Protein Engineering vol.9 no.11 pp.1063–1065, 1996

PROTOCOL

An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies

Lawrence A.Kelley[1], Stephen P.Gardner[2] and Michael J.Sutcliffe[1,3]

[1]Department of Chemistry, University of Leicester, Leicester LE1 7RH and [2]Oxford Molecular Ltd, The Medawar Centre, Oxford OX4 4GA, UK

[3]To whom correspondence should be addressed

Keywords: automated clustering/cluster analysis/multiple conformations/NMR spectroscopy/protein structure

Introduction

Unlike structures determined by X-ray crystallography, which are deposited in the Brookhaven Protein Data Bank (Abola *et al.*, 1987) as a single structure, each NMR-derived structure is often deposited as an ensemble containing many structures, each consistent with the restraint set used. However, there is often a need to select a single 'representative' structure, or a 'representative' subset of structures, from such an ensemble. This is useful, for example, in the case of homology modelling or when compiling a relational database of protein structures. It has been shown that cluster analysis, based on overall fold, followed by selection of the structure closest to the centroid of the largest cluster, is likely to identify a structure more representative of the ensemble than the commonly used minimized average structure (Sutcliffe, 1993).

Two approaches to the problem of clustering ensembles of NMR-derived structures have been described. One of these (Adzhubei *et al.*, 1995) performs the pairwise superposition of all structures using C_α atoms to generate a set of r.m.s. distances. After cluster analysis based on these distances, a user-defined cut-off is required to determine the final membership of clusters and therefore the representative structures. The other approach (Diamond, 1995) uses collective superpositions and rigid-body transformations. Again, the position at which to draw a cut-off based on the particular clustering pattern was not addressed.

Whenever fixed values are used for the cut-off in clustering, there is a danger of missing 'true' clusters under the threshold imposed by the rigid cut-off value. Considering the highly diverse nature of NMR-derived ensembles of proteins, it would seem most appropriate to avoid the use of predefined values for determining clusters. In fact, of the 302 ensembles we have studied, the average pairwise r.m.s. distance across an ensemble varied from 0.29 to 11.3 Å (mean value 3.0, SD 1.9 Å). Here we present an automated method for cut-off determination that avoids the dangers of using fixed values for this purpose.

We have developed a computer program that automatically, systematically and rapidly (i) clusters an ensemble of structures into a set of conformationally related subfamilies, and (ii) selects a representative structure from each cluster. The program uses the method of average linkage to define how clusters are built up, followed by the application of a penalty function that seeks to minimize simultaneously the number of clusters and the spread across each cluster. This program, known as NMRCLUST, is available via the World Wide Web (URL: http://neon.chem.le.ac.uk/) and by anonymous ftp from ftp.oxmol.co.uk. Although developed for the analysis of NMR-derived structures, the program can be used to automatically cluster any data set.

Materials and methods

An overview of the method is given in Figure 1.

Step 1. Distance determination

Clustering requires a set of 'distances' between members of an ensemble. When a PDB file containing an ensemble of structures is used as input, NMRCLUST derives these distances by superposing each member of the ensemble onto each of the other members of the ensemble in a pairwise manner (McLachlan, 1982); the corresponding r.m.s. value is determined. This superposition is carried out, by default, on all non-hydrogen atoms or, alternatively, on a user-defined set of atoms (see Materials and methods). For an ensemble with N members, this results in an $N \times N$ matrix of r.m.s. values. NMRCLUST can also accept a predetermined matrix of 'distances' as input. This is useful in cases where objects other than protein structures are to be clustered.

Step 2. Clustering

This distance matrix is used with the average linkage algorithm for hierarchical cluster analysis. The method of average linkage takes the distance between two clusters m and n to be:

$$\mathrm{dist}(m,n) = \frac{\left(\sum_{i=1}^{X} \sum_{j=1}^{Y} \mathrm{dist}(i,j) \right)}{XY}$$

where cluster m contains X members, and cluster n contains Y members; dist(i,j) is the r.m.s. distance between the two members, i and j, of clusters m and n, respectively, after superposition. At each stage of the clustering algorithm, a search is performed for the two nearest clusters; these are merged to form a single cluster.

At each stage of clustering, the 'spread' of each cluster is calculated. The spread of a cluster m containing N members is given by:

$$\mathrm{spread}_m = \frac{\left(\sum_{k=1}^{N} \sum_{i=1, i<k}^{N} \mathrm{dist}(i,k) \right)}{N(N-1)/2}$$

where i and k are members of cluster m. The average spread is then given by:

$$\mathrm{AvSp}_i = \frac{\sum_{m=1}^{\mathrm{cnum}_i} \mathrm{spread}_m}{\mathrm{cnum}_i}$$

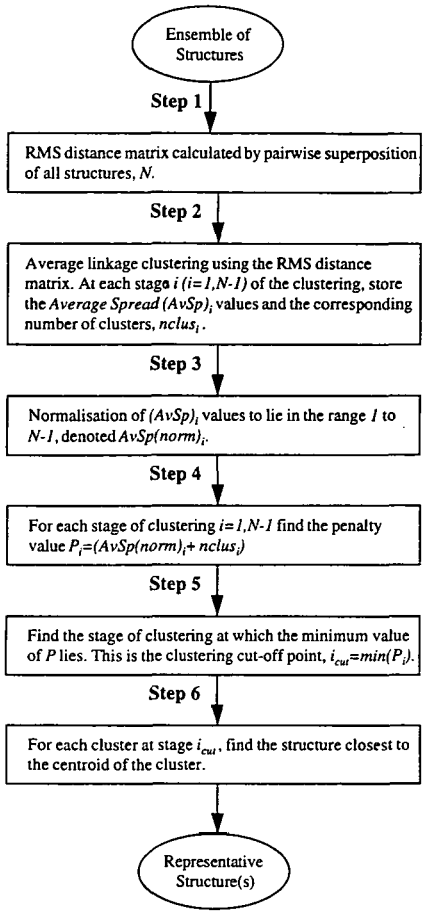

Fig. 1. Flow chart illustrating the progress of the NMRCLUST algorithm.

where cnum_i is the number of clusters at stage i of the clustering (excluding outlying points, i.e. clusters that contain only one member).

Step 3. Normalization of average spread

Once clustering is complete, the set of AvSp_i values is normalized to lie within the range 1 to $N-1$, where N is the number of structures in the original data set. Normalization is performed to give equal weight in the penalty function (Step 4) to the number of clusters and the average spread (a choice of relative weights which appears to work well).

$$\text{AvSp(norm)}_i = \left(\frac{N-2}{\text{Max(AvSp)} - \text{Min(AvSp)}}\right)(\text{AvSp}_i - \text{Min(AvSp)}) + 1,$$

where Max(AvSp) and Min(AvSp) are the maximum and minimum values respectively of average speed in the set $\{\text{AvSp}_1, \text{AvSp}_2, \ldots \text{AvSp}_{N-1}\}$.

Step 4. Penalty function

For each stage of clustering i, a penalty value, P_i, can now be calculated as:

$$P_i = \text{AvSp(norm)}_i + \text{nclus}_i,$$

where nclus_i is the total number of clusters at step i of the clustering (including outlying points).

Step 5. Defining the cut-off value

The minimum penalty value in the set $\{P_1, P_2, \ldots P_{N-1}\}$ is chosen as the cut-off level.

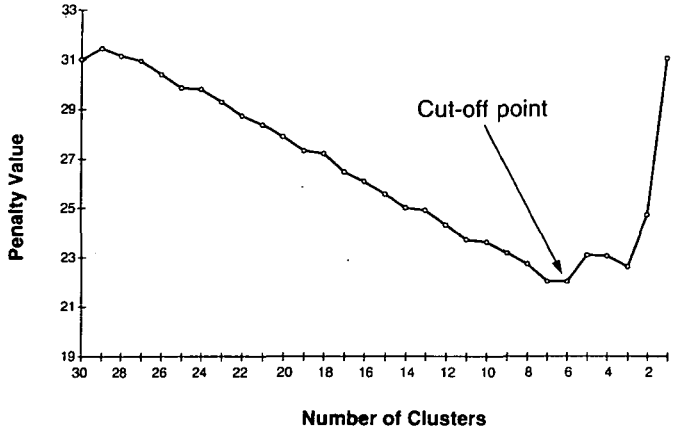

Fig. 2. Progress of the penalty function during clustering of ensemble 4HIR (Folkers *et al.*, 1989). The minimum value of the penalty function is chosen as the clustering cut-off point, as indicated.

$$P_{i\text{cut}} = \text{Min}(P)$$

Thus, the stage *i*cut represents a state where the clusters are as highly populated as possible, whilst simultaneously maintaining the smallest spread. The smaller the spread of the clusters, the more similar the conformations of its members; the greater the population of a cluster, the less likely is the chance of excluding a member of similar conformation.

Step 6. Representative structures

Once a cut-off value in the clustering has been determined in this way, Eigen analysis (Sutcliffe, 1993) is performed on each cluster at stage *i*cut. This allows for the determination of the structure within each cluster that is closest to the centroid of that cluster.

Example application

To illustrate the performance of the program, we present its application to hirudin (Folkers *et al.*, 1989; Protein Data Bank accession number 4HIR). In this example, all non-hydrogen atoms were used for the superposition. The penalty function arrives at a unique minimum value (Figure 2), which is chosen as the cut-off point for the clustering. It is interesting to note the correlation between the clusters and the conformation of hirudin (Figure 3). The four major clusters (i.e. excluding the two clusters containing only one member) correspond to different conformations of the region of the structure between residues Ser32 and Glu35. This observed lack of conformational order is consistent with the absence of any long-range nuclear Overhauser effects between this exposed 'finger' and the core region of the protein (Folkers *et al.*, 1989), as well as the alternative hydrogen bonding patterns known to exist in this region (Guntert *et al.*, 1995).

Flexibility of input

In addition to the automatically selected cut-off point, the program is able to accept a user-defined value for the minimum distance between representative structures. NMRCLUST can also superimpose the structures within the ensemble on the basis of a user-defined set of atoms. These are specified by the 'residue–residue:atom,atom' syntax. For example, to superimpose on all carbon atoms from residues 1 to 31 and residues 36 to 49, the syntax would be '1–31, 36–49:C*'. The atoms to be used for superposition may be determined, for instance, by using PROCHECK-NMR, the NMR version of the PROCHECK program (Laskowski *et al.*, 1993). This will

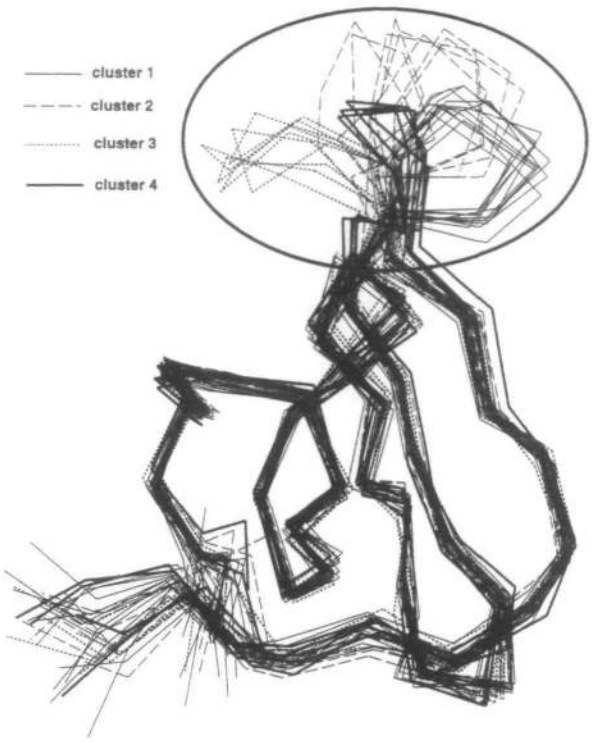

Fig. 3. The superimposed backbones of 28 hirudin (4HIR) structures. This illustrates the different conformations in the variable loop region between residues Ser32 and Glu35 (encircled). The application of NMRCLUST to this ensemble, with superposition on all non-hydrogen atoms, resulted in the four clusters shown and two outlying structures (not included for clarity). Different line styles indicate different cluster membership. (NB: Model 18 has been omitted from this analysis because of missing side-chain atoms on Gln49.)

allow the user to determine those residues that seem to be best defined in terms of backbone r.m.s. distance, side-chain r.m.s. distance and/or dihedral angle variability. Alternatively the user could use the technique described by Billeter (1992), which uses both backbone r.m.s. and all heavy atom r.m.s. values. (We are currently developing an automatic method for determining the optimum set of atoms to be used for superposition.) There is also the capability of performing the cluster analysis on a different, user-defined set of atoms to those used for superposition (e.g. Modi *et al.*, 1996).

Discussion

In this study, the decision to use the average linkage algorithm was based on an assessment of the value of Min(P) produced in 196 NMR-derived ensembles (available November 1994) using three different clustering algorithms: single linkage, complete linkage and average linkage. Of these three methods, average linkage performed best, producing the lowest average penalty value over the 196 ensembles. Another clustering algorithm commonly used with protein structures is the Jarvis–Patrick method (Allen and Doyle, 1991). However, this technique was not used in our studies because it requires a high level of user intervention: user-defined values for both the number of shared neighbours that two objects must possess to be in the same cluster (the commonality threshold, C_{JP}) and the number of nearest neighbours being considered for each cluster (K_{JP}).

A criticism has been raised against the technique described herein—the use of pairwise superposition followed by Eigen analysis can lead to negative Eigenvalues and hence information loss or distortion (Diamond, 1995). However, after running NMRCLUST on all 302 NMR-derived ensembles available in November 1995, no distortion of information above 10^{-6} Å (by comparing every distance in $N - 1$ dimensions to the original distance matrix) was found. Consequently, in practice negative Eigenvalues do not seem to be of particular concern. However, should a distortion of information occur that exceeds 10^{-5} Å, the program warns the user and, instead of determining the structure closest to the centroid of the cluster, selects the structure with the minimum average r.m.s. distance from all other cluster members (Adzhubei *et al.*, 1995). (The results of applying NMRCLUST to the 302 NMR-derived ensembles will be presented separately in a future paper.)

In conclusion, this method can be used to automatically cluster any data set (e.g. an ensemble of NMR-derived structures or an ensemble of homology models) rapidly and consistently, without the need for subjectively defined cut-offs. NMRCLUST will take a file in PDB format containing an ensemble of structures, and output the most representative structure from each of the resulting clusters. These representative structures can subsequently be used, for example, in homology modelling. Alternatively, NMRCLUST can take a predetermined matrix of 'distances' and automatically output the resulting clusters and their representative members. The program is freely available via both the World Wide Web (http://neon.chem.le.ac.uk/) and anonymous ftp (ftp.oxmol.co.uk).

Acknowledgements

We thank Roman Laskowski and Janet Thornton for useful discussions. L.A.K. is supported by a BBSRC CASE studentship, sponsored by Oxford Molecular Ltd. M.J.S. is a Royal Society University Research Fellow.

References

Abola,E.E., Bernstein,F.C., Bryant,S.H., Koetzle,T.F. and Weng,J. (1987) In Allen,F.H., Bergerhoff,G. and Sievers,R. (eds), *Crystallographic Databases—Information Content, Software Systems, Scientific Applications.* Data Commission of the International Union of Crystallography, Bonn, Germany, pp. 107–132.

Adzhubei,A.A., Laughton,C.A. and Neidle,S. (1995) *Protein Engng*, **8**, 615–625.

Allen,F.H. and Doyle,M.J. (1991) *Acta Crystallogr.*, **B47**, 41–49.

Billeter,M. (1992) *Quart. Rev. Biophys.*, **25**, 325–377.

Diamond,R. (1995) *Acta Crystallogr.*, **D51**, 127–135.

Folkers,P.J.M., Clore,G.M., Driscoll,P.C., Dodt,J., Kohler,S. and Gronenborn, A.M. (1989) *Biochemistry*, **28**, 2601–2617.

Guntert,P., Szyperski,T. and Wuthrich,K. (1995) *Protein Sci.*, **4**, 84.

Laskowski,R.A., MacArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) *J. Appl. Crystal.*, **26**, 283–291.

McLachlan,A.D. (1982) *Acta Crystallogr.*, **A38**, 871–873.

Modi,S., Paine,M.J., Sutcliffe,M.J., Lian,L.Y., Primrose,W.U., Wolf,C.R. and Roberts,G.C.K. (1996) *Biochemistry*, **35**, 4540–4550.

Sutcliffe,M.J. (1993) *Protein Sci.*, **2**, 936–944.

Received March 5, 1996; revised May 8, 1996; accepted May 15, 1996