Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Magnetic Resonance 172 (2005) 24–30
JMR Journal of Magnetic Resonance
www.elsevier.com/locate/jmr

Communication

The equilibrium unfolding of MerP characterized by multivariate analysis of 2D NMR data

Anders Berglund[a], Ann-Christin Brorsson[b], Bengt-Harald Jonsson[c], Ingmar Sethson[a,*]

[a] *Organic Chemistry, Department of Chemistry, Umeå University, Sweden*
[b] *Biochemistry, Umeå University, S-901 87 Umeå, Sweden*
[c] *Molecular Biotechnology/IFM, Linköping University, SE-581 83 Linköping, Sweden*

Received 1 July 2004; revised 27 August 2004

Abstract

A general problem when analysing NMR spectra that reflect variations in the environment of target molecules is that different resonances are affected to various extents. Often a few resonances that display the largest frequency changes are selected as probes to reflect the examined variation, especially in the case, where the NMR spectra contain numerous resonances. Such a selection is dependent on more or less intuitive judgements and relying on the observed spectral variation being primarily caused by changes in the NMR sample. Second, recording changes observed for a few (albeit significant) resonances is inevitably accompanied by not using all available information in the analysis. Likewise, the commonly used chemical shift mapping (CSM) [Biochemistry 39 (2000) 26, Biochemistry 39 (2000) 12595] constitutes a loss of information since the total variation in the data is not retained in the projection into this single variable. Here, we describe a method for subjecting 2D NMR time-domain data to multivariate analysis and illustrate it with an analysis of multiple NMR experiments recorded at various folding conditions for the protein MerP. The calculated principal components provide an unbiased model of variations in the NMR spectra and they can consequently be processed as NMR data, and all the changes as reflected in the principal components are thereby made available for visual inspection in one single NMR spectrum. This approach is much less laborious than consideration of large numbers of individual spectra, and it greatly increases the interpretative power of the analysis.

Keywords: Multivariate NMR data analysis; Protein folding; PCA; PLS; GuHCl

1. Introduction

An NMR resonance is sensitive to the surrounding environment of the nucleus to which it owes its origin. The NMR spectrum has therefore a potential to reflect intermolecular interactions since such interactions may induce spectral changes, either by inducing conformational changes in the molecule investigated or by creating direct contacts with the interacting molecule. In both cases the surroundings of the nucleus monitored is perturbed and the induced spectral changes can be used to monitor the molecular interactions. However, the spectral changes involved may often be difficult to interpret since the target molecule is often sizeable and accordingly generates complex NMR spectra. In addition, induced spectral changes are not easily identified and quantified since numerous resonances are affected by the molecular interactions and the changes are spread out over several different spectra.

The objective for applying multivariate data analysis (MVA) is twofold, firstly to be able to quantify how similar the different spectra are to one another and by that obtain a statistical representation of the spectral changes occurring. Second, to enable the representation of this

* Corresponding author. Fax: +46 90 13 88 85.
E-mail address: ingmar.sethson@chem.umu.se (I. Sethson).

1090-7807/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2004.09.010

variation in one single NMR spectrum obtained by the Fourier transform of the calculated principal components, which represents an image of changes occurring in several NMR spectra. The best approach to accomplish this would be to subject NMR experiments to multivariate data analysis using Principal Component Analysis (PCA)[1] [3,4] and partial least squares (PLS) [5] regression.

It has previously been shown that multivariate analysis can be applied to one-dimensional NMR free induction decays (FID) and that the principal components thereby obtained can be Fourier transformed to provide interpretable NMR spectra [6]. As expected, the results were found to be identical to those obtained from frequency domain NMR spectra. However, detailed interpretation of the features in such a 1D spectrum will be hampered by the restricted resolution, especially when analysing larger molecules. This suggests, therefore, that one should extend the multivariate analysis to 2 or 3D NMR to exploit the greater resolution of higher dimensional NMR. Another advantage of using 2D NMR is that the second frequency domain generally contains additional information, which improves the calculated models.

2. Methods

2.1. NMR

The NMR experiments were collected on an NMR sample that initially had a protein concentration 0.9 mM, a GuHCl concentration 0.35 M and a pH 6.5 (50 mM phosphate buffer). The increase in GuHCl concentration was obtained by adding small aliquots of 5.8 M GuHCl solution at pH 6.5 (50 mM phosphate buffer) to provide samples of the following denaturant concentrations: 0.35, 0.70, 1.05, 1.35, 1.55, 1.70, 1.90, 2.05, 2.35, 2.70, and 2.90 M GuHCl. The NMR data were collected as ^{1}H–^{15}N ge-HSQC spectra [7] consisting of $1024 * 200$ complex data points. Processing of the NMR data was conducted by using the NMR software package SwaN-MR [8].

2.2. Data pre-processing

The binary NMR data was transformed to ASCII format and thereafter each NMR 2D NMR experiment was converted to a row vector and normalized to length one. The row vectors for all NMR spectra was arranged into a matrix with 11 rows and 204,800 columns. The data were normalized by removing the mean from each column (time point).

[1] *Abbreviations used:* PLS, partial least squares, PCA, principal component analysis.

2.3. PCA and PLS

The PCA and PLS models where calculated and visualised using MATLAB (www.mathworks.com). PCA decomposes the X matrix, the collection of FID's, into score, T, and loading vectors, $P \cdot X = t_1 * p_1^T + t_2 * p_2^T + t_3 * p_3^T + \cdots + t_A * p_A^T + E$, where E are the residuals and A is the number of principal components. The first principal component is derived so it describes maximum variance in the X-matrix and is then removed from the X-matrix, $X = X - t_1 * p_1^T$. A second principal component can then be calculated from the deflated X matrix, and that component will be orthogonal to the first. This is repeated until all systematic variation is described. The scores are related to the observations, here denaturant concentration and will show similarities among they observations. The loadings, P, describe how and what variables, here time-points in the FID, that are responsible for the separation seen in the scores.

In PLS there is both a X-matrix and a Y-matrix; Y is in this case a vector with the concentration of the denaturant. In PLS, the X matrix is also decomposed into a set of orthogonal components, the difference is that instead of describing the maximum variance in X (PCA) they now describe the maximum covariance between X and Y (PLS). $X = t_1 * p_1^T + t_2 * p_2^T + t_3 * p_3^T + \cdots + t_A * p_A^T + E$ and $Y = t_1 * c_1^T + t_2 * c_2^T + t_3 * c_3^T + \cdots + t_A * c_A^T + F$. The scores, T, relate both X and Y to each other. For each component in PLS, Y is described as a linear combination of all the X variables. The weight for each component, $\mathbf{w}$, describes how important a certain variable is for describing the response.

For the PLS models the (GuHCl) concentration for each NMR spectrum was used as the response (inverse regression). Inverse regression uses less assumption about the X-matrix than direct regression, where the spectra would have been the response. For a more general discussion about this, see Wold and Josefson [9]. Only the first weight vector, $\mathbf{w}_1$, was used for the interpretation of the PLS model since this is the best estimate of how important a certain column is for describing the response [10]. The $\mathbf{w}_1$ vector was Fourier transformed as an ordinary NMR time-domain signal to obtain the frequency 2D NMR spectrum, representing the weight vector.

3. Results

The multivariate analysis of 2D NMR data was applied to the oxidised form of the mercury binding protein MerP at different folding conditions. The equilibrium folding properties have already been characterised by CD and fluorescence methods [11] and the assignment [12] and structure [13] of MerP have previously been determined (Fig. 1). The data used in the

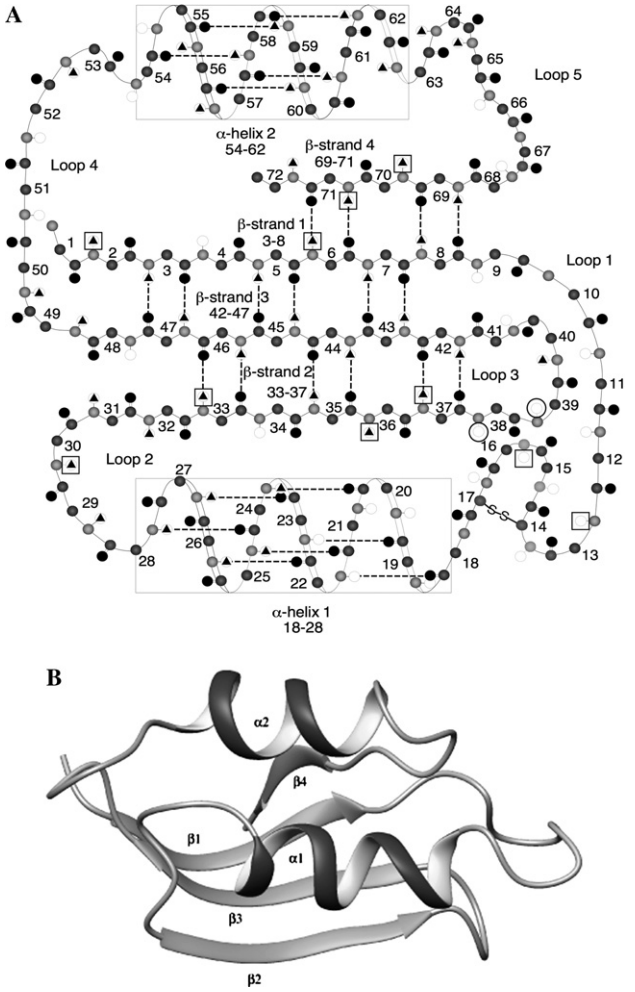

Fig. 1. The solution structure of the oxidized form of MerP presented as a two-dimensional cartoon (A) and as ribbon structure (B). The secondary structure elements are indicated in both representations. Selected amino acids are labelled as follows: circles—amino acids that change their proton chemical shifts up-field in the range 0.35–1.35 M GuHCl. Squares–amino acids that experience ^{15}N frequency shifts in the range 0.35–1.35 M GuHCl. Triangles—amino acids that are still observable at their native state frequencies in the range 2.05–2.90 M GuHCl.

analysis consisted of 11 ^{1}H–^{15}N ge-HSQC experiments [7], recorded at varying denaturant (GuHCl) concentrations. By stepwise increasing the denaturant concentration from 0.35 to 2.90 M, the transition from predominantly native to unfolded protein was induced. Interestingly, the native state and the unfolded ensemble of the protein coexist at all denaturant concentrations, due to the relatively low stability of the protein. Consequently, the addition of GuHCl induces spectral changes within the native state and the unfolded ensemble, as well as contributes to the global unfolding of native MerP molecules. The local unfolding processes in the native protein are within the fast exchange region of the NMR time scale, i.e., average NMR resonances are observed, whereas the major unfolding process is slow on the NMR time-scale and thus native and unfolded protein display separate resonances.

PCA calculations were performed on the original 2D NMR time-domain data. An inspection of the first three principal components revealed some discontinuity between the spectra obtained in the various NMR experiments. This occurs since the NMR signals obtained in different experiments are also affected by variations that are unrelated to the folding/unfolding of the protein, e.g., phase variation between the NMR transmitter/receiver. The unwanted phase variations were eliminated by rotating the data points in the NMR FIDs so that the 1st complex point coincided with the x axis, which guaranteed that the 0th order phase correction needed of the individual NMR spectra was identical. Using the same acquisition parameters in all NMR experiments ascertained that the needed 1st order phase correction did not vary between the experiments.

The first three principal components of the PCA analysis, after the data had been corrected for the phase and intensity variations, are shown in Fig. 2. The three-dimensional plot of the first three PCAs, explaining 59.2% (26.9, 19.7, and 12.6%, respectively) of the variance, provides an informative overview of the variation in the NMR data and shows how the different recorded 2D NMR spectra are related to one another. In this case, the addition of denaturant provides continuous changes between the different NMR experiments. Furthermore, the PCA model reveals that there are at least three main changes occurring, i.e., the addition of denaturant also induces spectral changes which do not correspond to the main two-stage unfolding process of the protein [11]. A Fourier transform of the loadings, i.e., how the various FIDs contribute to the principal components, would provide NMR spectra, enabling a visual inspection of all spectral changes caused by the addition of denaturant. However, these spectra would constitute a superposition of various effects, as revealed by the different directions of the changes in the score plot (Fig. 2). Consequently, a more interpretable representation of the induced spectral changes would be obtained by focussing on the GuHCl concentration ranges, where the PCAs vary in the same direction in the score plot. Therefore, three sub-models, covering the following denaturant concentrations ranges; 0.35–1.35 M, 1.55–2.05 M, and 2.05–2.90 M GuHCl; were subjected to a PLS analysis.

The denaturant concentration was used as the Y matrix, i.e., the response, to detect spectral changes related to the denaturant concentration. The Fourier transformation of the first PLS weight vector, $\mathbf{w_1}$, provided an NMR spectrum for each PLS model (Figs. 3A–C). These models explain 99% of the variation in the response, which is not surprising since there are so few samples (four) and so many variables. The correspond-

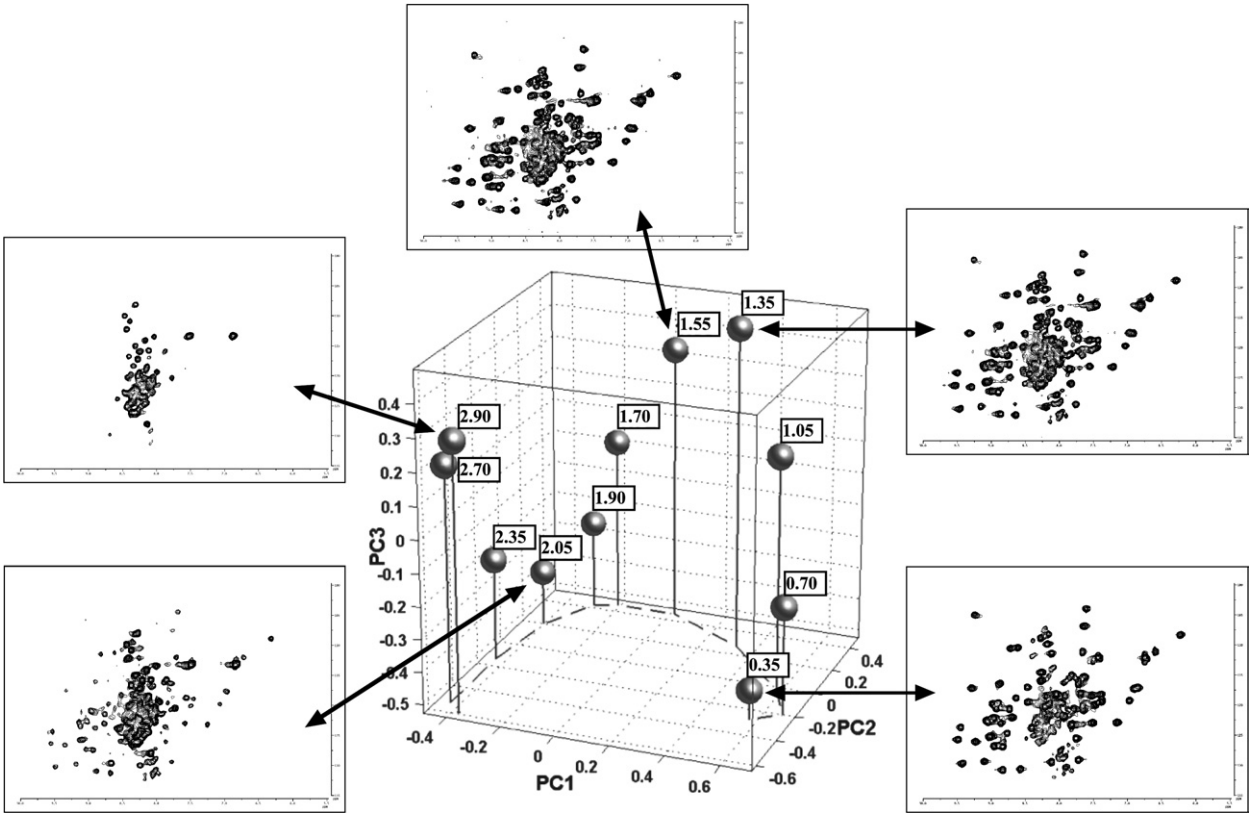

Fig. 2. The first three principal components representing 59.2% of the total variation in the spectra obtained in 11 2D NMR experiments are displayed. Each point in the plot is labelled with the GuHCl concentration used in the respective NMR experiment. ^{1}H–^{15}N HSQC-spectra are shown for the starting and ending points used in PLS-analysis.

ing cross-validated values for the models are: 74, 78, and 82%, respectively. This indicates that the models are true since they are able to predict samples that have been left out in the cross-validation scheme. Another important factor for validating the models is the fact that the Fourier transformed PLS weight vector actually resembles an NMR spectrum, which would not be the result in the case of a random model.

The PLS-derived 2D NMR spectra, hereafter referred to as PLS-spectra, reflect the changes occurring between the various recorded NMR experiments due to conformational changes induced by the denaturant, i.e., resonances that are invariant to the addition of denaturant do not show up. The visible resonances appear either as positive (blue), which intensities are positively correlated to the GuHCl concentration, and vice versa for the negative (red) resonances. Hence, such a PLS-spectrum offers a comprehensive overview of how the various resonances are affected by conformational changes that are correlated with the denaturant concentration. Moreover, only one PLS-spectrum needs to be analysed (i.e., integrated and evaluated for frequency changes etc.), rather than a whole range of individual 2D NMR spectra, as in a more traditional analysis, where the desired information would be quite tedious and difficult to extract. An alternative to our approach would be to use an ordinary difference spectrum between the starting and ending point of each identified denaturant concentration range. However, our approach with the PLS spectrum is superior in at least two aspects. First, all spectral changes that are not related to the denaturant concentration are filtered out. Second, since a PLS spectrum is calculated from all NMR experiments within a denaturant concentration range it is possible to distinguish between fast and slow NMR exchange processes (vide infra).

Comparing the three different PLS-spectra (Figs. 3A–C), the changes in the spectra for the pre-transition region (0.35–1.35 M GuHCl) and the post-transition region (2.05–2.90 M) are predominantly represented by frequency shifts (Figs. 3A and C) within the native and unfolded form, respectively. This is manifested by peaks having both negative and positive contributions. Such behaviour is expected for nuclei that are involved in conformational exchanges that are rapid on the NMR time-scale. The rapid exchange indicates that the populated conformations are easily accessible, and thus do not represent the major unfolding process. Therefore, the structural changes in these regions are mostly of local nature. Presumably, the bulk of these

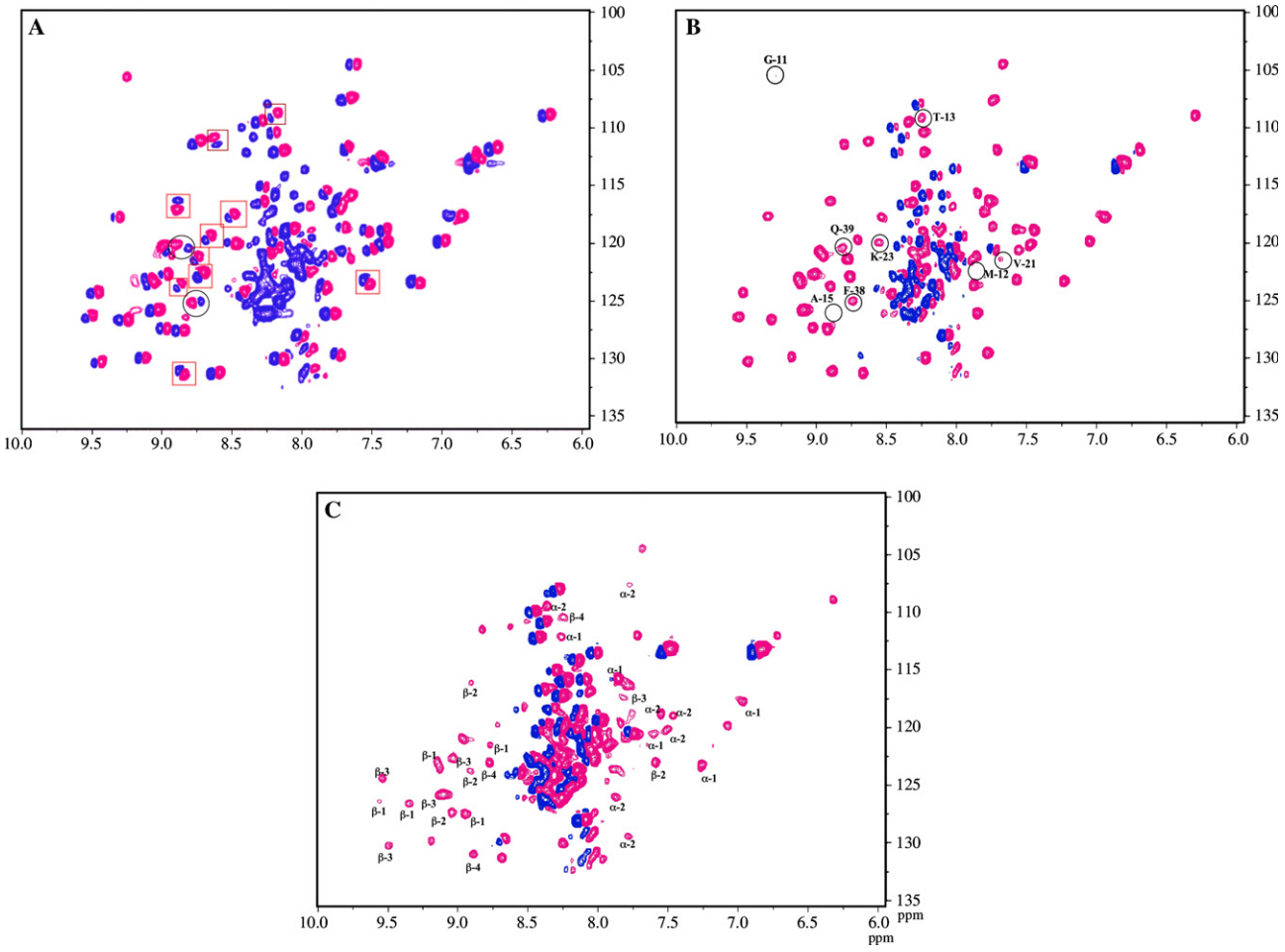

Fig. 3. 2D ^{1}H–^{15}N NMR spectra obtained by Fourier transformation of the first PLS weight vector for the three concentration ranges chosen. Positive peaks are coloured blue and negative peaks are coloured red. The peak volumes is a measure to what extent the GuHCl concentration affect the changes observed in the NMR spectra (A) 0.35–1.35 M GuHCl: resonances experiencing up-field proton chemical shifts upon addition of denaturant are indicated by circles, while those reflecting noticeable changes in ^{15}N frequencies are indicated by squares. (B) 1.55–2.05 M GuHCl: amino acids that display a substantial loss (integrals smaller than 50% of the most intense peaks) of their amide resonances in the native state are labelled (C) 2.05–2.90 M GuHCl: resonances still present at their native resonance frequencies are labelled with the secondary structure element, where they are located.

spectral changes are due to the already known affects of GuHCl on the resonance frequencies of unstructured peptides [14]. However, there are also a number of peaks that are either purely negative or positive in these spectra, suggesting that the resonances reflect processes that are involved in slow exchange on the NMR time-scale, due to a significant energy barrier between the conformational changes involved. These peaks represent inter-conversions between native and unfolded states of the protein and they are observable since a fraction of the unfolded state is present in the pre-transition PLS-spectrum and vice versa for the post-transition PLS-spectrum.

This slow exchange is the main feature of the PLS-spectrum based on the 1.6–2.1 M GuHCl region (Fig. 3B), in which there are large numbers of negative peaks representing the native state, while unfolded states of MerP appear as less well resolved positive peaks, mainly found in the centre of the spectrum. Hence, this range corresponds to the main transition from the native to the unfolded protein.

Closer inspections of the separate PLS-spectra reveal that, in addition to the general effect of changing the proton chemical shifts downfield by increasing GuHCl concentration, there are deviations from this behaviour. The PLS-spectrum for the pre-transition region, 0.35–1.35 M GuHCl (Fig. 3A), reveals that the proton chemical shifts of Phe-38 and Glu-39 decrease upon addition of denaturant since the positive contribution is located to the right of the negative part. Consequently, these two residues that are located in a loop connecting β-strand 2 and 3 (Fig. 1) adopt a non-native conformation prior to the main denaturation of the protein. The presence of such local conformational variations in this loop is not surprising since it has previously been reported that this region has different conformations in the oxi-

dized and reduced form of MerP [13,15], which implies that this loop can adopt different conformations of similar energy. There are also some resonances that display significant changes in their ^{15}N frequencies (indicated by squares in Figs. 1 and 3A). Interestingly, those amino acids for which the ^{15}N frequency most significantly decreases following addition of denaturant are located in β-strands 2 and 4 and hydrogen bonded to β-strands 3 and 1, respectively. However, most of the amino acids that show the opposite frequency shift are located in loops or at the edge of the β-sheet and not involved in hydrogen bonds. Hence, these ^{15}N frequency changes suggest that there are native-like conformations where the pattern of the hydrogen bonds between β-strands 1 and 4 and 2 and 3 have changed as compared to the native state of the protein. Interestingly, such frequency changes are not observed for the amide protons connecting β-strands 1 and 3, which appears to be part of the most stable region of the protein as shown by hydrogen exchange experiments [16].

The PLS-spectrum of the main transition (Fig. 3B) shows that all resonances of the native spectrum decrease in intensity during the transition. However, this reduction is substantially lower for eight amino acids (marked with names in Fig. 3B). Interestingly, six of these are found in loop regions located in close proximity to the disulfide bond. The remaining two are also located close to this region of the protein, while situated in the N-terminal part of the first α-helix. The reduction of the integral for these resonances in the PLS-spectrum indicate that this structural region of the protein has to a certain extent already adopted a non-native conformation during the pre-transition range. Further support for this assumption comes from the observation that these amino acids, along with Ala-16, are absent in the PLS-spectrum of the post-transition region (Fig. 3C).

In contrast, several amino acids can still be observed in the PLS-spectrum of the post-transition region (visualised in Fig. 1). A majority of these are located in the secondary structure (Fig. 3C). This indicates that these secondary structures are not completely ruptured even at the higher denaturant concentrations.

4. Conclusions

It has been shown that multivariate data analysis can be successfully applied to 2D NMR time-domain data. The PCA analysis showed how the NMR data varied as the denaturant was added to MerP and the score plot revealed the presence of three main principal spectral changes. Fourier transformations of the local PLS models provided NMR spectra in which merely the changes related to the addition of denaturant are present, thereby enabling substructures involved in local unfolding events to be identified. This information would hardly have been recognised without the availability of these PLS-spectra, which are far better suited for human interpretation than the analysis of peak picking tables extracted from the 11 recorded NMR spectra.

The methodology utilised in this work should be applicable to the analysis of any multidimensional NMR spectrum, where the objective is to obtain information about systematic variations in the data.

Acknowledgment

Dr J. Zdunek, Umeå University, is acknowledged for supplying C-programmes to convert NMR data from binary to ASCII format and vice versa.

References

[1] D.P. Meininger, M. Rance, M.A. Starovasnik, W.J. Fairbrother, N.J. Skelton, Characterization of the binding interface between the E-domain of staphylococcal protein A and an antibody Fv-fragment, Biochemistry 39 (2000) 26–36.
[2] A. Bergkvist, C. Johansson, T. Johansson, J. Rydström, B.G. Karlsson, Interactions of the NADP(H)-binding domain III of proton-translocating transhydrogenase from *Escherichia coli* with NADP(H) and the NAD(H)-binding domain I studied by NMR and site-directed mutagenesis, Biochemistry 39 (2000) 12595–12605.
[3] I.T. Joliffe, Principal Component Analysis, Springer, New York, 1986.
[4] J.E. Jackson, A Users Guide to Principal Components, John Wiley, New York, 1991.
[5] S. Wold, A. Ruhe, H. Wold, W.J. Dunn III, The collinearity problem in linear regression. the partial least squares approach to generalized inverses, SIAM J. Sci. Stat. Comput. 5 (1984) 735–743.
[6] J.B. Hauksson, U, Edlund, J. Trygg, NMR processing techniques based on multivariate data analysis and orthogonal signal correction. ^{13}C CP/MAS NMR spectroscopic characterization of softwood kraft pulp, Magn. Reson. Chem. 39 (2001) 267–275.
[7] L.E. Kay, P. Keifer, T. Saarinen, Pure absorption gradient enhanced heteronuclear single quantum correlation spectroscopy with improved sensitivity, J. Am. Chem. Soc. 114 (1992) 10663–10665.
[8] G. Balacco, SwaN-MR: from infancy to maturity, Mol. Biol. Today 1 (2000) 23–28.
[9] S. Wold, M. Josefson, Multivariate calibration of analytical chemistry, in: R.A. Meyers (Ed.), Encyclopedia of Analytical Chemistry, John Wiley, New York, 2000, pp. 9710–9736.
[10] D. Johansson, P. Lindgren, A. Berglund, A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription, Bioinformatics 19 (2003) 467–473.
[11] G. Aronsson, A.C. Brorsson, L. Sahlman, B.H. Jonsson, Remarkably slow folding of a small protein, FEBS Lett. 411 (1997) 359–364.
[12] L. Eriksson, B.E. Sandstrom, M. Sjostrom, M. Tysklind, S. Wold, Modeling the cytotoxicity of halogenated aliphatic-hydrocarbons—quantitative structure-activity-relationships for the Ic(50) to human Hela-cells, Quant. Struct. Act. Relat. 12 (1993) 124–131.
[13] H. Qian, L. Sahlman, P.O. Eriksson, C. Hambreus, U. Edlund, I. Sethson, NMR solution structure of the oxidized form of MerP, a

mercuric ion binding protein involved in bacterial mercuric ion resistance, Biochemistry 37 (1998) 9316–9322.

[14] K.W. Plaxco, C.J. Morton, B.S. Grimshaw, J.A. Jones, M. Pitkeathly, I.D. Campbell, C.M. Dobson, The effects of guanidine hydrochloride on the 'random coil' conformations and NMR chemical shifts of the peptide series GGXGG, J. Biomol. NMR 10 (1997) 221–230.

[15] R.A. Steele, S.J. Opella, Structures of the reduced and mercury-bound form of MerP, the periplasmic protein from the bacterial mercury detoxification system, Biochemistry 36 (1997) 6885–6895.

[16] A.C. Brorsson, A. Kjellson, G. Aronsson, I. Sethson, C. Hambraeus, B.H. Jonsson, The "Two-stateFolder" MerP forms partially unfolded structures that show temperature dependent hydrogen exchange, J. Mol. Biol. 340 (2004) 333–344.