Biophysical Journal Volume 74 June 1998 2876–2888

Multivariate Curve Resolution: A Possible Tool in the Detection of Intermediate Structures in Protein Folding

J. Mendieta, M. S. Díaz-Cruz, M. Esteban, and R. Tauler
Departament de Química Analítica, Universitat de Barcelona, 08028 Barcelona, Spain

ABSTRACT Different multivariate data analysis techniques based on factor analysis and multivariate curve resolution are shown for the study of biochemical evolutionary processes like conformational changes and protein folding. Several simulated CD spectral data sets describing different hypothetical protein folding pathways are analyzed and discussed in relation to the feasibility of factor analysis techniques to detect and resolve the number of components needed to explain the evolution of the CD spectra corresponding to the process (i.e., to detect the presence of intermediate forms). When more than two components (the native and unordered forms) are needed to explain the evolution of the spectra, an iterative multivariate curve resolution procedure based on an alternating least squares algorithm is proposed to estimate the CD spectrum corresponding to the intermediate form.

INTRODUCTION

The folding of a polypeptide chain in vitro into a native, biologically active conformation is apparently a self-assembly process (Anfinsen, 1973). Frequently, the folding occurs in a short period of time, which implies that this process is not a random search of all possible conformations but occurs along a defined pathway with structured intermediates (Creighton, 1985). The characterization of the folding intermediates is of fundamental importance to understand the mechanism of protein folding. However, the detection and characterization of these intermediate structures is not easy due to the highly cooperative process of protein folding, which makes the lifetime of these transient intermediates too short to be detected by the more commonly used experimental techniques. In some cases, transient structures (*molten globule states*) can be stabilized in solution using partially denaturing conditions (Ohgushi and Wada, 1983). These molten globule states keep the greater part of the secondary structure, but the ordered tertiary interactions are not present (Dolgikh et al., 1981). Unfolding of the molten globule state presents low cooperativity (Pfeil et al., 1986), which suggests that it can be smoothly transformed into the unfolded form by a gradual destruction of its secondary structure (Ptitsyn, 1987). Spectroscopic methods like CD and NMR are normally used to study protein refolding. The spectra obtained under partially denaturing conditions correspond normally to a mixture of the denatured polypeptide, the intermediate structures, and the native form, which makes the resolution and characterization of folding intermediates difficult.

Received for publication 1 December 1997 and in final form 16 February 1998.

Address reprint requests to J. Mendieta, Departament de Química Analítica, Universitat de Barcelona, Av. Diagonal 647, 08028 Barcelona, Spain. Tel.: 34-3-4021286; Fax: 34-3-4021233; E-mail: jesus@zeus.qui.ub.es.

0006-3495/98/06/2876/13 $2.00

With the explosive growth of chemometrics in recent years (Sharaf et al., 1986; Massart et al., 1988), a new general approach involving the identification of a model from numerical and statistical analysis of the data, without any a priori assumption about the nature or composition of the system under investigation, has been proposed to solve *the mixture analysis problem* (Liang et al., 1993; Brown and Bear, 1993). Mixture analysis implies the estimation of the number of chemical species simultaneously present in the mixture, the identification of these species, and the determination of their concentration. Among the computational and statistical methods used to solve mixture analysis problems, factor analysis (FA) (Malinowski, 1991), principal component analysis (PCA) (Wold et al., 1987a), and singular value decomposition (SVD) (Golub and Van Loan, 1989) techniques play a key role, especially in the estimation of the number of species contributing significantly to the experimental data variance. In chemistry, FA, PCA, and SVD are very similar techniques with slightly different formalisms for selection of dimensions and for changes of coordinate axes (rotations). Derived from factor analysis, evolving factor analysis (EFA) (Gampp et al., 1986; Maeder, 1987) and fixed-size moving window evolving factor analysis (FSMWEFA) (Keller and Massart, 1991) are two techniques that have been shown to be especially suitable for the study of evolutionary processes like those present in chromatography and in chemical reaction processes. By using EFA and/or FSMWEFA techniques, the evolution of the chemical contributions along a particular experiment can be mathematically estimated without any prior assumption about the nature of these contributions or any assumption about a chemical model. These pure mathematical solutions obtained by means of FA-derived methods can be transformed to physically meaningful solutions by means of multivariate curve resolution (MCR) methods (Tauler et al., 1995; Tauler, 1995). MCR has been already shown to be a powerful method for the study of the conformational changes in synthetic polynucleotides induced

by the pH and/or ion complexation using spectrophotometric techniques (Casassas et al., 1994, 1995). The MCR method has been also used in the study of metal-binding properties of peptides (Mendieta et al., 1996) and to solve mixture analysis problems in analytical chemistry (Tauler et al., 1993, 1994, 1996).

In the present work, several chemometric techniques successfully applied to the study of evolutionary processes in mixture analysis problems are extended to the study of protein folding and conformational protein changes in biochemical problems. First, the possibilities of PCA in the determination of the number of components are shown. Second, EFA and FSMWEFA methods are used to monitor the evolution of biochemical processes like protein folding. Third, MCR is used for the detection, identification, and quantitation of folding intermediates in a simulated data set representing a hypothetical folding pathway for polypeptides. Finally, the deconvolution of the pure secondary structure contributions from the CD-detected species spectra is shown by means of a least-squares data-fitting procedure. The folding pathways simulated in this work are deliberately simplified to facilitate the comprehension of the proposed approach and to validate the results obtained by using it. Real situations similar to those described here have been proposed in the gradual disorganization of the secondary structure from the molten globule state to the unordered form (Ptitsyn, 1987).

DATA MODEL

CD spectra corresponding to three different temperature-dependent folding pathways have been simulated using the reference spectra for α-helix, β-form, β-turn, and random-coil described by Chang et al. (1978). In all these cases the native form contains 40% α-helix, 40% β-form, 18% β-turn, and 2% random-coil (species spectrum 1). The unordered form contains 85% random-coil, but an important amount (15%) β-turn are preserved (species spectrum 2). In a first data set A, a transient structure containing 35% α-helix, 5% β-form, 16% β-turn, and 44% random-coil has been considered in the folding pathway (species spectrum 3). In a second data set B a intermediate structure containing 5% α-helix, 35% β-form, 16% β-turn, and 44% random-coil has been considered in the folding pathway (species spectrum 4). In data sets A and B the native form is generated from the previously formed transient. In the case of data set C, no transient form is involved in the folding pathway. Fig. 1 shows the relative proportion of the three forms present in the solution at different temperatures (these contributions will also be called concentration profiles of species 1, 2 and 3, or 4). Spectra corresponding to different temperatures were obtained by linear combination of the species spectra 1, 2, and 3 (Fig. 1), corresponding to each form with coefficients equal to their relative proportion at each temperature. A small random noise with a zero mean and a standard deviation equal to 0.005 signal units was added to the exact numerical values. Spectra for the three folding pathways, data sets A, B, and C, were ordered in three two-way data matrices, $\mathbf{D}_A$, $\mathbf{D}_B$, and $\mathbf{D}_C$. These matrices have nR rows, i.e., nR spectra at the different investigated chemical conditions (i.e., different temperatures, pH values, etc.) and nC columns, i.e., nC wavelengths, λ, measured spectrophotometrically, circular dichroism spectrometry, CD, in this case. Fig. 2 shows the plots corresponding to these three simulated data matrices obtained for the three different folding pathways. No significant differences can be observed by visual inspection of the data.

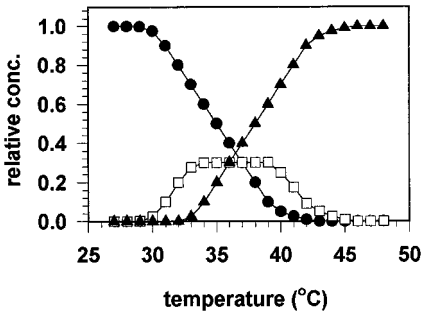

FIGURE 1 Relative proportions of the native (●), intermediate (□), and unordered (▲) forms at different temperatures used in the simulation of the data sets A and B.

The three data matrices $\mathbf{D}_A$, $\mathbf{D}_B$, and $\mathbf{D}_C$ can be analyzed individually, one by one, or simultaneously, two by two, $\mathbf{D}_A$ together with $\mathbf{D}_C$ and $\mathbf{D}_B$ together with $\mathbf{D}_c$ (Fig. 3). Simultaneous analysis of two or more correlated data matrices (multiway data analysis, Wold et al., 1987b; Smilde and Doornbos, 1991) is a very powerful approach to increase resolution of complex data systems. In this work, the simultaneous analysis of several data matrices is illustrated with the simultaneous study of a protein folding data matrix containing a transient structure (data sets A and B) together with a protein folding data matrix where the intermediate is absent (data set C). As previously described, in any of the two data sets, A and C or B and C, the native (species spectra 1) and unordered forms (species spectra 2) are common in the two matrices studied simultaneously, whereas the intermediate forms are not present in matrix C and they are different in both matrices A and B (species 3 in data set A and species 4 in data matrix B). For the simultaneous analysis of data matrices $\mathbf{D}_A$ and $\mathbf{D}_C$ and $\mathbf{D}_B$ and $\mathbf{D}_C$, two augmented columnwise matrices $[\mathbf{D}_A, \mathbf{D}_C]$ and $[\mathbf{D}_B, \mathbf{D}_C]$ are built (see Fig. 3). A more detailed description of the different possible data arrangements for simultaneous analysis of several correlated data matrices in the context of multivariate curve resolution has been given elsewhere (Tauler, 1995; Tauler et al., 1995).

FACTOR ANALYSIS AND PRINCIPAL COMPONENT ANALYSIS

Application of FA assumes that the experimental data are bilinear, i.e., that experimental data follow a linear additive model like that proposed by Beer's law for absorption

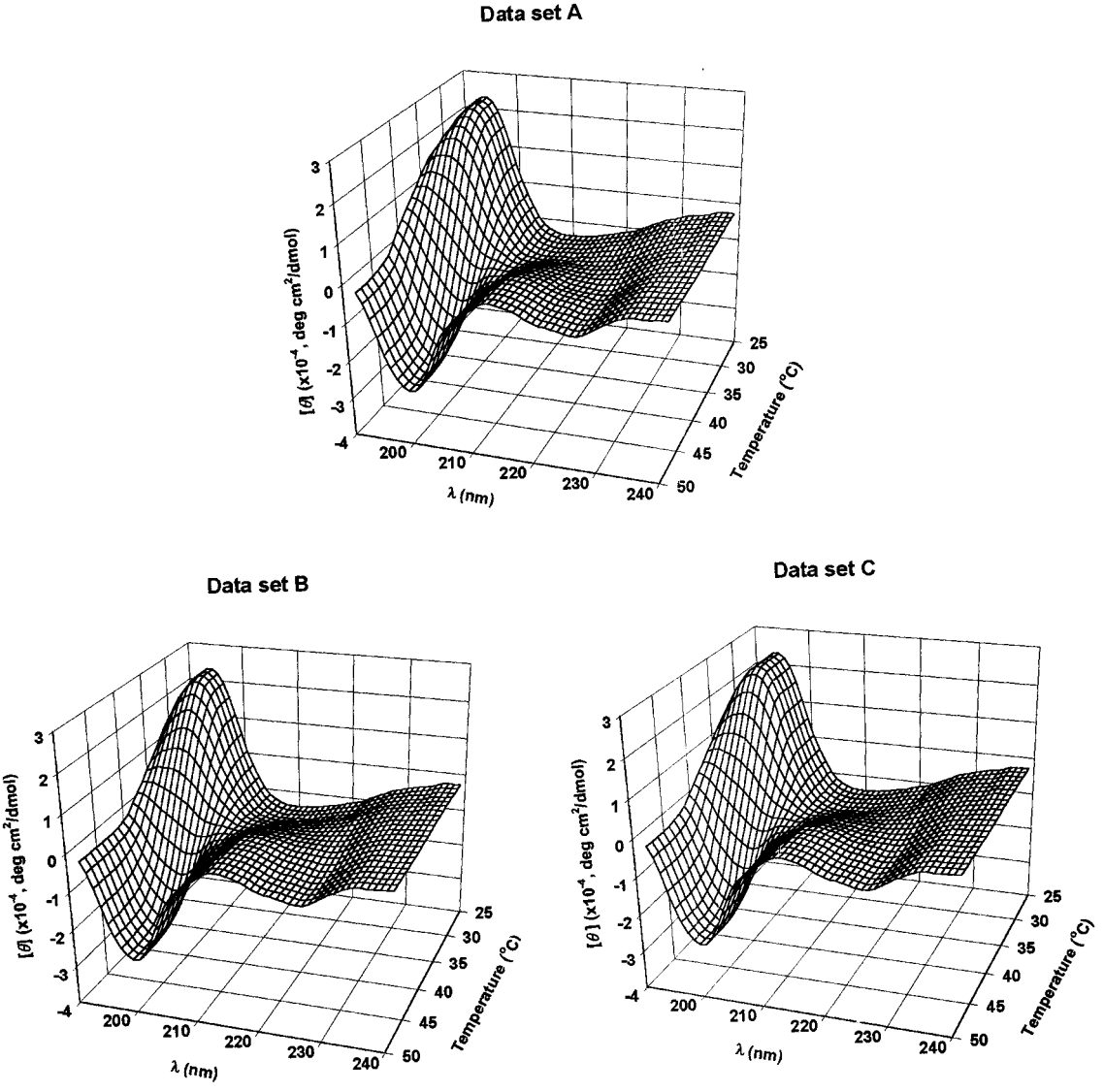

FIGURE 2 Evolution of the spectra corresponding to the three different simulated folding pathways versus the temperature.

spectroscopy. These conditions are usually satisfied for most of the spectrophotometric techniques normally used in the study of protein folding and conformational protein changes. A first basic goal of FA methods (Malinowski, 1991) is to mathematically decompose each experimental data matrix **D** ($\mathbf{D}_A$, $\mathbf{D}_B$, or $\mathbf{D}_C$) into a product of two abstract matrices, denoted as the scores matrix **Q**(*nR, nS*) and the loadings matrix $\mathbf{P}^T$(*nS, nC*), for a preselected number of components *nS*, contributing to the measured data. This can be expressed as

$$\mathbf{D} = \mathbf{Q}\mathbf{P}^T + \mathbf{E} \quad (1)$$

where **E** is a residual matrix containing the variance not explained by **Q** and $\mathbf{P}^T$.

Detection of the number of components by PCA

The number of chemical contributions, or pseudorank (mathematical rank in absence of noise), of the matrix **D**, *nS,* is chosen to minimize the residual data variance in **E**, leaving in it, if possible, only the experimental error or noise (Malinowski, 1991). There are many methods proposed for the selection of the number of components; most of them work well for uniformly distributed homocedastic noise in matrix **E**. However, when the noise is not uniformly distributed, most of the proposed approaches fail. Moreover, when instrumental and baseline contributions are present, the number of estimated components *nS* is higher than the number of real chemical contributions. On the contrary, in reaction-based systems, the opposite effect can be also observed, and the estimated pseudorank is lower than the real number of chemical contributions present in the system. This phenomenon has been called *rank deficiency* (Amrhein et al., 1996) and it is related to the fact that in reaction-based systems the pseudorank depends on the number of independent reactions present in the system. Resolution of rank-deficient systems is a matter of special interest at present (Amrhein et al., 1996; Izquierdo-Ridorsa et al., 1997).

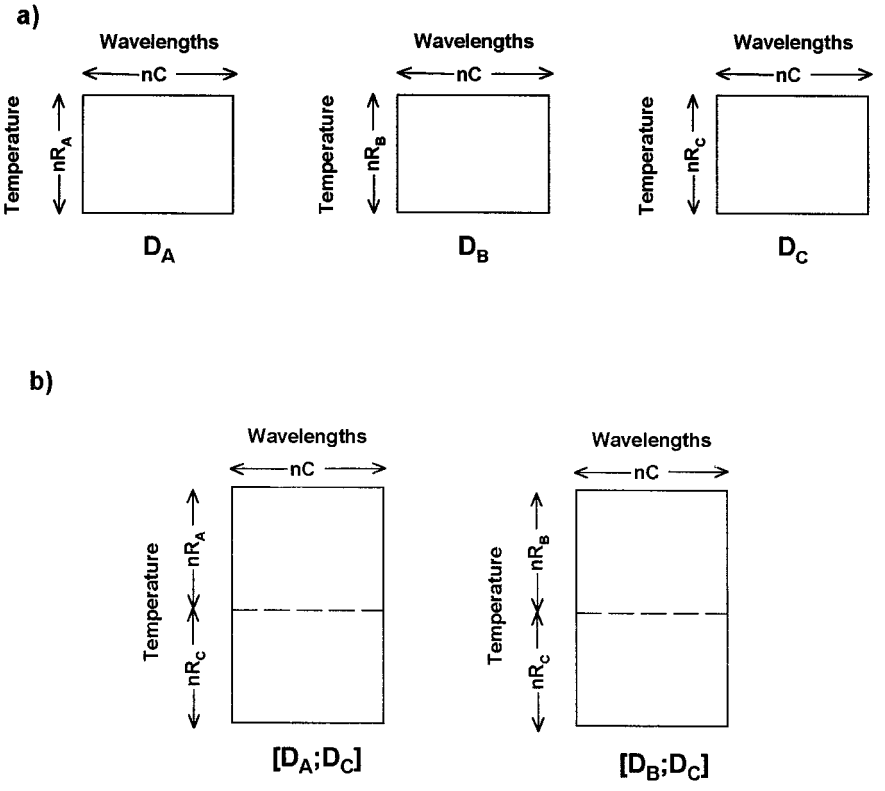

FIGURE 3 Data arrangement in single matrices $\mathbf{D}_A$, $\mathbf{D}_B$, and $\mathbf{D}_C$ (*a*) and augmented columnwise matrices $[\mathbf{D}_A, \mathbf{D}_C]$ and $[\mathbf{D}_B, \mathbf{D}_C]$ (*b*).

In this work the number of components is initially investigated by visual inspection of the magnitude of the singular values of matrix **D** (Golub and Van Loan, 1989) and also from the magnitude of residuals as a percentage of lack-of-fit or unexplained data variance in **E**, after a particular number of principal components has been extracted (Malinowski, 1991). Obviously the two approaches are closely correlated, presenting the same information in a different way. The basic assumption used by the two approaches is that the major components (major singular values) are associated with the chemical sources of data variation, which in the context of protein folding studied are interpreted as the different forms adopted by the polypeptidic chain (with different proportions of secondary structure) during the experiment.

Fig. 4 shows the normalized (divided by the largest one) singular values versus the number of considered components for the data sets previously described. In data sets A and B, three major components are clearly detected from these plots, since three singular values are much larger than the rest, which remain at the bottom of the plots at the noise level. On the contrary, in data set C, only two major components are clearly distinguished from the noise level.

These conclusions are also apparent from the results obtained in the PCA decomposition when different numbers of components are considered (Table 1). For data matrices $\mathbf{D}_A$ and $\mathbf{D}_B$, three components should be considered to leave the unexplained variance below 1%, whereas for data matrix $\mathbf{D}_C$ this is already achieved when only two components are considered. When more components are added, the unexplained data variance (lack-of-fit) decreases very slowly, since these additional components describe only the noise contribution. Thus, from a purely mathematical analysis of data, it is possible to estimate the number of components contributing to the data variance.

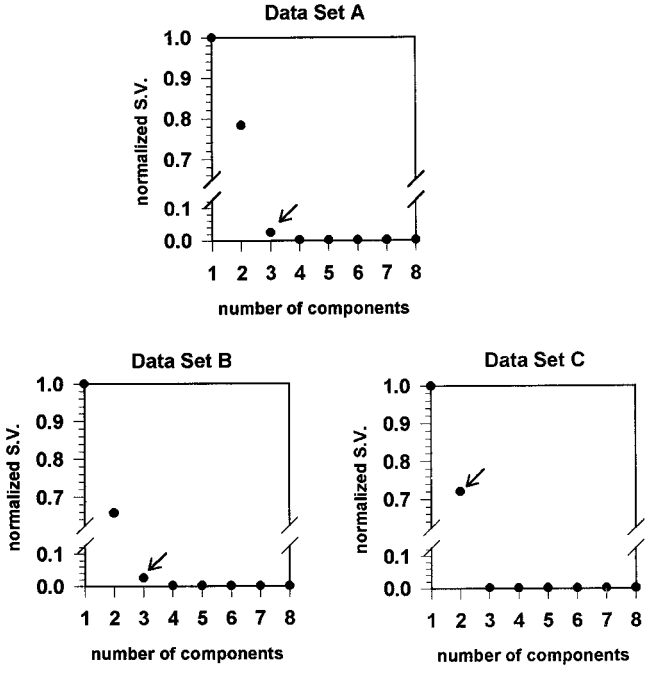

FIGURE 4 Normalized singular values corresponding to the data sets A, B, and C versus the number of considered components.

TABLE 1 Principal component analysis: percentage of lack of fit

	Data set*				
$n^{\#}$	A	B	C	[A, C]	[B, C]
1	61.7	55.1	58.5	60.1	56.9
2	2.02	2.21	0.53	1.89	1.89
3	0.53	0.55	0.51	0.54	0.55
4	0.49	0.50	0.47	0.52	0.53
5	0.46	0.47	0.47	0.50	0.51

Lack of fit = $100 \cdot \sqrt{[\Sigma(d_{ij} - d^*_{ij})^2]/\Sigma d^2_{ij}}$ where d_{ij} are the CD data at temperature i and wavelength j, and d^*_{ij} are the PCA recalculated data using the specified number of components.

*Data sets under study (see Data model section).

#Number of components considered in the PCA of the different data sets under study.

From these results, and in agreement with the proposed model, it is rightly concluded that all the spectra corresponding to the data set C are explained by a linear combination of two spectra, the one corresponding to the native form and the one corresponding to the unordered form. Instead, in the case of the data sets A and B, a third component is needed to explain the evolution of the spectra. From the only visual inspection of the data or from the single wavelength analysis of data sets A, B, and C, no significant differences could be detected between them (see Fig. 2), and therefore the presence of intermediates could not be detected.

In Table 1, also, the values of the percentage of lack-of-fit for the augmented columnwise data matrices, $[\mathbf{D}_A, \mathbf{D}_C]$ and $[\mathbf{D}_B, \mathbf{D}_C]$, are given. As it was mentioned in the previous section (Data model), the analysis of these augmented matrices implies the simultaneous analysis of two experiments. For columnwise augmented data matrices, $[\mathbf{D}_A, \mathbf{D}_C]$ and $[\mathbf{D}_B, \mathbf{D}_C]$, the total number of components needed to describe the data at the noise level is three, which is the same as the number of components needed to satisfactorily explain the individual data matrices $\mathbf{D}_A$ and $\mathbf{D}_B$. This means that columnwise matrix augmentation does not increase the rank, and therefore that the species in matrix $\mathbf{D}_C$ should have the same spectra as the corresponding species in matrices $\mathbf{D}_A$ and $\mathbf{D}_B$, which in this case is known to be true from the data simulation. For unknown systems, this study of the rank of augmented matrices is extremely helpful to check correspondence between species in different experiments.

Evolving factor analysis

Once the number of components is initially estimated by PCA or SVD, the changes and structure of the experimental data matrix can be analyzed by using EFA (Gampp et al., 1986; Maeder, 1987; Keller and Massart, 1991). This approach provides an estimation of the regions or windows where the concentration of the different components is changing or evolving and it also provides an initial estimation of how these concentration profiles change along the experiment. The EFA method is based on the evaluation of the magnitude of the singular values (or of the eigenvalues) associated with all the submatrices of a matrix **D** built up by adding successively one by one all the rows of the original data matrix. The calculations are performed in two directions: forward (in the same direction of the experiment), starting with the two first spectra, and backward (in the opposite direction of the experiment), starting with the last two spectra. In the forward direction, the detection of a new component is detected by the upsurging of a new singular value; in the backward direction, the disappearance of a component is detected by the upsurging of a new singular value. Singular values related with significant components become larger and clearly distinguished from the singular values associated with noise, in their graphical representation (EFA plots, Fig. 5). Singular values related with the noise are smaller and they are at the bottom of the EFA plots. Interpreting the EFA plots and appropriately joining the lines corresponding to forward and backward singular values (Gampp et al., 1986; Maeder, 1987) allow the estimation of the regions or windows of existence of each component and provide a first estimation of the abstract concentration profiles of the detected components. A more detailed description of the EFA plots can be found in previous works (Gampp et al., 1986; Maeder, 1987; Tauler and Casassas, 1988).

In Fig. 5 the EFA plots obtained in the analysis of matrices A–C are given. Two regions are clearly apparent in the three plots. One region at the bottom of these plots shows the evolution of lines describing the components related to the experimental noise. These lines do not supersede the limit of the noise level, approximately at $\log(0.005) = -2.3$, where 0.005 is the standard deviation of the noise. The second region shows the evolution of the lines describing the changes related in this case with the protein folding processes. When the analysis is performed in the forward direction (from lower to higher temperatures), three lines emerge for data sets A and B (1, 2, and 3), and two lines for data set C (1 and 2). In the three data sets, the first line (first singular value) is in the upper part of the plot ($\log SV > 2$) from the beginning of the process, increases a little more, and then it keeps constant. This line is showing the average absorption of the system when EFA is performed in the forward direction. Line 2 upsurges from the noise level when more spectra are included in the analysis, showing that at the corresponding temperatures, a new contribution not explained by the average absorption (line 1) becomes important and continues increasing steadily in the three data systems. For data sets A and B, a third line upsurges (line 3) from the noise level in the middle of the plot showing the evolution of a new contribution. This contribution does not appear in the analysis of data set C, in agreement with the fact that in this case no third species is present. As stated before, the analysis is also performed backward, giving lines 1′, 2′, and 3′, with identical meaning to before, but now looking at the experiment from higher to lower tem-

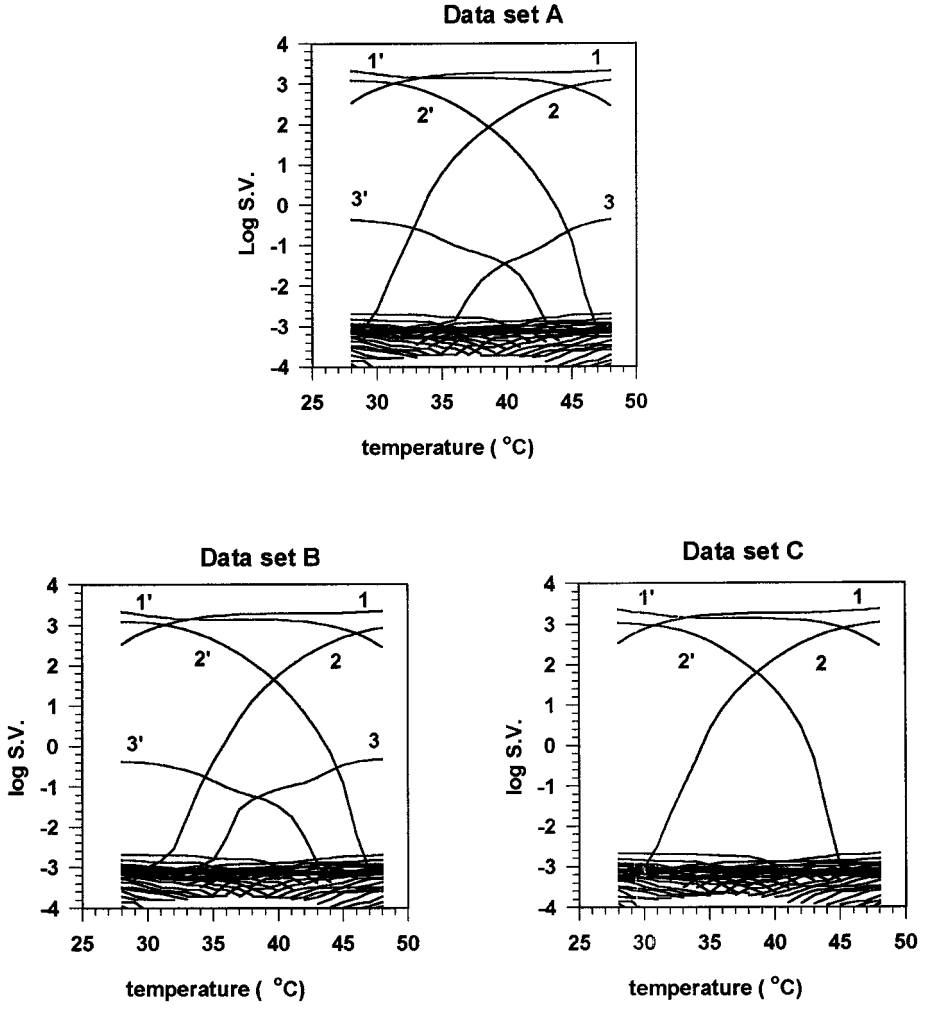

FIGURE 5 Evolving factor analysis plot corresponding to the simulated data sets A, B, and C.

peratures, i.e., increasing lines 1′, 2′, 3′ will be interpreted now as the disappearances of contributions. Now these two sets of lines can be interpreted simultaneously, and as proposed by Gampp et al. (1986), an initial estimation of the regions of existence of each species and also an initial estimation of the evolution of the concentration profiles of the species involved in the process can be derived. This interpretation will provide a first *abstract* plot of the evolution of the system when temperature is changed. This means that from a pure mathematical analysis of the data, not only the number of contributions can be deduced, but also, and even more importantly, how these contributions change along the experiment. The windows of existence of each species predicted by EFA for each experiment (Fig. 5) are coincident with those proposed in the simulation (Fig. 1). For real data with nonrandom noise contributions (like spectral baselines) the things become more complicated because these contributions can also appear as upsurging lines in the EFA plots (Tauler and Casassas, 1988). However, many times these lines stay at an intermediate region, between noise and chemical contributions, and can be safely distinguished. Nonlinear detector responses produce the appearance of extra lines not related to real chemical contributions. Additionally, weak chemical contributions and strongly correlated contributions are sometimes difficult to distinguish as new lines in the EFA plots. All this means that some practice and expertise is needed to extract definitive conclusions from the interpretation of EFA plots in the analysis of complex real systems with a large number of species. These problems, however, do not decrease the utility of EFA as an exploratory tool of those systems and also as a mathematical tool to provide an initial estimation of the evolution of the main chemical contributions.

Evolving factor analysis with a fixed-size moving window

A closely related and complementary method to EFA is the FSMWEFA method (Keller and Massart, 1991). In this case, the singular values are calculated for submatrices of equal size moving in the same direction as the experiment is performed. The size of the matrix is chosen to be slightly higher than the suspected number of components simultaneously present (overlapping) along the experiment and

kept constant. If this number is unknown, several sizes are attempted. The lower the size of the moving window, the better the local rank detection power; the larger the size of the window, the better resolution power between similar components. As with EFA, the appearance of a new component is distinguished with the upsurging of a new singular value. The interpretation of the FSMWEFA plots (Keller and Massart, 1991) allows the estimation of how many species coexist at the different stages of the experiment.

In Fig. 6 the FSWEFA plot obtained in the analysis of three data sets A, B, and C is given. The window size is five, i.e., five singular values are calculated from the submatrix corresponding to each fixed size moving window. In data set A, at the beginning, three of the five lines are at the bottom of the plot, at the noise level; the fourth line is increasing and the fifth line is at the top of the picture. This is interpreted as if two contributions are present at the beginning of the experiment (first five spectra), with the second one increasing in importance. Between the sixth and 16th window (see Fig. 6), a third singular value emerges two times from the noise level, showing the appearance and disappearance of a third contribution related to the intermediate form. Between these two peaks, the third singular value remains again at the noise level, showing a region were only two chemical species are changing their contribution in agreement with the constant concentration region of the intermediate form. Both EFA and WSFWEFA plots are only sensitive to changes and they are not sensitive to constant contributions. At the end of the experiment, again only two contributions are present. Similar patterns are present in the FSMWEFA plot of data set B with a more selective region at the beginning of the experiment where only one of the three contributions is clearly present for the window of the first five spectra. Finally, in the case of data set C (Fig. 6), only two contributions are present along the whole experiment, with selectivity (only one contribution) at the beginning and the end of the experiment.

The study of the mathematical structure of the data matrix by the two evolving factor analysis-related techniques (EFA and FSMWEFA) yielded a dynamic picture of the chemical process. The possibility to obtain this information from pure mathematical means can be important not only to understand the dynamic nature of the protein folding process, but also to outline the chemical and mathematical constraints to be applied in the resolution of the system by MCR (see next section).

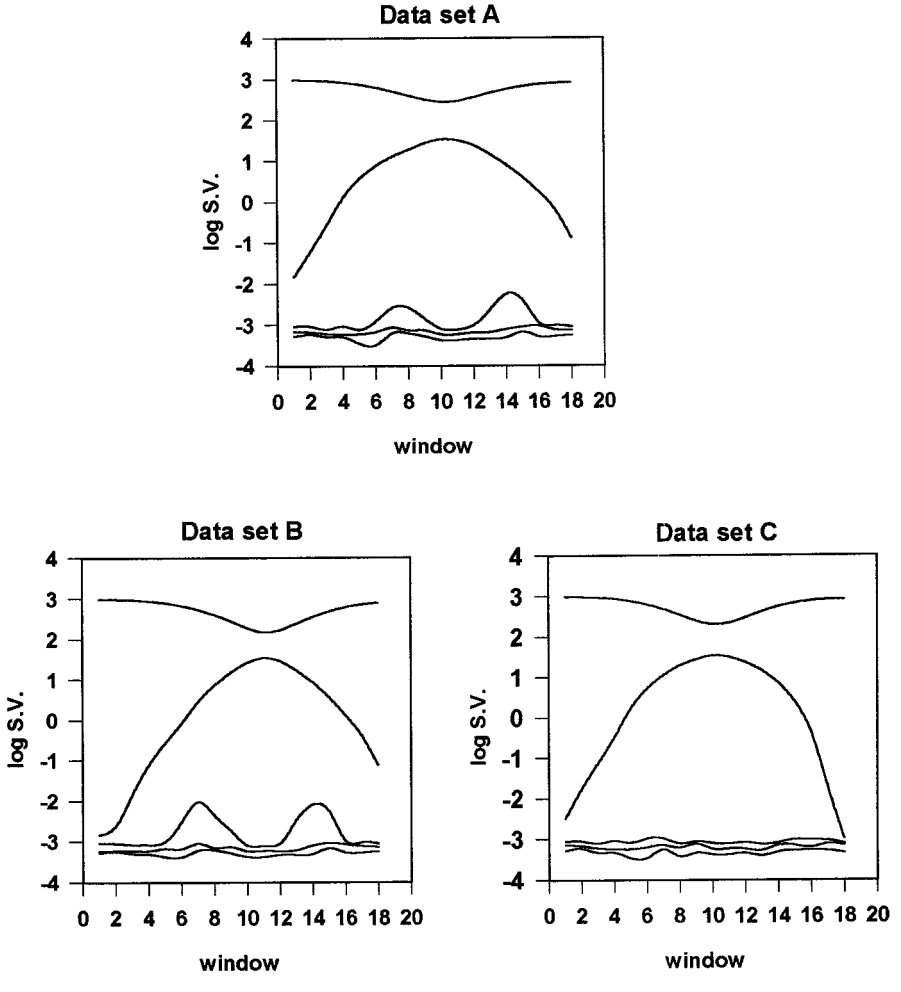

FIGURE 6 Fixed-size moving window evolving factor analysis plot with a window size of 5 corresponding to the simulated data sets A, B, and C.

MULTIVARIATE CURVE RESOLUTION

Multivariate curve resolution of a single data matrix

MCR (Lawton and Sylvestre, 1971; Martens, 1979) is a chemometric method included in the FA family of techniques (Malinowski, 1991). Its principal goals are the isolation, resolution, and quantitation of the sources of variation in a particular data set. The outstanding feature of this technique is that no a priori assumption about the contribution of the different factors in the global response are necessary. This feature can be of great importance in the study of complex problems such as protein folding. In previous works (Tauler et al., 1993–1996; Tauler and Casassas, 1992), MCR has been successfully applied to the study of other types of evolutionary chemical and analytical processes.

From the local rank analysis and initial estimations of evolving profiles derived from EFA and related methods, a constrained alternating least-squares (ALS) optimization is used to recover a physically meaningful set of concentration profiles and individual species spectra that best explain the observed data variance.

As in FA methods, this recovery is based on the assumption that the data matrix is bilinear, i.e., that it can be decomposed in the product of two matrices,

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} \tag{2}$$

In this equation $\mathbf{C}$ is the matrix whose columns describe how the chemical contributions (concentration profiles) change during the process; for the particular case of the analyzed data sets, the number of rows is equal to the number of temperatures included in the analysis or is equal to the number of spectra, and the number of columns is equal to the number of detected contributions (species). $\mathbf{S}^{\mathrm{T}}$ is the matrix whose rows are the pure individual spectra and with a number of columns equal to the number of wavelengths. Equations 1 and 2 show two possible ways of decomposition of the same data matrix. In fact, owing to the rotational and intensity factor analysis decomposition ambiguities, there are an infinite number of possible decompositions of the data matrix, reproducing it equally well. The conditions and constraints under which it is possible to recover the solutions of Eq. 2 for $\mathbf{C}$ and $\mathbf{S}^{\mathrm{T}}$ have been studied elsewhere (Tauler et al., 1995; Manne, 1995). When these conditions and the appropriate set of constraints are applied, the obtained solutions are very close or eventually equal to the true ones. Indeed, when the applied constraints are the mathematical expression of previously known chemical information, the recovered solutions are more easily interpreted from a chemical point of view. A short summary of the optimization procedure proposed to iteratively solve Eq. 2 is given.

When an initial estimation of the individual spectra is available, the best least-squares unconstrained solution of the concentration profiles is estimated from:

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

where $\mathbf{S}^{+}$ is the pseudoinverse (Golub and Van Loan, 1989) of $\mathbf{S}$ matrix.

If in contrast an initial estimation of the concentration profiles is available, the best unconstrained least-squares estimation of the spectroscopic contributions is estimated from:

$$\mathbf{S} = \mathbf{C}^{+}\mathbf{D} \tag{4}$$

where now $\mathbf{C}^{+}$ is the pseudoinverse of $\mathbf{C}$ matrix.

The least-squares solutions obtained in this way are pure mathematical solutions that probably will not be optimal from a chemical point of view. For instance, they can have negative concentrations. Therefore, an optimization procedure is started by iteratively resolving the two equations previously given and constraining, at each stage of the iterative optimization, the solutions to be non-negative (Lawson and Hanson, 1974; Bro and De Jong, 1997). Other constraints implemented during the ALS optimization can be the closure (sum of the concentration of all forms at different temperatures is equal to the total amount of protein), the unimodality (concentration profiles have unimodal peak or cumulative shapes), and the selectivity (at some temperatures only one form prevails). The selectivity constraint is very useful in protein folding because the native form is, by definition, the one present at physiological conditions, i.e., it is the form at the starting conditions of the experiment. Moreover, it can be supposed that at high denaturing conditions the only structure present in the solution is the unordered form. Local rank analysis by EFA is an idoneous method to test at which conditions the native forms or the unordered forms are the only forms present in the solution. As in the case of the non-negativity constraint, when the selectivity constraint is applied during the iterative optimization, the concentration value of only one component is allowed to be different from zero, at the temperature values where selectivity constraint is applied. This iterative procedure is carried out until the solutions and the data fitting do not improve. Details about the implementation of this method are described elsewhere and it has been applied to different types of chemical data (Casassas et al., 1994; Tauler and Casassas, 1992; Tauler et al., 1993–1996 and references therein).

The application of the ALS procedure to data sets A and B, using the non-negativity, closure, and selectivity constraints, allowed the estimation of the concentration profiles and spectra associated to each form of the protein. In Table 2 the lack-of-fit values obtained after the ALS optimization are given. They were obtained using a lack-of-fit converge criterion for the ALS optimization equal to 0.1% of difference between two consecutive iterations. A lower convergence criterion could be applied giving slightly lower lack-of-fit values, but considerably increasing the number of iterations. In all the cases, these lack-of-fit values are close

TABLE 2 Multivariate curve resolution results

Data sets*	(PCA)#	(ALS)§	S1¶	S2¶	S3¶	C1‖	C2‖	C3‖
A (3)	0.53	0.97	1.0000	0.9988	1.0000	0.9998	0.9997	0.9989
B (3)	0.55	0.70	1.0000	0.9597	0.9999	0.9997	0.9997	0.9991
C (2)	0.53	0.55	1.0000	—	1.0000	1.0000	—	1.0000
[A, C] (3)	0.54	0.62	1.0000	0.9999	1.0000	1.0000	0.9997	0.9999
						1.0000**	—	1.0000**
[B, C] (3)	0.55	0.64	1.0000	0.9997	1.0000	1.0000	0.9998	0.9999
						1.0000**	—	1.0000**

*Data sets under study (see Data section) by means of the ALS-MCR procedure. The number of components considered in the analysis is shown in parentheses.

#Lack-of-fit using PCA (see equation under Table 1).

§Lack-of-fit using ALS MCR procedure (see equation under Table 1, where d_{ij}^{*} is the reproduced data by means of the ALS MCR procedure).

¶Recovery of species spectra: S1, spectrum of the native form; S2, spectrum of the intermediate form; S3, spectrum of the unordered form. Recovery is measured by means of the correlation between the spectra used in the data simulation and those obtained using the MCR ALS procedure: $\sqrt{(s_i s_i^{*r})/(s_i s_i)(s_i^{*} s_i^{*r})}$ where s_i is the ith true spectrum (column vector) used in the data simulation and s_i^{*} is the corresponding spectrum calculated by means of the ALS MCR procedure. s_i is the same ith spectrum expressed as a transposed row vector to allow the dot product operation between vectors.

‖Recovery of concentration profiles: C1, concentration profile of the native form; C2, concentration profile of the intermediate form; C3, concentration profile of unordered form. Recovery is measured as for spectra by means of the same equation and substituting s_i (spectrum i) by c_i (concentration profile i).

Analysis of augmented matrices [A**, **C**] and [**B**, **C**] give two concentration profiles for the native and unordered forms presented in both matrices.

to those obtained by PCA and to the noise level. Also in Table 2 the similarities between the species profiles estimated by the ALS procedure and those used in the data simulation are given. These similarities are evaluated as the correlation between the recovered and the simulated concentration or spectra profiles. In the case of data matrix $\mathbf{D}_C$, there is a total agreement between the ALS recovered concentration profiles and those used in the data simulation. For matrices $\mathbf{D}_A$ and $\mathbf{D}_B$, the ALS recovered concentration profiles, although very similar to the theoretical (used in the simulation) ones, they are not exactly equal, i.e., some small rotational ambiguities persisted and they were not completely recovered by the ALS optimization. During the ALS optimization, the spectra at the two extreme temperatures of the study, 25° and 50°C, were considered to be pure, i.e., they were equal to the species spectra of the two extreme forms, the native and the unordered forms. Therefore, a selectivity constraint was applied at these two data points (see data treatment). All the other spectra at the different temperatures are considered to be mixture spectra, i.e., they are a linear combination of the spectra of the native, unordered, and intermediate forms. Accordingly to this, the pure (species) spectra of the two extreme forms, the native and the unordered, recovered by ALS are exactly equal to those used in the simulation (see Table 2). However, the species spectra of the intermediate forms in data matrices A and B, recovered by ALS, are slightly different from the theoretical ones. Since there is no temperature where the intermediate forms are the only species present and their concentration profiles are always totally embedded in the others, the rotational ambiguities cannot be totally solved and the spectra recovery is not perfect. Fig. 7 shows the recovered spectra corresponding to the intermediate forms recovered by the ALS procedure in the analysis of data sets A and B. In both cases the estimated spectrum approaches the spectrum of the intermediate form used in the data simulation, but it is not exactly equal to it. This is a consequence of the remaining unsolved rotational ambiguities present in the analysis of individual data matrices by the proposed MCR method. These remaining ambiguities can eventually be broken by using the simultaneous analysis of several data matrices by the proposed MCR method as it is shown in next section.

Multivariate curve resolution of a set of correlated data matrices

MCR can also be applied to the simultaneous analysis of several experiments, each one of them given an individual data matrix (Tauler and Casassas, 1992; Tauler et al., 1993–1996). In the data section it was already shown how different individual data matrices can be arranged to give an augmented columnwise data matrix. The two possible columnwise data matrices are written in a concise way (Fig. 3) as [$\mathbf{D}_A$, $\mathbf{D}_C$] and [$\mathbf{D}_B$, $\mathbf{D}_C$]. These two columnwise augmented matrices can be decomposed in the product of two matrices

$$[\mathbf{D}_A, \mathbf{D}_C] = [\mathbf{C}_A, \mathbf{C}_C]\mathbf{S}^T \quad (5)$$

and

$$[\mathbf{D}_B, \mathbf{D}_C] = [\mathbf{C}_B, \mathbf{C}_C]\mathbf{S}^T \quad (6)$$

where [$\mathbf{C}_A$, $\mathbf{C}_C$] or [$\mathbf{C}_B$, $\mathbf{C}_C$] are columnwise augmented concentration matrices and $\mathbf{S}^T$ is a nonaugmented species spectra matrix. In order to have a meaningful columnwise data augmentation, the spectra of the common contributions (species) in matrices $\mathbf{D}_A$ or $\mathbf{D}_B$ and $\mathbf{D}_C$ should be equal. This situation is quite common if the experimental conditions do not change between experiments (ionic strength, solvent, etc.). In the case of experiments changing the temperature, two effects can be observed. The first effect is a thermodynamic effect on the reaction equilibria, changing the relative concentrations of the different species (conformations,

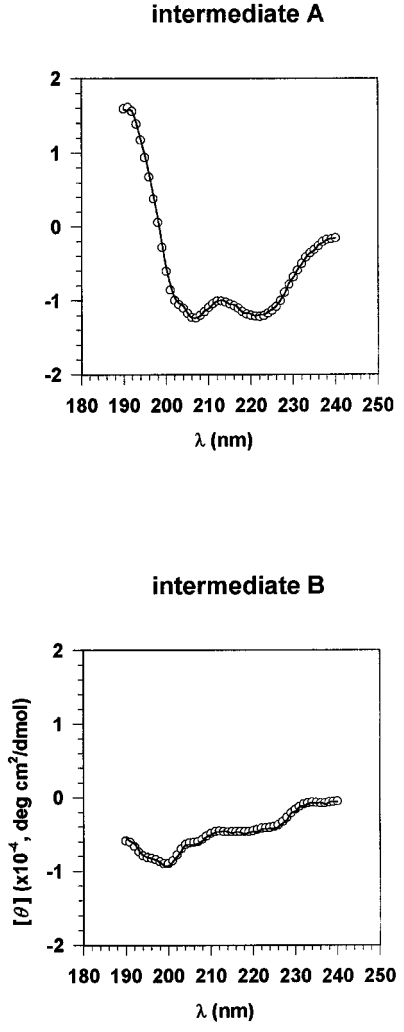

FIGURE 7 Comparison between the ALS recovered spectrum for intermediate form (*lines*) and the spectrum of the intermediate form used in the simulation (*symbols*) in data sets A and B.

forms) involved in the equilibria considered. A second possible effect of changes in temperature could be a physical effect on the shape and intensity of the spectra, usually shown as a peak broadening effect, especially important in the infrared region. For UV spectroscopy peak broadening effects caused by temperature changes are less important and can be considered negligible for narrow temperature changes. Therefore, the spectral changes in the UV observed in the study of proteins using circular dichroism when temperature is changed are mostly interpreted as changes caused by the equilibria between different conformations. On the contrary, when a columnwise data matrix is analyzed using the proposed procedure, the concentration profiles of the common species in matrices $\mathbf{C}_A$ or $\mathbf{C}_B$ and $\mathbf{C}_C$ (Eqs. 5 and 6) are allowed to be different in the different experiments (data matrices) simultaneously analyzed. This possibility is extremely important in the context of the protein folding studies, since different folding pathways are allowed to give different evolutions of the species concentrations.

The alternating least-squares multivariate curve resolution method previously described for individual data matrices can be easily extended to the analysis of the columnwise augmented data matrices. The ratio between number of unknowns and number of linear equations to be solved is drastically reduced when the augmented columnwise data matrix is analyzed with respect to when the individual data matrices are analyzed. Thus, the system of equations becomes more overdetermined and constrained. Also, in the simultaneous analysis of several experiments, additional constraints can be applied during the ALS procedure, apart from those used for the individual analysis. For instance, if a species is known not to be present in a particular data set, the appropriate column ranges in matrix $\mathbf{C}$ can be set to zero. On the contrary, species that are common in different experiments share their row spectra in matrix $\mathbf{S}^T$ and are on the same column of matrix $\mathbf{C}$. Further details about how these and other constraints are implemented for the simultaneous analysis of different data sets are given elsewhere. The simultaneous analysis of a set of correlated data matrices provides a powerful way to better solve the unavoidable FA ambiguities associated with the analysis of individual data matrices.

Simultaneous analysis of a set of correlated matrices falls under the discipline of multiway data analysis, a field of growing interest (Wold et al., 1987b). In particular, when a set of correlated data matrices or three-way data have a trilinear structure, the rotational ambiguities can be totally solved without ambiguities. In the case of the simultaneous analysis of k correlated matrices, the experimental data, d_{ijk}, have a trilinear structure if they can be decomposed by the following equation:

$$d_{ijk} = \sum t_{ks} c_{is} s_{js} \tag{7}$$

where t_{ks}, c_{is}, and s_{js} express the s profiles in the three orders of measurement: the row (descriptor i, i.e., number of temperatures), column (descriptor j, i.e., number of wavelengths), and tube (descriptor k, number of matrices) orders. Vectors $\mathbf{c}_s$ and $\mathbf{s}_s$ are, respectively, the pure species concentration and spectra profiles. Trilinear data, however, are rarely obtained in protein folding studies by means of CD, since it is difficult to have completely reproducible experiment conditions (i.e., temperature, pH, etc.) and the concentration profiles of the same species in the different k matrices will not be described by a single vector, $\mathbf{c}_s$, and they will evolve differently in the different experiments and they will have different shape. More commonly, the experimental data in the different simultaneously analyzed matrices do not have a trilinear structure, but they still have common profiles in the spectral order, since the CD spectra of the common species in the different data matrices are equal. Therefore, the analysis of columnwise augmented data matrices provides a powerful way of increasing the resolution of the system. For instance, if the resolution conditions (Manne, 1995) are achieved for one species in the individual analysis of one of the matrices simulta-

neously analyzed, then it is also possible to resolve the same species in the other matrices even if it was not possible when the matrices were analyzed individually.

In Table 2 the results achieved in the analysis of the columnwise augmented matrices $[\mathbf{D}_A, \mathbf{D}_C]$ and $[\mathbf{D}_B, \mathbf{D}_C]$ are given. Lack-of-fit values are also in these two cases very close to those obtained by means of PCA. Recovery of concentration profiles of the native unordered and intermediate forms in matrices $\mathbf{D}_A$ and $\mathbf{D}_B$ is improved significantly in relation to when these two matrices were analyzed individually. The similarity values of these two concentration profiles are practically equal to one, which means that the recovery was practically total. Only a very small ambiguity was not totally solved for the spectrum of the unordered form (second species) in matrix $\mathbf{D}_B$, although the remaining difference is so small that it has no practical importance. In Fig. 8 the concentration and spectra profiles achieved in the columnwise simultaneous analysis of data sets A and C are given. As a conclusion, although the data analyzed did not have a trilinear structure, the rotational ambiguities were finally practically completely solved when columnwise augmented matrices were analyzed by the proposed method. This situation can be easily extrapolated to most of the mixture analysis problems expected to be present in protein folding and conformational changes studied by means of spectrometric techniques.

The multivariate curve resolution method here described for both the analysis of a single data matrix or the analysis of a set of correlated data matrices, as well as for evolving factor analysis, has been implemented in a set of homemade MATLAB (Version 4.2, MathWorks Inc., Cochituate Place, MA, 1994) functions available upon request from one of the authors (R.T.).

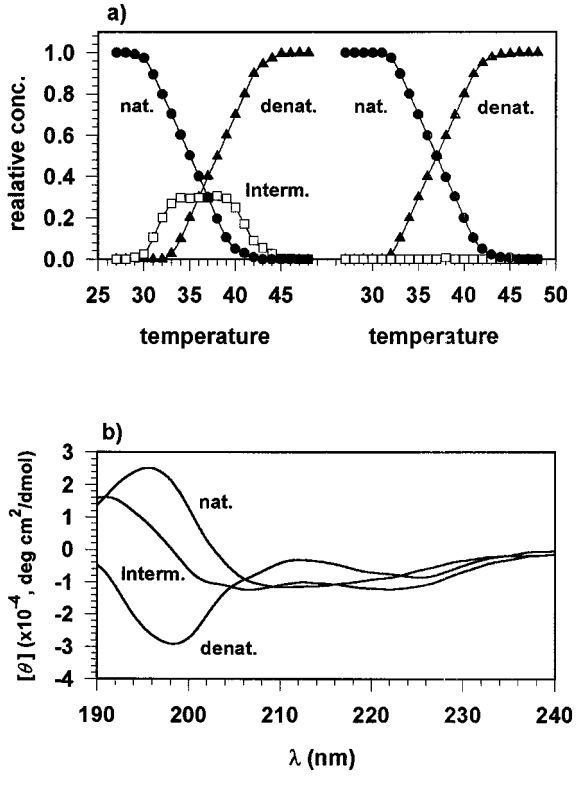

FIGURE 8 Concentration profiles (*a*) and individual spectra (*b*) recovered in the columnwise augmented simultaneous ALS analysis of data sets A and C.

Deconvolution of the CD spectra

In order to further analyze the results, an estimation of secondary structure of the intermediate forms is made by the deconvolution of the spectra recovered after their resolution. Since the data were simulated from basis spectra (see the model), the deconvolution was performed by simple least-squares data fitting using the Marquardt algorithm (Marquardt, 1963). In the case of real data, other more powerful methods must be applied to take into account the influence in the CD spectra of aromatics and other absorbing side chains, α-chain length, etc. (Compton and Johnson, 1986; Perczel et al., 1991).

Table 3 shows the comparison between the secondary structure present in the simulated intermediate and the secondary structure obtained after deconvolution of the ALS estimated intermediate spectrum for data sets A and B. The proportions of the different structural elements in the simulated intermediate form and in the estimated intermediate spectrum are very similar. These results confirm that the intermediate spectrum recovered by the proposed MCR-ALS procedure gave the same structural features than the one used in the data simulation.

The information provided by the study of the intermediate spectrum could not be obtained by the individual study of the spectra simulated at each temperature. For instance, in Table 3 the deconvolution of the simulated spectra at the temperatures 35°, 36°, 37°, and 38°C is shown. At these temperatures the concentration of the intermediate form should be maximal; therefore, the influence of the spectrum of the intermediate form in the simulated spectrum is also maximal. However, in both cases, data sets A and B, only a smooth and gradual destruction of the secondary structure could be observed with the increasing temperature, showing that the recovery of the information about the structural nature of the intermediate form is rather difficult by the study of individual spectra. Contrarily, the proposed multivariate approach allowed the recovery of the intermediate form spectrum and, therefore, allowed the study of its structural features.

CONCLUSIONS

In summary, this study showed that FA-derived methods like EFA and MCR methods can provide very powerful tools to analyze the number, nature, concentration, and evolution of the components needed to explain the spectra corresponding to a protein folding pathway. If more than two components (the native and unordered forms) are needed to explain the evolution of the spectra, the resolution

TABLE 3 Comparison between the structure present in the simulated intermediate and the secondary structure obtained after deconvolution of the estimated intermediate spectrum of both data sets, A and B. The deconvolution of spectra corresponding to the range of temperatures 35°–38°C are also shown

	Data Set A						Data Set B					
	Simulated Intermediate	Estimated Intermediate	38°	37°	36°	35°	Simulated Intermediate	Estimated Intermediate	38°	37°	36°	35°
α-Helix	35	35.08	18.45	22.40	26.50	30.52	5	5.63	9.45	13.38	17.50	21.54
β-Sheet	5	5.53	9.67	13.61	17.40	21.48	35	33.42	18.67	22.63	26.40	30.50
β-Turns	16	16.31	15.81	16.16	16.58	16.83	16	16.27	15.84	16.15	16.55	16.80
Random	44	43.08	56.06	47.83	39.52	31.18	44	44.68	56.03	47.82	39.52	31.21

Deconvolution was performed by simple least-squares data fitting using spectra described by Chang *et al.* (1978) as a basis.

of the spectrum corresponding to the intermediate form can be achieved using the MCR-ALS procedure proposed here. Posterior analysis of this spectrum yields an important amount of information about the secondary structure of the possible intermediates.

The total resolution of the intermediate forms by MCR is only possible if these intermediate forms exist in solution enough time to affect the response obtained by the techniques used in the structural analysis of proteins (CD, fluorescence spectroscopy, or NMR). However, transient structures can be stabilized in solution in partially denaturing conditions, allowing their detection and resolution by MCR even if the intermediates are present at low concentrations.

The authors gratefully acknowledge financial support from the Ministerio de Educación y Cultura (DGICYT, Projects PB96-0377 and PB96-0379). J. Mendieta also acknowledges financial support from the Programa de Acciones para la Incorporación de Doctores y Tecnólogos from the Ministerio de Educación y Cultura.

REFERENCES

Amrhein, M., B. Srinivasan, D. Bonvin, and M. M. Schumacher. 1996. On the rank deficiency and rank augmentation of the spectral measurement matrix. *Chemometrics Intell. Lab. Syst.* 33:17–33.

Anfinsen, C. B. 1973. Principles that govern the folding of protein chains. *Science.* 181:223–230.

Bro, R., and S. De Jong. 1997. A fast non-negativity constrained least squares algorithm. *J. Chemometrics*. 11:393–401.

Brown, S. D., and R. S. Bear, Jr. 1993. Chemometric techniques in electrochemistry: a critical review. *Crit. Rev. Anal. Chem.* 24:99–131.

Casassas, E., R. Gargallo, A. Izquierdo-Ridorsa, and R. Tauler. 1995. Application of a new multivariate curve resolution procedure to the study of the acid-base and copper(II) complexation equilibria of polycitidylic acid. *Reactive Polymers.* 27:1–14.

Casassas, E., R. Tauler, and M. Marques. 1994. Interactions of H^+ and Cu(II) ions with poly(adenilic) acid: study by factor analysis. *Macromolecules.* 27:1729–1737.

Chang, C. T., C. C. Wu, and J. T. Yang. 1978. Circular dichroic analysis of protein conformation: inclusion of β-turns. *Anal. Biochem.* 91:13–19.

Compton, L. A., and W. C. Johnson. 1986. Analysis of protein circular dichroism spectra for secondary structure using a simple matrix multiplication. *Anal. Biochem.* 86:155–167.

Creighton, T. E. 1985. The problem of how and why proteins adopt folded conformations *J. Phys. Chem.* 89:2452–2459.

Dolgikh, D. A., R. I. Gilmanshin, E. V. Brazhnikov, V. E. Bychkova, G. V. Semisotnov, S. Y. Venyaminov, and O. B. Ptitsyn. 1981. α-Lactalbumin compact state with fluctuating tertiary structure. *FEBS Lett.* 136: 311–315.

Gampp, H., M. Maeder, Ch. Meyer, and A. D. Zuberbuhler. 1986. Calculation of equilibrium constants from multiwavelength spectroscopic data model-free least-squares refinement by use of evolving factor analysis. *Talanta.* 33:943–951.

Golub, G. H., and Ch. F. Van Loan. 1989. Matrix computation. The Johns Hopkins University Press, Baltimore.

Izquierdo-Ridorsa, A., J. Saurina, S. Hernandez-Casou, and R. Tauler. 1997. Second order multivariate curve resolution applied to rank deficient data obtained from acid-base spectrophotometric titrations of mixtures of nucleic bases. *Chemometrics Intell. Lab. Syst.* 38:183–196.

Keller, H. R., and D. L. Massart. 1991. Peak purity control in liquid chromatography with photodiode array detection by fixed size moving window evolving factor analysis. *Anal. Chim. Acta.* 246:379–390.

Lawson, C. L., and R. J. Hanson. 1974. Solving Least Squares Problems. Prentice-Hall, Englewood Cliffs, NJ.

Lawton, W. H., and E. A. Sylvestre. 1971. Self-modeling curve resolution. *Technometrics.* 13:617–633.

Liang, Y. Z., O. M Kvalheim, and R. Manne. 1993. White, gray and black multicomponent systems. A classification of mixture problems and methods for their quantitative analysis. *Chemometrics. Intell. Lab. Syst.* 18:235–250.

Maeder, M. 1987. Evolving factor analysis for the resolution of overlapping chromatographic peaks. *Anal. Chem.* 59:527–530.

Malinowski, E. R. 1991. Factor Analysis in Chemistry. Wiley, New York.

Manne, R. 1995. On the resolution problem in hyphenated chromatography. *Chemometrics Intell. Lab. Syst.* 27:89–94.

Marquardt, D. W. 1963. An algorithm for least-squares estimation of non-linear parameters *J. Soc. Ind. Appl. Math.* 11:431–441.

Martens, H. 1979. Factor analysis of chemical mixtures. Non-negative factor solutions for spectra of cereal amino acids. *Anal. Chim. Acta.* 112:423–442.

Massart, D. L., B. G. M. Vandeginste, S. N. Deming, Y. Michotte, and L. Kaufman. 1988. Chemometrics: A Textbook. Elsevier, Amsterdam.

Mendieta, J., M. S. Diaz-Cruz, R. Tauler, and M. Esteban. 1996. Application of multivariate curve resolution to voltammetric data: study of metal-binding properties of the peptides. *Anal. Biochem.* 240:134–141.

Ohgushi, M., and A. Wada. 1983. Molten globule state. A compact form of globular proteins with mobile side chains. *FEBS Lett.* 164:21–24.

Perczel, A., M. Hollósi, G. Tusnády, and G. D. Fasman. 1991. Convex constraint analysis: a natural deconvolution of circular dichroism curves of proteins. *Protein Eng.* 4:669–679.

Pfeil, W., V. E. Bychkova, and O. B. Ptitsyn. 1986. Physical nature of the phase transition in globular proteins: calorimetric study of α-Lactalbumin. *FEBS Lett.* 198:287–291.

Ptitsyn, O. B. 1987. Protein folding: hypotheses and experiments. *J. Protein Chem.* 6:273–293.

Sharaf, M. A., D. L. Illman, and B. R. Kowalski. 1986. Chemometrics. John Wiley and Sons, New York.

Smilde, A. K., and D. A. Doornbos. 1991. Three-way methods for the calibration of chromatographic systems: comparing PARAFAC and three-way PLS, *J. Chemometrics.* 5:345–360.

Tauler, R. 1995. Multivariate curve resolution applied to second order data. *Chemometrics Intell. Lab. Syst.* 30:133–146.

Tauler, R., and E. Casassas. 1988. Principal component analysis applied to the study of successive complex formation data in Cu(II)-ethanolamine systems. *J. Chemometrics.* 3:151–161.

Tauler, R., and E. Casassas. 1992. Application of factor analysis to speciation in multiequilibria systems. *Analusis.* 20:255–268.

Tauler R., B. R. Kowalski, and S. Flemming. 1993. Multivariate curve resolution applied to process analysis. *Anal. Chem.* 65:2040–2047.

Tauler, R., S. Lacorte, and D. Barcelo. 1996. Application of multivariate curve self-modeling curve resolution for the quantitation of trace levels of organophosphorous pesticides in natural waters from interlaboratory studies. *J. Chromatogr.* 730:177–183.

Tauler, R., A. K. Smilde, J. M. Henshaw, L. W. Burguess, and B. R. Kowalski. 1994. Multicomponent determination of chlorinated hydrocarbons using a reaction-based chemical sensor. II. Chemical speciation using multivariate curve resolution. *Anal. Chem.* 66:3337–3344.

Tauler, R., A. K. Smilde, and B. R. Kowalski. 1995. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. *J. Chemometrics.* 9:31–58.

Wold, S., K. Esbensen, and P. Geladi. 1987a. Principal component analysis. *Chemometrics Intell. Lab. Syst.* 2:37–52.

Wold, S., P. Geladi, K. Esbensen, and J. Ohman. 1987b. Multi-way principal components and PLS-analysis. *J. Chemometrics.* 1:45–56.