternating Least Squares

ALS implements transformation 2 by the iterative updating of the pure component spectrum and concentration matrices $\mathbf{S}^0$ and $\mathbf{C}^0$ subject to non-negativity constraints, closure, unimodality, or model-specific hard constraints. The iterative process starts with setting all negative elements to zero in each of the abstract individual component and concentration matrices $\mathbf{S}^0$ and $\mathbf{C}^0$. Then, at each new iteration, the new pure component spectrum $\mathbf{S}_1$ and the concentration matrix $\mathbf{C}_1$ are calculated using the following equations:

$$\mathbf{S}_1 = (\mathbf{C}^{\mathrm{T}} \cdot \mathbf{C})^{-1} \cdot \mathbf{C}^{\mathrm{T}} \cdot \mathbf{D} \tag{4a}$$

$$\mathbf{C}_1 = \mathbf{D}(\mathbf{S}_1^{\mathrm{T}} \cdot (\mathbf{S}_1 \cdot \mathbf{S}_1^{\mathrm{T}})^{-1}) \qquad (4b)$$

where **C** is obtained from the previous ALS iteration and $\mathbf{C}_1$ and $\mathbf{S}_1$ are the new matrix estimations. The superscripts T and −1 denote transposition and inversion, respectively.

Application of PCA and ALS for protein structural studies is now well-established mainly due to Tauler and co-workers.[8,32,33] PCA has been employed to determine the number of conformers formed in the course of chymotrypsin inhibitor-2 and ubiquitin folding using ESI-MS data.[34] A specifically elaborated procedure of target factor analysis has allowed for resolving the mass spectra of individual conformers. The application of ALS for analyzing far-UV CD and near-UV CD spectra of α-chymotrypsin allowed for recovering the concentration profiles and the spectra of three different protein conformations. The reconstructed concentration profiles have been utilized for the quantitative analysis of changes in secondary and tertiary structure of α-chymotrypsin.[35] The number of conformers present at various stages of α-apolactalbumin folding has been estimated using chemometric analysis of fluorescence and CD spectroscopic data.[32] In particular, spectral signatures of various conformers and their concentrations in each sample have been obtained using the MCR-alternating least-squares (ALS) method. PCA combined with evolving factor analysis (EFA) of near-IR absorption data has allowed for characterizing the hydration of bovine serum albumin and its temperature-induced secondary structural rearrangements.[36] Statistical analysis of IR spectra has been used for evaluation of protein secondary structure.[37] In particular, the interval partial least-squares method (iPLS) has been applied for estimating the fractions of α-helices and *β*-sheets in proteins.[37] The PCA combined with two-dimensional correlation spectroscopy and dynamic hydrogen−deuterium exchange allowed for characterization of tertiary and quaternary structures of the hepatitis C virus protein.[38]

4.2.1. Application to Protein Folding Problems

Recently, the alternating least-squares method applied to IR data allowed for recovering concentration profiles and pure IR spectra of the species involved in the hydrolysis of bovine serum albumin with protease K.[39] The combination of principal component analysis, cluster analysis, and interval partial least-squares has been proposed for determining protein secondary structure based on the analysis of the IR and CD databases of proteins with known secondary structures.[40] ESI-MS spectra and concentrations of three *β*-lactoglobulin forms, i.e. a monomer in acidic form, a dimer, and a monomer in basic form, have been resolved in the study. Using augmented[41] ESI-MS-CD data matrices in ref 33 has permitted these authors to avoid the rotational ambiguity of ALS and to exclude the presence of artifacts generated during the ionization process. The PCA−varimax approach applied to simulated data has allowed for extracting useful information on the folding process and the number of involved intermediates.[42] See also refs 3, 4, and 11 in ref 43 for other applications of multivariate curve resolution methods applied to the analysis of protein structural rearrangements.

The simultaneous analysis of data obtained by different spectroscopic techniques has become increasingly popular, as it reduces the rotational ambiguity of MCR. In particular, the joint analysis of CD, IR data, and X-ray protein structural data using PCA and the iPLS method has been utilized to establish the IR/CD based method for protein secondary structural characterization.[44] The mechanism of the pH-induced conformation changes of bovine *β*-lactoglobulin[33] was elucidated by the simultaneous ALS analysis of ESI-MS and CD data.

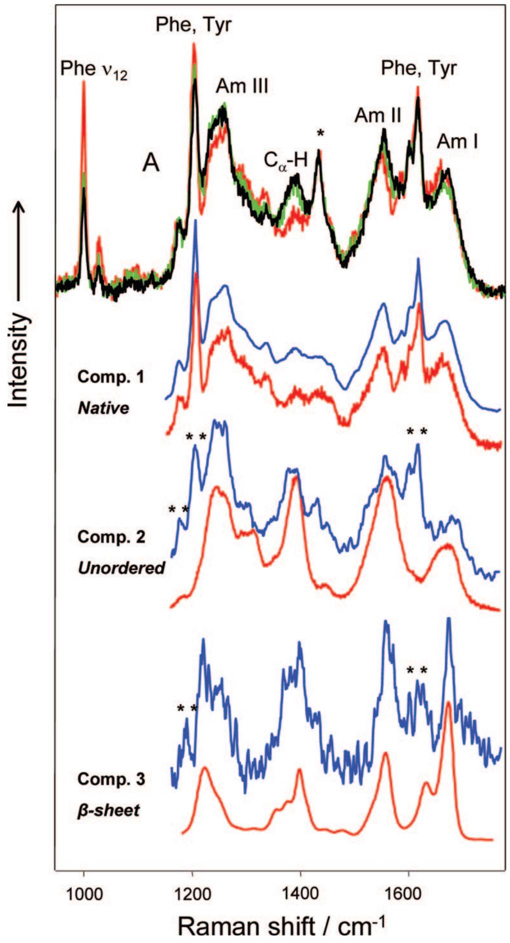

Figure 1. (A) 197-nm excited Raman spectra of native lysozyme (red) and the supernatant of the lysozyme solution incubated at 65 °C for 2 (green) and 8 days (black). ICA allows for extracting three independent components (blue curves), which show excellent agreement with Raman spectra of native lysozyme, unordered polypeptide (poly-L-lysine), and the *β*-sheet of the lysozyme fibril core.[43] A protein deep UV Raman spectrum is dominated typically by the contribution of the amide chromophore, the building block of a polypeptide backbone, and side chains of aromatic amino acids, mainly phenylalanine (Phe) and tyrosine (Tyr). The amide bands report on the protein secondary structure while the aromatic amino acid bands provide information about the local environment and the protein tertiary structure.[45] The amide I mode (Am I) consists of carbonyl C═O stretching, with a small contribution from C−N stretching and N−H bending. The amide II and amide III bands involve significant C−N stretching, N−H bending, and C−C stretching. The C_α−H bending vibrational mode involves C_α−H symmetric bending and C−C_α stretching. The asterisk (*) indicates the trifluoroacetate (internal standard) band. Double asterisk (**) indicates the contribution of aromatic amino acid residues.

Raman spectroscopy, resonance Raman in particular, has been demonstrated to be a powerful technique for biological studies.[45−73] In particular, Asher with co-workers[70] devised an approach for evaluating the protein secondary structure composition using basis spectra determined for pure secondary structural elements. The pure secondary structure spectra have been calculated using the least-squares analysis of 13 proteins with known secondary structure. Factor analysis has been employed to verify that deep UV resonance Raman

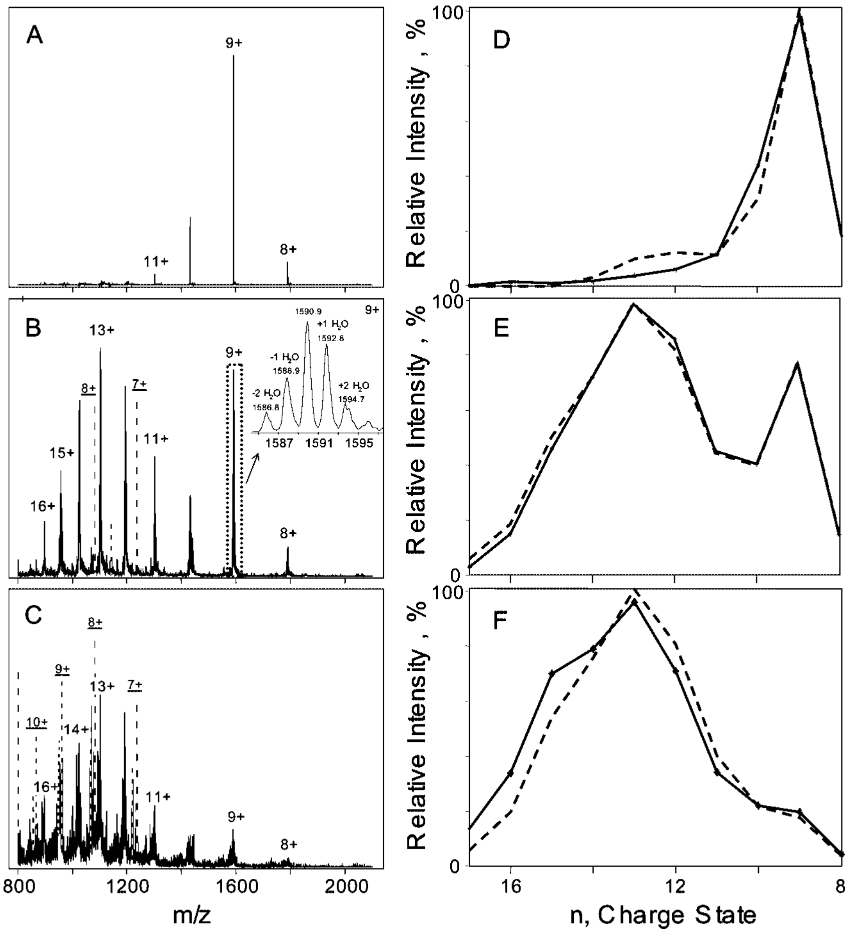

Figure 2. Electrospray ionization (ESI) mass spectra of a pH 2.0 lysozyme solution (14 mg/mL) incubated at 65 °C for 5 min (A), 14 h (B), and 96 h (C). The inset of part B shows the magnified view of the dotted rectangle, where the charge state "9+" includes peaks corresponding to different water adducts and dehydration products. The fittings of the charge state distribution (CSD) envelops for 5-min (D), 14-h (E), and 96-h (F) incubated lysozyme are shown on the right panel. The solid curves are the experimental CSD envelops, while the dashed curves are the two-component fittings. Major ion peaks that are derived from the fragments are drawn in dashed lines. The ion peaks labeled with the underlined charge state are derived from a fragment of 8644 Da. (Reprinted with permission from ref 14. Copyright 2007 The Protein Society).

(DUVRR) spectra of all proteins can be adequately fitted using three basis spectra, i.e. the DUVRR spectra of an α-helix, a β-sheet, and an unordered structure. However, the application of chemometric analysis of Raman spectra for characterizing protein structure is yet to be demonstrated.

Shashilov et al.[74] reported on the application of the ALS constrained to a kinetics model for the structural transition of lysozyme at the initial stages of fibril formation. All experimental Raman spectra of lysozyme recorded over the course of incubation were fitted with three pure component spectra, i.e. the spectra of a nucleus β-sheet and partially unfolded intermediate calculated by ICA, and the experimental spectrum of native lysozyme (Figure 1). A mixed soft−hard modeling approach[75] provided the refined DUVRR spectra of a β-sheet and partially unfolded intermediate, kinetic profiles for all three species, and the characteristic times for each step of lysozyme transformation. Namely, the algorithm was used to perform the following:

(i) calculate kinetic profiles by guessing the initial characteristic times;

(ii) fit experimental spectra using the kinetics profiles and pure components spectra; and

(iii) iterate over the characteristic time constants until the best fitting is achieved. The independence of the characteristic times on the protein concentration indicated that the early stages of lysozyme fibrillation, irreversible partial unfolding and nucleus β-sheet formation, were intramolecular processes.

We also applied ALS to fit the fluorescence, ESI-MS, and deep UV resonance Raman spectra of lysozyme at its different stages of irreversible unfolding with two significant components.[14] The perfect two-component fit of all charge distribution envelopes of ESI-MS spectra (Figure 2) served as a solid proof of the all-or-none mechanism of lysozyme partial unfolding. Furthermore, the ability to fit both aromatic and amide regions of DUVRR spectra with two pure component spectra further confirmed a 100% correlation in secondary and tertiary structural changes, which was additional evidence of an all-or-none transition mechanism.

4.2.2. When To Use Alternating Least-Squares

This method is well suited for all types of spectral data. It gives the best results when one or more pure component spectra or the concentration profiles of some individual components are known and can be fixed or constrained. Otherwise, the method often provides ambiguous results, especially when the individual component spectra overlap or the evolution profiles of the individual components are correlated. Caution should be exercised when assessing the correctness of the ALS model by the goodness of fit, as a meaningless model can provide an excellent fit to the data.

Another common mistake is to increase the number of principal components to minimize the residuals. In most cases, a better preprocessing of the data prior to the analysis, such as eliminating a baseline and shift of spectra along the wavelength axes, or proper subtraction of a background, will help improve the fitting.

4.3. Independent Component Analysis (ICA)

A new powerful latent variable approach called independent component analysis (ICA) has been developed in recent years.[76−78] A particular case of ICA is called the blind signal separation (BSS) method,[9,79] which allows for extracting spectra of individual components and their concentrations from the spectral sets of a multicomponent system without *a priori* information about the composition of the system. In contrast to PCA and MCR, ICA searches for *statistically independent*[80] pure component spectra rather than just uncorrelated ones.[76,81] For many applications, the ICA approach has been shown to be more powerful than PCA.[76,81,82] ICA has been applied for analyzing NMR,[83] X-ray crystallography,[84] electron energy loss,[85] and photoacoustic[86] spectroscopic data. Combination of ICA and the immune algorithm allowed for resolving overlapping chromatographic profiles.[78] Numerous applications of ICA for the advanced processing of images have also been reported.[87−89] The ICA algorithms using the maximum likelihood[80] method have been utilized for the resolution of overlapping Raman spectra.[90]

The main disadvantage of the general-purpose unconstrained ICA algorithms[79,91,92] was found to be the appearance of meaningless negative bands in the individual component spectra. To overcome this problem, a special class of ICA algorithms that deliberately search for non-negative individual components has been elaborated.[31,93−96]

The general approach of non-negative ICA can be formulated as follows. Assume that we have a sequence of observed signals X_1, X_2, ... X_n. Those signals could be arranged in the rows of the observation matrix $\mathbf{X}$. If each signal X_k is a linear combination of the original signals (independent components) S_k, then the observation matrix $\mathbf{X}$ can be represented as a product of two matrices[76]

$$\mathbf{X} = \mathbf{A} \cdot \mathbf{S} \tag{5}$$

where $\mathbf{A}$ is a mixing matrix and $\mathbf{S}$ is a matrix of original signals. The task of non-negative ICA is to determine the mixing matrix $\mathbf{A}$ and the matrix of non-negative signals $\mathbf{S}$ given just the observation matrix $\mathbf{X}$.

In our particular case,[43] matrix $\mathbf{X}$ was composed of ten DUVRR spectra of lysozyme incubated for various times (further designated as data matrix $\mathbf{D}$). Mixing matrix $\mathbf{A}$ was a concentration matrix $\mathbf{C}$ that contained the concentrations of each lysozyme conformer in various samples in a corresponding column. Finally, matrix $\mathbf{S}$ consisted of rows of DUVRR spectra of lysozyme conformers. Data matrix $\mathbf{D}$ was represented as a product of matrices $\mathbf{C}$ and $\mathbf{S}$ similarly to the case in eq 1.

The non-negative ICA starts off matrix $\mathbf{Z}$, composed of decorrelated abstract individual components. Matrix $\mathbf{Z}$ is calculated using the general prewhitening procedure for non-negative ICA,[31] such that prewhitening does not subtract mean data from matrix $\mathbf{Z}$ and, therefore, retains information about the non-negativity of the spectra.[97] We will describe two non-negative ICA algorithms: the algorithm using Fourier expansion over a geodesic[94] (hereinafter referred to as algorithm I) and the non-negative PCA algorithm (algorithm II).[31]

Given matrix $\mathbf{Z}$, algorithm I proceeds via iterative multiplication of $\mathbf{Z}$ by the orthonormal matrix $\mathbf{W}$ until the resolved spectra are non-negative. At each iteration, matrix $\mathbf{Z}$ was multiplicatively updated by some orthonormal rotation matrix $\mathbf{R}$. Either the Fourier or the Newton method can be used for updating the parameter of the rotation matrix $\mathbf{R(t)}$.[94] In the case of three independent components, the Rodrigues formula[98] can be utilized for calculating rotational matrix $\mathbf{R}$.

Algorithm II is a special case of the nonlinear PCA algorithm.[31] Similarly to algorithm I, dimension reduction and prewhitening were performed to calculate matrix $\mathbf{Z}$. Then matrix $\mathbf{Z}$ is iteratively multiplied by matrix $\mathbf{W}$, which is additively updated at each iteration using the non-negative PCA rule.

In many spectroscopic techniques, the spectrum of a protein is recorded in a complex matrix of compounds and then needs to be separated from strong background spectra. Independent component analysis has been used to separate water artifacts from 2-dimensional nuclear Overhauser enhancement (2D NOESY) spectroscopy signals of proteins dissolved in water.[99,100] Mantini et al. have applied ICA to the analysis of MALDI-TOF data[101] and demonstrated the capability of ICA to extract the protein signal from MALDI-TOF mass spectra. Chen et al.[102] have demonstrated the capability of ICA to resolve the NIR spectra of a mixture containing protein into the spectra of the components.

ICA algorithms treat noise in a spectral signal as an individual component and thus can be used for the denoising of spectra. The effective noise reduction algorithms applied to protein spectral data are reported by Gruber and co-workers.[100,103]

4.3.1. When To Use Independent Component Analysis

The separation of spectra in ICA is difficult to control, so ICA methods are preferred when no prior information about the spectra or concentrations of the mixture components is available. Blind ICA is often capable of extracting the spectra with unusual features such as sharp bands. The independent components obtained by ICA can be used as a starting point for the ALS fit. To the best of our knowledge, we have reported on the first application of ICA for characterizing protein structural changes. In our studies of lysozyme structural transitions at the early stage of fibrillation, a small contribution of the β-sheet conformation in the partially unfolded intermediate spectrum was evident in DUVRR spectra.[25] The PCA methods, which searched for the uncorrelated significant components, failed to separate newly formed β-sheet and unordered structures, since both of them appeared on the same time scale. In other words, the behaviors of unordered structure and β-sheet fractions were highly correlated. The term "unordered structure" refers the polyproline II helix or other unassigned protein structures. To deal with highly correlated spectral components, non-negative independent component analysis of DUVRR data of hen lysozyme was applied to recover a spectroscopic signature of the newly formed β-sheet as a third individual component.[43] The latter was overlooked by classical chemometric methods in ref 14 because of its small contribution and the high correlation of its development rate with that of the unordered structure.

Shashilov et al.[43,74] reported on the first application of joint diagonalization Jade,[104] second-order blind identification (SOBI) in the Fourier space data,[105] and second-order nonstationary source separation (SEONS)[106] ICA algorithms for the analysis of protein structural evolution. The joint diagonalization ICA methods were used to resolve sets of DUVRR spectra of lysozyme at the initial stage of fibril formation into pure spectra of native protein, partially unfolded intermediate, and nucleus β-sheet.[78] It is noteworthy that the DUVRR spectroscopic signature of the nucleus β-sheet is very close to the spectrum of the lysozyme fibril core isolated by using Bayesian source separation on DUVRR hydrogen−deuterium exchange data[107] and to the spectrum of the β-sheet recovered by non-negative ICA in our preceding study[43] (Figure 1).

Although ICA can provide a meaningful solution in some cases, we discourage using this method in the situations when a good initial guess about the spectra or concentrations of individual components is available. In our studies, ICA did not show satisfactory results when applied to CD or fluorescence data. More reliable algorithms developed specifically for the analysis of CD data are discussed in section 5.2.1.

4.4. Pure Variable Methods

Another possible alternative to MCR-ALS and ICA is pure variable methods, which extract the spectra of individual components by exploring the wavenumbers with the highest intensity variations in the data set.[108] The simple-to use interactive self-modeling mixture analysis (SIMPLISMA)[108] method has been employed for the analysis of IR, NMR, and Raman spectral data.[109,110] The method is especially efficient in analyzing spectra with both sharp and broad spectral bands.[111] We previously utilized a newly developed SMAC (stepwise maximum angle calculation)[112] algorithm for resolving the concentration profiles and DUVRR spectra of various complexes formed by lutetium and bicyclic diamide.[13] Our study on simulated Raman spectra[13] demonstrated that SMAC and SIMPLISMA were able to reconstruct spectra of components with complex and irregular concentration profiles. SMAC and SIMPLISMA performed well even in the cases when using EFA and MCR was completely impractical.

To illustrate the principle underlying the purest variable methods, consider a set of Raman spectra of a multicomponent mixture. Assume that the Raman spectrum of some individual component consists of several bands centered at certain wavenumbers. An increase in concentration of this component in the mixture will result in simultaneous rise of Raman intensities at the peak wavenumbers. And vice versa, a decrease in the component concentration will cause a drop of Raman intensities at corresponding wavenumbers. In other words, correlation between Raman intensities at various wavenumbers in the spectral set suggests that the intensities at those wavenumbers are mainly contributed by a single individual component. Vice versa, the least correlated wavenumbers in the spectrum set can be related to different individual components.

In addition, the band with the highest intensity variations is assumed to be contributed by a single individual component. In fact, if the intensity at some wavenumber is a sum of the intensities of various components, then the variation in the total intensity at this wavenumber will be relatively small. This is because the increase in the intensity due to the rise of concentrations of some components can be compensated by the decrease in the concentrations of other components, since their concentration behaviors are not correlated.

SIMPLISMA and SMAC methods identify the least correlated wavenumbers, which have the correspondingly largest intensity variations. Such wavenumbers are called the purest variables.[112] The purest variable is such a wavenumber at which the contribution of an individual component to the Raman intensity is maximal while the contributions from the other components are minimal. For a Raman spectrum of each sample, the intensity at a particular purest variable is approximated to be proportional to the concentration of the corresponding individual component in the sample. Consequently, the matrix of the Raman intensities at all purest variables $\mathbf{C_{int}}$ can be used as a concentration matrix $\mathbf{C}$ of various components.

Given the matrix $\mathbf{C_{int}}$ of Raman intensities at the purest variables, the spectra of individual components can be calculated by the method of least-squares

$$\mathbf{S} = \mathbf{D^T C_{int}}(\mathbf{C_{int}^T C_{int}})^{-1} \qquad (6)$$

The concentrations of the components are calculated based on the spectra obtained from

$$\mathbf{C} = \mathbf{DS}(\mathbf{S^T S})^{-1} \qquad (7)$$

SMAC and SIMPLISMA methods allow for estimating the number of spectroscopically distinguishable components in a mixture based on the residuals of fitting. The spectral set is consequently fitted with one, two, and more components until the fitting residual shows statistical properties of a random noise expected for the experimental spectra. The SIMPLISMA algorithm has been described in detail.[108,110] SMAC, a newly developed version of SIMPLISMA,[112] determines a purest variable based on the angles with respect to the space defined by the previously selected purest variables. The variable with the maximum angle with respect to the unit vector is set as the first purest variable.[112] The angle between a new variable and a previously selected one is calculated based on the projection of the new variable in the space of the previously selected variables.[112] The projection is calculated using the loadings of the singular value decomposition applied to the matrix of the previously selected variables.[13]

4.4.1. Using Second Derivative Spectra

Due to the overlap of spectra of individual components and the presence of a baseline, the intensities at the purest variables have contributions from more than one component. As a result, the calculated concentrations of each individual component have contributions from other components. In fact, the overlap of spectra results in under-resolved concentrations and over-resolved spectra.[113] To eliminate a baseline and lessen the overlap of spectra, inverted second derivative spectra are used instead of conventional, i.e. as recorded, spectra.[113] This is because the second derivative spectra usually have sharper and better-resolved peaks as compared to conventional Raman spectra.

4.4.2. Application to Protein Structural Characterization

Recently, a novel SIMPLISMA-based approach for the preselection of wavelengths, to be used in combination with partial least squares (PLS) or other multivariate regression techniques, was developed by Bogomolov and Hachey.[114,115] The method utilizes the purity function originally proposed for the SIMPLISMA algorithm. This new algorithm has shown excellent performance in the quantitative characterization of protein secondary structural content based on IR data. The combination of the neural network and the SIMPLISMA algorithm has been used as a novel classification tool in MALDI-MS studies.[116] In our studies, the application of pure variable methods enabled us to resolve the spectra of newly formed fibrillar β-sheet and partially unfolded intermediates into separate components. To the best of our knowledge, it was the first attempt to extract the spectroscopic signature of the fibrillation nucleus.[43]

4.4.3. When To Use Pure Variable Methods

SIMPLISMA and SMAC are specifically designed for vibrational spectra or any spectra with well resolved bands. The methods do not perform well on broad and featureless spectra such as UV−vis abrorption, fluorescence, or circular dichroism spectra. The spectra resolved by pure variable methods are usually refined by the ALS fit.

4.5. Bayesian Source Separation and Maximum Entropy Method (MEM)

The blind source separation algorithms such as ICA and pure variable methods do not easily allow for making use of prior information about the spectral features of individual components and the concentration matrix, while the latter can be readily anticipated in many studies. The Bayesian source separation by its nature is a prior information-based approach and thus can be used for solving extremely ill-posed MCR and prior-dominated problems.

The Bayes theorem for the MCR problem (eq 1) takes the form

$$P(C, S|\mathrm{Data, I}) \sim P(\mathrm{Data}|C, S, I) \cdot P(C|I) \cdot P(S|I) \tag{8}$$

where $P(\mathrm{Data}|C,S,I)$ is the likelihood of measuring the quality of data fitting and $P(S|I)$ and $P(C|I)$ are prior probabilities for individual component spectra and concentrations, respectively. In essence, the Bayesian theorem states that both the quality of data fitting $P(\mathrm{Data}|C,S,I)$ and the physical meaning of the resolved spectral and concentration matrices $P(S|I)$ and $P(C|I)$ must be used as the criteria of the correctness of the multivariate model.[6]

The vertical bar (|) marks the conditional probability, with, for example, $P(A|B)$ meaning the probability of A given B. Because finding either matrix C or S alone is enough for solving the problem in eq 1, the concentration matrix C is normally sought, since it contains far fewer elements. It was shown[117] that if the sources are independent and no prior information about the concentration matrix is available, the probability of the concentration matrix is given by

$$P(C|\mathrm{Data}, I) \sim \int \mathrm{d}s \prod_i \delta(\mathrm{Data}_i - C_{ik} \cdot S_k) \prod_i p_i(s_l) \tag{9}$$

In the case of noise-free data, eq 9 reduces to the logarithmic probability

$$P(C|\mathrm{Data}, I) = \log(\det(W)) + \sum_l \log(p_l(s_l)) \tag{10}$$

where W is the separation matrix such that $S = W \cdot \mathrm{Data}$.

MEM is closely related to the Bayesian statistics, as it provides the way to assign the probabilities in eqs 9 and 10 by maximizing the entropy S

$$S = -\sum P_i \ln(P_i) \tag{11}$$

subject to constraints based on the knowledge of the data.

4.5.1. Application to Protein Folding Kinetics

A hybrid algorithm combining the maximum entropy method (MEM) with nonlinear least-squares (NLS) fitting has been developed to interpret latent exponential analysis of multiexponential fluorescence decay.[118] The algorithm allowed for resolving five exponential decays and one exponential rise over the course of dihydrofolate reductase folding.

So far, only a few studies with the application of maximum entropy methods to the spectral data of biological samples have been reported, and the rigorous maximum entropy approach seems to have to be established and validated yet. Among those few reports is the recovery of the cross-β sheet DUVRR spectrum of lysozyme fibrils from the DUVRR hydrogen−deuterium exchange data set.[107] The spectrum of the fibrillar core cross-β sheet is, however, not directly observable in the Raman experiment because of significant interference with Raman signals from β-turns and unordered protein structures. We recently utilized the capacity of the novel Bayesian source separation algorithm to extract the DUVRR spectrum of the highly ordered cross-β sheet core of lysozyme fibrils and establish the relation between the spectral and structural features of the lysozyme fibrillar β-sheet.[107,119] We used hydrogen−deuterium exchange to substitute amide N−H protons in β-turns and unordered fragments of amyloid fibrils with deuterium and recover the spectral contribution of the hydrophobic cross-β sheet core that remained unaffected by hydrogen−deuterium exchange. *A priori* information about characteristic bands in the individual component spectra was incorporated via the Bayesian signal dictionary approach,[10] where individual components are presented as a linear combination of reference spectral bands. The concentration matrix was constrained to the fractions of protonated and deuterated species controlled in the experiment and to the fraction of the fibrillar core that was iteratively refined during the optimization by the genetic algorithm. The calculated spectrum of the cross-β sheet is shown in Figure 1. The well resolved amide III band enables us to estimate the facial Ψ angle of the fibrillar β-sheet using the approach by Asher and co-workers[120] and thereby prove the antiparallel organization of β-sheet strands in lysozyme fibrils. The maximum entropy analysis has been applied to deconvolute the electrospray spectra of protein mixtures,[121] which is the first application of the maximum entropy analysis in electrospray mass spectroscopy. Another application of the maximum entropy approach to separation ESI-MS spectra of protein and glycoprotein mixtures has been demonstrated.[122]

4.5.2. When To Use Bayesian Source Separation

Similarly to ALS, Bayesian methods are utilized when there is enough prior information about the pure component spectra and their evolution profiles. Compared to ALS, these methods provide more flexibility for setting constraints. One can constrain a portion of a pure spectrum or model certain bands by defining their shape, positions, and widths. The concentration matrix can also be constrained to a kinetic or equilibrium model. Unfortunately, the development of a Bayesian algorithm may be a laborious process and the convergence of the algorithm normally takes longer than in the case of the other methods reviewed in this section 5.2.1.

5. Calibration Methods

As discussed in section 4, the output of multivariate curve resolution methods is both spectra of individual components and their concentrations. Furthermore, the prior information about the contributions of the individual components in the data set spectra is not required by MCR methods. Multivariate calibration methods provide only information about the concentrations of pure components and require a training set of spectra measured on the samples with known concentrations of the component of interest. There are, however, two advantages of calibration algorithms over MCR that compensate for the limitations of the former. Calibration methods are (i) more accurate in estimating of the concentrations and (ii) more likely to provide a unique solution because they rely less on the analysist's intuition or the initial guess provided.

5.1. Univariate Calibration

Univariate calibration methods yield the concentration of the analyte by using the intensity at a single point in spectra. For example, the concentration of protein is routinely determined using the absorption of tyrosine and tryptophan at 282 nm. Univariate calibration can be used when the analyte has a spectral region that does not overlap the bands of other components in the system. For example, a 1000 cm^{-1} phenylalanine band in deep UV resonance Raman spectra has no overlap with any amide or amino acid side chain bands. The intensity of this band has been shown to respond to changes in protein tertiary structure and has been used for the quantitative characterization of tertiary structure evolution.[123]

5.1.1. When To Use Univariate Calibration

This calibration method is accurate and easy to use. Unfortunately, the spectra of complex biomolecules normally highly overlap, which makes the application of univariate methods impractical. The presence of a baseline and shift or change in the band shape are two most common sources of error in univariate calibration. Using proper baseline correction methods[124] and considering the area under the band instead of a single point intensity are recommended to improve the accuracy of the calibration.

5.2. Multivariate Calibration

Multivariate calibration methods use the entire spectrum to construct a prediction model.

5.2.1. Linear Ordinary or Classical Least Squares

In the simplest case, the spectra of individual components are known and the problem is to find the contribution of each component to the spectrum of an unknown sample. In this case, the contribution of each component can be found as a least-squares fit of the experimental spectrum with known pure component spectra. This approach is routinely used in the analysis of CD data where the CD spectrum is fitted by a set of basis spectra representing average spectra of protein secondary structures.[125]

Given the matrix **S** of individual spectra and the spectrum of a protein with unknown structure $\mathbf{D_1}$, the fractions of secondary structures can be evaluated as follows:

$$\mathbf{C}_1 = \mathbf{D}_1 \cdot (\mathbf{S}^{\mathbf{T}} \cdot (\mathbf{S} \cdot \mathbf{S}^{\mathbf{T}})^{-1}) \tag{12}$$

where $\mathbf{D}_1$ is a $1 \times v$ matrix and $\mathbf{C_1}$ is a $1\times n$ matrix, v is the number of spectral channels, and n is the number of secondary structure types to be evaluated, respectively.

Several software packages and methods are used for the analysis of CD data of proteins.[126,127] The most popular software packages are CONTIN,[128] SELCON3,[129] and CDSSTR.[130] Different programs reportedly provide inconsistent results and should be used with caution when assessing the secondary structure composition of proteins.[131] This discrepancy stems from different reference protein CD spectra as well as different calculation algorithms used by those programs. First, the standard error in the concentrations recovered by using eq 12 can be shown to be proportional to $\mathbf{S}\cdot\mathbf{S}^{\mathbf{T}}$, i.e. to the overlap of individual component spectra. Consequently, slight differences in broad and highly overlapping CD spectra of average secondary structures resulting from using different reference spectra can give rise to large uncertainties in the computed secondary structure composition.[132] Second, different calculation algorithms themselves can yield different fractions of secondary structures even when applied to the same reference CD data set. The latter is illustrated in Table 1 adopted from Sreerama and Woody.[133] Sreerama and Woody[133] compared the performance of three popular methods for estimating protein secondary structure fractions from CD spectra (implemented in the software packages CONTIN, SELCON3, and CDSSTR) and a variant of CONTIN, CONTIN/LL, that performs the variable selection in the locally linearized model in CONTIN. Although all packages used different calculation algorithms, the outputs of the programs were shown to be consistent in most tests. CDSSTR turned out to be more robust in cases with a small reference set of spectra and larger wavelength range. CONTIN/LL showed the best accuracy when applied to large reference sets and smaller wavelength ranges. The authors[133] recommend using all three methods to improve the reliability of predicted secondary structural fractions.

Keiderling and co-workers have established several numerical methods for the assessment of protein structure using vibrational circular dichoism (VCD).[134] VCD is sensitive to short-range order, allowing it to discriminate β-sheet and various helices as well as disordered structure. A new technique, which is a combination of the factor analysis and restricted multiple linear regression, has been elaborated and utilized to determine protein secondary structure content using FTIR and VCD amide I and amide II band intensities.[135] The predictive power of the method was assessed using a leave-one-out cross-validation. In this method, the dimension of the raw VCD spectral data is first reduced by

Table 1. Performance of SELCON3, CDSSTR, and CONTIN-LL Programs for Analyzing Protein CD Spectra for the 22-Protein Reference Set[a]

	α-helix		3/10 helix		β-sheet		turns		poly(Pro)II		unordered		overall	
method	δ_α	r_α	$\delta_{3/10}$	$r_{3/10}$	δ_β	r_β	δ_T	r_T	δ_{PII}	r_{PII}	δ_U	r_{PII}	δ	r
SELCON3	0.052	0.972	0.027	0.572	0.051	0.896	0.036	0.433	0.025	0.779	0.056	0.842	0.043	0.958
CDSSTR	0.038	0.986	0.026	0.635	0.044	0.928	0.041	0.419	0.027	0.759	0.047	0.899	0.039	0.965
CONTIN	0.048	0.976	0.026	0.641	0.060	0.843	0.046	0.351	0.029	0.710	0.044	0.906	0.044	0.956
CONTIN-LL	0.050	0.974	0.027	0.591	0.053	0.880	0.038	0.505	0.027	0.750	0.052	0.869	0.043	0.959

[a] δ is the root-mean-square deviation between the X-ray and CD estimated fractions of the secondary structure, and r is the correlation coefficient calculated using the actual and predicted fractions of secondary structures.

the principal component analysis and then the regression between the sets of PCA scores and the X-ray-determined secondary structure components has been performed.[136] Combining VCD and electronic CD data has been shown to improve the predictive power of the numeric method mainly to the presence of complementary information contained in protein spectra from both techniques. Another approach using the linear relationship between the fractional components of secondary structure for the protein set and the overlap integrals of the normalized spectra has been utilized to determine the secondary structure fractions in the unknown samples.[137]

Since no protein known so far consists of a single type of secondary structure, the average secondary structure spectra of proteins cannot be measured directly. The first quantitative methods for determining the set of basis spectra utilized the synthetic peptides that can adopt three conformations under different experimental conditions. For example, uncharged (Lys)n and (Glu)n were used as a model of the α-helix in CD studies.[132] The 21-residue alanine-based peptide AAAAA-(AAARA)3A (AP) has been shown to form pure PPII at elevated temperatures[138,139] Another example of the model PPII peptide is undecapeptide XAO, having the sequence XXAAAAAAAOO, where A is L-alanine, X is diaminobutyric acid, and O is ornithine.[120] The experimental deep UV resonance Raman spectra of the high temperature of AP and XAO can be regarded as pure spectra of unordered structure. The basis spectra extracted from model polypeptides can be hardly applicable to the analysis of protein spectra. First, CD, IR, and Raman spectra are side chain dependent. IR and Raman spectra have broader spectral bands because of a large number and variety of amino acid residues; the CD spectra of proteins and polypeptides have significant differences as well.[132] There is also an uncertainty in the choice of model polypeptides. For example, CD studies have shown that each of the secondary structures can be modeled by several polypeptides and have significantly different CD signatures.[132]

The alternative approach to determining basis spectra is to derive them from the spectral sets of proteins with known three-dimensional structures. Using X-ray data as reference, the average CD,[132] IR,[140] FR-IR,[141] and resonance Raman[70,142] pure secondary structure spectra have been calculated. The approach equivalent to using pure secondary structure spectra has been developed in NMR protein spectroscopy, where the average chemical shift of a particular nucleus in the protein backbone has been shown to correlate with protein secondary structure content.[143] Chemical shifts of more than 200 proteins have been used to calculate the average chemical shifts and establish correlations with the percentage of α-helix and β-sheet structures in the proteins under study.

The major steps of the classical least-squares approach are as follows:

The reference spectra with known three-dimensional structure are expressed using the matrix of pure component spectra **S** and the concentration matrix **C** as follows:

$$\mathbf{C} \cdot \mathbf{S} = \mathbf{D} \tag{13}$$

where the **C** matrix has dimension $N \times n$, the **S** matrix is $n \times v$, and **D** is $N \times v$, N is the number of spectra in the database, n is the number of pure secondary structural components, and v is the number of points in each spectrum. The pure secondary structure spectra **S** are obtained by multiplying the **D** matrix in eq 1 by the left pseudoinverse of **C**:

$$\mathbf{S} = (\mathbf{C}^{\mathrm{T}} \cdot \mathbf{C})^{-1} \cdot \mathbf{C} \cdot \mathbf{D} \tag{14}$$

where **C** and **S** are the matrices of the concentrations and the spectra of individual components, respectively.

Given matrix **S** from eq 2 and the spectrum of a protein with unknown structure $\mathbf{D}_1$, the fractions of secondary structures can be evaluated using eq 12.

5.2.2. Calculating CD Pure Secondary Structure Spectra

Saxena and Wetlaufer[144] pioneered the development of the least-squares approach by computing three basis spectra for α-helix, β-sheet, and unordered protein using the spectra of globular proteins. In 1978 Chang et al. used 15 global proteins and added the spectrum of the β-turn to the previously extracted spectra fo the α-helix, β-sheet, and unordered protein.[145] The reconstructed pure secondary component spectra have been shown to depend on the proteins used to build a regression model. Later, Bolotina et al.[146] proposed to treat parallel and antiparallel β-sheets as separate components. Several approaches for calculating basis CD spectra using multivariate and variable selection techniques have further been proposed, such as convex constraint analysis (CCA)[147] and a modified SELCON3 method.[133] In the latter work, a focus has been put on accurate characterization of denaturated and unordered proteins.

It is commonly accepted now that the estimation of protein secondary structure depends on the set of reference spectra, and therefore, there is no unified set of CD basis spectra.[132] See the review by Whitmore et al.[125] for other state-of-the-art methods for the analysis of protein CD data.

5.2.3. Calculation of Deep UV Resonance Raman Spectra of Pure Secondary Structures

The method of calculating the average secondary structure spectra of proteins based on the set of DUVRR spectra of proteins with known X-ray structure was established by Asher and co-workers.[70] The calculated basis DUVRR spectra of α-helix, β-sheet, and unordered structure have been a powerful tool for both quantitative and qualitative interpretation of protein DUVRR data.[148,149] A further development of this approach in Spiro's group has allowed for extracting four pure secondary structure DUVRR spectra for α-helix, β-sheet, unordered structure, and β-turns.[150] The DUVRR signature of β-turns was particularly valuable, as no DUVRR spectrum of β-turns from either proteins or model polypeptides had been available. Lednev and coworkers used the DUVRR signature of β-turns reported by Huang et al.[150] as a reference to extract and characterize the DUVRR spectrum of the β-turn in fibrils of genetically engineered polypeptides.[151]

The application of basis spectra obtained from either polypeptide or protein spectral data has its intrinsic limitations. Using peptide basis spectra for fitting DUVRR spectra of proteins is impractical, as homopolypeptide amide Raman bands are narrower than those of proteins.[14,123] The latter can be attributed to the greater structural inhomogeneity of proteins due to a large number and variety of amino acid residues. On the other hand, protein basis spectra once extracted from a particular training data set by means of the least-squares regression represent averaged spectra in the data set and thus may not be appropriate for other DUVRR protein spectra. In fact, variation in the dihedral angle of a secondary structure (parallel vs antiparallel β-sheet[120]), change in hydration (hydrated vs anhydrous helix[152]), and twist of the β-sheet[153] have been shown to alter the Raman spectra of proteins. Similar to IR and CD spectra of proteins that are sensitive to the number of strands in the β-sheet[154] and α-helix,[130] the deep UV resonance Raman cross section of the α-helix has been shown to decrease with the number of strands due to a hypochromic effect in UV absorption.[138,155] Furthermore, a deep UV resonance Raman cross section for the same type of secondary structure can be sequence dependent, as shown by Song et al. on the example of poly(L-glutamic) acid and poly(L-lysine) β-sheet.[155]

5.2.4. When To Use Classical Least-Squares

This method is fast and easy to use and gives good results when the spectra of individual components are known. In order to use this algorithm, one should know the spectra of all components in the system under study, which often makes its application to studying multicomponent biochemical systems impractical. Another potential pitfall in using the classical least-squares regression is large uncertainties in estimated concentrations caused by the multiple correlations of the spectral intensities at neighboring wavelengths or wavenumbers. More robust calibration methods described below can handle these multiple colinearities.

5.3. Partial Least Squares

Application of the PLS approach for elucidating protein secondary structure based on IR[37,156−161] and CD[33,44,162] data is well-established. Oberg et al. attempted to optimize the accuracy of spectroscopic protein secondary structure determination using 50 CD and IR protein spectra with known secondary structure.[163] The results demonstrated that no smaller subset of the database contains the necessary information to describe the entire set, and therefore, large protein databases are required for the accurate prediction of unknown proteins. Another important conclusion of the study is that independent analyses of CD and IR spectra should be done to verify the accuracy of prediction or to prove the failure to find the solution. Researchers in ref 126 compared several algorithms for predicting protein secondary structure using CD data including PLS, simultaneous PLS, support vector machines, principal component regression, and others. PLS and simultaneous PLS have provided consistently high prediction accuracy for all types of secondary structures. The PLS method has recently been applied for the analysis of deep UV resonance Raman data.[164,165] Partial least squares is a calibration latent variable method that builds the regression model to relate the reference matrix of protein spectra and the spectrum of the protein under study.[33,44] The latent variables are calculated so that they (i) capture maximum variance in the data set spectra **D** and (ii) correlate with the secondary structure composition in the concentration matrix **C** (see eq 1) constructed based on X-ray data.[12] PLS seeks a calibration model in the form

$$\mathbf{C} = \mathbf{D}\cdot\mathbf{B}_{\mathrm{PLS}} \tag{15}$$

where $\mathbf{B}_{\mathrm{PLS}}$ in eq 15 is the matrix of calibration or regression coefficients. Given the spectrum of a protein $\mathbf{D}_{\mathrm{new}}$ with unknown structure, the percentage of secondary structures is calculated as follows:

$$\mathbf{C}_{\mathrm{new}} = \mathbf{D}_{\mathrm{new}}\cdot\mathbf{B}_{\mathrm{PLS}} \tag{16}$$

PLS performs by factorizing both the concentration matrix **C** and the data matrix **D** into score and loading matrices as follows:

$$\mathbf{D} = \mathbf{T}\cdot\mathbf{P}_{\mathbf{D}} \tag{17a}$$

$$\mathbf{C} = \mathbf{T}\cdot\mathbf{P}_{\mathbf{C}} \tag{17b}$$

where **T** is a score matrix common for **D** and **C**, and $\mathbf{P}_{\mathbf{D}}$ and $\mathbf{P}_{\mathbf{C}}$ are the matrices of loadings for the **D** and **C** accordingly. The regression coefficient matrix $\mathbf{B}_{\mathrm{PLS}}$ is then given by the product

$$\mathbf{B}_{\mathrm{PLS}} = \mathbf{P}_{\mathrm{D}}^{+}\cdot\mathbf{P}_{\mathrm{C}} \tag{18}$$

where the superscript + stands for pseudoinversion.

PLS has been demonstrated to outperform classical least-squares in solving ill-posed calibration problems,[19] mainly due to its higher predictive accuracy. In fact, in a typical training data set,[70,150] the number of variables (wavenumbers or wavelengths) (~1000) is much larger than the number of spectra (~10) or the number of pure component spectra (3 or 4) and spectral intensities at nearby wavenumbers are often correlated. As a consequence, the $(\mathbf{S}\cdot\mathbf{S}^{\mathrm{T}})^{-1}$ (see eq 12) matrix cannot be accurately calculated, which ultimately leads to large uncertainties in the predicted fractions of secondary structures. PLS deals with the problem of ill-conditioned matrices by substituting orthogonal principal component loadings for the matrix **D**.

Another advantage of the PLS approach over classical least-squares lies in its ability to handle spectral data where there are extra components besides those under study. In our Raman and IR spectra of proteins, for example, bands from aromatic amino acid residues take on a significant portion of variance in the data set but do not relate to the secondary structure composition of a protein. To minimize interference from the bands of aromatic amino acids in the classical least-squares analysis, their spectral contribution is numerically subtracted from experimental spectra

Table 2. Standard Errors of Prediction and Model Frequencies[141]

	rms error	FTIR frequencies	model → % of the secondary structure[a]
α-helix	21.8	1545−1655−1613	$-179.45 + 2.030A_{1545} + 0.431A_{1655} + 0.828A_{1613}$
β-sheet	17.6	1656−1634−1691	$26.31 - 0.451A_{1656} + 0.335A_{1634} + 0.586A_{1691}$
β-turns	4.2	1677−1528−1577	$-30.49 + 0.428A_{1677} + 0.322A_{1528} + 0.141A_{1577}$
random	11.4	1544−1627−1691	$178.56 - 1.578A_{1544} - 0.332A_{1627} - 0.735A_{1691}$
3/10 helix	3.2	1631−1694−1625	$5.077 - 0.198A_{1631} + 0.440A_{1694} + 0.160A_{1625}$
random + turns + 3/10 helices	12.8	1549−1629−1513	$224.544 - 2.010A_{1549} - 0.345A_{1629} - 0.775A_{1513}$

[a] A_ω refers to the absorption at frequency ω in the area-normalized FTIR spectra of a protein. Area normalization was performed so that the area between the spectrum and the baseline drawn between the spectrum points at 1700 cm^{-1} and 1500 cm^{-1} was equal to 10,000.

and corresponding spectral regions are omitted from IR and Raman spectra. This procedure suffers from large uncertainties because the shapes and positions of aromatic amino acid bands can be slightly different in different proteins. Using PLS bypasses the necessity of numerical subtraction and thus increases the accuracy of calibration.

Nowadays, a typical spectroscopic instrument used in protein studies records hundreds or thousands of variables for a single spectrum. When working with high resolution spectral data, one should be aware of the potential pitfalls caused by multiple linear correlations at adjacent variables. It has now become a common practice to convert raw instrumental output to low resolution spectra in order to reduce the amount of redundant information contained therein. Although there are a variety of methods used for variable selection,[19] the genetic algorithm seems to be the most trusted and commonly used one. Tauer and co-workers have applied the iPLS method to establish the relationships between selected amide regions of IR protein spectra and protein secondary structure.[37,44] Researchers in ref 141 have shown that the absorbances at three distinct frequencies in FT-IR spectra contain all nonredundant information about protein secondary structure. In other words, among the 3200 frequencies in the region from 4000 cm^{-1} to 800 cm^{-1}, only three frequencies in the amide I and amide II regions contain all the information required to predict the secondary structure of proteins and none of the other data points provide additional information once those three characteristic frequencies are included into the model (Table 2, adopted from Goormaghtigh et al.[141]). Proteins with known secondary structure composition have been used to construct a predictive model for assessment of secondary structure composition of proteins with unknown structure.[141]

5.3.1. When to Use PLS

PLS is used when the spectra of individual components are not known. The calibration or training data set composed of spectra of samples with known composition is required for PLS. The algorithm is sensitive and selective to the components of interest and can be applied for detection and quantification of analytes in complex matrices. Unfortunately, the quality of PLS calibration often relies on the number of principal components or latent variables in the model. In order to find the optimal number of components, cross-validation and proper data scaling are performed prior to the analysis.

5.4. Least Squares Support Vector Machines

LS-SVM is a novel nonlinear multivariate calibration method and an extension of traditional support vector machines (SVMs).[166] Similarly to PLS, LS-SVM can deal with ill-posed calibration problems and often yields a unique solution.[167] In addition, LS-SVM has been shown to produce robust models in the case of spectral variations due to nonlinear interferences.[167]

LS-SVM minimizes the cost function:

$$Q_{\text{LS-SVM}} = \frac{1}{2}||\mathbf{w}||^2 + \gamma||\mathbf{e}||^2 \qquad (19)$$

where **w** is the matrix of regression coefficients, "$|| ||^2$" stands for the Frobenius norm, γ is the relative weight of the regression error, and $\mathbf{e} = \mathbf{C} - \mathbf{w}^{\mathrm{T}}\cdot\mathbf{D} - \mathbf{B}$, where **B** is the offset matrix.

The constrained optimization problem (eq 19) is solved by using Lagrange multipliers in the form

$$Q_{\text{LS-SVM}} = \frac{1}{2}||\mathbf{w}||^2 + \gamma||\mathbf{e}||^2 - \sum_{i=1}^{n}\alpha_i(\mathbf{w}_i^{\mathrm{T}}\cdot D_i + B_i + e_i - C_i) \qquad (20)$$

where α_i are Lagrange multipliers, n is the number of spectra in the data set D, and index i refers to the i^{th} row of the matrices. The well-elaborated procedure[167,168] further reduces eq 20 to the following:

$$C_i = \sum_{i=1}^{n}\alpha_i\cdot D_i^{\mathrm{T}}D + B_i = \sum_{i=1}^{n}\alpha_i\cdot\langle D_i, D\rangle + B_i \qquad (21)$$

The nonlinearity in the LS-SVM approach is taken into account by replacing the inner product $\langle D_i,D\rangle$ by a kernel function that is typically a polynomial function

$$K(D_i, D_j) = (D_i^{\mathrm{T}}\cdot D_j + t)^d \qquad (22a)$$

or a radial basis function (RBF)

$$K(D_i, D_j) = \exp\left(-\frac{||D_i - D_j||^2}{\sigma^2}\right) \qquad (22b)$$

with the polynomial parameters t and d and the Gaussian variance σ^2 correspondingly. Recently, Wu and co-workers applied LS-SVMs to develop a method for assessing protein content in milk powder using infrared data.[169]

5.5. Artificial Neural Networks

Artificial neural networks are a *computational model* that emulates a biological neural system. It consists of an interconnected group of *artificial neurons*. Each neuron receives activation signals through multiple pathways, with the strength of each activation depending on other neurons to which the neuron is connected. In most cases, an artificial neural network is an *adaptive system* that changes its structure during the learning or training phase. The strength of neural networks lies in their ability to model complex nonlinear relationships where most linear statistical methods fail. In least-squares support vector machines (section 5.4), the nonlinear dependence between the intensities in spectra and the percentage of secondary structures is modeled by nonlinear kernel functions. In order to achieve an accurate prediction in LS-SVMs, one should select the appropriate function with right parameters such as the polynomial parameters t and d and the Gaussian variance σ^2 correspondingly in eqs 22a and 22b. In neural networks, no explicit assumption about the form of a nonlinear relationship is required. The latter made artificial networks a powerful tool for the analysis of spectral data, where, for example, multiple structural motifs contribute to the spectrum in a nonlinear fashion.[170,171] Böhm and co-workers have demonstrated a new method based on the backpropagation network model for the analysis of protein far UV CD spectra.[172] The method was able to predict the content of five secondary structure fractions: helix, parallel and antiparallel β-sheet, β-turn, and unordered structure. The best performance has been achieved when a separate neural network model has been applied to each wavelength region. A database of 18 protein FTIR spectra has been used to train a multilayer feed-forward neural network approach using an enhanced "resilient backpropagation" learning algorithm.[173] In this study, one region of the amide I band was found to provide the best prediction accuracy.

5.6. Comparison of Calibration Algorithms

A large group of algorithms developed for the analysis of protein spectral data, and IR data in particular, is based on multiple linear regression (MLR).[174] The MLR regression, however, can often yield models with lower prediction capacity because of the correlations among the spectral channels of protein spectra.[175] We recommend to, first, perform variable selection prior to MLR modeling to limit the number of variables and then use stepwise variable selection at each MLR iteration. Multiple linear regression and partial least-squares are more robust in dealing with multiple correlations among spectral channels and normally provide more accurate estimation of the secondary structural content. A word of caution should be advised when using neural networks for the analysis of biospectral data. While neural networks in general provide a good fit to the training data set, the prediction accuracy of the method is normally lower compared to that by PCR, PLS, or MLR.[175] Goormaghtigh and co-workers[163] compared the accuracy of several multivariate methods [SELCON3,[129] PCA followed by multiple regression (PCA-MR), PLS, and weighted PLS (PLS-1)[163]] applied to IR and CD data of 50 proteins with known secondary structure. Table 3, adopted from Oberg et al.,[163] summarizes the accuracy of the prediction of α-helix, β-sheet, and turn content in terms of the determination of root-mean-squared deviations of the calculated fractions from those from X-ray data (rms), correlation coefficients (R), and the information content score ξ. The ξ is the ratio of the standard deviation of the predicted secondary structure fractions to the root-mean-squared error (rms). As seen from the table, the PLS-1 method was the most accurate in predicting both α-helix and β-sheet content for all three, IR, CD, and the combined CD-IR, data sets. The SELCON3 method[129] outperformed the others in estimating the fraction of turns based on IR and CD-IR data whereas PLS and PLS-1 were the most accurate to predict the fraction of turns from the CD spectra. SELCON3 was found to be the least accurate when estimating α-helix and β-sheet fractions for IR, CD, and CD-IR data. The second least accurate algorithm for predicting α-helix and β-sheet fractions was PCA, followed by multiple regression (see Table 2).

To ensure higher accuracy of calibration methods, the authors[163] encourage using data from multiple spectroscopy techniques simultaneously. They also describe a procedure for the proper construction of the training data set aimed at eliminating anomalous spectra from the model.

6. Classification Methods

Classification methods are utilized for assigning a spectrum of protein to one of the classes. The classification is often used to assign the protein to one of the types of tertiary structure using, for example, CD data.[176] An efficient approach of classification of protein powders based on Raman spectra has been established and proposed as a routine screening test for the pharmaceutical industry.[177]

Table 3. Performance Comparison of Different Analysis Algorithms with the RaSP50 Set

		α-helix (H)			β-sheet (E)			turn (T)			Σother (C + G + B + S)		
data	algorithm[a]	rms	ζ	R^b	rms	ζ	R^b	rms	ζ	R^b	rms	ζ	R^b
IR (amide I + II)	SELCON 3	8.5	2.6	0.92	7.73	2.34	0.90	3.52	1.23	0.38	11.58	1.04	0.26
	PCA-MR	6.91	3.2	0.95	7.64	2.35	0.91	4.38	0.99	0.22	9.27	1.3	0.64
	PLS	7.29	3.03	0.94	7.58	2.37	0.91	4.36	1	0.21	9.48	1.27	0.62
	PLS-1	7.16	3.09	0.95	7.36	2.44	0.91	4.31	1.01	0.13	9.49	1.27	0.62
IR + CD	SELCON 3	7.57	2.91	0.939	7.97	2.27	0.90	3.97	1.09	0.36	10.30	1.17	0.47
	PCA-MR	6.83	3.24	0.95	7.23	2.48	0.92	4.23	1.02	0.29	9.26	1.3	0.64
	PLS	6.8	3.25	0.95	6.97	2.58	0.92	4.3	1.01	0.27	9.06	1.33	0.66
	PLS-1	6.73	3.28	0.95	6.68	2.69	0.93	4.45	0.97	0.03	9.16	1.31	0.66
CD	SELCON 3	8.15	2.71	0.91	10.43	2.73	0.82	4.74	0.91	0.00	9.70	1.24	0.55
	PCA-MR	7.97	2.77	0.93	9.37	1.92	0.85	4.55	0.95	0.14	8.93	1.35	0.68
	PLS	7.72	2.86	0.94	9.47	1.9	0.85	4.47	0.97	0.14	8.96	1.34	0.67
	PLS-1	7.7	2.87	0.94	9.22	1.95	0.89	4.47	0.97	0.00	9.03	1.33	0.67

[a] PCA-MR, principal component analysis followed by multiple regression, constrained to a 100% total; PLS, simultaneous partial least-squares analyses of all structure classes, constrained to a 100% total; PLS-1, separate partial least-squares analyses of each structure type with the use of weighting during the spectral decomposition step. SELCON3 is described in detail in ref 129. [b] The correlation coefficient (R) between the determined and actual fractional composition for the full set of protein spectra.

6.1. Cluster Analysis and Unsupervised Classification

6.1.1. Cluster Analysis

Cluster analysis is the most commonly used unsupervised classification approach. In unsupervised classification, the classes of samples are indentified during the analysis and, therefore, no prior information about the number of clusters and the sample membership is available. Most cluster analysis methods assume that samples with similar spectra belong to the same group or cluster, and Euclidian or Mahalanobis distances are most often used to measure similarity. In the simplest case of a Euclidian distance, the distance d_{ij} is calculated as

$$d_{ij} = \sqrt{(S_i - S_j) \cdot (S_i - S_j)^{\mathrm{T}}} \qquad (23)$$

where S_i and S_j are the spectra of two samples. Preprocessing, such as autoscaling, is often used prior to calculating the distance. Another option is to calculate the distance using the PCA scored in place of experimental spectra in eq 23. If t_i and t_j are the score vectors, the distance d_{ij} will be defined as

$$d_{ij} = \sqrt{(t_i - t_j) \cdot (t_i - t_j)^{\mathrm{T}}} \qquad (24)$$

In the case of a Mahalanobis distance, the distance is weighted by inverse eigenvalues. A Mahalanobis distance is a good measure of dissimilarity in cases where the variation along one of the principal components is larger and therefore more important than the distance in other directions. The Mahalanobis distance is computed as

$$d_{ij} = \sqrt{(t_i - t_j) \cdot \lambda^{-1} \cdot (t_i - t_j)^{\mathrm{T}}} \qquad (25)$$

where λ^{-1} are reciprocal eigenvalues.

In cluster analysis, each sample is assumed to belong to a separate class. After the distances between all pairs of samples are calculated using one of the equations (eqs 23−25), the samples with shortest distances are linked together to form clusters.

Cluster analysis techniques were used to examine a set of Fourier transform infrared (FT-IR) spectra of bovine serum albumin (BSA) in the adsorbed and nonadsorbed states. The region from 1480 to 1600 cm^{-1}, comprising the amide II band, was used, and the single linkage method of cluster analysis was applied. The spectra of adsorbed and nonadsorbed BSA have been found to fall into two distinct clusters. This work illustrated the value of using cluster analysis in the FT-IR study of proteins as a complement to other data analysis methods.[178] Venyaminov and Vassilenko have applied cluster analysis to develop the approach for the characterization of protein tertiary structure using CD spectroscopy.[176] Fifty-three circular dichroism (CD) spectra consisting of the spectra of 46 native proteins, 3 denatured proteins, and 1 oligopeptide were investigated in order to examine the correlation between the shape of the CD spectrum and the tertiary structure class of the protein. To account for the effect of temperature on CD signal in the region 190−236 nm, the spectra of two denatured proteins and an oligopeptide were taken at two different temperatures. Five classes were considered: all-α, all-β, α + β, α/β, and denatured proteins. The cluster analysis was able to divide spectra into several compact groups with good correlation with protein tertiary structure. To further improve the separation of protein classes, the decision function was derived and applied to separate groups of patterns corresponding to each of the tertiary structures. The accuracy of the method has been checked by using the leave-one-out cross-validation, where one spectrum at a time was removed from the training set and the remaining spectra were used to determine the class of the excluded protein. The proposed test gave 100% accuracy for all-α, α/β, and denatured proteins; 85% for α + β; and 75% for all-β proteins. The unsupervised cluster analysis in that study allowed separation of control hemoglobin from β-thalassemia hemoglobin spectra, based mainly on differences in protein secondary structure.[179]

6.1.2. Principal Component Analysis

Principal component analysis is a visual tool for a spectrum classification. Scores and loadings scatter plots in PCA help observe the similarities and natural grouping of proteins as well as the correlation and importance of the spectral channels in protein spectra, respectively. Principal component analysis is a multivariate technique to reduce matrices to their lowest orthogonal space. PCA assumes a bilinear model to explain the observed data variance using a reduced number of factors (principal components):

$$\mathbf{D} = \mathbf{U} \cdot \mathbf{V}^{\mathrm{T}} + \mathbf{E} \qquad (26)$$

where **D** is the data matrix, **U** is the scores, **V** is the loadings matrix, and **E** is a matrix containing the contribution of factors not included into the model. This matrix decomposition (eq 26) is usually performed using the singular value decomposition. The selection of the number of principal components to retain in the model is performed using one of the methods described in section 4.1.1, Deducing the Number of Individual Components in the Data Set. As in the case of multivariate curve resolution, the number of principal components to retain in the model is a trade-off between model simplicity, maximum variance explained by the model, and chemical interpretability.[44] The scores matrix **U** gives information about the samples (proteins) distribution, and the loadings matrix $\mathbf{V}^{\mathrm{T}}$ gives information about relevant variables (wavelengths or wavenumbers). The combination of the information of score and loading plots identifies the structural nature of the different detected protein groups.[44,180] Barron and co-workers have established a PCA-based approach for the analysis of structural relationships between proteins using Raman optical activity data.[153,181] In particular, using a data set of 75 ROA protein spectra, McColl et al.[153] identified seven major classes of proteins based on using the two-dimensional PCA map. The elaborated PCA-based approach allowed the researchers[153] to determine the structural differences between native ovine PrP and reduced ovine PrP. Virkler and Lednev have recently devised a method to identify traces of body fluids found on a crime scene using a combination of the near-infrared Raman and PCA cluster analysis.[182−186] They, for example, demonstrated a complete separation of clusters corresponding to human, canine, and feline blood samples on two- and three-dimensional PCA score scatter plots.[187] The revealed differences among human, canine, and feline blood samples are primarily due to the difference in the relative concentrations of blood proteins across the three groups of sample. The disadvantage of the PCA classification is that the method attempts to capture and explain the gross overall variance in the data set and therefore can often provide a poor representation of individual samples.[188]

6.1.3. When To Use Unsupervised Classification

These methods are good when no information about class membership is available. For example, a PCA model that identifies the direction of the largest variation will not necessarily be a good model for predicting the classes of new protein samples. This is because no information about the classes is incorporated into the PCA model. However, unsupervised methods can provide a basis for the application of supervised methods discussed in the following sections.

6.1.4. Nonlinear Mapping

Nonlinear mapping (NLM)[189,190] can be considered as an alternative to PCA classification. In this approach, multivariate data is normally projected into the two-dimensional space. The points that are close to or distant from each other on a two-dimensional plot represent, respectively, similar or dissimilar protein structures.[191] The PCA transformation searches for directions of the greatest variation of the data in the original multidimensional space but does not preserve the distances in the original space. Unlike PCA, NLM tends to preserve the distance structure of the original multidimensional space[190] and thus provides a more accurate representation of similarities and dissimilarities among spectra.[191] In NLM methods, the following goodness-of-fit function is minimized:

$$\varphi = \sum_{i,j} [d_{ij} - f(\delta_{ij})]^2 \qquad (27)$$

where d_{ij} are the model distances and δ_{ij} stands for the distances in the original space. The function $f(\delta_{ij})$ defines a monotonic transformation of the distances from the original to the model space. The monotonic property of this function ensures that dissimilar spectra are represented by distant points on the two-dimensional chart or, in the mathematics language, the conservation of the rank-ordering of input distances. The Shepard scatter diagram that plots the model distances d_{ij} versus the original similarities between spectra is used to assess the goodness of the transformation (eq 27).

Barron and co-workers have reported several applications of the NLM clustering approach to the vibrational Raman optical activity (ROA) spectra of proteins.[191−193] The two-dimensional NLM plot of a set of 80[188,193] and 85[191] polypeptide, protein, and virus ROA spectra showed excellent clustering corresponding to different types of protein structure. The observed positioning and grouping of proteins on the NLM plot agreed well with X-ray or NMR data for proteins with well-defined native folds and UVCD and/or NMR data for polypeptides and natively unfolded proteins. The orientation of the NLM map axes was chosen such that the *x*-axis represented the steepest change in the α-helix and *β*-sheet content and the *y*-axis aligned with the change in the unordered structure. Such orientation of axes simplified the interpretation of the plot and allowed the researchers, for example, to correlate the polyproline II content of alanine peptides with their chain lengths or reveal the correlation between polyproline II and *β*-sheet contents in the series of alanine peptides.[191] The similarity in the positions of α-sinuclein and its pathogenic mutants on the NLM maps has reinforced the suggestion that the pathogenic properties of the mutants could unlikely be explained by the structural changes caused by the mutations.

6.2. Supervised Classification

In contrast to cluster analysis, supervised classification methods require a training data set containing the spectra of samples with known membership. For example, in ref 177 the Raman spectra of human, bovine, porcine, and other types of insulin have been used as a training data set and the spectrum of the unknown sample was then assigned to one of the classes by using partial least-squares discriminant analysis (PLS-DA) and linear discriminant analysis (LDA).

6.2.1. Linear Discriminant Analysis

This method maximizes the ratio of between-class variance to the within-class variance in any particular data set, thereby guaranteeing maximal separation.[194] LDA along with cluster analysis have been applied to the classification of IR hemoglobin spectra of healthy people and *β*-thalassemia patients.[179] The supervised LDA method provided 100% classification accuracy for the training set and 98% accuracy for the validation set in partitioning control and *β*-thalassemia samples. The IR spectra revealed changes in the secondary structure of hemoglobin from *β*-thalassemia patients compared to those taken from healthy individuals. In pathological samples, a decreased α-helix content, an increased content of parallel and antiparallel *β*-sheets, and changes in the tyrosine ring absorption band have been found. The hemoglobin from *β*-thalassemia patients also showed an increase in the intensity of the IR bands from the cysteine -SH groups.

6.2.2. SIMCA Classification

There is some disagreement about the source of the acronym (Soft Independent Method of Class Analogy or Standard Isolinear Method of Class Assignment). SIMCA can be regarded as a supervised version of the PCA classification (see section 6.1.2).[195] In PCA, class information is not used in the construction of the model and a PCA model just attempts to describe the overall variation in the data.[19] SIMCA uses principal component analysis but incorporates the information about the classes contained in a training data set. In SIMCA, a PCA is performed on each class in the data set, and a sufficient number of principal components are retained to account for most of the variation within each class. The number of principal components retained for each class is usually different. Deciding on the number of principal components that should be retained for each class is important, as retention of too few components can result in loss of useful information while using too many principal components increases the fraction of noise in the model. Cross-validation is often used to determine the optimal number of principal components (see section 4.1.1: Deducing the Number of Individual Components in the Data Set). SIMCA classification has some attractive features that have made the method popular among spectroscopists. First, raw spectral data with thousands of variables in SIMCA is mapped on a PCA map, with the dot or marker representing each sample. Thus, the analysis has a visual tool for assigning the unknown sample to a class. Besides the visual method of classification, SIMCA provides a rigorous quantitative classification based on the residual variance for the unknown sample. If the residual variance of a sample exceeds the upper limit for every class in the data set, the sample would not be assigned to any of the classes and treated as an outlier. In addition, most statistical packages, such as the PLS Toolbox by Eigenvector, Inc., SIMCA P+ by Umetrics, and Unscrambler from CAMO, allow evaluating the importance of each sample in the modeling of variation and discriminating among samples. Samples and variables with low modeling power are usually deleted from the model because they account for noise in the data. Furthermore, SICMA requires as few as ten samples per class and thus can be applied in the studies where only a limited number of spectra can be collected.

6.2.3. Partial Least-Squares Discriminant Analysis

While SIMCA is a very useful classification tool, it does have drawbacks. The main one is that the PCA submodels in SIMCA are computed to capture maximum variation within each class. No attempt is made to identify directions in the data space that discriminate classes directly.[19] In contrast to SIMCA, PLS-DA is specifically looking for directions that ensure the best separation or discrimination between the classes. PLS-DA is very similar to the linear discriminant analysis discussed in section 6.2.1. PLS-DA combines the discriminative power of LDA with robustness to noise and multiple correlations between spectral channels inherited from PLS.[196] PLS-DA is a variant of classical PLS regression where the response variable is categorical rather than numeric. The *y*-block of dummy variables for each sample is created to establish the class membership so that 1 indicates that the sample belongs to a class and 0 shows that it does not. The PLS model yields the numeric value for each class, and the value is compared with the threshold estimated using the Bayesian approach or cross-validation. For example, if the calculated value for a sample is 0.45 and the class threshold is 0.4, the sample will be assigned to this class.

Application of classification methods to spectral data of proteins is not well-established. Navea and co-workers[44] tested the suitability of CD and MIR spectra, much simpler than X-ray diffraction or NMR experimental measurements, for protein classification and elucidation of protein secondary structure using unsupervised and supervised classification followed by iPLS modeling. The researchers applied unsupervised pattern recognition methods, such as PCA and cluster analysis, to explore the natural distribution of proteins into different groups on the basis of their CD and MIR spectra. Later, protein classification has been performed with PLS-DA. The study attempted to establish a method of assigning the unknown protein to one of the following classes: all-α (mainly α-helical), all-β (mainly β-pleated sheet), $\alpha-\beta$ (separate α-helix and β-sheet regions and intermixed α-helices and β-sheet regions), and random (predominantly unordered).

The combined use of chemometric tools and CD and IR spectroscopies has been proven to be a good alternative for protein class assignment to the main all-α and all-β protein classes. Intermediate protein classes could not be explicitly modeled, even using combined CD/IR measurements. A recent study by Berman et al.[197] compares the performance of five chemometric methods [PCA, LDA, PLS-DA, soft independent modeling of class analogy (SIMCA), and decision tree] in the analysis of time-of-flight secondary ion mass spectrometry (ToF-SIMS) of complex biological samples including proteins. For both pure protein and complex protein mixture samples, LDA, PLS-DA, and SIMCA all produced excellent classification. PLS-DA has been shown to provide a more accurate classification of tissue samples compared to SIMCA and decision tree.[197]

What Is the Best Classification Method? Linear discriminant analysis is easy to use and should suffice in most cases. Unfortunately, the accuracy of the algorithm can be affected by multiple colinearities in spectral data, which is often the case. SIMCA and PLS-DA are more robust to correlations in the data, and PLS-DA has a higher discriminative power.

7. Feature Extraction and Database Search Algorithms

7.1. Feature Extraction

Feature extraction is a broad class of algorithms that help identify relevant peaks or regions in protein spectra that can then be used for classification or calibration.[198] We have already mentioned the genetic algorithm for variable selection and interval partial least-squares when discussing PLS in section 5.3. Both these examples pertain to feature extraction. This section provides a brief overview of feature selection methods used in mass spectroscopy proteomics studies where feature extraction is a routine tool for the analysis of high throughput mass spectral data.[199] Any further progress in feature extraction methods applicable to MS proteomics data will be highly recognized, since the fidelity of many proposed clinical tests largely relies on the accuracy of feature extraction and classification methods.[200,201] Over the past decade, a lot of effort has been put into developing mass spectroscopy methods for the diagnostics of various types of cancer.[202,203] Unfortunately, the complexity of mass spectra and the large amount of redundant information make the comparison of the spectra from healthy individuals and cancer patients extremely difficult, if not impossible. At the stage of method development, two groups of spectra—one from healthy individuals and one from cancer patients—are recorded. Feature extraction algorithms are utilized to find peaks called features such that the variation in those peaks between the two groups of spectra is significantly larger than the variation within either group. Normally the selected spectral features are associated with a protein biomarker whose concentration is higher in samples from cancer patients. Considering the complexity of mass spectra and the enormously large number of features, several selection algorithms have been elaborated.[203,204] Most selection algorithms can be classified as filters, wrappers, or embedded methods. Filters select one feature at a time and then rank features based on their classification power.[204] An example of a filter selection approach is a *t*-test. The genetic algorithm is normally used as a search method in order to find the optimal subset of features in wrapper methods. Wrapper methods search for subsets of features that yield the best classification performance. Decision trees are examples of embedded search methods.[205] No one feature selection method is superior to all the others; every method has its own advantages and disadvantages. For example, filters usually run faster than wrappers, but wrappers are more likely to select features that can produce better classification results than filters.[205] Feature extraction algorithms aim at finding the minimum number of spectral features that result in reliable classification.

To distinguish between healthy women and those afflicted with cancer,[206] principle component analysis (PCA) for dimensionality reduction and linear discriminant analysis have been applied. In ref 201, the researchers compared two feature extraction algorithms together with several classification approaches on MALDI TOF acquired data. The Student *t*-test was used to rank features in terms of their relevance. Support vector machines, random forests, linear/quadratic discriminant analysis (LDA/QDA), *k*-nearest neighbors, and bagged/boosted decision trees were subsequently used to classify the data. The studies by Ilya Levner[199,203] examined the performance of the nearest centroid classifier coupled with the following feature selection algorithms. The Student *t*-test, Kolmogorov−Smirnov test, and P-test are univariate statistics used for filter-based feature ranking. For other applications of feature extraction methods in cancer research, see refs 207 and 208. A novel algorithm called a guilt-by-association feature selection has been proposed by Shin et al.[205] A good review of feature selection methods which use NMR and MS case examples is provided in ref 198.

7.2. Database Search Methods

This group of algorithms is used for quick matching of the spectrum of an unknown sample with a spectrum in a reference database. The examples of a reference spectral database are SP175, which consists of more than 70 CD protein spectra,[125] with known X-ray secondary structures, FTIRsearch.com, with FTIR and Raman spectra, and reference libraries of protein sequences in tandem mass spectroscopy (for example, http://gpmdb.thegpm.org/).

7.2.1. Classical Search Methods

Classical methods of spectral search are generic, and their application is not limited to a specific type of spectral data. We will briefly describe correlation, Euclidian distance, absolute value correlation search, and least-squares search algorithms.[209,210] All these algorithms calculate a measure of similarity between the experimental and reference spectrum called the hit quality index (HQI).[209] The database spectra are ranked relative to their hit quality indices: the closer the HQI to 1, the better the match between the library spectrum and the unknown spectrum. The difference in these search methods lies in the equation that each of the methods is using to calculate the HQI. Each of the classical search methods can be applied to both raw spectra and their first derivatives. Calculating the first derivative eliminates the effect of a baseline and exaggerates small shift differences in peak positions and thus helps achieve better selectivity. Furthermore, first derivative search methods put more emphasis on peak positions than on peak intensities. Unfortunately, differentiation normally yields very noisy spectra even with a moderate level of noise in the original spectra. Smoothing methods such as fast Fourier transform, Savitzky−Golay, or wavelet denoising are usually applied prior to differentiation. Classical algorithms constitute the body of the search engine of many software packages, e.g. Thermo Scientific and spectral banks such as FTIRsearch.com.

7.2.1.1. Correlation and Euclidean Distance Search Algorithms. The correlation is very similar to the Euclidean distance search. Correlation search is performed on mean-centered data, and this makes this method robust to the presence of negative peaks or negative regions caused by improper baseline correction. This algorithm is a good choice in searching spectra with low signal-to-noise ratio. The main disadvantage of the method is that it is slower than most other methods. Euclidean distance is faster than correlation search but is more sensitive to a baseline and negative peaks. It is a good option when a baseline is properly subtracted.

7.2.1.2. Absolute Value and Least Squares Search. In the absolute value search method, the absolute differences between the unknown spectrum and reference spectra are calculated. This method is sensitive to small differences between the submitted unknown and reference spectra. The disadvantage of the absolute value method is its low selectivity. This algorithm may retrieve absolutely different spectra with the same HQI, especially in the case of low hit performance indices. The least squares method finds the difference between the unknown and reference spectra at each spectral channel and computes the hit quality index as the average mean least-squares difference between the spectra. This method gives large weights to strong peaks and can be a good choice in the case of noisy spectra. The advantage of classical methods is that they are fairly simple and generic so that the analyst is not required to fine-tune the search algorithm in each particular case. Unfortunately, classical methods rely on computer brute force and can be very slow when applied to fairly large databanks. More efficient search algorithms have been developed to compensate for the shortcomings of classical methods.

7.2.1.3. What Is the Best Classical Search Algorithm? As we discussed above, different algorithms differ in how they weigh the importance of peak positions and relative intensity in matching spectra.[209] Correlation and Euclidean distance methods are a good choice when both band intensities and positions matter whereas all derivative methods exaggerate peak positions.

7.2.2. Search in the Fourier and Wavelet Domains

Both fast Fourier transform (FFT)[211] and wavelet transform (WT)[212] are routinely used for data compression. For example, the IR spectrum containing 4000 points can be reconstructed using a few hundreds of Fourier or wavelet decomposition coefficients without loss in quality.[212] Transforming data into Fourier or wavelet domains serves two purposes: first the storage space required for a spectral library is significantly reduced; second, the speed of search algorithms is dramatically increased due to smaller size and complexity of data.

Researchers in ref 213 demonstrated a technique for the efficient searching of mid-infrared spectral libraries. The algorithm has been used for the analysis of the composition of complex mixtures by decomposing the spectra of mixtures using the spectra in the reference library. Both library and mixture spectra have been converted into the Fourier domain to enhance the searching performance. The algorithm further invokes principal component analysis to generate an orthonormal reference library and to compute the projections or scores of a mixture spectrum onto the principal space spanned by the orthonormal set. Calibration coefficients calculated from library scores have been used to predict the mixture composition. Leung et al.[214] compared search algorithms based on Fourier and wavelet transforms. In this paper, the fast wavelet transform (FWT) and its derivative, the wavelet packet transform (WPT), were applied to compress the infrared (IR) spectrum for storage and spectral searching. The authors have concluded that algorithms using a wavelet transform are faster compared to FFT algorithms.

7.2.3. Methods Using Principal Component Decomposition

Principal component analysis (PCA) is another powerful technique used for searching spectral libraries.[215] Researchers in ref 216 applied principal component decomposition to the library of UV−vis spectra to find pure compounds. Then the scores of the unknown spectra were converted to the concentration fractions of individual components. Researchers used adaptive filtering by repeating principal component regression (PCR) on the subset of the library spectra having the best match with the unknown spectra and eliminating the contribution of the identified compound at each subsequent PCR step. Recently, Gianella et al.[217] proposed a method for the search for an infrared spectral library using principal component regression and adaptive spectral filtering. The latter subtracts the contribution of each identified compound from both the unknown spectra at each iteration of principal component regression, thus sorting the identified compounds according to their spectral fraction in the submitted unknown spectra. Additional improvements have enhanced the robustness of the algorithm in cases when the unknown spectrum has contributions from compounds not contained in a database.

A number of search algorithms have been specifically developed for tandem mass spectroscopy applications.[218,219] The goal of the tandem mass spectrometry is to find the best amino acid sequence that matches the spectrum. Matching the mass spectra with a low signal-to-noise ratio can often be a nontrivial task, and multiple search algorithms are used in such cases. A review by Sadygov et al. provides classification, detailed explanation, and use examples of state-of-the-art methods in the field.[220] A detailed description of the algorithms for searching MS libraries is also provided in refs 221 and 222.

8. Further Trends in the Development of Statistical and Numerical Methods in the Context of (Bio)-chemistry and Spectroscopy Data

The continuing progress in spectroscopic technology and increased volumes of data generated by instruments calls for the development of dimension reduction[223] and feature extraction[224] methods such as nonlinear PCA, self-organizing maps, and multidimensional scaling in the context of biospectroscopy data. Increasing accuracy of spectroscopic measurements will also allow taking into account and correcting for the nonlinearity of the spectroscopic response. The development of hyphenated analytical techniques such as LC-MS, GS-MS, and CE-MS, which produce a matrix of data for each sample and a 3D matrix for the data set, paves the way for further development of multiway algorithms[225] such as Multiway PCA, PARAFAC, Tri Linear Decomposition, and others.

The majority of methods discussed in this paper relies on the assumptions that the intensity of signal is (i) proportional to the concentration of the component and (ii) is not affected by the presence of other components. These, however, are not always the case. We envision an increasing number of applications of support vector machines[226] and least-squares support vector machines[168] where the nonlinear response can be modeled by the appropriate kernel function.

The complexity of protein spectra hampers the application of most multivariate curve resolution methods. Namely, resolving broad spectra with multiple overlapping bands is error-prone and may often provide ambiguous results. There is a high demand for interactive methods and algorithms that allow for incorporating maximum prior information about the spectra and concentrations. The most promising methods providing a quantitative way to incorporate prior information into the model are based on the Bayesian formalism.[227]

Over the recent years, a lot of effort has been invested into the development of wavelet-based methods for the analysis of complex mixtures. New powerful techniques such as fractional wavelet transform, the combined used of the wavelet transform with zero-crossing, ratio spectra-zero crossing, the generalized mean value approach, and double divisor approaches for overlapping signals may find their applications in biomedical and protein studies.[228,229]

The content of electronic spectral libraries has been increasing in size exponentially over the past decade. In order to cope with large amounts of data, the selectivity and the speed of database search algorithms should be optimized. In general, the increase in selectivity can be achieved by using probabilistic approaches for feature selection and incorporating a maximum of prior information about the spectra in order to direct the database search.[230] For example, Razor Spectrometry software,[231] which is now available as part of GRAMS AI by Thermo Scientific, utilizes Bayesian algorithms[232] for denoising and feature selection.

There is no doubt that the size and the complexity of a typical spectral data set collected in the biochemical experiment will be increasing in the upcoming years, so that the data analysis may soon become impractical without intelligent informatics methods. The number of available statistical algorithms as well as their sophistication will also be growing high, which leaves the experimentalist with a hard choice as to which method to choose and how to use it properly. Although discoveries in mathematical sciences has always been ahead of the time, vendors of statistical software have already begun to commercialize many of the methods described here and will continue to do so in the future. For those readers who would like to begin using advanced statistical methods, we strongly recommend to start with validated out-of-shelf statistical products such as PLS_Toolbox by Eigenvector, SIMCA P+ by Umetrics, and Unscrambler by CAMO. These companies provide cheep academia licenses as well as flexible training and consulting if needed. In addition, these programs have user-friendly GUI and interactive result interpreters, which newcomers into the field will find especially useful. We would advise beginners against using custom data analysis programs available on some university or personal Web sites. Although in most cases those are excellent programs, they are written for expert level users and are poorly documented, which increases the risk of getting erroneous results from using such homemade programs.

9. List of Abbreviations

AFA	abstract factor analysis
ALS	alternating least squares
BSA	bovine serum albumin
BSS	blind signal separation
CCA	convex constraint analysis
CD	circular dichroism
CSD	charge state distribution
DUVRR	deep UV resonance Raman
EFA	evolving factor analysis
ESI-MS	electrospray ionization mass spectrometry
FFT	fast Fourier transform

FT-IR	Fourier transform infrared
FWT	fast wavelet transform
HQI	hit quality index
ICA	independent component analysis
iPLS	interval partial least squares
LDA	linear discriminant analysis
LDA/QDA	linear/quadratic discriminant analysis
LS-SVM	least squares support vector machines
MCR	multivariate curve resolution
MEM	maximum entropy method
MLR	multiple linear regression
NLM	nonlinear mapping
NLS	nonlinear least squares
PCA	principal component analysis
PCR	principal component regression
PLS	partial least squares
PLS-DA	partial least squares discriminant analysis
PLS-1	weighted PLS
RBF	radial basis function
rms	root-mean-squared error
ROA	Raman optical activity
SIMCA	soft independent method of class analogy
SIMPLISMA-	simple-to-use interactive self-modeling mixture analysis
SEONS	second-order nonstationary source separation
SMAC	stepwise maximum angle calculation
SOBI	second-order blind identification
SVM	support vector machines
ToF-SIMS	time-of-flight secondary ion mass spectrometry
VCD	vibrational circular dichoism
WPT	wavelet packet transform
WT	wavelet transform
2D NOESY	2D nuclear Overhauser enhancement

10. Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant CHE-0809525 (I.K.L.).

11. References

(1) Brereton, R. G. *Chemometrics. Data Analysis for the Laboratory and Chemical Plant*; John Wiley & Sons: New York, 2003.

(2) *Protein Structure, Stability and Folding*; Murphy, K. P., Ed.; Humana Press Inc.: Totowa, NJ, 2001.

(3) Permyakov, E. *Luminescent Spectroscopy of Proteins*; CRC Press: Boca Raton, FL, 1993.

(4) Lakowicz, J. *Principles of Fluorescence Spectroscopy*, 3rd ed.; Plenum Press: New York, 2006.

(5) Bretthorts, G. L.; Hutton, W. C.; Garbow, J. R.; Ackerman, J. J. H. *Concepts Magn. Reson.* **2005**, *27A*, 55.

(6) Sivia, D. S.; Skilling, J. *Data Analysis: A Bayesian Tutorial*, 2nd ed.; Oxford University Press: Oxford, 2006.

(7) Sivia, D. S.; Carlile, C. J. *J. Chem. Phys.* **1992**, *96*, 170.

(8) Juan, A. d.; Tauler, R. *Crit. Rev. Anal. Chem.* **2006**, *36*, 163.

(9) Cichocki, A.; Amari, S.-i. *Adaptive Blind Image and Signal Processing*; John Wiley & Sons: New York, 2002.

(10) Zibulevsky, M.; Pearlmutter, B. A. *Neural Comput.* **2001**, *13*, 863.

(11) Widjaja, E.; Li, C.; Chew, W.; Garland, M. *Anal. Chem.* **2003**, *75*, 4499.

(12) Malinowski, E. R. *Factor Analysis in Chemistry*, 3rd ed.; John Wiley & Sons, Inc.: New York, 2002.

(13) Shashilov, V. A.; Ermolenkov, V. V.; Lednev, I. K. *Inorg. Chem.* **2006**, *45*, 3606.

(14) Xu, M.; Shashilov, V. A.; Ermolenkov, V. V.; Fredriksen, L.; Zagorevski, D.; Lednev, I. K. *Protein Sci.* **2007**, *16*, 815.

(15) Svensson, O.; Josefson, M.; Langkilde, F. W. *Chemom. Intell. Lab. Syst.* **1999**, *49*, 49.

(16) Cooper, J. B. *Chemom. Intell. Lab. Syst.* **1999**, *46*, 231.

(17) Armenta, S.; Garrigues, S.; de la Guardia, M. *Anal. Chim. Acta* **2004**, *521*, 149.

(18) Gallagher, N. B.; Blake, T. A.; Gassman, P. L.; Shaver, J. M.; Windig, W. *Appl. Spectrosc.* **2006**, *60*, 713.

(19) Wise, B. M.; Gallagher, N. B.; Bro, R.; Shaver, J. M.; Windig, W.; Koch, R. S. *PLS_Toolbox 4.0 for Use with Matlab*; Eigenvector Research, Inc.: Manson, WA, 2006.

(20) Rencher, A. C. *Multivariate Statistical Inference and Apllications*; Jonh Wiley & Sons: New York, 1998.

(21) Tomisic, V.; Simeon, V. *Phys. Chem. Chem. Phys.* **2000**, *2*, 1943.

(22) Keller, H. R.; Massart, D. L. *Chemom. Intell. Lab. Syst.* **1992**, *12*, 209.

(23) Thibault, C.; Huguet, P.; Sistat, P.; Pourcelly, G. *Desalination* **2002**, *149*, 429.

(24) de Juan, A.; Tauler, R. *Anal. Chim. Acta* **2003**, *500*, 195.

(25) Xu, M.; Ermolenkov, V. V.; He, W.; Uversky, V. N.; Fredriksen, L.; Lednev, I. K. *Biopolymers* **2005**, *79*, 58.

(26) Garrido, M.; Lazaro, I.; Larrechi, M. S.; Rius, F. X. *Anal. Chim. Acta* **2004**, *515*, 65.

(27) Gemperline, P. J.; Cash, E. *Anal. Chem.* **2003**, *75*, 4236.

(28) Jaumota, J.; Gargalloa, R.; de Juan, A.; Tauler, R. *Chemom. Intell. Lab. Syst.* **2005**, *76*, 101.

(29) van Zomeren, P. V.; Darwinkel, H.; Coenegracht, P. M. J.; de Jong, G. J. *Anal. Chim. Acta* **2003**, *487*, 155.

(30) Potyrailo, R. A. *Trends Anal. Chem.* **2003**, *22*, 374.

(31) Plumbley, M. D.; Oja, E. *IEEE Trans. Neural Network* **2004**, *15*, 66.

(32) Navea, S.; de Juan, A.; Tauler, R. *Anal. Chem.* **2002**, *74*, 6031.

(33) Navea, S.; Tauler, R.; Juan, A. *Anal. Chem.* **2006**, *78*, 4768.

(34) Mohimen, A.; Dobo, A.; Hoerner, J. K.; Kaltashov, I. A. *Anal. Chem.* **2003**, *75*, 4139.

(35) Borges, A.; Tauler, R.; de Juan, A. *Anal. Chim. Acta* **2005**, *544*, 159.

(36) Yuan, B.; Murayama, K.; Wu, Y.; Tsenkova, R.; Dou, X.; Era, S.; Ozaki, Y. *Appl. Spectrosc.* **2003**, *57*, 1223.

(37) Navea, S.; Tauler, R.; de Juan, A. *Anal. Biochem.* **2005**, *336*, 231.

(38) Rodríguez-Casado, A.; Molina, M.; Carmona, P. *Proteins: Struct., Funct., Bioinf.* **2006**, *66*, 110.

(39) Domıńguez-Vidal, A.; Saenz-Navajas, M. a. P.; Ayora-Cañada, M. J.; Lendl, B. *Anal. Chem.* **2006**, *78*, 3257.

(40) Tauler, S. N. R.; Goormaghtigh, E.; Juan, A. d. *Proteins: Struct., Funct., Bioinf.* **2006**, *63*, 527.

(41) Tauler, R.; Marques, I.; Casassas, E. *J. Chemometrics* **1998**, *12*, 55.

(42) Pan, P. W.; Gordon, H. L.; Rothstein, S. M. *J. Chem. Phys.* **2006**, *124*, 024905.

(43) Shashilov, V. A.; Xu, M.; Ermolenkov, V. V.; Lednev, I. K. *J. Quant. Spectrosc. Radiat. Transfer* **2006**, *102*, 46.

(44) Navea, S.; Tauler, R.; Goormaghtigh, E.; Juan, A. *Proteins: Struct., Funct., Bioinf.* **2006**, *63*, 527.

(45) Lednev, I. K. In *Protein Structures, Methods in Protein Structures and Stability Analysis*; Uversky, V. N., Permyakov, E. A., Eds.; Nova Science Publishers, Inc.: New York, 2007; p 1.

(46) Asher, S. A. In *Handbook of Vibrational Spectroscopy*; John Wiley & Sons: New York, 2001; Vol. 1, p 558.

(47) Hudson, B.; Mayne, L. *Methods Enzymol.* **1986**, *130*, 331.

(48) Carey, P. R. *Molecular Biology: Biochemical Applications of Raman and Resonance Raman Spectroscopies*; Academic Press: New York, 1982.

(49) *Infrared and Raman Spectroscopy of Biological Materials*; Gremlich, H.-U., Yan, B., Eds.; Marcel Dekker: New York, 2001; Vol. 24.

(50) Tu, A. T. In *Spectroscopy of Biological Systems*; Clark, R. J. H., Hester, R. E., Eds.; John Wiley & Sons: Chichester, U.K., 1986; Vol. 13.

(51) *Biological Applications of Raman Spectroscopy. Vol. 1: Raman Spectra and the Conformations of Biological Macromolecules*; Spiro, T. G., Ed.; John Wiley and Sons: New York, 1987; Vol. 1.

(52) Asher, S. A. *Anal. Chem.* **1993**, *65*, 59A.

(53) Harada, I.; Takeuchi, H. In *Advances in Spectroscopy, Vol. 13: Spectroscopy of Biological Systems*; Clark, R. J. H., Hester, R. E., Eds.; John Wiley & Sons: Chichester, U.K., 1986.

(54) Ahmed, Z.; Beta, I. A.; Mikhonin, A. V.; Asher, S. A. *J. Am. Chem. Soc.* **2005**, *127*, 10943.

(55) Wu, Q.; Li, F.; Wang, W.; Hecht, M. H.; Spiro, T. G. *J. Inorg. Biochem.* **2002**, *88*, 381.

(56) Vaillancourt, F. H.; Barbosa, C. J.; Spiro, T. G.; Bolin, J. T.; Blades, M. W.; Turner, R. F.; Eltis, L. D. *J. Am. Chem. Soc.* **2002**, *124*, 2485.

(57) Juszczak, L. J.; Fablet, C.; Baudin-Creuza, V.; Lesecq-Le Gall, S.; Hirsch, R. E.; Nagel, R. L.; Friedman, J. M.; Pagnier, J. *J. Biol. Chem.* **2003**, *278*, 7257.

(58) Samuni, U.; Dantsker, D.; Khan, I.; Friedman, A. J.; Peterson, E.; Friedman, J. M. *J. Biol. Chem.* **2002**, *277*, 25783.

(59) Hashimoto, S.; Sasaki, M.; Takeuchi, H.; Needleman, R.; Lanyi, J. K. *Biochemistry* **2002**, *41*, 6495.

(60) Lin, S. W.; Kochendoerfer, G. G.; Carroll, K. S.; Wang, D.; Mathies, R. A.; Sakmar, T. P. *J. Biol. Chem.* **1998**, *273*, 24583.

(61) Rodriguez-Casado, A.; Thomas, G. J., Jr. *Biochemistry* **2003**, *42*, 3437.

(62) Serban, D.; Arcineigas, S. F.; Vorgias, C. E.; Thomas, G. J., Jr. *Protein Sci.* **2003**, *12*, 861.

(63) Wu, Q.; Hamilton, T.; Nelson, W. H.; Elliott, S.; Sperry, J. F.; Wu, M. *Anal. Chem.* **2001**, *73*, 3432.

(64) Wu, Q.; Nelson, W. H.; Treubig, J. M., Jr.; Brown, P. R.; Hargraves, P.; Kirs, M.; Feld, M.; Desari, R.; Manoharan, R.; Hanlon, E. B. *Anal. Chem.* **2000**, *72*, 1666.

(65) Mukerji, I.; Williams, A. P. *Biochemistry* **2002**, *41*, 69.

(66) Maiti, N. C.; Tomita, T.; Kitagawa, T.; Okamoto, K.; Nishino, T. *J. Biol. Inorg. Chem.* **2003**, *8*, 327.

(67) Nagatomo, S.; Nagai, M.; Shibayama, N.; Kitagawa, T. *Biochemistry* **2002**, *41*, 10010.

(68) Okada, A.; Miura, T.; Takeuchi, H. *Biochemistry* **2003**, *42*, 1978.

(69) Couling, V. W.; Fischer, P.; Klenerman, D.; Huber, W. *Biophys. J.* **1998**, *75*, 1097.

(70) Chi, Z.; Chen, X. G.; Holtz, J. S.; Asher, S. A. *Biochemistry* **1998**, *37*, 2854.

(71) Schulze, H. G.; Greek, L. S.; Barbosa, C. J.; Blades, M. W.; Gorzalka, B. B.; Turner, R. F. *J. Neurosci. Methods* **1999**, *92*, 15.

(72) Clarkson, J.; Smith, D. A. *FEBS Lett.* **2001**, *503*, 30.

(73) Ianoul, A.; Mikhonin, A.; Lednev, I. K.; Asher, S. A. *J. Phys. Chem. A* **2002**, *106*, 3621.

(74) Shashilov, V.; Xu, M.; Ermolenkov, V. V.; Fredriksen, L.; Lednev, I. K. *J. Am. Chem. Soc.* **2007**, *129*, 6972.

(75) Amigoa, J. e. M.; Juan, A. d.; Coello, J.; Maspocha, S. *Anal. Chim. Acta* **2006**, *567*, 245.

(76) Hyvarinen, A.; Karhunen, J.; Oja, E. *Independent component analysis*; John Wiley & Sons: New York, 2001.

(77) Hyvarinen, A. *Neural Comput. Surveys* **1999**, *2*, 94.

(78) Shao, X.; Wang, G.; Wang, S.; Su, Q. *Anal. Chem.* **2004**, *76*, 5143.

(79) Bell, A. J.; Sejnowski, T. J. *Neural Comput.* **1995**, *7*, 1129.

(80) Hyvarinen, A.; Oja, E. *Neural Networks* **2000**, *13*, 411.

(81) Kano, M.; Hasebe, S.; Hashimoto, I.; Ohno, H. *Comput. Chem. Eng.* **2004**, *28*, 1157.

(82) Gustafsson, M. G. *J. Chem. Inf. Model.* **2005**, *45*, 1244.

(83) Szabo de Edelenyi, F.; Simonetti, A. W.; Postma, G.; Huo, R.; Buydens, L. M. C. *Anal. Chim. Acta* **2005**, *544*, 36.

(84) Ladisa, M.; Lamurab, A.; Nicoc, G.; Siliq, D. *Physica A* **2005**, *349*, 571.

(85) Bonnet, N.; Nuzillard, D. *Ultramicroscopy* **2005**, *102*, 327.

(86) Pichler, A.; Sowa, M. G. *J. Mol. Spectrosc.* **2005**, *229*, 231.

(87) Alrubaiee, M.; Xu, M.; Gayen, S. K.; Brito, M.; Alfano, R. R. *Appl. Phys. Lett.* **2005**, *87*, 191112.

(88) Chen, Y.-W.; Han, X.-H.; Nozaki, S. *Rev. Sci. Instrum.* **2004**, *75*, 3977.

(89) Kopriva, I.; Du, Q.; Szu, H.; Wasylkiwskyj, W. *Opt. Commun.* **2004**, *233*, 7.

(90) Chung, S. H.; Park, C. S.; Park, K. S. *Proc. SPIE* **2005**, *5702*, 168.

(91) Nielsen, H. B. UCMINF—an Algorithm for Unconstrained, Nonlinear Optimization; IMM, Technical University of Denmark: 2001.

(92) Opper, M.; Winther, O. *Phys. Rev. Lett.* **2001**, *86*, 3695.

(93) Oja, E.; Plumbley, M. *Neural. Computat.* **2004**, *19*, 1811.

(94) Plumbley, M. D. *International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2004)*, Granada, Spain, 2004; p *49*.

(95) Yuan, Z.; Oja, E. *International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2004)*, Granada, Spain, 2004; p 1.

(96) Plumbley, M. D. *IEEE Trans. Neural Network* **2003**, *14*, 534.

(97) Oja, E.; Plumbley, M. *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2003)*, Nara, Japan, 2003; p 11.

(98) Painter, T. H.; Paden, B.; Dozier, J. *Rev. Sci. Instrum.* **2003**, *74*, 5179.

(99) Gruber, P.; Stadlthanner, K.; Tomé, A. M.; Teixeira, A. R.; Theis, F. J.; Puntonet, C. G.; Lang, E. W. In *Independent Component Analysis and Blind Signal Separation*; Springer: Berlin/Heidelberg, 2004; Vol. 3195.

(100) Gruber, P.; Theis, F. J.; Stadlthanner, K.; Lang, E. W.; Tome, A. M. Teixeira. In *2004 IEEE International Joint Conference on Neural Networks*; 25−29 July, 2004.

(101) Mantini, D.; Petrucci, F.; Boccio, P. D.; Pieragostino, D.; Nicola, M. D.; Lugaresi, A.; Federici, G.; Sacchetta, P.; Ilio, C. D.; Urbani, A. *Struct. Bioinf.* **2008**, *24*, 63.

(102) Chen, J.; Wang, X. Z. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 992.

(103) Gruber, P.; Stadlthanner, K.; Böhm, M.; Theis, F. J.; Lang, E. W.; Tomé, A. M.; Teixeira, A. R.; Puntonet, C. G.; Saéz, J. M. G. *Neurocomputing* **2006**, *69*, 1485.

(104) Cardoso, J.-F. *Neural Comput.* **1999**, *11*, 157.

(105) Nuzillard, D.; Nuzillard, J.-M. *Signal Process.* **2003**, *83*, 627.

(106) Choi, S.; Cichocki, A.; Belouchrani, A. J. *VLSI Signal Process.* **2002**, *32*, 93.

(107) Xu, M.; Shashilov, V.; Lednev, I. K. *J. Am. Chem. Soc.* **2007**, *129*, 11002.

(108) Windig, W.; Guilment, J. *Anal. Chem.* **1991**, *63*, 1425.

(109) Ferrasse, J. H.; Chavez, S.; Arlabosse, P.; Dupuyc, N. *Thermochim. Acta* **2003**, *404*, 97.

(110) Windig, W. *Chemom. Intell. Lab. Syst.* **1997**, *36*, 3.

(111) Windig, W.; Antalek, B.; Lippert, J. L.; Batonneau, Y.; Bremard, C. *Anal. Chem.* **2002**, *74*, 1371.

(112) Windig, W.; Gallagher, N. B.; Shaver, J. M.; Wise, B. M. *Chemom. Intell. Lab. Syst.* **2005**, *77*, 85.

(113) Windig, W.; Stephenson, D. A. *Anal. Chem.* **1992**, *64*, 2735.

(114) Bogomolov, A.; Hachey, M. *Chemom. Intell. Lab. Syst.* **2007**, *88*, 132.

(115) Hachey, M.; Bogomolov, A.; Boruta, M. *Easter Analytical Symposium and Exhibit (EAS)*; Somerset, NJ, 2006.

(116) Buxton, T. L. Solving problems in ion mobility measurements of forensic samples with thermal desorption and dynamics. Ph.D. Dissertation, Ohio University, Athens, OH, 2002.

(117) Knuth, K. *SPIE'98 Proceedings: Bayesian Inference for Inverse Problems*; SPIE: San Diego, 1998; p 147.

(118) Steinbach, P. J.; Ionescu, R.; Matthews, C. R. *Biophys. J.* **2002**, *82*, 2244.

(119) Shashilov, V. A.; Lednev, I. K. *AIP Conf. Proc.* **2007**, *954*, 450.

(120) Mikhonin, A. V.; Bykov, S. V.; Myshakina, N. S.; Asher, S. A. *J. Phys. Chem. B* **2006**, *110*, 1928.

(121) Ferrige, A. G.; Seddon, M. J.; Green, B. N.; Jarvis, S. A.; Skilling, J.; Staunton, D. J. *Rapid Commun. Mass Spectrosc.* **2005**, *6*, 707.

(122) Green, B. N.; Hutton, T.; Vinogradov, S. N. In *Protein and Peptide Analysis by Mass Spectrometry*; Humana Press: Totowa, NJ, 1996; Vol. 61.

(123) Xu, M.; Ermolenkov, V. V.; Uversky, V. N.; Lednev, I. K. *J. Biophotonics* **2008**, *1*, 215.

(124) Shaver, J. M. In *Handbook of Raman Spectroscopy: From the Research Laboratory to the Process Line*; Lewis, I. R., Edwards, H. G. M., Eds.; Marcel Dekker: New York, 2001; p 275.

(125) Whitmore, L.; Wallace, B. A. *Biopolymers* **2007**, *89*, 392.

(126) Lees, J. G.; Miles, A. J.; Janes, R. W.; Wallace, B. A. *BMC Bioinf.* **2006**, *7*, 507.

(127) Sreerama, N.; Woody, R. W. *Methods Enzymol.* **2004**, *383*, 318.

(128) Provencher, S. W.; Glockner, J. *Biochemistry* **1981**, *20*, 33.

(129) Sreerama, N.; Woody, R. W. *Anal. Biochem.* **1993**, *209*, 32.

(130) Sreerama, N.; Venyaminov, S. Y.; Woody, R. W. *Protein Sci.* **1999**, *8*, 370.

(131) Sreerama, N.; Woody, R. W. *Anal. Biochem.* **2000**, *287*, 252.

(132) Venyaminov, S.; Yang, J. T. In *Circular Dichroism and Conformational Analysis of Biomolecules*; Fasman, G. D., Ed.; Plenum: New York, 1996.

(133) Sreerama, N.; Venyaminov, S.; Robert, W. *Woody Anal. Biochem.* **2000**, *287*, 243.

(134) Keiderling, T. A. *Curr. Opin. Chem. Biol.* **2002**, *6*, 682.

(135) Baumruk, V.; Pancoska, P.; Keiderling, T. A. *J. Mol. Biol.* **1996**, *259*, 774.

(136) Pancoska, P.; Bitto, E.; Janota, V.; Urbanova, M.; Gupta, V. P.; Keiderling, T. A. *Protein Sci.* **1995**, *4*, 1384.

(137) Baello, B. I.; Pancoska, P.; Keiderling, T. A. *Anal. Biochem.* **1997**, *250*, 212.

(138) Asher, S. A.; Mikhonin, A. V.; Bykov, S. *J. Am. Chem. Soc.* **2004**, *126*, 8433.

(139) Lednev, I. K.; Karnoup, A. S.; Sparrow, M. C.; Asher, S. A. *J. Am. Chem. Soc.* **1999**, *121*, 8074.

(140) Vedantham, G.; Sparks, H. G.; Sane, S. U.; Tzannis, S.; Przybycien, T. M. *Anal. Biochem.* **2000**, *285*, 33.

(141) Goormaghtigh, E.; Ruysschaert, J.-M.; Raussens, V. *Biophys. J.* **2006**, *90*, 2946.

(142) Huang, C.-Y.; Balakrishnan, G.; Spiro, T. G. *J. Raman Spectrosc.* **2006**, *37*, 277.

(143) Sibley, A. B.; Cosman, M.; Krishnan, V. V. *Biophys. J.* **2003**, *84*, 1223.

(144) Saxena, V. P.; Wetlaufer, D. B. *Proc. Natl. Acad. Sci. U.S.A.* **1971**, *68*, 969.

(145) Chang, C. T.; Wu, C.-S. C.; Yang, J. T. *Anal. Biochem.* **1978**, *91*, 13.

(146) Bolotina, I.; Chekhov, V.; Lugauskas, V.; Ptitsyn, O. *Mol. Biol. (Mosk.)* **1981**, *15*, 167–75.

(147) Perczel, A.; Park, K.; Fasman, G. D. *Proteins: Struct., Funct., Bioinf.* **2005**, *13*, 57.

(148) Balakrishnan, G.; Hu, Y.; Bender, G. M.; Getahun, Z.; DeGrado, W. F.; Spiro, T. G. *J. Am. Chem. Soc.* **2007**, *129*, 12801.

(149) Dehring, K. A.; Smukler, A. R.; Roessler, B. J.; Morris, M. D. *Appl. Spectrosc.* **2006**, *60*, 366.

(150) Huang, C.-Y.; Balakrishnan, G.; Spiro, T. G. *J. Raman Spectrosc.* **2006**, *37*, 277.

(151) Sikirzhytski, V.; Topilina, N. I.; Higashiya, S.; Welch, J. T.; Lednev, I. K. *J. Am. Chem. Soc.* **2008**, *130*, 5852.

(152) Balakrishnan, G.; Hu, Y.; Case, M. A.; Spiro, T. G. *J. Phys. Chem. B* **2006**, *110*, 19877.

(153) McColl, I.; Blanch, E.; Gill, A.; Rhie, A.; Ritchie, M.; Hecht, L.; Nielsen, K.; Barron, L. *J. Am. Chem. Soc.* **2003**, *125*, 10019.

(154) Kubelka, J.; Keiderling, T. A. *J. Am. Chem. Soc.* **2001**, *123*, 12048.

(155) Song, S.; Asher, S. A. *J. Am. Chem. Soc.* **1989**, *111*, 4295.

(156) Sarver, R. W., Jr.; Krueger, W. C. *Anal. Biochem.* **1993**, *212*, 519.

(157) Bruun Susanne, W.; Sondergaard, I.; Jacobsen, S. *J. Agric. Food Chem.* **2007**, *55*, 7234.

(158) Cai, S.; Singh, B. R. *ACS Symp. Ser.* **2000**, *750*, 117.

(159) Cai, S.; Singh, B. R. *Biochemistry* **2004**, *43*, 2541.

(160) Dousseau, F.; Pezolet, M. *Biochemistry* **1990**, *29*, 8771.

(161) Jiang, H.; Song, Z.; Liu, B.; Liu, C.; Sun, M. *Guangpuxue Yu Guangpu Fenxi* **1996**, *16*, 29.

(162) Ogawa, M.; Horimoto, Y.; Durance, T.; Nakai, S. *Book of Abstracts, 215th ACS National Meeting, Dallas, March 29−April 2* **1998**, ANYL-064.

(163) Oberg, K. A.; Ruysschaert, J.-M.; Goormaghtigh, E. *Eur. J. Biochem.* **2004**, *271*, 2937.

(164) Shashilov, V. A. Development of mathematical methods for quantitative resonance Raman spectroscopy. Ph.D. Thesis, University at Albany, State University of New York, 2007.

(165) Shashilov, V. A.; Sikirzhytski, V.; Popova, L. A.; Lednev, I. K. *Methods* **2010**, doi:10.1016/j.ymeth.2010.05.004.

(166) Vapnik, V. *Statistical Learning Theory*; John Wiley & Sons: New York, 1998.

(167) Thissen, U.; Üestuen, B.; Melssen, W. J.; Buydens, L. M. C. *Anal. Chem.* **2004**, *76*, 3099.

(168) Suykens, J. A. K.; Gestel, T. V.; Brabanter, J. D.; Moor, B. D.; Vandewalle, J. *Least Squares Support Vector Machines*; World Scientific: Singapore, 2002.

(169) Wu, D.; He, Y.; Feng, S.; Sun, D.-W. *J. Food Eng.* **2008**, *84*, 124.

(170) Andrade, M. A.; Chacón, P.; Merelo, J. J.; Morán, F. *Protein Eng.* **1993**, *6*, 383.

(171) Severcan, M. *J. Mol. Struct.* **2001**, *565*−*566*, 383.

(172) Böhm, G.; Muhr, R.; Jaenicke, R. *Protein Eng.* **1992**, *5*, 191.

(173) Hering, J. A.; Innocent, P. R.; Haris, P. I. *Spectrosc. Int. J.* **2002**, *16*, 53.

(174) Vedantham, G.; Sparks, H. G.; Sane, S. U.; Tzannis, S.; Przybycien, T. M. *Anal. Biochem.* **2000**, *285*, 33.

(175) Clementi, M.; Clementi, S.; Cruciani, G.; Pastor, M.; Davis, M. A.; Flower, R. D. *Protein Eng.* **1997**, *10*, 747.

(176) Venyaminov, S. Y.; Vassilenko, K. S. *Anal. Biochem.* **1994**, *222*, 176.

(177) Gryniewicz, C. M.; Reepmeyer, J. C.; Kauffman, J. F.; Buhse, L. F. *J. Pharm. Biomed. Anal.* **2009**, *49*, 601.

(178) Lipkus, A. H.; Lenk, T. J.; Chittur, K. K.; Gendreau, R. M. *Biopolymers* **2004**, *27*, 1831.

(179) Liu, K.-Z.; Tsang, K. S.; Li, C. K.; Shaw, R. A.; Mantsch, H. H. *Clin. Chem.* **2003**, *49*, 1125.

(180) Navea, S.; Tauler, R.; Goormaghtigh, E.; de Juan, A. *Proteins: Struct., Funct., Bioinf.* **2006**, *63*, 527.

(181) Zhu, F.; Isaacs, N.; Hecht, L.; Barron, L. *Structure* **2005**, *13*, 1409.

(182) Sikirzhytski, V.; Virkler, K.; Lednev, I. K. *Sensors* **2010**, *10*, 2869.

(183) Virkler, K.; Lednev, I. K. *Forensic Sci. Int.* **2008**, *181*, e1.

(184) Virkler, K.; Lednev, I. K. *Forensic Sci. Int.* **2009**, *193*, 56.

(185) Virkler, K.; Lednev, I. K. *Anal. Bioanal. Chem.* **2010**, *396*, 525.

(186) Virkler, K.; Lednev, I. K. *Analyst* **2010**, *135*, 512.

(187) Virkler, K.; Lednev, I. K. *Anal. Chem.* **2009**, *81*, 7773.

(188) Barron, L. D.; Zhu, F.; Hecht, L.; Tranter, G. E.; Isaacs, N. W. *J. Mol. Struct.* **2007**, *834*−*836*, 7.

(189) Krzanowski, W. *Principles of multivariate analysis, a user's perspective*; Oxford University Press: Oxford, 1998.

(190) Azuaje, F.; Wang, H.; Chesneau, A. *BMC Bioinformatics* **2005**, *6*, 1.

(191) Zhu, F.; Kapitan, J.; Tranter, G. E.; Pudney, P. D. A.; Isaacs, N. W.; Hecht, L.; Barron, L. D. *Proteins* **2008**, *70*, 823.

(192) Barron, L. D.; Zhu, F.; Hecht, L.; Tranter, G. E.; Isaacs, N. W. *J. Mol. Struct.* **2006**, *834*−*836*, 7.

(193) Zhu, F.; Tranter, G. E.; Isaacs, N. W.; Hecht, L.; Barron, L. D. *J. Mol. Biol.* **2006**, *363*, 19.

(194) Todorov, V.; Pires, A. M. *REVSTAT*−*Statist. J.* **2007**, *5*, 63.

(195) Lindgren, F.; Geladi, P.; Berglund, A.; Sjöstrom, M.; Wold, S. *J. Chemometrics* **1995**, *9*, 331.

(196) Barker, M.; Rayens, W. *J. Chemometrics* **2003**, *17*, 166.

(197) Berman, E. S. F.; Wu, L.; Fortson, S. L.; Kulp, K. S.; Nelson, D. O.; Wu, K. J. *Surf. Interface Anal.* **2009**, *41*, 97.

(198) Bryan, K.; Brennan, L.; Cunningham, P. *BMC Bioinformatics* **2008**, *9*, 470.

(199) Levner, I.; Bulitko, V.; Lin, G. In *Feature Extraction, Foundations and Applications*; Springer: Berlin/Heidelberg, 2006.

(200) Tibshirani, R.; Hastie, T.; Narasimhan, B.; Chu, G. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *99*, 6567.

(201) Wu, B.; Abbott, T.; Fishman, D.; McMurray, W.; Mor, G.; Stone, K.; Ward, D.; Williams, K.; Zhao, H. *Bioinformatics* **2003**, *19*, 1636.

(202) Tibshirani, R.; Hastiey, T.; Narasimhanz, B.; Soltys, S.; Shi, G.; Koong, A.; Le, Q. *Bioinformatics* **2004**, *20*, 3034.

(203) Levner, I. *BMC Bioinformatics* **2005**, *6*, 68.

(204) Guyon, I.; Elisseeff, A. *J. Mach. Learn. Res.* **2003**, *3*, 1157.

(205) Shin, H.; Sheu, B.; Joseph, M.; Markey, M. K. *J. Biomed. Inf.* **2008**, *41*, 124.

(206) Lilien, R. H.; Farid, H.; Donald, B. R. *J. Comput. Biol.* **2003**, *10*, 925.

(207) Petricoin, E. F.; Ardekani, A. M.; Hitt, B. A.; Levine, P. J.; Fusaro, V. A.; Steinberg, S. M.; Mills, G. B.; Simone, C.; Fishman, D. A.; Kohn, E. C.; Liotta, L. A. *Lancet* **2002**, *359*, 572.

(208) Adam, B.-L.; Qu, Y.; Davis, J. W.; Ward, M. D.; Clements, M. A.; Cazares, L. H.; Schellhammer, O. J. S. P. F.; Yasui, Y.; Feng, Z.; George, L.; Wright, J. *Cancer Res.* **2002**, *62*, 3609.

(209) Smith, B. C. *Fundamentals of Fourier Transform Infrared Spectroscopy*; CRC Press: Boca Raton, FL, 1996.

(210) Loudermilk, J. B.; Himmelsbach, D. S.; Barton, F. E.; de Haseth, J. A. *Appl. Spectrosc.* **2008**, *62*, 661.

(211) Smith, S. W. *Digital Signal Processing: A Practical Guide for Engineers and Scientists*; Newnes: Burlington, MA, 2002.

(212) Chan, F.; Liang, Y.; Gao, J.; Shao, X. *Chemometrics: from Basics to Wavelet Transform*; Hoboken: NJ, 2004.

(213) Lo, S.-C.; Brown, C. W. *Appl. Spectrosc.* **1992**, *46*, 790.

(214) Leung, A. K.-m.; Chau, F.-t.; Gao, J.-b.; Shih, T.-m. *Chemom. Intell. Lab. Syst.* **1998**, *43*, 69.

(215) Bjerga, J. M.; Small, G. W. *Anal. Chem.* **1990**, *62*, 226.

(216) Brown, C. W.; Okafor, A. E.; Donahue, S. M.; Lo, S.-C. *Appl. Spectrosc.* **1995**, *49*, 1022.

(217) Gianella, M.; Sigrist, M. W. *Appl. Spectrosc.* **2009**, *63*, 261.

(218) Boutilier, K.; Ross, M.; Podtelejnikov, A. V.; Orsi, C.; Taylor, R.; Taylor, P.; Figeys, D. *Anal. Chim. Acta* **2005**, *534*, 11.

(219) Craig, R.; Cortens, J. C.; Fenyo, D.; Beavis, R. C. J. *Proteome Res.* **2006**, *5*, 1843.

(220) Sadygov, R.; Cociorva, D.; Yates, J. R., III *Nat. Methods* **2004**, *1*, 195.

(221) Kapp, E.; Schütz, F. *Curr. Protoc. Protein Sci.* **2007**, *49*, 25.

(222) Eng, J. K.; Fischer, B.; Grossmann, J.; MacCoss, M. J. *J. Proteome Res.* **2008**, *7*, 4598.

(223) Gorban, A. N. *Principal Manifolds for Data Visualization and Dimension Reduction*; Springer-Verlag: New York, 2007.

(224) Lipo Wang; Fu, X. *Data Mining with Computational Intelligence*; Springer: Birkhäuser, 2005.

(225) Wise, B. M.; Gallagher, N. B.; Butler, S. W.; White, D.; Barna, G. G. *J. Chemom.* **1999**, *13*, 379.

(226) Schlkopf, B.; Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*; MIT Press: Boston, 2001.

(227) Rowe, D. B. *Multivariate Bayesian Statistics: Models for Source Separation and Signal Unmixing*; Chapman & Hall/CRC: Boca Raton, FL, 2003.

(228) Uğurlu, G.; Özaltın, N.; Erdal, D. *Rev. Anal. Chem.* **2008**, *27*, 218.

(229) Dinç, E.; Kaya, S.; Doganay, T.; Baleanu, D. *J. Pharm. Biomed. Anal.* **2007**, *44*, 991.

(230) Sadygov, R. G.; Good, D. M.; Swaney, D. L.; Coon, J. J. *J. Proteome Res.* **2009**, *8*, 3198.

(231) Razor Spectrometry Software. http://www.spectrumsquare.com, 2009.

(232) DeNoyer, L. K.; Dodd, J. G. In *Handbook of Vibrational Spectroscopy*; John Wiley & Sons: New York, 2002; Vol. 3.

CR900152H