

Structural bioinformatics

Bio3d: an R package for the comparative analysis of protein structures

Barry J. Grant^{1,*}, Ana P. C. Rodrigues², Karim M. ElSawy³, J. Andrew McCammon^{1,4} and Leo S. D. Caves³

¹Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA 92093, USA,

²The Burnham Institute for Medical Research, La Jolla, CA 92037, USA, ³Department of Biology, University of York, York YO10 5YW, UK and ⁴Howard Hughes Medical Institute, University of California, San Diego, La Jolla, CA 92093, USA

Received on July 18, 2006; revised on August 22, 2006; accepted on August 23, 2006

Advance Access publication August 29, 2006

Associate Editor: Anna Tramontano

ABSTRACT

Summary: An automated procedure for the analysis of homologous protein structures has been developed. The method facilitates the characterization of internal conformational differences and inter-conformer relationships and provides a framework for the analysis of protein structural evolution. The method is implemented in bio3d, an R package for the exploratory analysis of structure and sequence data.

Availability: The bio3d package is distributed with full source code as a platform-independent R package under a GPL2 license from: <http://mccammon.ucsd.edu/~bgrant/bio3d/>

Contact: bgrant@mccammon.ucsd.edu

1 INTRODUCTION

The detailed comparison of homologous protein structures can be used to infer pathways for evolutionary adaptation and, at closer evolutionary distances, mechanisms for conformational change. Traditionally, such investigations have involved careful visual inspection combined with structural alignment methods. These procedures are both time consuming and labor intensive, and require expert insight into the systems studied. With the growing number of determined protein structures, the availability of automatic procedures for analyzing the differences and similarities between structures becomes increasingly desirable.

The bio3d package contains utilities to process, organize and explore structure and sequence data. Features include the ability to read and write structure, sequence and dynamic trajectory data, perform atom summaries, atom selection, re-orientation, superposition, rigid core identification, clustering, distance matrix analysis, structure and sequence conservation analysis, and principal component analysis (PCA). Bio3d takes advantage of the extensive graphical and statistical capabilities of the R environment (R development core team, 2006; <http://www.R-project.org>), and thus represents a useful framework for exploratory analysis of structural data.

2 COMPARATIVE ANALYSIS OF PROTEIN STRUCTURES WITH BIO3D

The bio3d package employs refined structural superposition and PCA to examine the relationship between different conformers. Conventionally, structural superposition of protein structures minimizes the root mean square difference between their full set of equivalent residues. However, for the current application such a superposition procedure can be inappropriate. For example, in the comparison of a multi-domain protein that has undergone a hinge-like rearrangement of its domains, standard ‘all atom’ superposition would result in an underestimate of the true atomic displacement by attempting superposition over all domains (whole structure superposition). A more appropriate and insightful superposition would be anchored at the most invariant region and hence more clearly highlight the domain rearrangement (sub-structure superposition). To avoid such problems, the current protocol includes an iterated superposition procedure, where residues displaying the largest positional differences are excluded at each round until only the invariant ‘core’ residues remain (Gerstein and Altman, 1995).

Following core identification and subsequent superposition, PCA is employed to examine the relationship between different conformers/structures based on their equivalent residues. The application of PCA to both distributions of experimental structures and Molecular Dynamics trajectories, along with its ability to provide considerable insight into the nature of conformational differences in a range of protein families and other biomolecules, has been discussed previously (Abseher *et al.*, 1998; Caves *et al.*, 1998; ElSawy *et al.*, 2005; van Aalten *et al.*, 1997). Briefly, the resulting principal components (orthogonal eigenvectors) describe the axes of maximal variance of the distribution of structures. Projection of the distribution onto the subspace defined by the largest principal components results in a lower dimensional representation of the structural dataset. The percentage of the total mean square displacement (or variance) of atom positional fluctuations captured in each dimension is characterized by their corresponding eigenvalue. Experience suggests that 3–5 dimensions are often sufficient to capture over 70% of the total variance in a given family of structures. Thus, a handful of principal components are sufficient to

*To whom correspondence should be addressed.

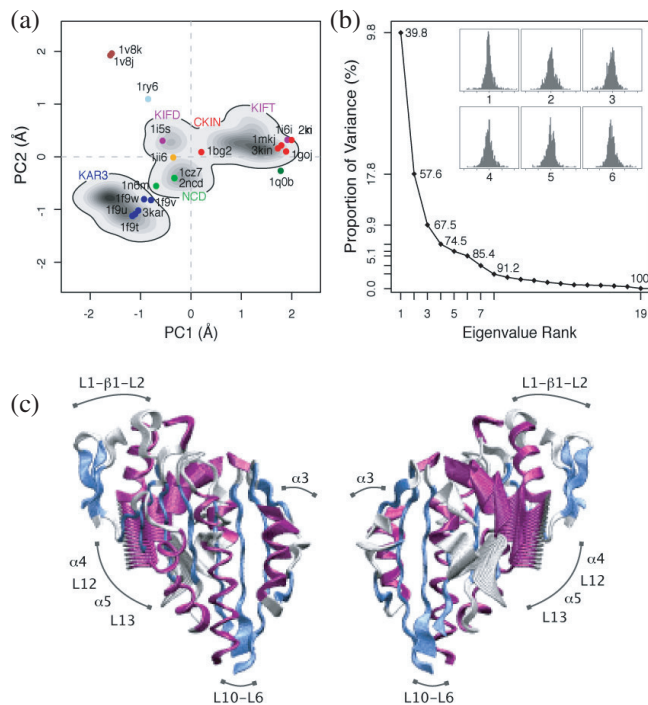


Fig. 1. Results of PCA on the kinesin molecular motor using standard Euclidean distance. **(a)** Conformer plot: Projection of the kinesin X-ray structures (circles) and transient MD conformers (shaded density contours) onto the principal planes obtained from analysis of all kinesin X-ray structures. **(b)** Eigenvalue spectrum: Results obtained from diagonalization of the atomic displacement correlation matrix of $C\alpha$ atom coordinates from the kinesin crystal structures. Inset shows histograms for the projection of the distribution of structures onto the first six principal components. **(c)** Interpolation: Front and back views of the kinesin motor domain, with the first principal component represented as equidistant atomic displacements from the mean structure. Displacements are scaled by two times the standard deviation of the distribution along the first principal component. Molecular figure was generated using VMD (Humphrey *et al.*, 1996).

provide a useful description while still retaining most of the variance in the original distribution. These low-dimensional representations, here termed ‘conformer plots’, succinctly display the relationships between different conformers, highlight the major differences between structures and enable the interpretation and characterization of multiple interconformer relationships (see example conformer plot, Fig. 1).

To further aid interpretation, a graphic ‘trajectory’ can be produced that interpolates between the most dissimilar structures in the distribution, as determined from the conformer plots. This involves dividing the difference between the conformers into a number of evenly spaced steps along the principal components, forming the frames of the trajectory. Such trajectories can be directly visualized in a molecular graphics program, such as VMD (Humphrey *et al.*, 1996). Furthermore, the interpolated structures can be analyzed for possible domain and shear movements with the DynDom package (Hayward and Berendsen, 1998), or used as initial seed structures for more advanced reaction path refinement methods such as Conjugate Peak Refinement (Fischer and Karplus, 1992).

3 SUMMARY

The bio3d comparative analysis results are in good agreement with descriptions established by human experts (Grant, 2004). In addition, the tools provide quantitative and visual information allowing for a more complete appreciation of interconformer relationships. Access to the open source software, full documentation, quick start guide and example data are available at <http://mccammon.ucsd.edu/~bgrant/bio3d/>

4 CONCLUSIONS AND PERSPECTIVES

The structure comparison procedures described here should facilitate the examination of diverse protein families, helping to identify common structural and dynamic features. Such analysis of structural homologues can provide invaluable conformational landmarks useful for assessing both new crystallographic structures and the results of theoretical methods. More generally, the current analysis methods may prove valuable to any study where knowledge of backbone flexibility must be modeled. For example, in flexible protein–protein docking and the generation of homology models where sampling along identified principal components may generate plausible alternative conformations. Another important area of research is deciphering possible networks of communication within proteins and, in particular, understanding allosteric mechanisms that appear to be preserved in distant relatives. Theoretical studies combined with comparative analysis of structural homologues are an initial step in this direction.

ACKNOWLEDGEMENTS

We would like to thank members of the Caves and McCammon groups for fruitful and entertaining discussions. This work was supported in part by the National Institutes of Health, National Science Foundation, the Howard Hughes Medical Institute, the National Biomedical Computation Resource and the National Science Foundation Center for Theoretical Biological Physics. Funding to pay the Open Access publication charges was provided by The Howard Hughes Medical Institute.

Conflict of Interest: none declared.

REFERENCES

- Abseher, R. *et al.* (1998) Essential spaces defined by NMR structure ensembles and molecular dynamics simulation show significant overlap. *Proteins*, **31**, 370–382.
- Caves, L.S.D. *et al.* (1998) Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein Sci.*, **7**, 649–666.
- Elsawy, K.M. *et al.* (2005) The physical determinants of the DNA conformational landscape. *Nucleic Acids Res.*, **33**, 5749–5762.
- Fischer, S. and Karplus, M. (1992) Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom. *Chem. Phys. Lett.*, **194**, 252–261.
- Gerstein, M. and Altman, R.B. (1995) Average core structures and variability measures for protein families: application to the immunoglobulins. *J. Mol. Biol.*, **251**, 161–175.
- Grant, B.J. (2004) Kinesin sequence, structure and dynamics. PhD Thesis. University of York, York, UK.
- Hayward, S. and Berendsen, H. (1998) Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme. *Proteins*, **30**, 144–154.
- Humphrey, W. *et al.* (1996) VMD: visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38.
- R Development Core Team (2006) R: a language and environment for statistical computing. Vienna, Austria.
- van Aalten, D.M.F. *et al.* (1997) Protein dynamics derived from clusters of crystal structures. *Biophys. J.*, **73**, 2891–2896.