Volume 219, number 1, 11–16 FEB 04819 July 1987

A simple method for delineating well-defined and variable regions in protein structures determined from interproton distance data

Michael Nilges, G. Marius Clore and Angela M. Gronenborn

Max-Planck-Institut für Biochemie, D-8033 Martinsried bei München, FRG

Received 25 March 1987

A simple method is described for identifying well-defined regions in a set of protein structures calculated from experimental interproton distance restraints. Two different functions, one based on the mean global rms difference, the other on the distance variation between equivalent atoms in different residues, are used to distinguish 'variable' from 'well-defined' regions. These functions are calculated in an iterative manner. The method is also capable of identifying several locally well-defined regions whose relative positions are not well-defined globally.

Protein structure; 3D solution structure; NMR; Nuclear Overhauser effect; Interproton distance

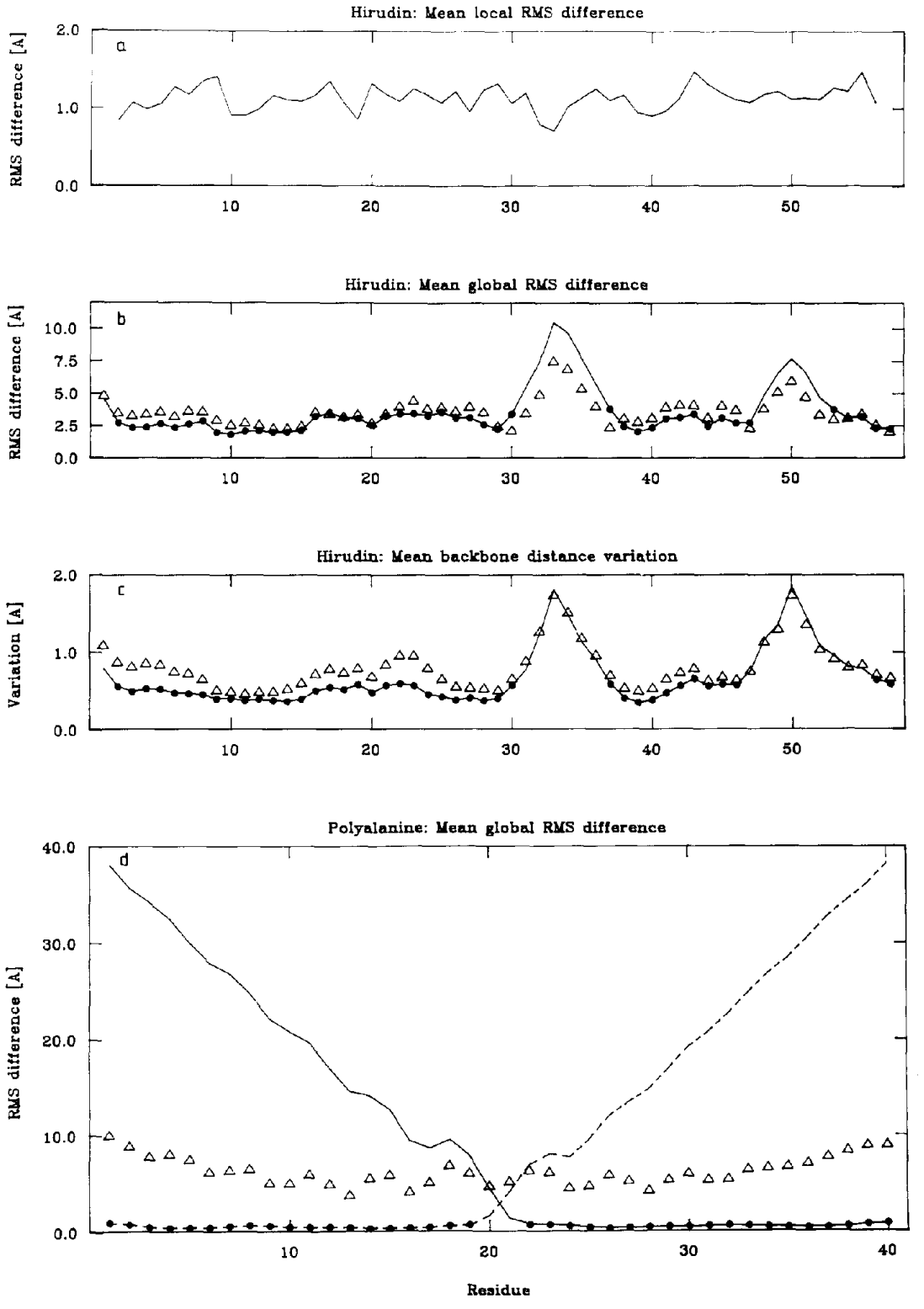
1. INTRODUCTION

In recent years considerable progress has been made towards determining three-dimensional structures of proteins in solution using NMR spectroscopy [1–9]. This generally involves three stages: (i) the sequential assignment of proton resonances by means of two-dimensional NMR techniques identifying through-bond and through-space (<5 Å) connectivities [10]; (ii) the assignment of as many cross-peaks as possible in the NOESY spectra in order to obtain a large set of approximate interproton distance restraints; and (iii) the determination of a 3D-structure on the basis of these restraints using a suitable computer algorithm, for example metric matrix distance geometry [11–14], restrained least squares minimization in torsion angle space with a variable target function [15] and restrained molecular dynamics [5,16–18]. In order to obtain a measure of the region of conformational space over which the interproton distances can be satisfied, several structures are always calculated and the degree of convergence examined. In this way, incorrectly folded structures that fail to satisfy the experimental restraints are easily identified. In many cases, however, it is found that the experimental restraints, although sufficient to determine the approximate overall polypeptide fold, may not be sufficient to determine the structure completely such that the lack of experimental information produces a certain amount of variability in the structures. Thus, for example, some parts of the structure may be better defined than others. Considering the ill-defined regions, two cases can be distinguished. The ill-defined residues can be fully disordered. Alternatively, they can be locally well-defined but globally ill-defined such that the position of a whole group of residues varies from structure to structure with respect to the remainder of

Correspondence address: M. Nilges or G.M. Clore, Max-Planck-Institut für Biochemie, D-8033 Martinsried bei München, FRG

Abbreviations: NMR, nuclear magnetic resonance; NOE, nuclear Overhauser effect; NOESY, two-dimensional NOE spectroscopy; rms, root mean square

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/87/$3.50


Hirudin: Mean local RMS difference
RMS difference [A]
a
2.0
1.0
0.0
10
20
30
40
50
Hirudin: Mean global RMS difference
RMS difference [A]
b
10.0
7.5
5.0
2.5
0.0
10
20
30
40
50
Hirudin: Mean backbone distance variation
Variation [A]
c
2.0
1.0
0.0
10
20
30
40
50
Polyalanine: Mean global RMS difference
RMS difference [A]
d
40.0
30.0
20.0
10.0
0.0
10
20
30
40
Residue

the protein. The recently determined structure of the protein hirudin [8] provides a typical example of the latter case. Hirudin, a 63 residue protein, has a well defined core and two minor domains consisting of a finger of antiparallel β-sheet (residues 31–36) and an exposed loop (residues 47–55). The two minor domains are locally well determined (see fig.1 of [8]) but their relative positions with respect to the central core could not be determined as no long range (i.e. $|i-j|>5$) interproton distance restraints between the two minor domains and the core could be identified.

In this paper we present a simple method for delineating such 'fixed' and 'variable' regions in structures generated on the basis of interproton distance data and for ascertaining whether the variable regions are locally well-defined or disordered.

2. RESULTS AND DISCUSSION

The simplest approach that one could use to identify fixed and variable regions is to display the set of computed structures on an interactive graphics system. In practice, however, this simple approach may be extremely difficult. The reason for this is that in order to compare the structures it is necessary to best fit them first. The larger the size of the variable regions, however, the worse will be the best fit of the fixed regions. Moreover, it becomes increasingly more difficult and time consuming to distinguish visually fixed from variable regions as the number of structures and the size of the protein increases.

An alternative approach involves making use of a function which is calculated from the structures themselves, and which has higher values for the variable regions than for the fixed ones.

The mean of the local atomic rms differences of sequential tripeptide segments between all pairs of structures [14] is one such function but is only able to detect very local variability. Thus, as shown in fig.1a, it would not have been possible to identify the variable regions in hirudin on the basis of this function. Indeed, the tip of the finger of antiparallel β-sheet around residue 33 has the lowest value of the mean local rms difference in the whole protein, and not even at the 'hinges' connecting the minor domains to the main core are the values increased. It is clear, therefore, that more global aspects of the structure have to be taken into account.

The mean of the global rms differences between all pairs of structures is a function which incorporates global features but suffers from the same drawback as the superposition on a graphics display. Namely, a best fit of the structures has to be performed with respect to all residues and, consequently, the result will be distorted when the variable regions of the protein are large. An obvious way to solve this problem is provided by a modification in which the best fitting is done only with respect to the well-defined regions such that:

$$f(i) = \frac{1}{n_{\text{pair}}} \sum_{j<k}^{n_{\text{struct}}} \text{RMS}_{jk}(i) \qquad (1)$$

where $\text{RMS}_{jk}(i)$ is the rms difference of structures j and k at residue i:

$$\text{RMS}_{jk}(i) = \left\{ \frac{1}{N_i} \sum_{l=1}^{N_i} \sum_{m=1}^{3} [X_{jlm}(i) - X_{klm}(i)]^2 \right\}^{1/2} \qquad (2)$$

Fig.1. (a) Mean backbone local atomic rms difference for hirudin. The hirudin structures are from [8] and the values plotted represent best fit atomic rms differences for successive tripeptide segments along the chain as a function of the sequence number of the middle residue. (b) Mean global rms difference for hirudin. The triangles (△) and solid line (——) represent the function f (eqn 2) before and after residues have been excluded, respectively. The solid circles (●) indicate the residues comprising the fixed region of hirudin. The overall mean global rms difference of the fixed region is 3.0 Å. (c) Weighted mean backbone distance variation g (eqn 3) in hirudin. Symbols as in b. The mean of g in the fixed region is 0.5 Å. (d) Mean global rms difference in the polyalanine test case in which residues 1–19 and 22–40 are helical and the conformation of residues 20–21 is varied in a totally random manner from structure to structure in a set of 10 structures. The solid line represents residues 22–40 (helix II) as the fixed region. The dashed line represents residues 1–19 as the fixed region. Other symbols as in b. The overall mean global rms difference in the fixed regions is 0.6 Å.

and n_{struct} is the number of calculated structures, n_{pair} is $[n_{struct}(n_{struct}-1)]/2$, and N_i is the number of atoms included at residue i. The prime indicates that structure k has been best fitted to structure j with respect to the fixed residues only.

An alternative function that can be used to identify variable regions in a protein structure is based on distances between equivalent atoms in all residue pairs. Distances between atoms in the variable regions of the structure and the fixed region also vary. A measure of this variation for a particular distance is its standard deviation. If only Cα atoms are considered for simplicity, a standard deviation of the Cα-Cα distance for every residue pair is obtained giving an $N \times N$ (where N is the number of residues) symmetric matrix of Cα-Cα distance standard deviations, similar to a Cα-Cα distance plot. To get a single value for each residue i, the values over all other residues j in the fixed region are averaged (i.e. each row in the standard deviation matrix is averaged). Thus, the function is

$$g(i) = \frac{1}{n_{fix}} \sum_{j \in F} w_{ij} \mathrm{STDEV}\{D(i,j)\} \qquad (3)$$

where $D(i,j)$ is the distance between the Cα atoms in residue i and j, STDEV the usual standard deviation taken over the distances obtained from all calculated structures, F the set of residues comprising the fixed region of the protein, n_{fix} the number of residues in F, and w_{ij} a weighting factor used in the averaging. If more than just the Cα atoms are to be included, the arithmetic means of the standard deviations of the distances between equivalent atoms are simply taken in order to get a single value for each residue pair.

In the case of both functions f and g reference is made to the residues comprising the fixed region of the protein, so that at first sight it would seem that one has to know a priori which residues comprise the fixed region, F. This, however, is not the case as both functions can be calculated in an iterative manner. Initially all residues are taken to lie in the fixed region, the 'worst' residues are then excluded, and f or g recalculated, until no further changes in the values of f or g occur and all ill-defined residues are excluded. The procedure is illustrated by the flow chart in fig.2, and the manner in which it is carried out is explained below for the function f and applies equally to the function g by analogy.

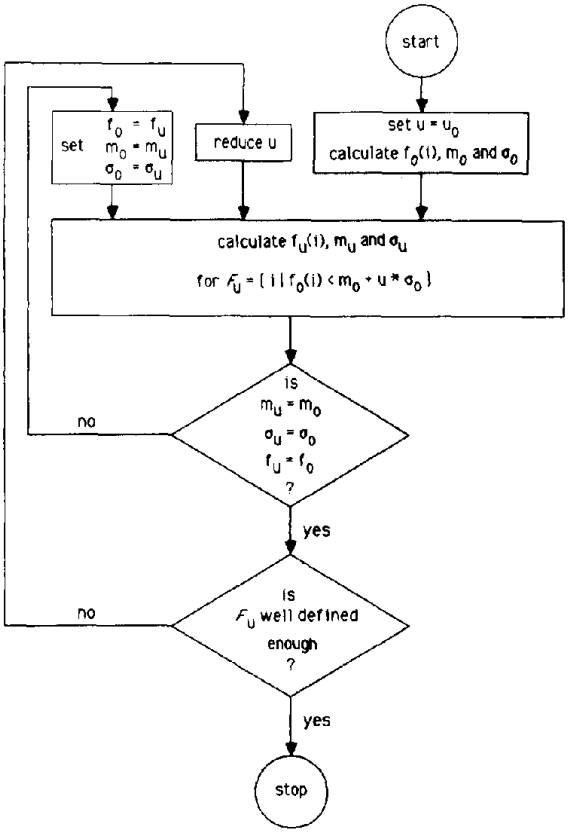


Fig.2. Flow chart diagram of the algorithm used to identify and delineate fixed and variable regions in a set of protein structures computed on the basis of interproton distance data.

A residue i is considered to lie in F_u, the fixed region of the structure, if the value of $f_u(i)$ lies below $m_u + u \cdot \sigma_u$, such that

$$F_u = \{ i \mid f_u(i) < m_u + u \cdot \sigma_u \} \qquad (4)$$

where m_u and σ_u are the self-consistently obtained limits

$$m_u = \frac{1}{N_u} \sum_{i=1}^{n_{res}} f_u(i) \cdot \theta(m_u + u \cdot \sigma_u - f_u(i)) \qquad (5)$$

$$\sigma_u = \{ [\sum_{i=1}^{n_{res}} (f_u(i))^2 \cdot \theta(m_u + u \cdot \sigma_u - f_u(i)) - m_u^2/N_u] / (N_u - 1) \}^{1/2} \qquad (6)$$

and

$$N_u = \sum_{i=1}^{n_{res}} \theta(m_u + u \cdot \sigma_u - f_u(i)) \quad (7)$$

n_{res} is the number of residues in the protein, $\theta(x)$ the Heaviside step function which is defined as $\theta(x) = 1$ for $x > 0$, $\theta(x) = 0$ otherwise. That is to say m_u and σ_u are the mean and standard deviation of f_u for all residues in F_u. The best fitting necessary to calculate f_u is performed with respect to the residues in F_u only. u is a positive and real scaling factor, which is initially set to a value for which only one or a few residues are excluded from the fixed region F_u.

For a given u, residues are excluded from the fixed part until there are no further changes in f_u, σ_u and m_u. Thus a group of residues is excluded generally for each value of u. Generally, f_u changes every time a residue is excluded so that the criterion for excluding residues itself changes. An unambiguous decision on which residues to exclude can therefore only be made when a self-consistent limit of f_u, m_u, and σ_u has been reached. Then, the remaining fixed region F_u is checked if it is fixed enough; for example, if it has the desired mean global rms difference. In this case, the algorithm stops; otherwise u is reduced and the next limits of f_u, σ_u and m_u are calculated. In this manner it is possible to identify directly a region that has a required mean global rms difference.

Once the main fixed region of the structure has been determined the algorithm can then proceed to search the remaining residues (i.e. the variable regions) for further well-defined regions. Thus the method is able to decompose a given structure into several parts each of which has a specified rms difference.

The algorithm has been tested with the structures of α_1-purothionin [6], phoratoxin [7] and hirudin [8] determined on the basis of experimental interproton distance restraints as well as with the structures of crambin derived on the basis of model interproton distance data [18], and it has been used to identify variable regions in the recently determined solution structure of the globular domain of histone H5 [9]. Both functions f and g (eqns 1 and 3) give similar results as illustrated by fig.1b and c in the case of hirudin. The function g, which is based on distances, has the advantage that no best fitting is necessary, a feature that speeds up the calculations. The rms difference, on the other hand, is the usual measure for comparing different structures, and computing times are still in the range of a few minutes on a VAX 11/780, so it is preferable to work with the function f. The best fitting is performed with the least-squares technique of Kabsch [19].

In all cases, the algorithm detected the previously determined variable regions. In the case of hirudin, the effect of excluding ill-defined regions from the best fitting procedure to calculate the mean global rms difference can be seen in fig.1b. The triangles represent f calculated with all residues included in the best fit, the solid line after residues had been excluded so that the overall mean global rms of the remaining fixed region was below 3 Å. The fixed region is indicated by closed circles on the solid line. Not surprisingly, the 'peaks' of f around residues 33 and 50, the tips of the finger of antiparallel β-sheet and of the exposed loop, respectively, are increased, while f is reduced at the other residues. Thus, residues 31 and 36, which belong to the finger of antiparallel β-sheet, and 48 and 52, which belong to the exposed loop, can be excluded from the fixed region. This would not have been possible on the basis of f calculated with all residues. For comparison, fig.1c shows the function g before and after exclusion of the ill-defined residues. Here the weighting factor w_{ij} (cf. eqn 3) has been set to the average distance between residues i and j divided by the maximum average distance, in order to emphasize global features.

The changes in the functions f and g resulting from the exclusion of ill-defined residues are important, but not very large in the case of hirudin, as only comparatively few residues constitute the variable region of the hirudin structure. This situation is entirely different when the size of the variable portion approaches that of the fixed part. This is illustrated by the test case shown in fig.1d. A total of 10 polyalanine structures were generated, each containing two regular α-helices of equal length from residues 1 to 19 (helix I) and 22 to 40 (helix II); the values for the backbone torsion angles were varied randomly within $\pm 5°$ around the ideal values. The conformation of residues 20 and 21, however, was varied in a completely random manner from structure to structure. Thus,

there are two locally well-defined regions whose relative position (i.e. the angle between the two helices) is completely undefined. The triangles in fig.1d show once again the mean global rms difference before the exclusion of any residues. Clearly, the mean global rms difference does not reveal anything of the true situation. Residues are then excluded from the fixed part, starting with residue 1, which has the highest value of the function f. As a result, helix II receives more weight in the best fitting than helix I. The values of f for all residues in helix I are therefore increased so that the next excluded residues also belong to helix I. Should a residue in helix II happen to be excluded in one iteration, it could reenter the fixed part at any time later, as the cutoff criterion (cf. eqn 4) is evaluated for all atoms in every iteration. In this manner, all residues in helix I and residues 20 and 21 are excluded, and helix II is identified as a well-defined region. The solid line in fig.1d shows the resulting function f and the closed circles represent the included residues. The excluded residues 1–21 are then searched for another well defined region, and helix I is found (dashed line).

It should be noted that if the situation had been exactly symmetric, it would not have been possible to identify the two well defined regions in this test case, as f would then also have been exactly symmetric, and residues would have been excluded in pairs, one from helix I, and one from helix II. Obviously such a situation is unlikely to occur in reality.

Finally, the iterative procedure presented here for identifying variable and fixed regions in a set of protein structures computed on the basis of interproton distance restraints can also be used to compare structures of two different proteins with the same number of amino acids. A generalization of the algorithm would have to be able to handle the general case where deletions and insertions occur in the proteins being compared.

ACKNOWLEDGEMENTS

This work was supported by the Max-Planck-Gesellschaft, grant no.321/4003/0318909A from the Bundesministerium für Forschung und Technologie and grant no.Cl186/1-1 from the Deutsche Forschungsgemeinschaft (G.M.C. and A.M.G.).

REFERENCES

[1] Braun, W., Wider, G., Lee, K.H. and Wüthrich, K. (1983) J. Mol. Biol. 169, 921–948.
[2] Williamson, M.P., Havel, T.F. and Wüthrich, K. (1985) J. Mol. Biol. 182, 295–315.
[3] Kline, A.D., Braun, W. and Wüthrich, K. (1986) J. Mol. Biol. 189, 377–382.
[4] Braun, W., Wagner, G., Wörgötter, E., Vassak, M., Kagi, J.H.R. and Wüthrich, K. (1986) J. Mol. Biol. 187, 125–129.
[5] Kaptein, R., Zuiderweg, E.R.P., Scheek, R.M., Boelens, R. and Van Gunsteren, W.F. (1985) J. Mol. Biol. 182, 179–182.
[6] Clore, G.M., Nilges, M., Sukumaran, D.K., Brünger, A.T., Karplus, M. and Gronenborn, A.M. (1986) EMBO J. 5, 2729–2735.
[7] Clore, G.M., Sukumaran, D.K., Nilges, M. and Gronenborn, A.M. (1987) Biochemistry 26, 1732–1745.
[8] Clore, G.M., Sukumaran, D.K., Nilges, M., Zarbock, J. and Gronenborn, A.M. (1987) EMBO J. 6, 529–537.
[9] Clore, G.M., Gronenborn, A.M., Nilges, M., Sukumaran, D.K. and Zarbock, J. (1987) EMBO J., in press.
[10] Wüthrich, K., Wider, G., Wagner, G. and Braun, W. (1982) J. Mol. Biol. 155, 311–319.
[11] Crippen, G.M. and Havel, T.F. (1978) Acta Crystallogr. A34, 282–284.
[12] Havel, T.F., Kuntz, I.D. and Crippen, G.M. (1983) Bull. Math. Biol. 45, 665–720.
[13] Havel, T.F. and Wüthrich, K. (1984) Bull. Math. Biol. 46, 673–698.
[14] Havel, T.F. and Wüthrich, K. (1985) J. Mol. Biol. 182, 281–294.
[15] Braun, W. and Go, N. (1985) J. Mol. Biol. 186, 611–626.
[16] Clore, G.M., Gronenborn, A.M., Brünger, A.T. and Clore, G.M. (1985) J. Mol. Biol. 185, 435–455.
[17] Brünger, A.T., Clore, G.M., Gronenborn, A.M. and Karplus, M. (1986) Proc. Natl. Acad. Sci. USA 83, 3801–3805.
[18] Clore, G.M., Brünger, A.T., Karplus, M. and Gronenborn, A.M. (1986) J. Mol. Biol. 191, 523–551.
[19] Kabsch, W. (1976) Acta Crystallogr. A32, 922–923.