J Biomol NMR (2014) 58:61–67
DOI 10.1007/s10858-013-9805-z

ARTICLE

# A robust method for quantitative identification of ordered cores in an ensemble of biomolecular structures by non-linear multi-dimensional scaling using inter-atomic distance variance matrix

**Naohiro Kobayashi**

Received: 13 September 2013 / Accepted: 10 December 2013 / Published online: 3 January 2014


**Abstract** Superpositioning of atoms in an ensemble of biomolecules is a common task in a variety of fields in structural biology. Although several automated tools exist based on previously established methods, manual operations to define the atoms in the ordered regions are usually preferred. The task is difficult and lacks output efficiency for multi-core proteins having complicated folding topology. The new method presented here can systematically and quantitatively achieve the identification of ordered cores even for molecules containing multiple cores linked with flexible loops. In contrast to established methods, this method treats the variance of inter-atomic distances in an ensemble as information content using a non-linear (NL) function, and then subjects it to multi-dimensional scaling (MDS) to embed the row vectors in the inter-atomic distance variance matrix into a lower dimensional matrix. The plots of the identified atom groups in a one or two-dimensional map enables users to visually and intuitively infer well-ordered atoms in an ensemble, as well as to automatically identify them by the standard clustering methods. The performance of the NL-MDS method has been examined for number of structure ensembles studied by nuclear magnetic resonance, demonstrating that the method can be more suitable for structural analysis of multi-core proteins in comparison to previously established methods.

**Keywords** Ensemble · Overlay · Core · Inter-atomic distance variance matrix · Multi-dimensional scaling

**Abbreviations**

| | |
|---|---|
| IVM | Inter-atomic distance variance matrix |
| NL-MDS | Non-linear multi-dimensional scaling |

## Introduction

Three-dimensional structures of biological macromolecules are widely used for studies like protein functions, drug design, and evolutional relationships. Structural comparisons play an important role in studying the dynamic properties of biomolecules during molecular dynamics simulations, structural sequence alignments in homology modeling, and superposition of nuclear magnetic resonance (NMR) models. For instance, functional analysis using an ensemble of NMR structures and the determination of atoms in ordered cores, can provide important aspects of structural features, which can be a first step to infer the function of the target molecules. Analyzing structural ensembles based on the similarity or heterogeneity of the atomic coordinates using computational methods and algorithms has been debated from a long time. This issue can be crucial, especially in NMR structure analysis, because typical final results represent an ensemble of many structures, each of which has to be consistent with the experimentally determined NMR constraints. The simplest and easiest way to find the ordered cores is iteratively "removing one atom and superimposing models" until satisfying specified sizes of cores and root-mean-squared

**Electronic supplementary material** The online version of this article (doi:10.1007/s10858-013-9805-z) contains supplementary material, which is available to authorized users.

N. Kobayashi (✉)
Institute for Protein Research, Osaka University,
Yamadaoka 3-2, Suita, Osaka 565-0871, Japan
e-mail: naohiro@protein.osaka-u.ac.jp

distances (RMSDs) of atom pairs. This has been achieved with some software programs currently available such as MolMol (Koradi et al. 1996). Another preferably used way is filtering atoms with dihedral angle order parameters on backbone atoms less than a specified cut-off (Nilges et al. 1987), and setting boundary residues on detected secondary structures (Kabsch and Sander 1983). These two methods were the most popular ways to define ordered core regions in ensembles, however, they cannot be easily applied for ensembles containing multiple cores with a complicated folding topology. In 1997, Kelly et al., developed a more systematic method using the inter-atomic distance variance matrix (IVM) and released a program called "NMRCore" (Kelley et al. 1997). Although a variety of other methods and tools have been established to date (Diamond 1995; Kelley et al. 1996; Schneider 2000; Snyder and Montelione 2005; Hirsch and Habeck 2008; Mechelke and Habeck 2010), many scientists still prefer to manually identify the cores by visual inspection on a molecular viewer. Recently, Kirchner and Güntert (2011) have demonstrated that the program "Cyrange" using IVM is sufficiently robust in identifying the most representative core in a structure ensemble. Figure 1a shows a 2D IVM map for a typical example of a protein containing two tandem cores (Kainosho et al. 2006). The distinct cores are easily identified by Cyrange, each of which can be observed by the wire-frame representation of superimposed Cα atoms in the identified residue ranges (see Fig. 1b, c). The 2D map produced by IVM is obviously not straightforward for the users to confirm the identified cores in the ensemble, even though they have been correctly identified in an automated way. Another problem in the previous methods using IVM is its less mathematical relevance when the row vector of a certain atom is merely applied to the clustering. For example, if there are several cores with largely different convergence, a direct comparison of variance can be misleading when gauging proximities between atom pairs.

In this study, a new method is presented that greatly improves the analysis of structure core in a more quantitative manner by combining IVM with non-linear multi-dimensional scaling (NL-MDS). It is demonstrated that a program using NL-MDS in a fully automated identification of defined cores performs better than programs FindCore and Cyrange for a number of structure ensembles studied using NMR.

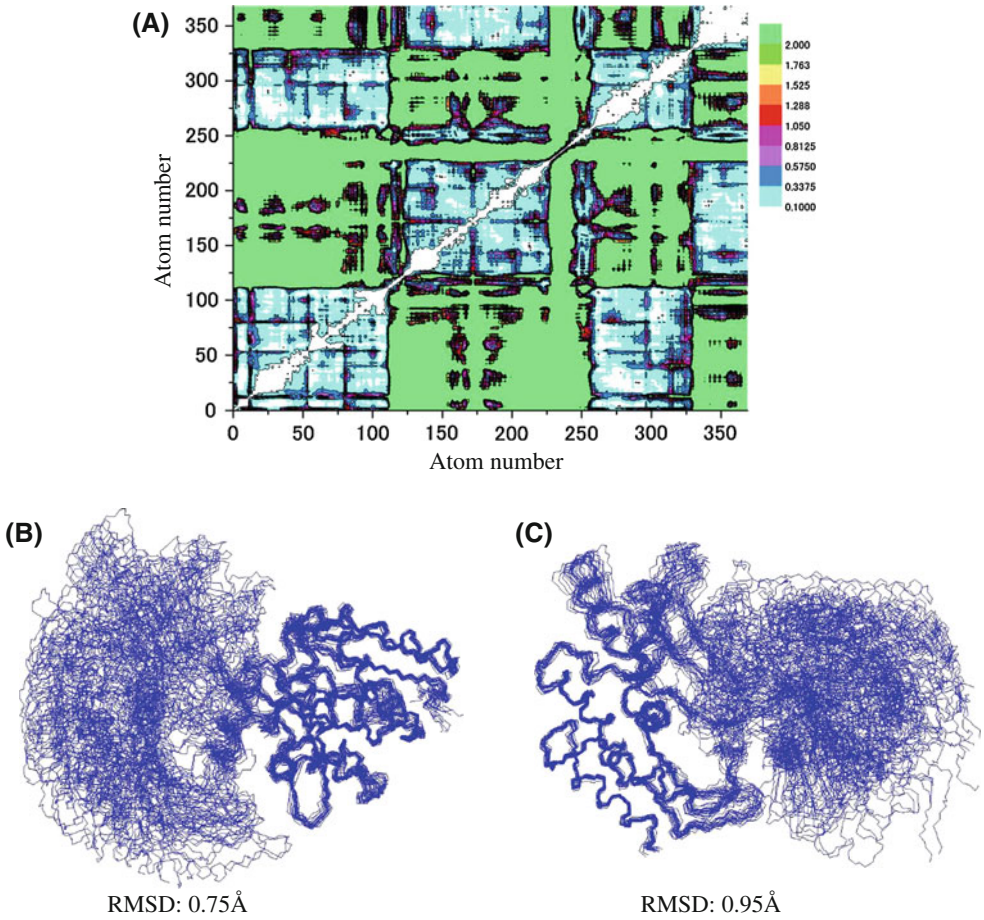


**Fig. 1** Color-coded contour plot of the inter-atomic distance variance matrix (IVM) calculated for the two-core protein, 2D21.pdb (**a**). The structural ensemble contains 20 models that were determined in an NMR study (Kainosho et al. 2006). X- and Y-axes corresponding to the amino acid residue number and colored by *cyan* to *light green* from 0.10 to 2.0 Å$^2$ as indicated by the *right gradation bar*. Residues with variance greater than 2.0 Å$^2$ are shown as *green* and those less than 0.1 Å$^2$ are shown as *white*. **b** and **c** depict two different Cα traces overlaid at atoms identified by the program Cyrange. RMSD values were calculated for the Cα atom coordinates in the ensemble

## Methods

### Non-linear scaled inter-atomic distance variance matrix

The IVM, $V$ for $N$ target atoms in an ensemble containing $M$ models can be given by;

$$V = \begin{bmatrix} 0 & \sigma_{12}^2 & \cdots & \sigma_{1N}^2 \\ \sigma_{21}^2 & 0 & \cdots & \sigma_{2N}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1}^2 & \sigma_{N2}^2 & \cdots & 0 \end{bmatrix} \quad (1)$$

$$\sigma_{ab}^2 = \frac{1}{M}\sum_{i=1}^{M}\{x_i(a,b) - x_{ave}(a,b)\}^2$$

where the variance $\sigma_{ab}^2$ can be obtained using the distance $x_i(a,b)$ between atoms $a$ and $b$ in the $i$th model of the ensemble and the average $x_{ave}(a,b)$ across the entire ensemble. In this study, to quantitatively scale the inter-atomic distance variance, linear and non-linear functions are applied to obtain a scaled matrix $H$:

$$H = \begin{bmatrix} 0 & h_{12} & \cdots & h_{1N} \\ h_{21} & 0 & \cdots & h_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ h_{N1} & h_{N2} & \cdots & 0 \end{bmatrix}, \quad (2)$$

$$\begin{cases} linear: \ h_{ij} = k_{off} + \sigma_{ij}^2 \\ non{-}linear: \ h_{ij} = \log(1 + \sigma_{ij}^2/k_{off}) \end{cases}$$

where $k_{off}$ is the offset value of the distance variance. In the above matrix, each row vector, $\boldsymbol{h}_j = (h_{j1}, \ h_{j2} \ \ldots \ h_{jN})$, $h_{jj} = 0$ can be used to represent the proximity relationship between target atoms $j$ and all other atoms (1,2, … $N$). The established methods directly use the matrix for clustering the row vectors to identify ordered cores in a structure ensemble.

### Multi-dimensional scaling

Prior to multi-dimensional scaling of the matrix $H$, the norm of the difference vectors between the row vectors $\boldsymbol{h}_j$ and $\boldsymbol{h}_i$ are calculated to give the matrix, $D$:

$$D = \begin{bmatrix} 0 & d_{12} & \cdots & d_{1N} \\ d_{21} & 0 & \cdots & d_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & \cdots & 0 \end{bmatrix}, \quad d_{ij} = \left\| h_i - h_j \right\| \quad (3)$$

where $d_{ij}$ quantitatively describes the distance between the row vectors $\boldsymbol{h}_j$ and $\boldsymbol{h}_i$ in the $N$-dimensional space. Next, the matrix $D$ can be embedded in a lower dimensional space using multi-dimensional scaling (MDS) whose theory and algorithms has been used for distance geometry calculations (Havel et al. 1983). According to the Young-Householder theorem (Young and Householder 1938), the Gram matrix $G$ can be obtained from the matrix $D$ using a centering matrix $Z$:

$$G = -\frac{1}{2} Z D Z^T,$$

$$Z = \begin{bmatrix} 1 - 1/N & -1/N & \cdots & -1/N \\ -1/N & 1 - 1/N & \cdots & -1/N \\ \vdots & \vdots & \ddots & \vdots \\ -1/N & -1/N & \cdots & 1 - 1/N \end{bmatrix} \quad (4)$$

In the matrix $Z$, each diagonal component is $1 - \frac{1}{N}$ whereas all others are $-\frac{1}{N}$. Because $G$ is an $N \times N$ symmetric matrix, it has an eigenvalue–eigenvector decomposition given by a similarity transform

$$G = X \Lambda X^T \quad (5)$$

where the eigenvalues of $\Lambda$ are ordered such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$. To generate lower dimensional plots, for instance, the two-dimensional matrix $X'$ is constructed;

$$X' = \left(\sqrt{\lambda_1} p_1, \sqrt{\lambda_2} p_2\right) \quad p_i = (p_{i1}, p_{i2} \cdots p_{iN}) \quad (6)$$

where $\lambda_i$ and $p_i$ corresponding to the eigenvalue and eigenvector, respectively, of the obtained matrix $X$ at the $i$th component.

### Clustering atoms in the lower dimensional matrix

The atoms associated with the lower-dimensional matrix $X'$ undergo standard clustering, to identify the ordered core atoms in the ensemble. In this study, the centroid-based method was used to build the hierarchical clustering dendrogram. In brief, a cluster of atoms is represented by a central vector, which starts from the $N$ clusters for each atom. At the $N$th level, each cluster is joined together with the closest vector to generate a new cluster for $(N - 1)$th level. The task is repeated until clustering reaches the 1st level, or the distance between the cluster vectors is greater than the cutoff value $k_{clust}$.

### Fully automated identification of ordered cores

The method presented here and automated tasks have been deployed in a single package, FitRobot, compiled with GCC version 3.4 on a 64-bit Linux system (CentOS ver. 5.7). Figure 2 shows the workflow for the identification of atoms in ordered cores in an ensemble. In the first step, the residues having low order parameter ($S_\phi < 0.8$ or $S_\psi < 0.8$) were eliminated to discard largely flexible regions such as N- and C-terminal tails and long loops. The

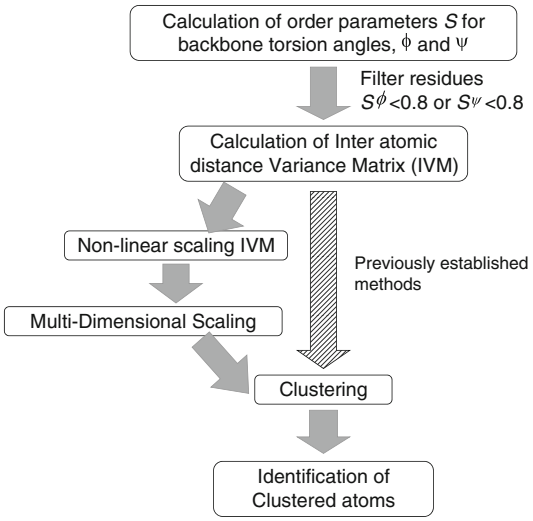


**Fig. 2** Workflow illustrating the progress of the fully automated algorithm that has been developed in this study. In the new methods, non-linear scaling and multiple dimensional scaling stages are inserted between the order parameter filtering step and the clustering row vectors in the IVM matrix step

angular order parameter for a certain residue in an ensemble with $N$ structure models can be calculated by;

$$S_\theta = \frac{1}{N}\left[\left(\sum_{i=1}^{N}\cos\theta\right)^2 + \left(\sum_{i=1}^{N}\sin\theta\right)^2\right]^{0.5} \quad (7)$$

where $\theta$ is the target dihedral angle. In the next step, the IVM, $V$, is calculated to determine the Gram matrix $G$ mentioned above. The eigenvalue–eigenvector decomposition is applied to obtain the eigenvalues $\Lambda$ and eigenvectors $X$ of matrix $G$. In this study, the QR algorithm is used for the decomposition step, because it is widely known to be more stable and faster than other methods to solve the eigenvalue–eigenvector problem by diagonalization of the symmetrical matrix. FitRobot automatically determines the optimal dimension to analyze how similar the variance vectors reduced to the lowest dimension, $n$, are according to the reliability of the MDS, $\eta$ (the default value is 0.95) given by;

$$\eta = \sum_{i=1}^{n}\lambda_i^2 \Big/ \sum_{i=1}^{N}\lambda_i^2 \quad (8)$$

The vectors of the thus-determined dimensionality undergo the above-mentioned clustering to identify residue ranges for RMSD fitting. The ensemble coordinates are overlaid according to the identified residues in each cluster, and then the RMSD of the Cα atoms for the first model coordinates is calculated. To represent the identified cores as simply as possible, the coverage of the atom groups in the coordinates is slightly extended to the residue range with the RMSD value multiplied by an extension factor $k_{ext}$. The segments with short contiguous residue range are eliminated based on the cutoff residue length, $k_{short}$. The identified ensemble sets including redundantly selected residues are filtered based on the identity of selected residues, $k_{red}$. At the end of the clustering stage, if the number of remaining residues is greater than the minimum length of cluster $k_{short}$, the cluster level is raised. The final identified residue ranges are used for fitting the target ensemble to generate final coordinate files.

## Benchmarks compared with state-of-the-art methods

Benchmarks were performed on the same system compiled for FitRobot, equipped with a Core i7 920 (2.66 GHz) processor (see Fig. 5). In this study, the program tools, FindCore and Cyrange, which have similar calculation schemes as previously reported by Snyder et al. Snyder and Montelione (2005) and Knichner and Güntert (2011) were used as benchmarks to compare with FitRobot. NMR structure coordinates for the benchmarks were obtained from the Protein Data Bank (PDB); 90 proteins with single core and 54 with multi-cores of between 2 and 4 cores (see the Supplemental material; Tables S3 and S4 list all PDB-ID for the benchmarks). For the benchmarks, the reference subset of atoms in the ensemble that were overlaid based on the ordered region were determined through careful inspection of the molecular structures by an expert in NMR studies using MolMol (Koradi et al. 1996). For complicated domain structures, the function "CalcMatch" in molmol was repeatedly used to automatically define the ordered core and then manually refined. Prior to the assessments, each of the structure coordinates in an ensemble was randomly rotated about the x-, y- and z-axes on the molecular frame. For the benchmarks using FitRobot, the parameter $k_{clust}$, $k_{short}$ and $k_{red}$, was set at 2.5, 8, 0.6, respectively, unless stated otherwise. The structural ensembles were overlaid using the standard RMS fitting method by generating a rotation matrix using quaternion methods (Coutsias et al. 2004) with respect to the atoms identified in the core regions. With this approach the program exports several sets of structure files, labeled with the PDB-ID followed by the number of clustering levels and core-ID. For example, in a three-core protein, 2K6B, three files, 2K6B_lev_2_0.pdb, 2K6B_lev_2_1.pdb and 2K6B_lev_2_2.pdb are generated. In this case, three cores (0, 1, and 2) were detected at the clustering level number 2. To compare the performance of the programs, the following judging protocols were used: (1) The total number of residues, $N_{total}$, providing the RMSD of the Cα atom coordinates in the ensemble less than $RMSD_{min}$, was counted; (2) The number of correctly identified residues $N_{correct}$ showing the difference between the RMSD of a targeted atom and a reference atom for each residue less than the value $E_{diff}$ was counted; (3) The content of the correctly identified residues

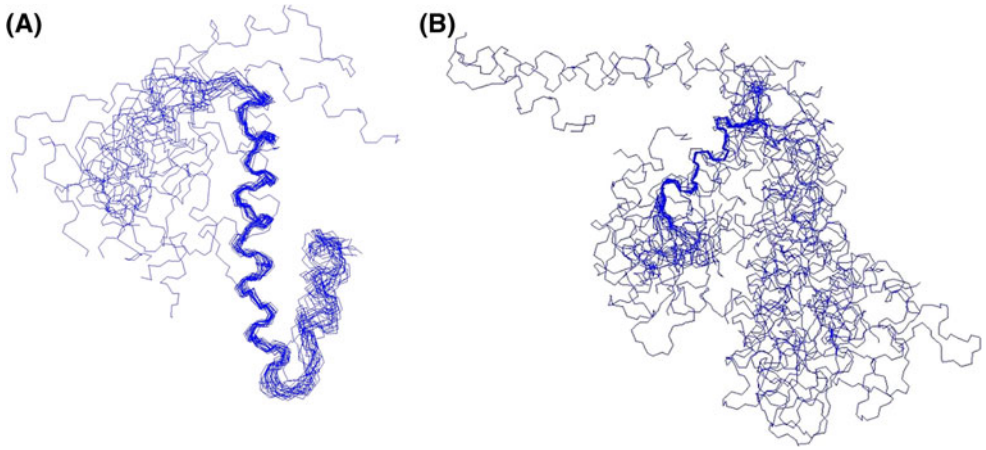


**Fig. 3** Typical case showing how difficult to identify cores in structure ensemble of multi-core protein. FitRobot, the new method using NL-MDS, identified two cores (**a** and **b**), however, FindCore and Cyrange missed the second core

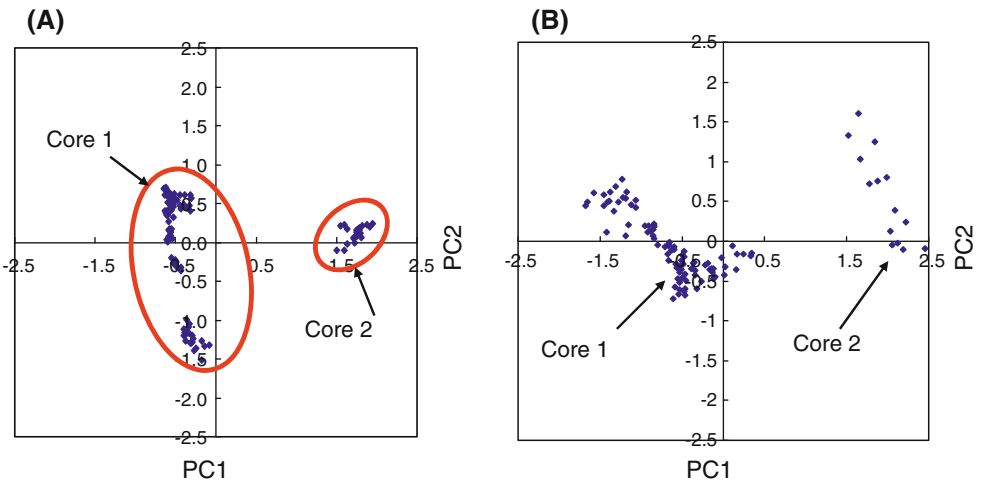


**Fig. 4** Two dimensional plots of eigenvalues derived from the multiple dimensional scaling analysis. The x- and y-axes correspond to the 1st and 2nd components of the factor loading (see main text). The maps derived from the NL-MDS method show obvious clustering of three cores (**a**), whereas those applying a linear function produce a first and second core scattered (**b**)

$T_{correct} = 100.0 \times N_{correct}/N_{total}$ greater than 50 % supposing the target core has been correctly identified. Then, if there is at least one generated ensemble found to be "overlaid with correctly identified atoms", the number of correctly identified cores was counted. In this study, $RMSD_{min}$ and $E_{diff}$ were set at 1.5 and 0.5 Å, respectively. The parameters used to operate FitRobot and the optimized values are summarized in supplemental Table S1.

Availability

The precompiled programs, source codes, and documentation for the tool used for this study are available from http://bmrbdep.protein.osaka-u.ac.jp/en/nmrtoolbox.

## Results and discussion

To quantitatively determine the variance of atoms in the IVM, the value of the element in the matrix would be more relevantly treated as information content in accordance with Shannon's entropy theorem. Nabuurs et al. has originally applied the idea to define a measure of uncertainty using distance constraints experimentally determined by NMR analysis (Nabuurs et al. 2003). Converse to this study, assuming the probability distribution of atoms $a$ and $b$ can be found equally in a certain distance range from $-D_{ab}/2$ to $D_{ab}/2$, the uncertainty in the distance between the atoms can be given by

$$h_{ab} = -\int_{D_{ab}/2}^{-D_{ab}/2} \frac{1}{D_{ab}} \log\left(\frac{1}{D_{ab}}\right) dx = \log D_{ab} \tag{9}$$

As the above equation only depends on the distance range $D_{ab}$, the same idea can be simply applied to the distance variance $\sigma_{ij}^2$ for atoms $i$ and $j$ to generate a value $h_{ij}$ as a scalable measure using the non-linear function $h_{ij} = \log\left(1 + \sigma_{ij}^2/k_{off}\right)$. The greater the value, the less information there is to restrict the atoms in distance range

from $-D_{ab}/2$ to $D_{ab}/2$, which can imply non-bond interactions such as Lennard-Jones interaction, hydrogen bond, electrostatic interaction and user-defined restraints such as dihedral restraints and NOE-derived distance restraints. NMR studies are especially suited to identify ordered cores in an ensemble, because atoms in short distance proximity can be more interesting. This distance is approximately from 2.0 to 5.0 Å corresponding to the lower limit for the van der Waals boundary of the atoms and the upper limit for distance constraints. The logarithmic function is suitable in emphasizing relatively small distance variance while suppressing variances with very large values. In this study, the offset value $k_{off}$, was set at 0.01 Å$^2$ to maintain uncertainty $h_{ij}$ positive. Any factor multiplying the value $h_{ij}$ in the non-linear function has no effect on the identification of cores in the NL-MDS and clustering stages (data not shown).

**Table 1** Summary of benchmark results performed for the programs FitRobot, FindCore, and Cyrange

| Examined proteins (number of cores to be identified) | Correctly identified ratio (%) (correctly identified cores) | | |
|---|---|---|---|
| | FitRobot | FindCore | Cyrange |
| Single-core proteins (90) | 100.0 (90) | 100.0 (90) | 97.8 (88)[a] |
| 1st core of multi-core proteins (54) | 100.0 (54) | 85.2 (46) | 96.3 (52) |
| 2nd core of multi-core proteins (54) | 98.1 (53) | 42.6 (23) | 75.9 (41) |
| 3rd and 4th cores of multi-core proteins (23) | 91.3 (21) | 26.1 (6) | 56.5 (13) |

[a] Two ensembles 2K8M and 2OYW failed because of segmentation error and the default limit setting for the chain length, respectively. Detailed results can be found in Supplemental Tables S3, S4 and S5

As mentioned above, the 2D IVM map itself (see Fig. 1a) is not intuitive enough to assess how correct the atoms are identified by the conventional methods compared with the 1D histogram of NL-MDS as shown in Supp-Fig. S1. For the small multi-core protein, 2JV5, two distinctive cores can be observed (see Fig. 3a, b). The parameter $k_{clust}$ was set at 2.5, the program FitRobot correctly identified the two cores, however, FinCore and Cyrange recognized the protein as a single-core. This result illustrates well the limits of the method directly using IVM. As shown in Fig. 4 with the 2D map derived from NL-MDS using the logarithmic function, the residues in the two cores are obviously recognized as distinctive clusters. In contrast, the MDS using the linear function (Fig. 4b) reveals more scattered plots in the 2D map, making the residues involved in the second core difficult to recognize.

The $k_{clust}$ value is the most crucial in the MDS analysis, as it determines the sensitivity in identifying the residues involved in multiple cores. First, to optimize the $k_{clust}$ for the NL-MDS, the benchmarks with the 90 single-core and 54 multi-core ensembles were performed and compared with the results using the linear functions, which varied from 0.5 to 4.5. The optimized values for the non-linear and linear functions were 2.5 and 3.5, respectively. As shown in Table 1 and Supplement Table S2, the non-linear function gave better results than the linear function with the same method. The benchmark results were also compared with the other programs, Findcore and Cyrange. In the summary of the benchmark results (Table 1), all three programs correctly identified the core in the single-core

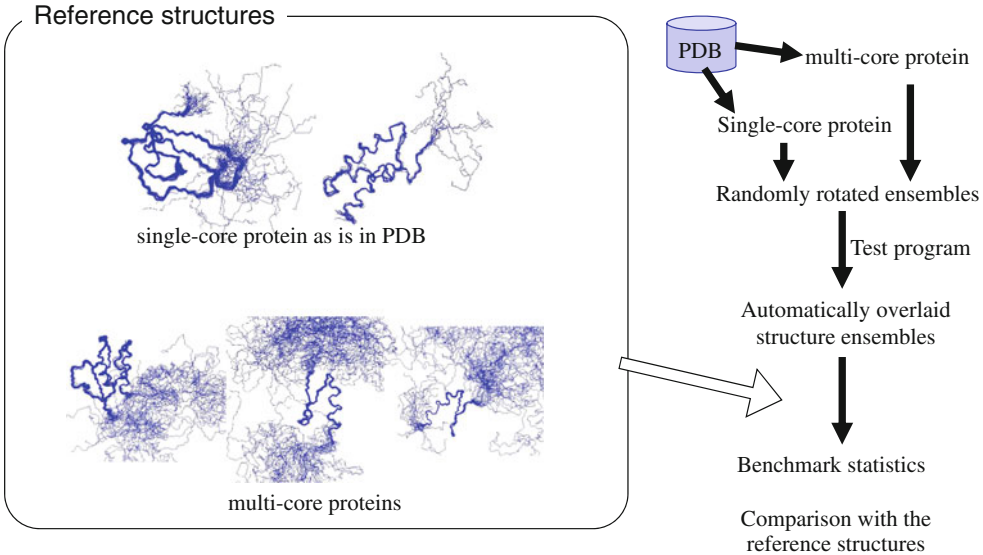


**Fig. 5** Benchmark scheme used to assess core identifications in ensembles targeted to reference proteins in a comparison of previously reported methods

proteins. In contrast, the benchmark results for the multi-core proteins show that FitRobot can detect 98 % of secondary domains but Cyrange and FindCore missed 24 and ~57 % of the domains, respectively. Cyrange and FindCore have the tendency to identify fewer cores than does FitRobot (see Table S3 and S4), which is understandable because the programs are strongly aimed at identifying representative cores of the targeted protein ensemble rather than discovering sub-structures. The program FitRobot has not only shown better results than established programs, but also, discovered small and meaningful cores in ensembles. Supplement Figure S2 shows typical cases of multi-domain proteins which have not been identified by Cyrange and Findcore. According to the report by Kirchner and Güntert (2011) on the intensive application of Cyrange to more than 6,000 entries in the PDB, 94 % of structures were considered to be single-core proteins. From this study, it has been suggested that more minor but actual sub-structures can be potentially found in the PDB entries. FitRobot algorithms can assist the user to inspect the distribution of sub-structures in structure ensembles with better quantitative evaluations. The scale of the 2D NL-MDS mapping is less dependent on the shape size and structural convergence of the proteins.

To summarize, there are two distinct advantages in the method presented over the conventional ones; (1) Multi-dimensional scaling using a non-linear function applied to the inter-atomic distance variance matrix IVM demonstrated greater capability in identifying cores in the structure ensemble of multi-core proteins. (2) By developing lower dimensional plots of NL-MDS, the analysis provides a more intuitive approach under visual inspection. Using standard software such as Microsoft Excel that can display 1D histograms or 2D scattering plots, one can easily find the clustered residues interactively.

## Conclusion

Using NL-MDS, the method presented here is useful in automatically identifying atoms involved in ordered cores of macromolecules among structural ensembles. The output of the program as 1D histograms or 2D scattering plots help in the inspection and refinement of the identified cores in a more systematic and intuitive manner. The method not only automates and standardizes identification of the representative cores in structure ensembles, but can be used also for careful inspection of sub-structures associated with over-constrained or less converging regions arising from NMR restraints.

**Acknowledgments** I would like to thank Prof. Toshimichi Fujiwara, Prof. Junichi Higo (Institute for Protein Research, University Osaka, Japan) and Prof. Daron M. Standley (Immunology Frontier Research Center, University Osaka, Japan) for helpful discussions. Dr. David A. Snyder and Prof. Gaetano T. Montelione, and Prof. Peter Güntert are greatly acknowledged for kindly providing their respective programs FindCore and Cyrange. I also acknowledge Mr. Bikash Ranjan Shahoo for proofreading the manuscript. This work was supported by National Bioscience Database Center (NBDC) in Japan Science and Technology Agency (JST) and also by JSPS KAKENHI Grant Number 80272160.

## References

Coutsias EA, Seok C, Dill KA (2004) Using quaternions to calculate RMSD. J Comput Chem 25(15):1849–1857

Diamond R (1995) Coordinate-based cluster analysis. Acta Crystallogr D Biol Crystallogr 51(Pt 2):127–135

Havel TF, Kuntz ID, Crippen GM (1983) The theory and practice of distance geometry. Bull Math Biol 45(5):665–720

Hirsch M, Habeck M (2008) Mixture models for protein structure ensembles. Bioinformatics 24(19):2184–2192

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12):2577–2637

Kainosho M, Torizawa T, Iwashita Y, Terauchi T, Mei Ono A, Güntert P (2006) Optimal isotope labelling for NMR protein structure determinations. Nature 440(7080):52–57

Kelley LA, Gardner SP, Sutcliffe MJ (1996) An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. Protein Eng 9(11):1063–1065

Kelley LA, Gardner SP, Sutcliffe MJ (1997) An automated approach for defining core atoms and domains in an ensemble of NMR-derived protein structures. Protein Eng 10(6):737–741

Kirchner DK, Güntert P (2011) Objective identification of residue ranges for the superposition of protein structures. BMC Bioinformatics 12(170):1471–2105

Koradi R, Billeter M, Wüthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. J Mol Graph 14(1):51–55

Mechelke M, Habeck M (2010) Robust probabilistic superposition and comparison of protein structures. BMC Bioinformatics 11: 363

Nabuurs SB, Spronk CA, Krieger E, Maassen H, Vriend G, Vuister GW (2003) Quantitative evaluation of experimental NMR restraints. J Am Chem Soc 125(39):12026–12034

Nilges M, Clore M, Gronenborn A (1987) A simple method for delineating well-defined and variable regions in protein structures determined from interproton distance data. Bioinformatics 219(1):11–16

Schneider TR (2000) Objective comparison of protein structures: error-scaled difference distance matrices. Acta Crystallogr D Biol Crystallogr 56(Pt 6):714–721

Snyder DA, Montelione GT (2005) Clustering algorithms for identifying core atom sets and for assessing the precision of protein structure ensembles. Proteins 59:673–686