ELSEVIER

Analytica Chimica Acta 446 (2001) 187–197

ANALYTICA CHIMICA ACTA

www.elsevier.com/locate/aca

Three-way data analysis applied to multispectroscopic monitoring of protein folding

Susana Navea, Anna de Juan*, Romà Tauler

Department of Analytical Chemistry, Universitat de Barcelona, Chemometrics Group, Diagonal 647, 08028 Barcelona, Spain

Received 15 November 2000; received in revised form 17 January 2001; accepted 1 February 2001

Abstract

Multivariate curve resolution-alternating least squares (MCR-ALS) is proposed as a three-way analysis method to deal with multispectroscopic monitoring of protein folding. MCR-ALS provides the concentration profiles associated with the different protein conformations occurring during the process and their related spectra. The concentration profiles describe the folding mechanism and the spectra provide the structural information of the conformations involved. Analysis either of the protein folding process monitored with different techniques (i.e. a row-wise augmented data matrix) or of several experiments done in different conditions using the same technique (i.e. a column-wise augmented matrix) or both possibilities at the same time (i.e. a row- and column-wise augmented matrix), can be performed. Thermal unfolding and refolding of α-lactalbumin, monitored using far- and near-UV circular dichroism, fluorescence and UV spectrometry, is shown as example. Information related to changes in the tertiary and the secondary structure of the protein, to the presence of intermediates along the protein folding process and to the reversibility of the thermal process can be obtained.

Keywords: Protein folding; Multivariate curve resolution; Three-way data analysis; Circular dichroism; α-Lactalbumin

1. Introduction

The relation between the structure of a protein and its biological activity is a major reason to understand folding and unfolding protein processes [1,2] but, far from being a study of exclusive academic or biochemical interest, the recent advents in protein engineering have made it also a practical work for producing safe and economic biotechnology products [3,4]. Synthesising any protein in an insoluble, unfolded, inactive and useless form is no longer a problem, refolding these denatured compounds into active organised entities is actually the interesting point. Therefore, finding out how this protein folding takes place in natural products is undoubtedly useful to design artificially homologous processes.

All proteins show different organisation levels [5], namely

1. Primary structure: the amino acid sequence in the polypeptide chain.
2. Secondary structure: regular spatial arrangements of the backbone of the polypeptide chain stabilised by hydrogen bonds between peptide amide and carbonyl groups. The most usual arrangements are either helical or flat structural motifs, such as α-helices or β-sheets. Less frequently found are other elements like β-turns. Disordered zones of the polypeptide chain are said to have a random coil structure.

* Corresponding author. Tel.: +34-93-4034445; fax: +34-93-4021233.
E-mail address: annaj@apolo.qui.ub.es (A. de Juan).

0003-2670/01/$ – see front matter
PII: S0003-2670(01)00912-6

3. Tertiary structure: the spatial arrangement of secondary structure motifs (α-helix, β-sheet, etc., ...) due to the presence of disulfide bridges (covalent interactions) and other weak interactions among the side chains of further amino acids. These interactions give rise to the so-called domains, which can show either a fibrillar or a globular shape. The tertiary structure of a protein determines its biological activity and it is known as the 'native structure' of the protein.

In a protein unfolding process, the complete loss of the secondary and tertiary structure is undergone. Because of the breaking of disulfide bridges and the weakening of the other between amino acid interactions, the protein acquires its unfolded/denatured state. Protein unfolding can be promoted by a variety of denaturant agents, like urea or guanidine hydrochloride, or simply by modifying in a suitable way experimental variables, such as temperature, pH or pressure. Protein unfolding processes can take place through one or more steps, depending on the nature of the protein. In a one-step process, both secondary and tertiary structures are lost simultaneously. In more complex examples, the process goes through one or more intermediates, for which the tertiary structure is absent or very loose and there is a partial or a total native-like secondary structure (molten globule state) [1,2].

The mechanism and the identity of the protein conformations involved in a protein folding process can be studied by monitoring spectrometrically changes on the tertiary and secondary structure [6–9]. UV, fluorescence and near-UV circular dichroism (CD) are techniques proposed to monitor changes on tertiary structure. All these techniques work in a wavelength range higher than 250 nm. UV and near-UV CD responses record absorption or ellipticity variations caused by changes in orientation of the aromatic rings present in the side chain of certain amino acids, like tryptophan, tyrosine and phenylalanine. Fluorescence is mainly sensitive to changes in the environment around the tryptophan aromatic system, which is more solvent-exposed in unfolded states. Far-UV CD serves to monitor modifications of the secondary structure, indeed, the working wavelength range (195–250 nm) is sensitive to rotations and changes of peptide bonds, which have specific angle values for each of the secondary structural motifs.

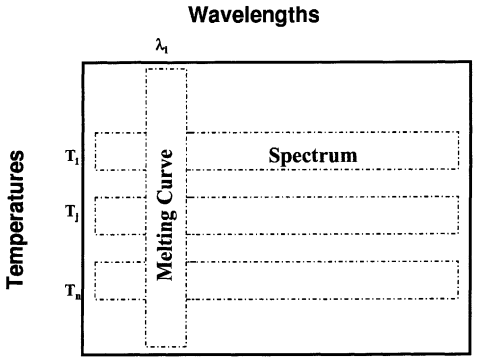

Fig. 1. Information used in the study of a thermal-dependent folding protein process according to classical procedures (dashed lines) and to the multivariate approach (solid lines).

Protein folding processes have been classically studied through the examination of melting curves, i.e. evolution of one-wavelength response values (e.g. absorbances, ellipticities, ...) as a function of the variation of denaturant agent [10,11]. Large response variations are supposed to be linked to protein conformation changes. Rough qualitative information is afterwards obtained by looking at some complete spectra recorded in the experimental conditions defined by some strategic points in the melting curve (see Fig. 1). This procedure is strongly wavelength-dependent and there is no way to ensure that a single spectrum is exclusively representative of a certain protein conformation. The inherent uncertainty of this approach may be solved if the complete spectrum at each stage of the process is recorded and all the related information is appropriately used. Multivariate data analysis provides the perfect tools for this purpose [12]. Specifically, curve resolution methods are focused on handling multivariate mixed responses to extract the pure contributions related to each of the components in the system [13,14].

The wealth of mechanistic and structural information that can be obtained from the multispectroscopic monitoring of a protein folding process is shown through the application of multivariate curve resolution-alternating least squares (MCR-ALS) [14,15] to the thermal-induced protein folding of α-lactalbumin, a globular metalloprotein present in milk with a predominant α-helix secondary structure (see Fig. 2).

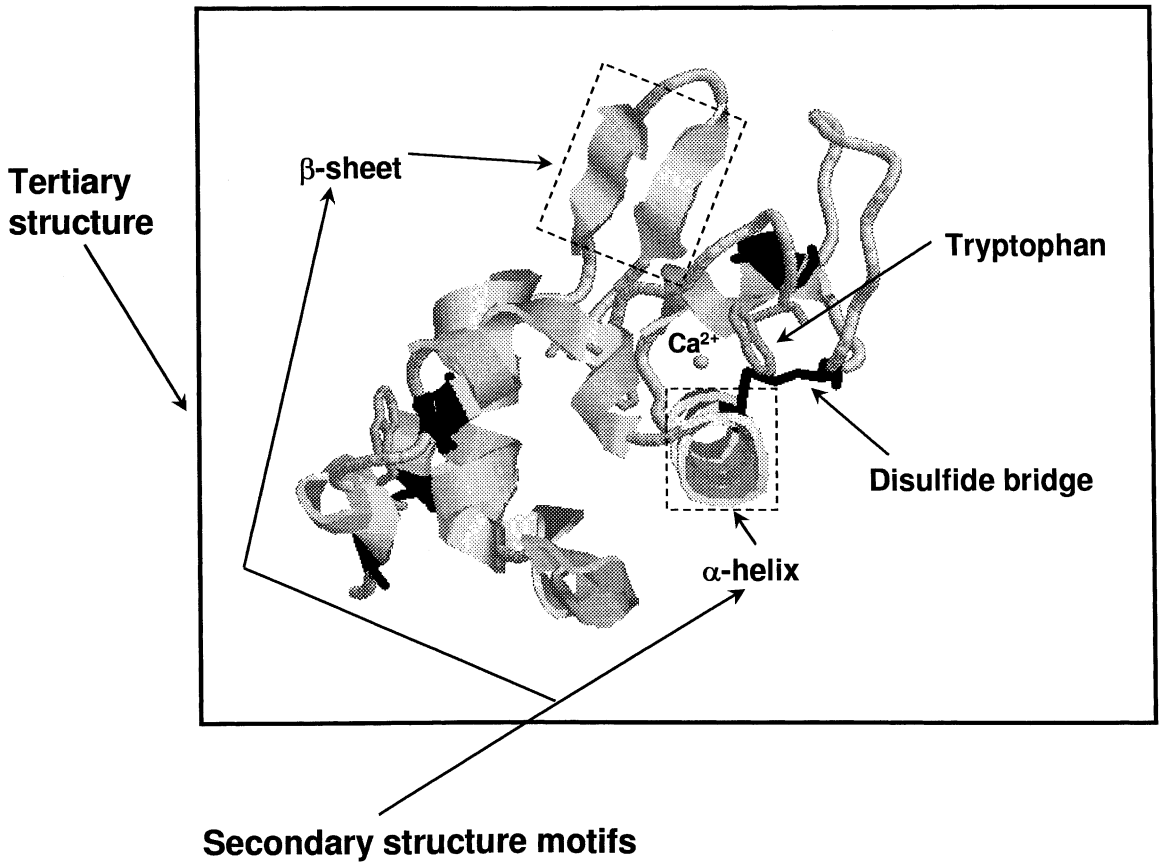

Fig. 2. Sketch of the native structure of α-lactalbumin, a protein formed by 123 amino acids with a molecular weight equal to 14 kDa. The solid line encloses the globular tertiary structure. Dashed lines surround protein zones with α-helix and β-sheet structural motifs. Black zones in protein chain indicate the four disulfide bridges and grey aromatic rings belong to the three tryptophan residues.

2. Experimental procedures

2.1. Chemicals

α-Lactalbumin from bovine milk was purchased from Sigma and sodium chloride (a.r.) from Merck. Both products were used without further purification. The protein was dissolved in bidistilled water. Protein solutions of 0.1 mg/ml were used to record UV, fluorescence and far-UV circular dichroism spectra. Near-UV circular dichroism spectra were measured in 1 mg/ml protein solutions. All the solutions had an ionic strength of 0.15 M in NaCl.

2.2. Instrumentation

UV spectra were recorded with a Perkin-Elmer λ-19 spectrometer equipped with a Peltier-type thermostatted cell holder. Absorbance lectures were measured from 250–350 nm and the wavelength step was 1 nm. A 10 mm pathlength closed quartz cell was used.

Fluorescence emission spectra were obtained with an Aminco Bowman AB-2 fluorimeter, equipped with a cell holder thermostatted by a water circulation bath. Emission spectra were measured between 300 and 430 nm and fluorescence intensities were recorded every 1 nm. The excitation wavelength was

285 nm. A 10 mm pathlength closed quartz cell was used.

Far- and near-UV CD spectra were obtained using a Jasco spectropolarimeter (model J-720) equipped with a cell holder thermostatted by a water circulation bath. Far- and near-UV CD spectra covered 195–250 and 250–330 nm wavelength ranges, respectively, and the wavelength step between ellipticity readings was 1 nm. A 10 mm pathlength closed quartz cell was used.

Data acquisition and storage was performed using the specific software of each instrument.

2.3. Experiments

2.3.1. Protein unfolding experiments

Suitable spectra are recorded at temperature ranges enclosed between 5 and 85°C. Temperature intervals among consecutive spectra increase from 2 to 5°C, depending on the spectral variation.

2.3.2. Protein refolding experiments

Suitable spectra are recorded at temperature ranges enclosed between 85 and 5°C. Temperature intervals among consecutive spectra decrease from 2 to 5°C, depending on the spectral variation.

In both kinds of experiments, a 2 min stabilisation time is waited before recording the spectrum at a given temperature. This period of time is fixed to avoid the presence of uncontrolled time-dependent effects on the evolution of the monitored processes. This working procedure is, therefore, adequate to detect protein conformations at equilibrium [1], but does not give information about transient kinetic intermediates, which evolve in a millisecond time scale [2].

3. Datasets and data analysis

The spectra recorded during the monitoring of a thermal-dependent protein folding process are organised in a data matrix $\boldsymbol{D}$, whose rows are the spectra collected at each temperature and whose columns depict the melting curves (absorbance versus temperature profiles) for each wavelength (see Fig. 1). This dataset contains all the necessary information to describe quantitatively the evolution of the different protein conformations involved in the process and to identify structurally each of these species. The task of the data analysis should then be unravelling the pure contributions associated with each protein conformation along the monitored process from the mixed raw experimental response [16,17].

In chemometrics, mixture analysis is often tackled using resolution techniques [12–14], which decompose the raw dataset, $\boldsymbol{D}$, according to the general expression

$$\boldsymbol{D} = \boldsymbol{C}\boldsymbol{S}^{\mathrm{T}} + \boldsymbol{E} \quad (1)$$

where $\boldsymbol{C}$ and $\boldsymbol{S}^{\mathrm{T}}$ are small matrices, which contain profiles related to the evolution along the rows and along the columns of each pure component in $\boldsymbol{D}$, respectively, and $\boldsymbol{E}$ is the error matrix, i.e. the residual variation of the dataset that is not related to any chemical contribution. In the context of protein folding, columns in matrix $\boldsymbol{C}$ are the concentration profiles of each protein conformation as a function of the temperature and rows in matrix $\boldsymbol{S}^{\mathrm{T}}$ are their related pure spectra.

Among the available resolution methods, multivariate curve resolution-alternating least squares (MCR-ALS) has been selected because of the versatility shown to adapt to very diverse chemical datasets and the possibility to handle either two-way datasets (single data matrices) or three-way datasets [16–23]. Analysable three-way datasets can be either row-wise augmented matrices, column-wise augmented matrices or row- and column-wise augmented matrices (see Fig. 3). Note that in the case of either row- or column-wise augmented data arrays, only one direction of the array (rows and columns, respectively) needs to be common to all the appended data matrices. This is a very flexible feature, hardly found in other three-way data analysis methods [12,24]. The general steps followed in the application of MCR-ALS to any kind of dataset are listed below

1. Determination of the number of compounds in $\boldsymbol{D}$ (e.g. by singular value decomposition (SVD) [25]).
2. Building initial estimates of the $\boldsymbol{C}$ matrix (e.g. using evolving factor analysis (EFA) [26]) or the $\boldsymbol{S}^{\mathrm{T}}$ matrix (e.g. using methods to select the purest spectra in $\boldsymbol{D}$ [27]).
3. Given $\boldsymbol{D}$ and $\boldsymbol{C}$, least-squares calculation and constraint of $\boldsymbol{S}^{\mathrm{T}}$ [12,14,15,24].

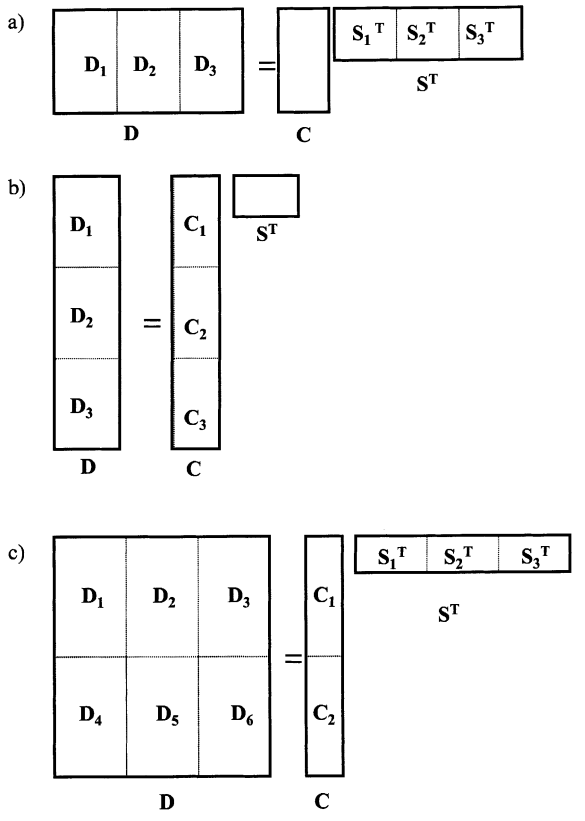

Fig. 3. MCR-ALS data decomposition of: (a) a row-wise augmented data matrix $[\boldsymbol{D}_1\ \boldsymbol{D}_2\ \boldsymbol{D}_3] = \boldsymbol{C}[\boldsymbol{S}_1^{\mathrm{T}}\ \boldsymbol{S}_2^{\mathrm{T}}\ \boldsymbol{S}_3^{\mathrm{T}}]$; (b) a column-wise augmented data matrix $[\boldsymbol{D}_1;\ \boldsymbol{D}_2;\ \boldsymbol{D}_3] = [\boldsymbol{C}_1;\ \boldsymbol{C}_2;\ \boldsymbol{C}_3]\boldsymbol{S}^{\mathrm{T}}$ and (c) a row- and column-wise augmented data matrix $[\boldsymbol{D}_1\ \boldsymbol{D}_2\ \boldsymbol{D}_3;\ \boldsymbol{D}_4\ \boldsymbol{D}_5\ \boldsymbol{D}_6] = [\boldsymbol{C}_1;\ \boldsymbol{C}_2][\boldsymbol{S}_1^{T}\ \boldsymbol{S}_2^{T}\ \boldsymbol{S}_3^{T}]$. MATLAB notation is used to design the three-way arrays (i.e. blanks between matrices for row-wise augmentation and semicolons for column-wise augmentation).

4. Given $\boldsymbol{D}$ and $\boldsymbol{S}^{\mathrm{T}}$, least-squares calculation and constraint of $\boldsymbol{C}$ [12,14,15,24].
5. Go to step 3 until convergence is achieved.

In any iterative resolution method, the appropriate application of constraints is crucial to drive on the optimisation to the right solution. A constraint forces a profile to be shaped so as it fulfils a systematic feature in the dataset, either mathematical or chemical. MCR-ALS allows for a very flexible application of constraints. Thus, $\boldsymbol{C}$ and $\boldsymbol{S}^{\mathrm{T}}$ can be differently constrained and, within each of these matrices, all or some of the profiles can be forced to obey a selected constraint. The same policy holds for three-way datasets, where each submatrix in either an augmented $\boldsymbol{C}$ or $\boldsymbol{S}^{\mathrm{T}}$ matrix can be treated according to their particular characteristics [12,14,15,28–30].

Focusing on the protein folding problem, the concentration profiles related to the different protein conformations have been constrained to be non-negative and to form a closed system. Unimodality (i.e. presence of one maximum per profile) has also been applied due to the monotonic variation of temperature along the experiments performed, either gradually increasing or decreasing. Such a variation induces either the transformation of organised conformations into disordered structures or vice-versa, but never both ways in a single experiment. Selectivity has also been introduced at the lowest temperature, where only the native (organised) conformation is said to be present by definition.

Spectra profiles have been constrained according to the properties of their related techniques. Thus, UV and fluorescence spectra are forced to be positive, whereas CD spectra remain unconstrained because of the natural occurrence of negative ellipticies in these measurements. No selective wavelengths have been detected for any of the techniques used.

The convergence criterion in the MCR-ALS optimisation is based on the comparison of the lack of fit obtained in two consecutive iterations. When the relative difference in fit is below a threshold value, the optimisation is finished. The lack of fit is calculated according to the expression

$$\text{lack of fit (\%)} = 100 \times \sqrt{\frac{\sum_{i,j} r_{ij}^2}{\sum_{i,j} d_{ij}^2}} \qquad (2)$$

where d_{ij} designs an element of the raw data matrix $\boldsymbol{D}$ and r_{ij} its related residual. A maximum number of iterative cycles may be also used as stop criterion.

4. Results and discussion

The experiments performed to monitor unfolding and refolding of α-lactalbumin are shown in Table 1. Experiment codes in this table will be used later on in this section. To describe the datasets analysed, MATLAB notation is adopted (see caption of Fig. 3 for illustrative examples).

Table 1
Experiment coding used along the work

Data matrix	Temperature variation (°C)	Spectrometric technique	Process monitored
$\boldsymbol{D}_1$	13 → 75	UV	Protein unfolding
$\boldsymbol{D}_2$	13 → 75	Fluorescence	
$\boldsymbol{D}_3$	5 → 85	Near-UV CD	
$\boldsymbol{D}_4$	5 → 85	Far-UV CD	
$\boldsymbol{D}_5$	75 → 13	UV	Protein refolding
$\boldsymbol{D}_6$	75 → 13	Fluorescence	
$\boldsymbol{D}_7$	85 → 5	Near-UV CD	
$\boldsymbol{D}_8$	85 → 5	Far-UV CD	

The experimental data obtained in this work contain all the information necessary to describe properly the thermal-induced unfolding and refolding of α-lactalbumin. However, not all the experiments focus on the same aspects related to these processes. Therefore, taking either one experiment or combining several of them to form the adequate three-way dataset, i.e. a row-wise, a column-wise or a row- and column-wise augmented matrix, different questions may be answered.

Table 2 summarises all the MCR-ALS results obtained in this work. For each data analysis, the purpose of the study is described and the conditions of application of MCR-ALS are specified. The lack of fit is also included as a quality measure of the results. All the analysis shown used EFA initial estimates, though similar results were obtained working with spectral estimates based on pure variable selection methods. The sections below describe in detail how different issues related to a protein folding process have been addressed.

4.1. Description of changes in the protein secondary structure

As pointed out in the introduction, far-UV CD is the spectral technique specifically sensitive to changes in the protein secondary structure. Experiments $\boldsymbol{D}_4$ and $\boldsymbol{D}_8$ monitor these variations in the protein unfolding and refolding process, respectively. In both cases, SVD analysis detected two components, i.e. the presence of two distinct secondary structures. Fig. 4 shows the concentration profiles and spectra related to these structures during the α-lactalbumin unfolding process. The very noisy concentration profiles and the related lack of fit are due to the typical high noise level associated with CD measurements performed at far-UV wavelength ranges, where some interferences derived from the instrumental measure cannot be completely suppressed. The spectra obtained for the native and the denatured/unfolded secondary structures show the typical spectral features described in the literature as representative of the α-helix motif (an intense

Table 2
MCR-ALS results related to α-lactalbumin unfolding and refolding processes[a]

Dataset[b]	Phenomenon studied	Constraints[c]	Lack of fit (%)
$\boldsymbol{D}_4$	Changes in secondary structure	[1, 2, 3, 4/0]	5.4
$\boldsymbol{D}_8$			4.7
$[\boldsymbol{D}_1 \quad \boldsymbol{D}_2 \quad \boldsymbol{D}_3]$	Changes in tertiary structure	[1, 2, 3, 4/1|1|0]	3.0
$[\boldsymbol{D}_4 \quad \boldsymbol{D}_5 \quad \boldsymbol{D}_6]$			4.9
$[\boldsymbol{D}_1 \quad \boldsymbol{D}_2 \quad \boldsymbol{D}_3 \quad \boldsymbol{D}_4; \quad \boldsymbol{D}_5 \quad \boldsymbol{D}_6 \quad \boldsymbol{D}_7 \quad \boldsymbol{D}_8]$	Temperature-dependence of folding process	[1, 2, 3, 4/1|1|0|0]	7.2

[a] According to SVD results, all datasets have been resolved using two components.

[b] Data matrices are coded in Table 1.

[c] [Concentration submatrices/spectra submatrices]. Solid lines are used to separate information associated with the constraints applied to each submatrix in $\boldsymbol{C}$ or $\boldsymbol{S}^{\mathrm{T}}$ only when these submatrices are constrained differently. Coding of constraints: 0 → unconstrained, 1 → non-negativity, 2 → unimodality, 3 → closure, 4 → selectivity.

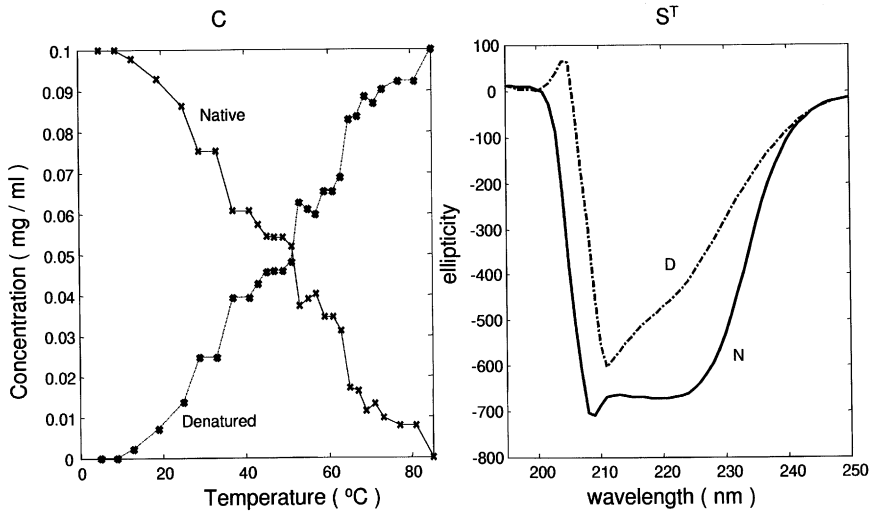

Fig. 4. MCR-ALS results related to the study of changes in secondary structure motifs. Concentration profiles and spectra from the single matrix $\boldsymbol{D}_4$ (see Table 1 for matrix coding description).

negative band with two shoulders located around 220 and 210 nm) and the random coil (a sharp negative band at short wavelengths and weaker features around 220 nm), respectively [31,32].

4.2. Description of changes in the protein tertiary structure

UV absorption, fluorescence and near-UV CD are spectral techniques known to be sensitive to changes in the tertiary structure of proteins. Therefore, analysis of the row-wise augmented data matrix formed by the datasets coming from each of these three techniques will provide an accurate description of the evolution of this organisation level of the protein along the process of interest.

MCR-ALS analysis of the three-way datasets $[\boldsymbol{D}_1 \quad \boldsymbol{D}_2 \quad \boldsymbol{D}_3]$ and $[\boldsymbol{D}_5 \quad \boldsymbol{D}_6 \quad \boldsymbol{D}_7]$ is performed to describe changes in tertiary structure linked to α-lactalbumin unfolding and refolding processes, respectively. Fluorescence spectra (in experiments $\boldsymbol{D}_2$ and $\boldsymbol{D}_6$) have been normalised to suppress random intensity changes unrelated to the monitored processes. Data matrices within a three-way dataset have been conveniently scaled, i.e. multiplied by a certain factor, to balance the different scale units of the combined techniques. As shown in Fig. 3, MCR-ALS analysis of each row-wise augmented matrix provides a matrix $\boldsymbol{C}$, common to all experiments treated together, and a row-wise augmented $\boldsymbol{S}^{\mathrm{T}}$ matrix, formed by the $\boldsymbol{S}_i^{\mathrm{T}}$ submatrices that contain the pure spectra obtained by the technique used in the related $\boldsymbol{D}_i$ matrix.

SVD of both $[\boldsymbol{D}_1 \quad \boldsymbol{D}_2 \quad \boldsymbol{D}_3]$ and $[\boldsymbol{D}_5 \quad \boldsymbol{D}_6 \quad \boldsymbol{D}_7]$ datasets detected two components. Fig. 5 shows the results related to the α-lactalbumin unfolding process. The spectra shown have been rescaled to the original units. Working with the augmented dataset instead of analysing individually each of the matrices results in a decrease of ambiguity in the final solutions and, as a consequence, in an improvement of the profile shapes obtained. Notably, the concentration profiles in Fig. 5 are much smoother than those from individual matrix analysis (see Fig. 4, for instance). This smoothness is due to the fact that the $\boldsymbol{C}$ single matrix shows concentration profiles that describe variations common to all matrices in the augmented dataset and this removes noisy patterns or small distortions specific of a certain technique. In protein folding processes, these better defined profiles allow for a more accurate definition of the melting point (T at which 50% of the native protein has been denatured, i.e. crossing point in the concentration profile plot). The benefit of using a row-wise dataset does not restrict to the definition of concentration profiles. Pure spectra related to techniques with

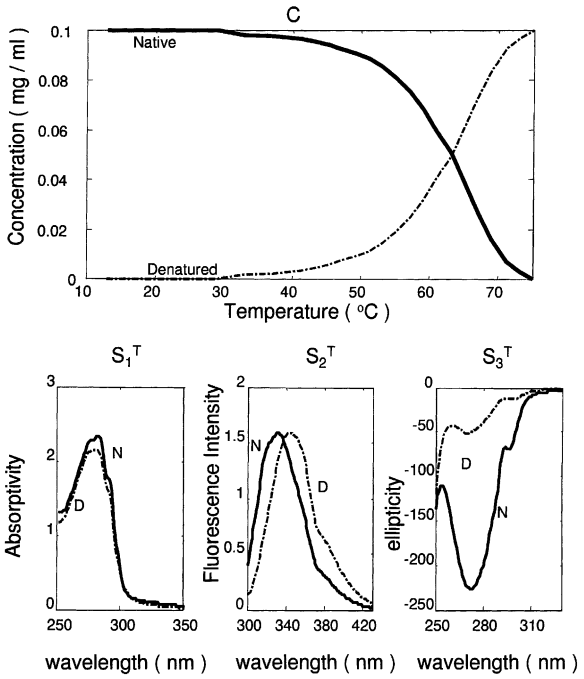

Fig. 5. MCR-ALS results related to the study of changes of tertiary structure. Concentration profiles and spectra from the row-wise augmented matrix [$\boldsymbol{D}_1$ $\boldsymbol{D}_2$ $\boldsymbol{D}_3$] (see Table 1 for matrix coding description).

weak variations between native and disordered structures, like UV, are more easily differentiated than when treated alone.

It is worth noting that the melting point obtained from MCR-ALS analysis stems from the whole spectral information contained in the multivariate dataset and it is, therefore, much more robust than the equivalent parameter obtained with univariate wavelength-dependent classical approaches. The pure CD spectra agree with literature references associating unfolding processes with a strong loss of intensity in the CD absorption band. Loss of tertiary structure shows up also through a clear band shift to longer wavelengths in fluorescence spectra and a disappearance of spectral features on top of the UV absorption band.

4.3. Mechanism of the process: detection of intermediates

Protein unfolding and/or refolding can be one- or multistep processes. Detection of intermediate species requires the simultaneous study of the evolution of protein secondary and tertiary structures. Indeed, although separate studies on each of these structure levels may show the same number of species, this fact does not mean that they evolve identically as a function of the process-controlling variable. When losses or gains of secondary and tertiary structure (linked to protein unfolding and refolding, respectively) do not take place simultaneously, i.e. temperature-dependent concentration profiles related to both kinds of structure are shifted from one another, there are intermediates. Because of the difference between the melting temperatures for α-lactalbumin in Fig. 4 (changes in secondary structure) and Fig. 5 (changes in tertiary structure), the presence of an intermediate cannot be discarded. However, this visual comparison is not conclusive due to the high noise associated with the data matrix obtained with far-UV CD and to the possible resolution ambiguities related to the analysis of this single data matrix.

A more robust and general strategy to find out the presence of intermediates relies on the comparison of ranks related to SVD analysis of single data matrices linked to changes in secondary or in tertiary structure with the rank of a row-wise augmented dataset formed by matrices that result from monitoring changes at both structure levels. In this work, near- and far-UV CD matrices have been taken to form the row-wise augmented matrix. Circular dichroism measurements are specially suitable for this purpose. They are very sensitive to protein conformation changes and have the great value of showing two separate wavelength regions (far- and near-UV) where changes in secondary and tertiary structure are specifically monitored, respectively. Moreover, since the measurements compared are done with the same instrument, undesirable contributions to the rank due to different instrumental backgrounds and to slightly different calibrations of the temperature control devices used in each spectrometric technique are avoided. In a row-wise augmented matrix including both far- and near-UV CD measurements, the pure spectrum of a protein conformation is actually a long spectrum formed by the suitable appended near- and far-UV spectra. Bearing this in mind, the pure spectra related to native, intermediate (when existing) and denatured species are extremely different. Thus, the native and intermediate protein conformations have a very similar secondary structure (far-UV spectrum), but the tertiary structure (near-UV

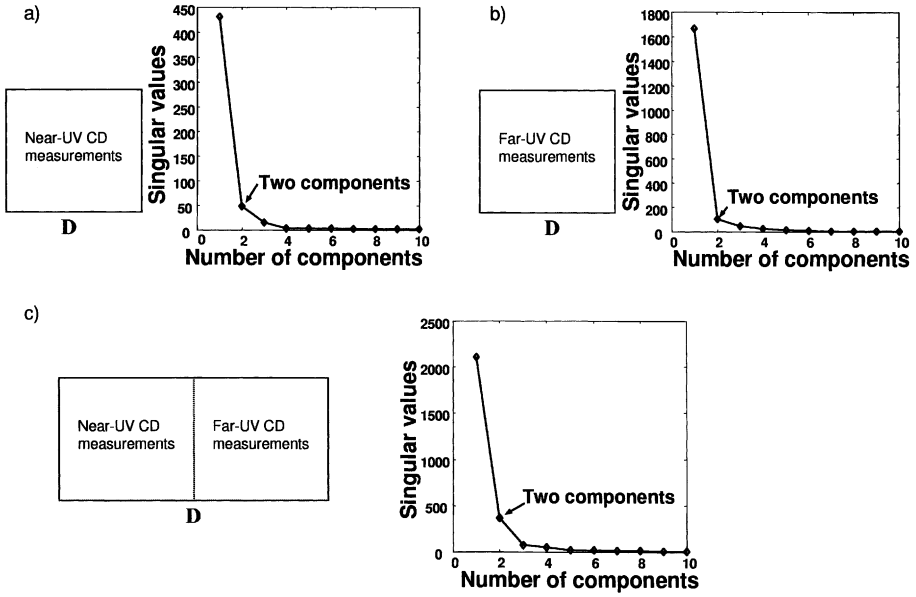

Fig. 6. Application of SVD to the detection of intermediates in a protein folding process: (a) rank analysis of the near-UV CD matrix ($\boldsymbol{D}_7$) related to tertiary structure changes; (b) rank analysis of the far-UV CD matrix ($\boldsymbol{D}_8$) related to secondary structure changes; and (c) rank analysis of the row-wise augmented matrix [$\boldsymbol{D}_7$ $\boldsymbol{D}_8$]. Rank [$\boldsymbol{D}_7$ $\boldsymbol{D}_8$] = rank $\boldsymbol{D}_7$ = rank $\boldsymbol{D}_8$ indicates absence of intermediates (see Table 1 for matrix coding description).

spectrum) is completely different (ordered in the native and disordered in the intermediate), whereas the intermediate and the denatured conformations have an almost identical tertiary structure, but their secondary structures differ dramatically from one another (ordered in the intermediate and disordered in the denatured). Therefore, the pure 'long' spectra of the three conformations are perfectly distinguishable. The suitability of the CD measurements in both near- and far-UV regions to find intermediate species has been proven in studies performed to find kinetic transient species and protein conformations related to intermediate species at equilibrium [16,17,33].

Taking α-lactalbumin refolding process as example, SVD analyses of single matrices $\boldsymbol{D}_7$ and $\boldsymbol{D}_8$ show the presence of two components (see Fig. 6(a) and (b)). When these two matrices are row-wise appended, the rank should not increase if losses of both structures are synchronised. If there is a temperature shift in these structure changes, three components would be detected: the first would have a spectrum with both disordered structures, the second would show a spectrum with a partially or totally ordered secondary structure and no tertiary structure and the third would have a spectrum with both structure levels ordered. As shown in Fig. 6(c), there is no rank increase in the augmented dataset and, therefore, α-lactalbumin refolds in a one-step process at the working conditions and time scale used. Similar conclusions are obtained for the analogous unfolding process.

4.4. Comparison of thermal-dependence of unfolding and refolding protein processes

In Sections 4.1–4.3, α-lactalbumin unfolding and refolding processes have been proven to occur as one-step processes from completely ordered to completely disordered structures or vice-versa. However, this information is not enough to answer whether both processes have the same temperature-dependence and whether they share structurally identical ordered and disordered protein forms.

A row- and column-wise augmented dataset formed by all the experiments performed and organised as [$\boldsymbol{D}_1$ $\boldsymbol{D}_2$ $\boldsymbol{D}_3$ $\boldsymbol{D}_4$; $\boldsymbol{D}_5$ $\boldsymbol{D}_6$ $\boldsymbol{D}_7$ $\boldsymbol{D}_8$] is used to answer these queries. In this data array, matrices

related to the same technique should be scaled using the same factor. According to the architecture of this dataset (see Fig. 3(c) for comparison), MCR-ALS decomposition provides a column-wise augmented data matrix $\boldsymbol{C}$ of concentration profiles formed by submatrices $\boldsymbol{C}_1$ (related to protein unfolding experiments $\boldsymbol{D}_1$–$\boldsymbol{D}_4$) and $\boldsymbol{C}_2$ (related to protein refolding experiments $\boldsymbol{D}_5$–$\boldsymbol{D}_8$) and a row-wise augmented $\boldsymbol{S}^{\mathrm{T}}$ matrix formed by submatrices $\boldsymbol{S}_1^{\mathrm{T}}$ (UV spectra related to $\boldsymbol{D}_1$ and $\boldsymbol{D}_5$), $\boldsymbol{S}_2^{\mathrm{T}}$ (fluorescence spectra from $\boldsymbol{D}_2$ and $\boldsymbol{D}_6$), $\boldsymbol{S}_3^{\mathrm{T}}$ (near-UV CD spectra from $\boldsymbol{D}_3$ and $\boldsymbol{D}_7$) and $\boldsymbol{S}_4^{\mathrm{T}}$ (far-UV CD spectra from $\boldsymbol{D}_4$ and $\boldsymbol{D}_8$). Working with this large array provides more reliable estimates of $\boldsymbol{C}_i$ and $\boldsymbol{S}_i^{\mathrm{T}}$ matrices. Indeed, each $\boldsymbol{C}_i$ submatrix is defined by the common information contained in all the experiments performed to monitor the same protein process (either unfolding or refolding) and each $\boldsymbol{S}_i^{\mathrm{T}}$ matrix has the common spectral information included in all the experiments carried out using the same spectral technique.

SVD of this large array showed the presence of only two components. This means that the two protein species found in both unfolding and refolding experiments can be described as combinations of a pure ordered and a pure disordered protein conformation. Fig. 7 shows the concentration profiles and spectra related to the analysis of this array. The lack of fit obtained (see Table 2) is not very large considering the presence of the very noisy far-UV CD matrices in the array and the possible error that could come from slight inaccuracy in the closure constants (protein concentrations) applied during the MCR-ALS optimisation.

This augmented dataset is the most suitable to compare the thermal-dependence of unfolding and refolding of α-lactalbumin because the concentration profiles for both kind of processes ($\boldsymbol{C}_1$ and $\boldsymbol{C}_2$, respectively) are referred to the same pure spectra. As can be seen in Fig. 7, these pure spectra agree with the features described in previous sections for ordered and disordered protein conformations. Examining the concentration profiles plots, different melting points correspond to unfolding and refolding and neither the maximum disorder is reached in the protein unfolding process nor the native structure is totally recovered in the protein refolding experiments. Therefore, it can be concluded that at the experimental conditions and the time scale used, α-lactalbumin does not show neither

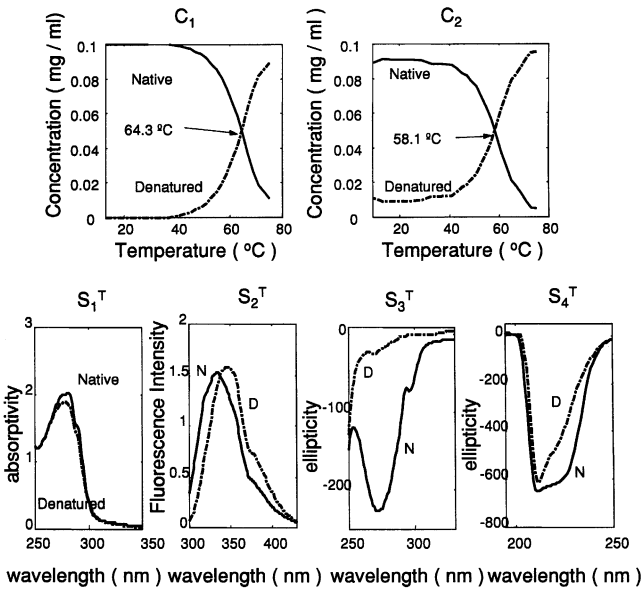

Fig. 7. MCR-ALS results related to the study of the thermal-dependence of protein folding and unfolding processes. Concentration profiles and spectra from the row- and column-wise augmented matrix [$\boldsymbol{D}_1$ $\boldsymbol{D}_2$ $\boldsymbol{D}_3$ $\boldsymbol{D}_4$; $\boldsymbol{D}_5$ $\boldsymbol{D}_6$ $\boldsymbol{D}_7$ $\boldsymbol{D}_8$] (see Table 1 for matrix coding description).

the same temperature-dependence nor structurally identical ordered and disordered conformations in unfolding and refolding processes. It must be noted that this conclusion stems from the experimental datasets analysed and no generalisation of the unequal temperature-dependence in unfolding/refolding processes for α-lactalbumin for any working condition is allowed, e.g. it cannot be discarded that longer stabilisation times before recording the spectra at each temperature led to different results. Because of the lack of previous knowledge about the time-dependence of protein folding in α-lactalbumin and to avoid as much as possible confounding temperature and time effects on this process, one of these factors (time) was kept constant throughout the experiments. A complete description of this protein folding would, however, need a clarification of the kinetic side of this protein conformational transition process.

5. Conclusions

The use of curve resolution techniques is shown to be a powerful tool to address the problem of qualitative and quantitative description of protein folding. Thus, curve resolution techniques can handle easily large and complete multitechnique spectral outputs and can extract the pure contributions due to different protein conformations from the variation in the mixed spectra obtained during the monitoring of a protein folding process. Within this group of techniques, MCR-ALS can analyse simultaneously experiments of different origin taking advantage of the specific features of each of them through a flexible application of constraints. The adaptation of this method to handle three-way datasets of very diverse structure puts no limit to the possibilities of combination of experiments, which can be appropriately designed to answer different possible mechanistic and structural queries related to protein folding processes.

Acknowledgements

This work has been financially supported by the Spanish and Catalan governments through projects BQU2000-0788 and 1999-SGR-00048, respectively. S. Navea also acknowledges the Universitat de Barcelona for a PhD scholarship.

References

[1] R.H. Pain, Mechanisms of Protein Folding, Oxford University Press, 1994.
[2] B. Nölting, Protein Folding Kinetics, Springer, Berlin, 1999.
[3] R. Hlodan, S. Craig, R.H. Pain, Biotechnol. Genet. Eng. Rev. 9 (1991) 47.
[4] M.L. Anson, Adv. Protein Chem. 2 (1945) 361.
[5] G.E. Schulz, R.H. Schirmer, Principles of Protein Structure, Springer, Berlin, 1979.
[6] J.T. Pelton, L.R. McLean, Anal. Biochem. 277 (2000) 167.
[7] M.R. Eftink, Methods Enzimol. 259 (1995) 487.
[8] K. Fujiwava, M. Arai, A. Shimizu, M. Ikeguchi, K. Kuwajima, S. Sugai, Biochemistry 38 (1999) 4455.
[9] L.K. Cramer, Biochemistry 34 (1995) 7170.
[10] G.A. Manderson, L.K. Creamer, M.J. Hardman, J. Agric. Food Chem. 47 (1999) 4557.
[11] M. Svensson, H. Sabharwal, A. Hakanson, A.-K. Mossberg, P. Lipniunas, H. Leffler, C. Svanborg, S. Linse, J. Biol. Chem. 274 (1999) 6388.
[12] A. de Juan, E. Casassas, R. Tauler, Encyclopedia of Analytical Chemistry: Instrumentation and Applications. Soft-Modeling of Analytical Data, Wiley, NY, 11 (2000) 9800.
[13] J. Craig Hamilton, P.J. Gemperline, J. Chemometr. 4 (1990) 1.
[14] R. Tauler, A.K. Smilde, B.R. Kowalski, J. Chemometr. 9 (1995) 31.
[15] R. Tauler, Chemom. Intell. Lab. Sys. 30 (1995) 133.
[16] J. Mendieta, H. Folqué, R. Tauler, Biophys. J. 76 (1999) 451.
[17] J. Mendieta, M.S. Díaz-Cruz, M. Esteban, R. Tauler, Biophys. J. 74 (1998) 2876.
[18] A. de Juan, A. Izquierdo-Ridorsa, R. Tauler, G. Fonrodona, E. Casassas, Biophys. J. 73 (1997) 2937.
[19] M. Vives, R. Gargallo, R. Tauler, Anal. Chem. 71 (1999) 4328.
[20] J. Saurina, S. Hernández-Cassou, R. Tauler, Anal. Chem. 69 (1997) 2329.
[21] M. Esteban, C. Ariño, J.M. Díaz-Cruz, M.S. Díaz-Cruz, R. Tauler, Trends Anal. Chem. 19 (2000) 49.
[22] J.S. Salau, R. Tauler, J.M. Bayona, I. Tolosa, Environ. Sci. Technol. 37 (1997) 3482.
[23] S. Nigam, A. de Juan, R.J. Stubbs, S.C. Rutan, Anal. Chem. 72 (2000) 1956.
[24] A. de Juan, R. Tauler, J. Chemometr. (2001), in press.
[25] G.H. Golub, C. Reinsch, Numer. Math. 14 (1970) 403.
[26] M. Maeder, Anal. Chem. 59 (1987) 527.
[27] W. Widing, J. Guilment, Anal. Chem. 63 (1991) 1425.
[28] A. de Juan, Y. Vander Heyden, R. Tauler, D.L. Massart, Anal. Chim. Acta 346 (1997) 307.
[29] R. Bro, S. de Jong, J. Chemometr. 11 (1997) 393.
[30] R. Bro, N.D. Sidiropoulos, J. Chemometr. 12 (1998) 223.
[31] S.Y. Venyaminov, I.A. Baikalov, C.-S.C. Wu, J.T. Yang, Anal. Biochem. 198 (1991) 250.
[32] N.J. Greenfield, Anal. Biochem. 235 (1996) 1.
[33] K. Fujiwara, M. Arai, A. Shimizu, M. Ikeguchi, K. Kuwajima, S. Sugai, Biochemistry 38 (1999) 4455.