Protein Science (1997), 6:2628–2630. Cambridge University Press. Printed in the USA.

FOR THE RECORD

OLDERADO: On-line database of ensemble representatives and domains

LAWRENCE A. KELLEY AND MICHAEL J. SUTCLIFFE

Department of Chemistry, University of Leicester, Leicester, LE1 7RH, United Kingdom

(RECEIVED July 2, 1997; ACCEPTED August 1, 1997)

Abstract: In cases where the structure of a single protein is represented by an ensemble of conformations, there is often a need to determine the common features and to choose a "representative" conformation. This occurs, for example, with structures determined by NMR spectroscopy, analysis of the trajectory from a molecular dynamics simulation, or an ensemble of structures produced by comparative modeling. We reported previously automatic methods for (1) defining the atoms with low spatial variance across an ensemble (i.e., the "core" atoms) and the domains in which these atoms lie, and (2) clustering an ensemble into conformationally related subfamilies. To extend the utility of these methods, we have developed a freely available server on the World Wide Web at http://neon.chem.le.ac.uk/olderado/. This (1) contains an automatically generated database of representative structures, core atoms, and domains determined for 449 ensembles of NMR-derived protein structures in the Protein Data Bank (PDB) in May 1997, and (2) allows the user to upload a PDB-formatted file containing the coordinates of an ensemble of structures. The server returns in real time: (1) information on the residues constituting domains; (2) the structures that constitute each conformational subfamily; and (3) an interactive java-based three-dimensional viewer to visualise the domains and clusters. Such information is useful, for example, when selecting conformations to be used in comparative modeling and when choosing parts of structures to be used in molecular replacement. Here we describe the OLDERADO server.

Keywords: comparative modeling; database; domains; ensemble analysis; molecular replacement; protein cores; World Wide Web

The analysis is performed in three stages: (1) definition of the core atoms; (2) definition of the rigid body(ies) (domains) in which these lie; and (3) definition of conformationally related subfamilies. Our core-defining procedure takes a Protein Data Bank (PDB; Bernstein et al., 1977; Abola et al., 1987) formatted coordinate file as input, and uses dihedral angle order (Hyberts et al., 1992) followed by application of a penalty function to define atoms as "core" or "non-core" (Kelley et al., 1997). Pairwise interatom distance variances are then automatically clustered to define sets of core atoms that lie in rigid bodies or "domains." Core definition is based purely on structural variance across the ensemble, and no average structures are computed. The structures are then superposed in a pairwise manner on the largest domain identified by the procedure above, and the RMSD is calculated. The resulting RMSD matrix is used as the basis for clustering in conformational space, and a penalty function is used to determine automatically the clustering cut-off. Each resulting cluster comprises ensemble members that lie in the same conformational subfamily (Kelley et al., 1996).

We have applied this procedure to 449 ensembles of NMR-derived protein structures in the PDB as of May 1997, and deposited the results in our database, known as OLDERADO (On Line Database of Ensemble Representatives And DOmains). This service has some similarity to the "internet library of protein family core structures" (LPFC; Schmidt et al., 1997; http://www-smi.stanford.edu/projects/helix/LPFC/), in that the LPFC provides information about core structures but based on different members of a protein family rather than an ensemble of conformations representing a single protein. OLDERADO affords access to the results of these analyses (as tabular data or interactively in a java-based three-dimensional viewer; Fig. 1) as well as the original programs used to generate them. In addition, if working on their own ensemble of protein structures, users may upload their coordinates for real-time analysis over the internet. Other applications of OLDERADO include selecting conformations to be used in comparative modeling and choosing parts of structures to be used in molecular replacement.

Results and discussion: To demonstrate the ease of analysis of ensembles using OLDERADO, we illustrate its application to the set of structures determined for troponin C (PDB accession code 1TNW; Slupsky & Sykes, 1995). Once the "SEARCH" button (Fig. 1A) has been selected, there are two options. If the PDB accession code is known, this can be entered directly, and the results returned. Alternatively, the PDB accession code can be determined via the link to the PDB's 3DB Browser (http://www.pdb.bnl.gov/pub-bin/pdbmain; the PDB has integrated links to OLDERADO into their

Reprint requests to: Michael J. Sutcliffe, Department of Chemistry, University of Leicester, Leicester, LE1 7RH, United Kingdom; e-mail: sjm@le.ac.uk.

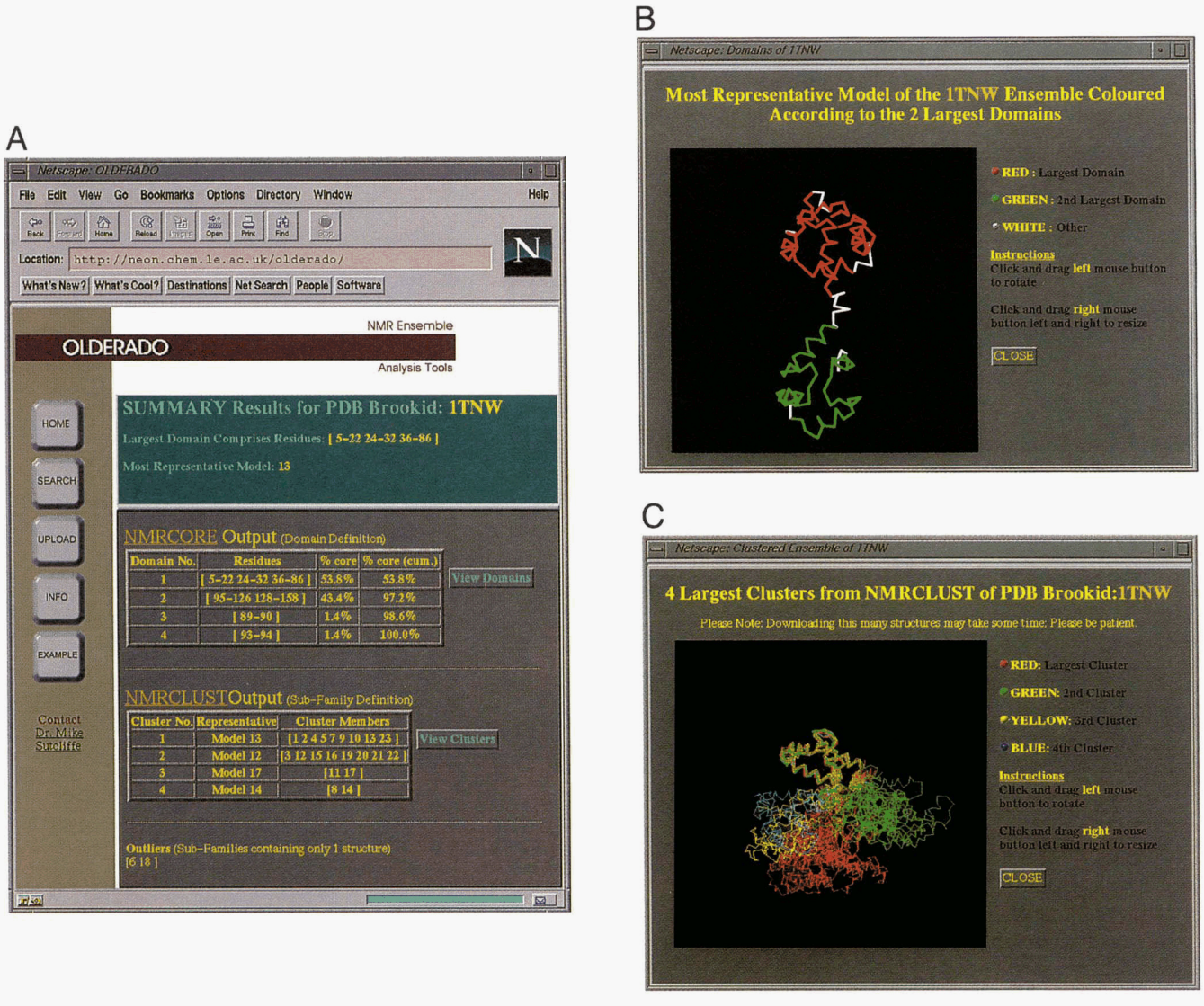

Fig. 1. OLDERADO browser interface, showing the results of a search for PDB accession code 1TNW (Slupsky & Sykes, 1995). **A:** Main browser window showing the domains and clusters in tabular form, and the Java 3D viewing mode for (**B**) domains and (**C**) clusters, respectively. These three-dimensional viewing windows allow real-time rotation and scaling of the protein structure.

3DB Browser). A table of results is then returned (Fig. 1A). At the top is a summary that defines the largest domain and most representative model. Under this, there are two tables—the first detailing (in order of domain size) the core and domain(s), and the second (in order of cluster size), the representative structure(s) and cluster membership. Two major domains are identified for troponin C: an N-terminal domain (residues 5–22, 24–32, 36–86) and a C-terminal domain (residues 95–126, 128–158).

This domain composition is in close agreement with the authors' definition of domains (Slupsky & Sykes, 1995), which states that the N-terminal domain comprises residues 5–85 and the C-terminal domain, residues 95–158. These two domains are not simultaneously superposable because they are connected by a flexible linker region (residues 86–94). Based on superposition of residues in the largest domain, four clusters are defined—each of which corresponds to a different orientation of the C-terminal domain with respect to the N-terminal domain. Both the defined domains and the clusters can be viewed interactively in three dimensions via the "View Domains" and "View Clusters" buttons, respectively (Fig. 1B,C).

This analysis was performed automatically without any user-intervention or pre-programmed parameters for core, domain, or cluster definition. All of these essentially "subjective" criteria are determined by the programs for each individual ensemble.

Supplementary material in Electronic Appendix: The supplementary material includes a list of PDB accession numbers in the file 449.ENS. The title is: List of PDB accession numbers of the 449 NMR-derived ensembles of protein structures in the OLDERADO database.

Acknowledgments: We thank Roman Laskowski, Janet Thornton, and Stephen Gardner 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 EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. 1987. Protein Data Bank. In: Allen FH, Bergerhoff G, Sievers R, eds. *Crystallographic databases—Information content, software systems, scientific applications.* Bonn/Cambridge/Chester: Data Commission of the International Union of Crystallography. pp 107–132.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanovichi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 186*:611–626.

Hyberts SG, Goldberg MS, Havel TF, Wagner G. 1992. The solution structure of Eglin-c based on measurements of many NOEs and coupling-constants and its comparison with X-ray structures. *Protein Sci 1*:736–751.

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*: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*:737–741.

Schmidt R, Gerstein M, Altman RB. 1997. LPFC: An internet library of protein family core structures. *Protein Sci 6*:246–248.

Slupsky CM, Sykes BD. 1995. NMR solution structure of calcium-saturated skeletal-muscle troponin-C. *Biochemistry 34*:15953–15964.