Journal of Biomolecular NMR, **20:** 61–70, 2001.
KLUWER/ESCOM
© 2001 *Kluwer Academic Publishers. Printed in the Netherlands.*

Principal components analysis of protein structure ensembles calculated using NMR data

Peter W.A. Howe
Analytical Sciences, Syngenta, Jealott's Hill International Research Centre, Bracknell, Berkshire RG42 6EY, U.K.
E-mail: peter.howe@syngenta.com

Received 15 December 2000; Accepted 8 March 2001

Key words: essential dynamics, NMR structure calculation, principal components analysis, protein structure, structure refinement

Abstract

One important problem when calculating structures of biomolecules from NMR data is distinguishing converged structures from outlier structures. This paper describes how Principal Components Analysis (PCA) has the potential to classify calculated structures automatically, according to correlated structural variation across the population. PCA analysis has the additional advantage that it highlights regions of proteins which are varying across the population. To apply PCA, protein structures have to be reduced in complexity and this paper describes two different representations of protein structures which achieve this. The calculated structures of a 28 amino acid peptide are used to demonstrate the methods. The two different representations of protein structure are shown to give equivalent results, and correct results are obtained even though the ensemble of structures used as an example contains two different protein conformations. The PCA analysis also correctly identifies the structural differences between the two conformations.

Abbreviations: NOE, nuclear Overhauser effect; PC, principal component; PCA, principal components analysis; RMSD, root mean square deviation.

Introduction

Protein structure determination by NMR spectroscopy is an iterative process (Wüthrich, 1986; Neuhaus and Williamson, 2000). To determine the structure of a protein, a large number of restraints are identified from NMR spectra of the protein, and are used to calculate preliminary structures. From these, a converged subset of structures is selected and analysed to resolve ambiguous restraints and identify incorrect ones. Inspection of the structures on a graphics workstation may highlight regions of the structure which are poorly defined, so that efforts to obtain more constraints can be concentrated in those regions. The improved restraint list is then used for another calculation, and the process is repeated a number of times until resolution cannot be improved any further or until the desired resolution is reached. The NMR spectroscopist carrying out the structure determination then selects some of the final structures to represent the protein structure. This selection is usually made by inspecting the RMSDs and calculated energies of the structures (calculated energy reflects the agreement of the protein structure with ideal covalent constraints such as bond lengths and bond angles, and with experimental constraints such as dihedral angles and distances derived from NOE peaks).

This process would be relatively straightforward if all the structures resulting from the structure calculation satisfied all the experimental and ideal constraints. Unfortunately, they do not. The currently available protein structure calculation methods produce some structures where most of the constraints are satisfied, some where a large number of constraints are violated and some in between. The exact cause of this has not been rigourously investigated, but it is thought to be because some structures become trapped in false energy minima during calculations so cannot reach a conformation where constraints are all satisfied before the calculation finishes. This suggested cause has led to the use of the terms 'converged' structures and 'outlier' structures. Converged structures are those where the structure calculation has completed and produced a structure where constraints are satisfied, and outlier structures are those where the calculation has become trapped in a local energy minimum and has not resulted in a structure where the constraints are satisfied. The problem for the NMR spectroscopist is to distinguish the two types of structure. This is important both during the calculation, where initial structures are used to resolve ambiguous constraints, and in the final reporting of the structure where the selection of structures will affect the reported RMSD. Distinguishing converged and outlier structures is a challenging problem. The population of calculated structures may contain a high proportion of outlier structures and may contain sub-groups (for example, structures with different conformations of a loop). Visualising proteins is difficult because they have many degrees of conformational freedom, so their conformation can only be characterised by a large number of variables. Also, these variables are likely to be highly correlated because changes in protein conformation tend to be regional movements. These four characteristics of structure ensembles make them very difficult to analyse.

One method suggested for identifying converged structures is the use of the energy ordered RMSD profile (Widmer et al., 1993), but this has not become widely used. This approach has the disadvantage that it requires manual inspection of the results. An automatic method would save time, and would help avoid subjective judgements. This paper describes how Principal Components Analysis (PCA) has the potential to identify converged structures automatically. As will be described below, PCA has been used to relate the variability in protein structures to biological function, but it is also suitable for classifying protein structure ensembles.

Principal components analysis

PCA was first introduced to statistics about 70 years ago, but only recently has it become widely used through the availability of desktop computers and of computational procedures for handling large matrices (Manly, 1986; Eriksson et al., 1999). PCA has two important characteristics:

1. It can be applied to data sets where the number of variables exceeds the number of samples.
2. It produces valid results with data sets containing highly correlated variables.

In fact, PCA exploits these two characteristics. If some of the variables in the data set are highly correlated, then it will be possible to replace them by a smaller number of latent variables, referred to as Principal Components (PCs). This data model can be represented as:

$$\begin{pmatrix} x_{11} \ldots x_{1k} \\ x_{21} \ldots x_{2k} \\ \ldots \\ x_{n1} \ldots x_{nk} \end{pmatrix} =$$

$$\begin{pmatrix} t_1 \\ t_2 \\ \ldots \\ t_k \end{pmatrix} \times (p_1 \ldots p_n) + \begin{pmatrix} \epsilon_{11} \ldots \epsilon_{1k} \\ \epsilon_{21} \ldots \epsilon_{2k} \\ \ldots \\ \epsilon_{n1} \ldots \epsilon_{nk} \end{pmatrix}$$

The original data consists of n samples, each of which is represented by k variables. The observed variables (x_{nk}) can be represented as the product of the PC scores (p_n) and the PC loadings (t_k) plus an error term. The PC loadings are constant for all the samples in the population, but are selected to minimise the sum of the error terms across all samples and all variables ($\Sigma\epsilon_{nk}$). If the variables present in the population are highly correlated, then the PC model can replace a large number of variables (x_{nk}) by one variable per sample (p_k) with only a small disagreement between the observations (x_{nk}) and those calculated from the PC score ($p \times t_k$).

Only one PC is shown in this example. However, once the initial PC has been calculated, it can be subtracted from the data and another PC can be calculated on the matrix of error terms. This process can then be repeated until the error term reaches zero, which is when the number of PCs equals the number of samples in the population.

Three important features of PCs are that:

1. They allow data reduction: a large amount of variance can often be explained by a small number of PCs.
2. The first PC explains the highest proportion of variance, and subsequent PCs explain decreasing proportions of the variance.
3. All PCs are uncorrelated with one another (i.e. they are orthogonal).

An additional feature is that the PC loadings (t_k) are interpretable. Correlated variables which vary across the population will have high PC loadings, while uncorrelated variables have low loadings.

An important application of PCA is the identification of outliers within a multivariate population (Egan and Morgan, 1998; Eriksson et al., 1999). This can be done using cross-validated PCA, where the PCA calculation is repeated several times with a proportion of the data left out. The samples that were left out are then fitted to the model, and the process is repeated until all samples have been left out at least once. This process produces confidence limits for the PC scores and the error terms which can be used to test whether an individual sample is a member of that population or not. A detailed discussion of the different methods for outlier detection is outside the scope of this paper.

An important consideration in outlier detection with NMR structure ensembles is that a large proportion of the structures could be outliers, and the structures may be sub-grouped. With such data, many methods of outlier detection fail both by not identifying outliers and by incorrectly identifying population members as outliers. These failures are due to distortions in the centre of the population, so the methods which are least susceptible to a high proportion of outliers are based around selecting a subset of data and using the subset to test the remainder of the population.

Essential dynamics

The first suggestion that PCA is a suitable method for analysing ensembles of protein structures was by Amadei et al. (1993), who applied it to the results of molecular dynamics and named the analysis 'Essential Dynamics'. Molecular dynamics calculations produce large numbers of structures, which may vary considerably in conformation, so analysing them is a challenging problem. Amadei et al. realised that applying PCA to the ensemble of structures would highlight the regions of the protein which were undergoing the correlated movements that they thought would be the most significant for biological function. These regions could be identified from the PC loadings rather than by inspecting the structures manually. Since the initial suggestion, Essential Dynamics has been applied to molecular dynamics results from a range of different proteins (for example, Mello et al., 1998; Chau et al., 1999), has been used to compare the results of molecular dynamics with structures determined by NMR (Abseher et al., 1998), and has also been used to identify biologically important variability in a converged ensemble of NMR structures (van Aalten et al., 1998; O'Donoghue et al., 2000). However, this paper appears to be the first to use PCA to categorise an ensemble of protein structures into converged ones and outliers.

Representing protein structures

To apply PCA to protein structures, a format for representing the structures as variables has to be developed. Although it would be possible to use the co-ordinates of every atom in the protein, this would produce a huge data matrix which would take very long times to analyse and the variation would be dominated by the movements of surface side-chains. To avoid these problems, this paper only considers two different representations of the protein backbone and ignores protein side chains. The first representation is very similar to that used in Essential Dynamics and it will be referred to as 'deviation in Cα position'. To calculate this representation, each structure is first superimposed onto the lowest energy structure. Then for every structure, the deviation in the x-, y- and z- co-ordinates of each Cα atom from its position in the lowest energy structure is calculated. Each structure is then represented by the x-, y- and z-deviations of the Cα atoms of all its residues. The main difference between Essential Dynamics and this approach is that Essential Dynamics uses the mean structure rather than the lowest energy one. The lowest energy structure is preferable because the mean structure will be distorted by outliers and may well be a poor structure in terms of covalent contacts, bond lengths and such. This last problem is commonly recognised in NMR structure determinations, where ensembles are usually represented by the structure with the lowest energy, or the one closest to the mean, rather than the mean structure itself.

As well as this approach, this paper also uses the Cα-Cα distance matrix to represent protein structures (Phillips, 1970). The Cα-Cα distance matrix contains each and every Cα-Cα distance within the structure. Cα-Cα distance matrices have the advantage that they can be calculated without superposition (which is computationally expensive) and are simple to interpret, but they do contain a large number of variables; this problem is considered in the discussion. There are several examples of analysis of NMR structure ensembles using Cα-Cα distance matrices to identify

Table 1. Structural statistics for the 37 lowest energy structures as calculated by the program X-PLOR. The RMSD values are the average RMSD to the average structure

Backbone RMSD (Å)	0.68 ± 0.06
All heavy-atom RMSD (Å)	1.63 ± 019
Bond length RMSD (Å)	0.0030 ± 0.0001
Bond angles RMSD (°)	0.60 ± 0.01
Improper angle RMSD (°)	0.32 ± 0.02
NOE restraint RMSD (Å)	0.044 ± 0.002

structured subdomains within proteins (Schwabe et al., 1993; Kundrot, 1996).

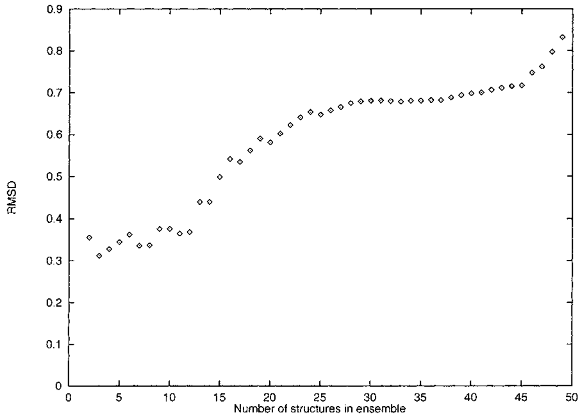

Figure 1. The energy-ordered RMSD profile of the 49 structures showing the RMSD of an ensemble plotted against the number of structures within it. Structures were added to the ensemble in order of their total calculated X-PLOR energy.

Materials and methods

All data were calculated using EF40, a 28 amino acid disulphide-linked protein. The structure determination of this protein will be reported in full elsewhere but the structural statistics of the 37 lowest-energy structures are summarised in Table 1. Forty-nine structures were calculated from 25 different random starting structures by simulated annealing within the program X-PLOR 3.1 (Brünger, 1992) running on an IBM AIX workstation. X-PLOR was also used to calculate the RMSD profiles of the structure ensemble, the deviations in Cα position from the mean structure and the Cα-Cα distance matrices. Principal Components Analysis was done using the program Pirouette v 2.03 (Infometrix Inc., Seattle) running on a 400 MHz Pentium PC computer. Prior to PCA, the mean and the standard deviation for each variable across all structures were calculated. The mean was subtracted from each variable, and the residual was divided by the standard deviation. This is usually referred to as 'mean centring/variance scaling' data and it ensures that larger and more variable values do not dominate during analysis. Ten PCs were calculated, leaving out 10–20% of the data for cross-validation. This calculation took less than 2 min. For further analysis, sufficient PCs to explain 75% of the variation in the population were retained.

Outlier detection

As pointed out above, NMR structure ensembles may be clustered and may contain a very high proportion of outliers. The Smallest Half Volume method (SHV) is a robust way of identifying outliers in populations which have these characteristics (Egan and Morgan, 1998). In the SHV method, a proportion of the population is taken as the core. The remainder of the population is then tested against this core using a χ^2-test of the Mahalanobis distance. This use of the Mahalanobis distance means that the SHV method has much in common with clustering methods which use the Mahalanobis distance to group structures.

The SHV method will work in any population where the core population is correctly identified. The original implementation of SHV defined the core population as the 50% of the population clustered most tightly. To apply the method to NMR structure ensembles, two modifications were necessary. First, only 20% of structures were used to define the core; this reduces the dependence of the results on the proportion of outliers. Second, the 20% of the population selected as the core were selected as the 20% with lowest total energy. Selecting the structures based on energy ensures that the core population is made up of the structures that agree best with experimental and theoretical data. This alteration to the method is necessary because NMR structure ensembles could contain clusters of high energy structures, particularly in the initial stages of automated structure calculation. In such cases, the most tightly clustered structures could be ones that agree poorly with experimental and theoretical data. This makes it unsatisfactory to select the core population on tight clustering alone. Once the core had been defined, other structures were tested against this set using the supplemented Mahalanobis distance as implemented in the program Pirouette.

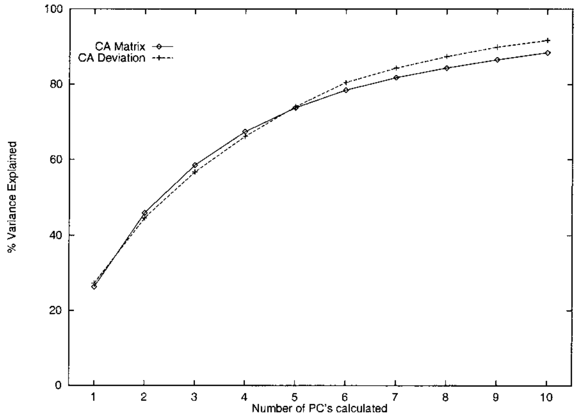

Figure 2. The cumulative variance explained by Principal Components. The percentage of variance within the ensemble explained by a PC model is plotted against the number of PCs in that model.

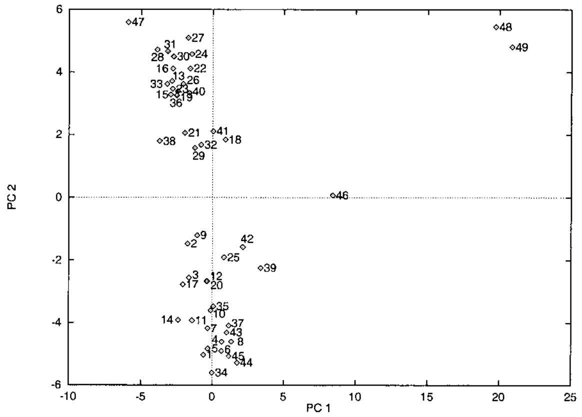

Figure 3. The PC scores of the 49 structures in the ensemble, calculated from the deviation in Cα position. Structures are numbered in order of increasing energy and are plotted at the intersection of their PC 1 score (horizontal axis) and their PC 2 score (vertical axis). PC 1 and PC 2 together represent 47% of the variation in the ensemble.

Results

Quality of the structures

Figure 1 shows the energy-ordered RMSD profile of the 49 structures. For the 12 lowest-energy structures, the backbone RMSD is 0.38 Å while for the 37 lowest-energy structures, the RMSD of backbone atoms is 0.68 Å, showing that the structure is well defined. Table 1 shows the structural statistics for the 37 lowest-energy structures.

Principal Component Analysis: All 49 structures

To demonstrate the effectiveness of PCA in representing protein structures, 10 PCs were calculated for all 49 structures using both different structure representations. The graph in Figure 2 shows the cumulative variance explained by these PCs. The first PC (PC1) accounts for 28% of the variability across the 49 structures. In other words, 28% of the variation across the 49 structures can be explained by one latent variable which is a product of all the original variables. Six PCs together explain over 75% of the variability within the 49 structures. PC scores explain approximately the same proportion of variance, regardless of which of the two different structure representations is used.

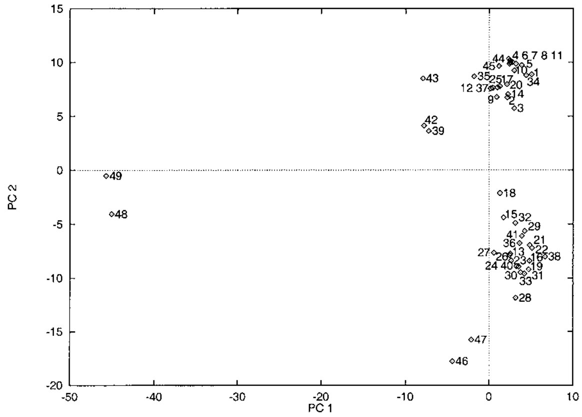

Figure 4. The PC scores of the 49 structures in the ensemble calculated from the Cα-Cα distance matrix.

Figures 3 and 4 plot the first two PC scores of all the structures in the ensemble as calculated using the two different structure representations. As these graphs plot the first two PCs, they visualise 47% of the total variation in the population of 49 structures and so provide a simple overview of the composition of the population. The graphs also show that the two data representations give very similar results. The axes of the graphs are inverted relative to one another, so almost all the structures are in similar positions relative to the other structures.

Most of the structures are divided into two distinct clusters, suggesting that the population contains two different conformations. Although this was not immediately apparent from the RMSD profile (Figure 1), it can be verified by calculating the RMSD of the two clusters separately and together. The RMSD of the 17 structures in the low-energy cluster is 0.381 Å; the RMSD of the 19 structures in the higher-energy cluster is 0.413 Å; and the RMSD of both clusters together is 0.68 Å. This clearly demonstrates that the two clusters represent different conformations of the protein.

As well as the two main clusters, there are also some structures which do not group with either cluster (structure numbers 46, 48 and 49). These are among the highest energy structures in the ensemble. In the different data representations, their positions vary relative to the main clusters, showing that the two structure representations are not entirely equivalent.

Interpreting principal components in structural terms
The presence of two different conformations in the population raises the question, what is the difference between them? This can be answered from the loadings that make up the PC that separates the clusters (Manly, 1986; Eriksson et al., 1999). If a variable has a loading near zero, it does not contribute to the PC, but if it has a magnitude near 1 then it contributes considerably to the PC. The sign of variable loading is also important. Variables with loadings of the same sign are positively correlated (when one increases, so does the other) while variables with opposite sign loadings are negatively correlated (one decreases when the other increases). Examination of both the PC plots shows that the two clusters have very similar PC 1 values, and are separated by PC 2. Therefore, examining the loadings of PC 2 will reveal which variables contribute to PC 2, and so identify the variables which result in differences between the clusters.

Figure 5 plots the loadings of PC 2 for the deviation in Cα position versus residue number. Some residues have low loadings, so do not vary greatly between the two conformations (e.g. residues 11–16), but others have much higher loadings, so are different between the two clusters. The region with the highest loadings is the loop involving residues 21–25, but there is also some variation involving the regions around residue 4, residue 9, residue 19 and at the N-terminus. The directions of the loadings give information about the direction of the movements. For example, the x- and z- co-ordinates of residue 24 have large negative loadings, implying that residue 24 has lower x- and z- co-ordinates in one conformation than in the other. In contrast, residue 21 has high x-, y- and z- loadings, so this moves the opposite way (these movements are relative to the co-ordinate system of the lowest energy structure).

The loadings of PC 2 calculated using the Cα-Cα distance matrix are plotted as a contour plot in Figure 6. Each variable in the matrix represents the distance between two Cα atoms and the three axes of the contour plot are therefore the originating Cα atom, the destination Cα atom and the loading, with the loading as the z-axis or height. The area above and right of the diagonal has been left blank because the Cα-Cα distance matrix is entirely symmetrical. The upper left portion of the graph contains no regions of high loadings, implying they do not contribute to PC 2 so do not vary consistently between the two clusters of structures. Most of the high loadings of PC 2 occur on distances involving residues 20–25. For example, high negative loadings of PC 2 are observed for the distances between residue 21 and residues 8 to 11. In contrast, the loadings of the distances between residue 21 and residues 2 to 6 are high but positive. This implies that high values of PC 2 are associated with a decrease in the distances between residue 21 and residues 8 to 11 and an increase in the distances between residue 21 and residues 2 to 6. Figure 4 shows that structures with higher values of PC 2 have lower energies, suggesting that the conformation with shorter distances between residue 21 and residues 8 to 11 is the more favourable. Other consistent distance variations can also be identified, such as the distances between residue 23 and residues 25 to 28, which also appear to be shorter in the low energy structures.

This brief analysis demonstrates how to analyse the Cα-Cα distance matrix to get structural information without having to examine the structures using interactive molecular graphics. Simple visual inspection of the PC contour plot (Figure 6) provides considerable clues about the differences between clusters of structures. The results also explain the observations made from analysing the deviation in Cα position (Figure 5). The residues with the highest deviation in Cα position are residues 20–25, while the other regions with high deviations move relative to residues 20–25 (residues 1–5, 9–10 and 17–19).

The differences identified using the PC scores can be checked by inspecting the structures. Figure 7 shows the superposition of the lowest energy structures from each of the two clusters (structures 1 and 13). It is clear from this figure that most of the protein backbone can be closely superimposed, but that the loop between residues 21 and 25 is in very different conformations in the two structures. This is exactly as predicted from the analysis of PCA loadings. Closer investigation shows that the observed structural differences are consistent with the PCA loadings. For example, the distance between the Cα atoms of residues 8 and 21 is shorter in the low energy structure than in the higher one (7.9 Å compared with 11.0 Å), as is the distance between the Cα atoms of residues 23 and 28 (13.6 Å compared with 15.3 Å). This demonstrates that the structural information provided by PCA is equivalent to that obtained by manually analysing the structures.

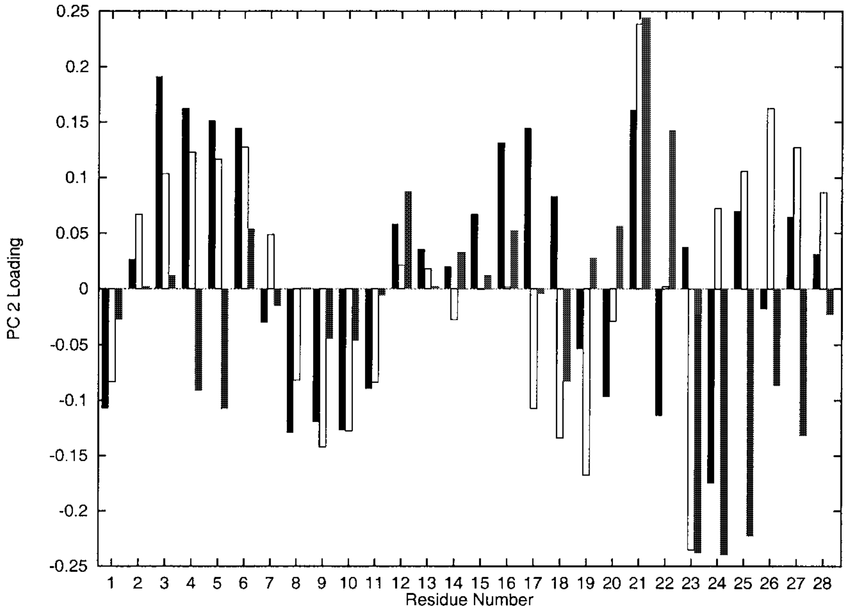

Figure 5. The loadings of PC 2 from the PCA model calculated for the 49 structures in the ensemble using the deviation in Cα position. The loading of a variable is plotted against the residue number of the Cα atom the variable refers to. The three values for each residue refer to the x- (black), y- (white) and z- (grey) deviations of the Cα atom of that residue.

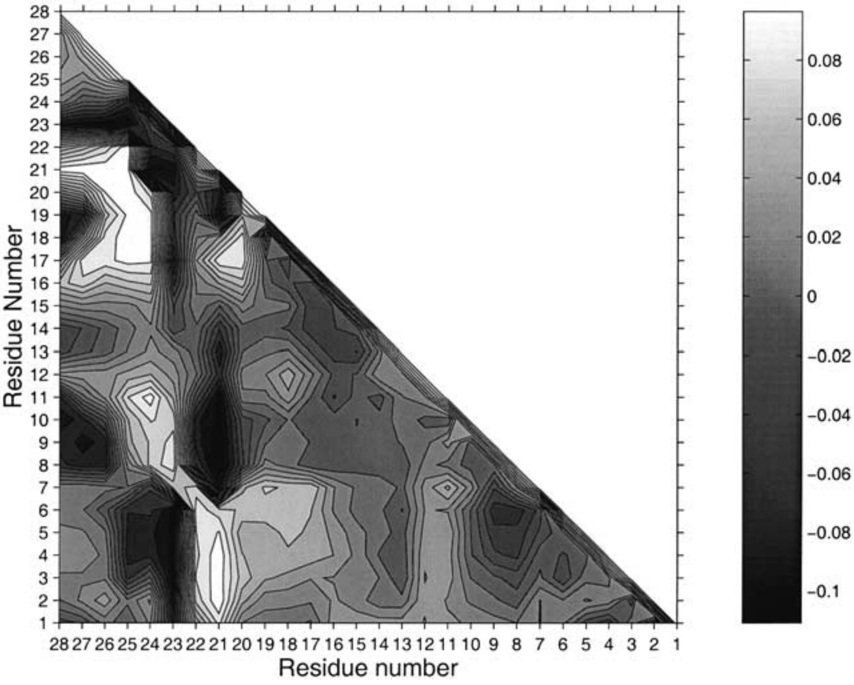

Figure 6. The loadings of PC 2 for the PCA model calculated for the 49 structures in the ensemble using the Cα-Cα distance matrix. The horizontal and vertical axes are the residue numbers that the Cα-Cα distance is between, and the z-axis (height) is the loading of that distance.

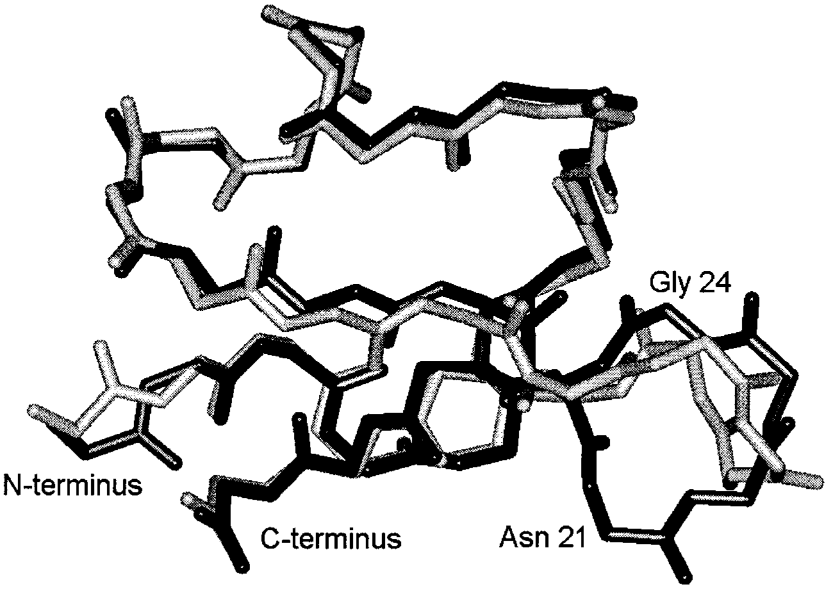

Figure 7. Superposition of the lowest-energy structures from the two clusters that PCA identifed. Structure 1 (grey) is from the low-energy cluster and structure 13 (black) is from the high-energy cluster.

Detecting outliers

The 10 structures with lowest energy were used to generate a cross-validated PCA model. The results of testing the entire population against this model are shown in Figures 8 and 9. The two structure representations consistently identify the same structures as being members of the core population. These are structures 1–12, 14, 17, 20, 25, and 34 and one structure (37) which is identifed as being a member of the core population by one representation and an outlier by the other. This apparent inconsistency is explained by examining Figures 8 and 9, which show that structure 37 is close to the 95% confidence limit in the two data representations.

Comparison with the results of PCA on the whole population shows that all the converged structures are from one of the two conformational clusters that are present (the lower cluster in Figure 3). This suggests that structures with the second conformation should not be selected to represent the structure, but it would be more thorough to acquire further NMR data to attempt to confirm that the protein does adopt the suggested loop conformation.

Converged structures can also be identified from the RMSD profile (Figure 1). Structures 1–12 are clearly a well-defined core population, but the population RMSD begins to increase when structure 13 is added. However, the increase is not continuous; when structures 14, 17, 20 and 25 are added the RMSD decreases or levels off. This can be explained from the PCA results, which show that these structures are from the first protein conformation, while the remaining structures are from the second alternative conformation. Thus, the RMSD increases when a structure from the second conformation is added, but decreases when one from the first is. This difference was very difficult to identify from the RMSD profile, but is readily identified by PCA – either visually using the PCA plot, or automatically using outlier detection.

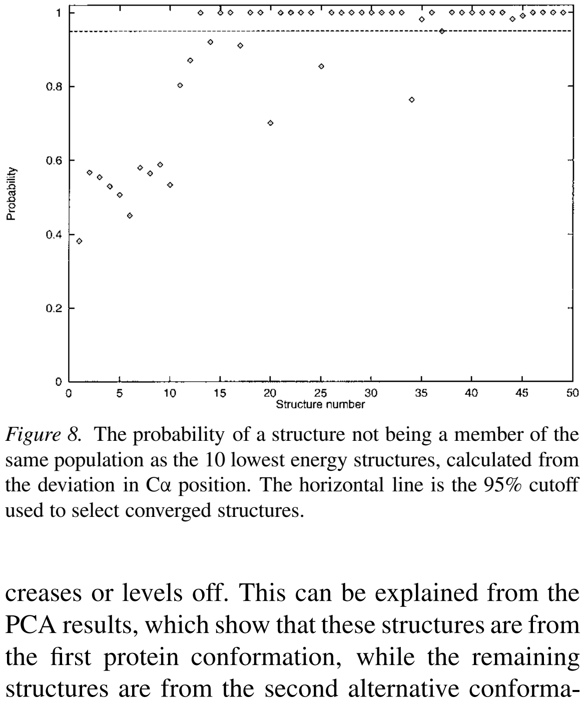

Figure 8. The probability of a structure not being a member of the same population as the 10 lowest energy structures, calculated from the deviation in Cα position. The horizontal line is the 95% cutoff used to select converged structures.

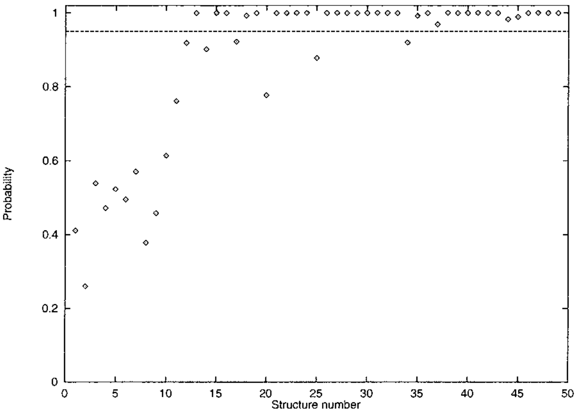

Figure 9. The probability of a structure not being a member of the same population as the 10 lowest energy structures, calculated from the Cα-Cα distance matrix.

Discussion

It must be stressed that the methods described here are not aimed at validating NMR structures. Validation can only be carried out by comparing the results of a structure calculation with data which was not used during the calculation – for example, by cross-validation of the constraints themselves (Brünger et al., 1993), by comparison with NMR observables not used in the calculation, or by comparison of the observed NOE cross peaks with those expected from the calculated structure (Doreleijers, 1999). Although the core population is selected based on total energy, once that has been done then other structures are classified purely by whether or not they are within the structural variability of the core population. This characteristic distinguishes the approach from methods based solely on analysis of total energy, RMSD or other means such as Ramachandran angles.

It could be argued that structural variability alone should be used as the criterion for identifying converged structures. However, as pointed out in the Methods section, this has the risk of selecting structures which are very similar to one another but which agree poorly with the experimental data. This problem was observed with the structures used here; some of the higher energy structures were more tightly clustered than the low energy ones, so selection on structural features alone would have been misleading (results not shown).

The results here show that PCA can be usefully applied to ensembles of protein structures. One particular success in this example was that PCA identified clustering of the structures into two different conformations. The outlier detection method coped successfully with this clustered data, while the RMSD calculations were difficult to interpret. An additional advantage of PCA is that the results can be analysed to explain the difference in structure between clusters. In this case, it is a difference in the conformation of the loop involving residues 20 to 25. This paper has only shown the application of PCA to one protein, but previous applications to structure ensembles suggest that the approach will be generally applicable. For example, in the ensemble of 40 insulin monomers, a PCA model containing five PCs explained 75% of the variance in the data set, showing that the protein structure ensemble contains considerable correlated variation (O'Donoghue et al., 2000).

The main disadvantage of using cross-validated PCA is that specialist software programs are required. Several dedicated programs for multivariate statistical analysis are available (Pirouette, used here, SIMCA-P and Unscrambler), but it can also be implemented in most specialist statistical programs (such as SAS and SPSS) and in some mathematical ones (MATLAB) which are more commonly available.

Two different representations of protein structures were used here, and both of them gave very similar results, showing that the assumptions made are valid. The two different represenions have different characteristics. The Cα-Cα distance matrix does not require superposition, or an average structure, and it is very easy to interpret without the use of molecular graphics. Its big disadvantage is the large number of variables in the matrix. For a protein of length n it is $(n^2 - n)/2$, so that, for example, the matrix of a 50 amino acid protein would contain 1250 variables. PCA can still be applied to such large data sets, but calculations take longer and the results are harder to interpret. Work is now underway to investigate how the size of the matrix could be reduced for larger proteins without loss of information, and how Cα-Cα distance matrices could be effectively compressed before PCA calculations. The alternative representation of protein structure (deviation in Cα position) is much more compact (a data size of $3n$ for an n-residue protein). However, calculating it requires superposition to an average structure and the results can only be interpreted with reference to the average structure, which will normally require interactive molecular graphics. Despite these disadvantages, it appears to give very similar results to the Cα-Cα distance matrix, so it is the method of choice for larger proteins until more efficient methods of analysing Cα-Cα distance matrices are developed.

Finally, possible broader applications of these methods should be mentioned. Although this paper has shown the application of PCA to one ensemble of protein structures, the methods could also be used to compare protein structures from different sources. For example, the crystal structure of a protein could be tested against a PCA model derived from an NMR structure ensemble to see if it falls within the variation of the population of NMR structures or is significantly different.

Conclusions

The method suggested here for analysing NMR structure ensembles has two main advantages. First, the use of cross-validated PCA provides a relatively non-subjective method of selecting which structures in an ensemble are atypical so can be discarded. PC calculation can be fully automated, making it an especially useful method for combined structure calculation/assignment protocols. The second advantage is that the results are easy to interpret in structural terms. The calculated PCs identify which regions of the protein structure are undergoing correlated movement, so these areas can be targeted to obtain extra restraints.

Acknowledgements

The protein EF40 was provided by Mr David Acland and Dr Jane Cayley of Syngenta. I would like to thank Dr Alex Breeze and Ms Sandra Donoghue of AstraZeneca Pharmaceuticals for access to the copy of X-PLOR used to run these calculations. I would also like to thank Professor Jens Jørgen Led and Dr Chang Xiaoqing at the University of Copenhagen for introducing me to the use of Essential Dynamics for analysing NMR structure ensembles and I am especially grateful to Dr Stephen Hole of Syngenta for many useful discussions concerning multivariate statistics and assistance in preparing figures.

References

Abseher, R., Horstink, L., Hilbers, C.W. and Nilges, M. (1998) *Proteins Struct. Funct. Genet.*, **31**, 370–382.

Amadei, A., Linssen, A.B.M. and Berendsen, H.J.C. (1993) *Proteins Struct. Funct. Genet.*, **17**, 412–425.

Brünger, A. (1992) *X-PLOR Version 3.1. A System for X-ray Crystallography and NMR,* Yale University Press, Boston, MA.

Brünger, A., Clore, G.M., Gronenborn, A.M., Saffrich, R. and Nilges, M. (1993) *Science*, **261**, 328–331.

Chau, P.-L., van Aalten, D.M.F., Bywater, R.P. and Findlay, J.B.C. (1999) *J. Comput.-Aided Mol. Design*, **13**, 11–20.

Doreleijers, J.F., Raves, M.L., Rullman, T. and Kaptein, R. (1999) *J. Biomol. NMR*, **14**, 123–132.

Egan, W.J. and Morgan, S.L. (1998) *Anal. Chem.*, **70**, 2372–2379.

Eriksson, L., Johansson, E., Kettaneh-Wold, N. and Wold, S. (1999) *Introduction to Multi- and Megavariate Data Analysis using Projection Methods*, UMETRICS AB, Umeå, Sweden.

Kundrot, C.E. (1996) *J. Am. Chem. Soc.*, **118**, 8725–8726.

Manly, B. (1986) *Multivariate Statistics – A Primer*, Chapman & Hall.

Mello, V.C., van Aalten, D.M.F. and Findlay, J.B.C. (1998) *Biochemistry*, **37**, 3137–3142.

Neuhaus, D. and Williamson, M.P. (2000) *The Nuclear Overhauser Effect in Stereochemical and Conformational Analysis*, 2nd ed., Wiley, New York, NY.

O'Donoghue, S.I., Chang, X., Abseher, R., Nilges, M. and Led, J.J. (2000) *J. Biomol. NMR*, **16**, 93–108.

Phillips, D.C. (1970) *Biochem. Soc. Symp.*, **30**, 11–28.

Schwabe, J.W.R., Chapman, L., Finch, J.T., Rhodes, D. and Neuhaus, D. (1993) *Structure*, **1**, 187–204.

van Aalten, D.M.F., Grotewold, E. and Joshua-Tor, L. (1998) *Methods*, **14**, 318–328.

Widmer, H., Widmer, A. and Braun, W. (1993) *J. Biomol. NMR*, **3**, 307–324.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids,* Wiley, New York, NY.