Bioinformation
www.bioinformation.net

open access
Web Server

CARON – Average RMSD of NMR structure ensembles

Kresimir Sikic[1,2*], Oliviero Carugo[1,3]

[1]Departement of Structural and Computational Biology, Max F. Perutz Laboratories, Vienna University, 1030 Vienna, Austria; [2]Faculty of Electrical Engineering and Computing, University of Zagreb, 10000 Zagreb, Croatia; [3]Departement of General Chemistry, University of Pavia, I-27100 Pavia, Italy; Kresimir Sikic – E-mail: kresimir.sikic@univie.ac.at; *Corresponding author

Received May 13, 2009; Accepted June 15, 2009; Published September 30, 2009

Abstract:
The NMR protein structures are often deposited in the Protein Data Bank as ensembles of models that agree with the experimental restraints. Information about stereochemical variability and the molecular flexibility can be obtained by systematic comparison of all models. Here we describe CARON, a software that allows the computation of the root-mean-square-distances between equivalent atoms and residues in all models and introduces these values into the occupancy and the B-factor fields of PDB-formatted files. This tool allows the user to both get a quantitative estimation of the conformational homogeneity of the models and to exploit this information in common computer graphics programs.

Key words: CARON, Bioinformatics software, conformational homogeneity, superposition, Root-Mean-Square Deviation, Protein Data Bank, NMR spectroscopy.

Availability: The source code and the manual are free and available at www.univie.ac.at/biolchem/

Background:

Despite a number of protein three-dimensional structures is determined with NMR spectroscopic methods, the format of the Protein Data Bank (PDB) files **[1, 2]** was designed to account for crystallographic analyses. The crystallographic occupancy and the atomic displacement parameters (adp; often referred to as B-factors) are not computed in NMR structure determinations despite they are declared in PDB formatted files. Here we present a computer program that replaces the occupancy and adp fields with quantities that describe the conformational homogeneity of the models deposited in the PDB file. NMR protein structures are deposited in the PDB as ensembles of models. If there are N models, it is thus possible to superpose the N(N-1)/2 unique pairs of models and to compute distances D_i between the same atoms in all the superposed model pairs. These distances are then used for calculating (i) the RMSD of each atom and (ii) the average RMSD between the equivalent atoms of the same residue in all the superposed models. These two variables provide a quantitative measure of the spatial dispersion of each atom and residue. By inserting the atom RMSD and the average residue RMSD values in the occupancy and B-factor fields it is then possible to provide the quantitative information in a format compatible with molecular graphics computer programs and, for example, color the molecule. These computations can be performed with CARON, a program that uses the Ying-Hunk et al. **[3]** implementation of Kabsch algorithm **[4]** for superposing pairs of molecules.

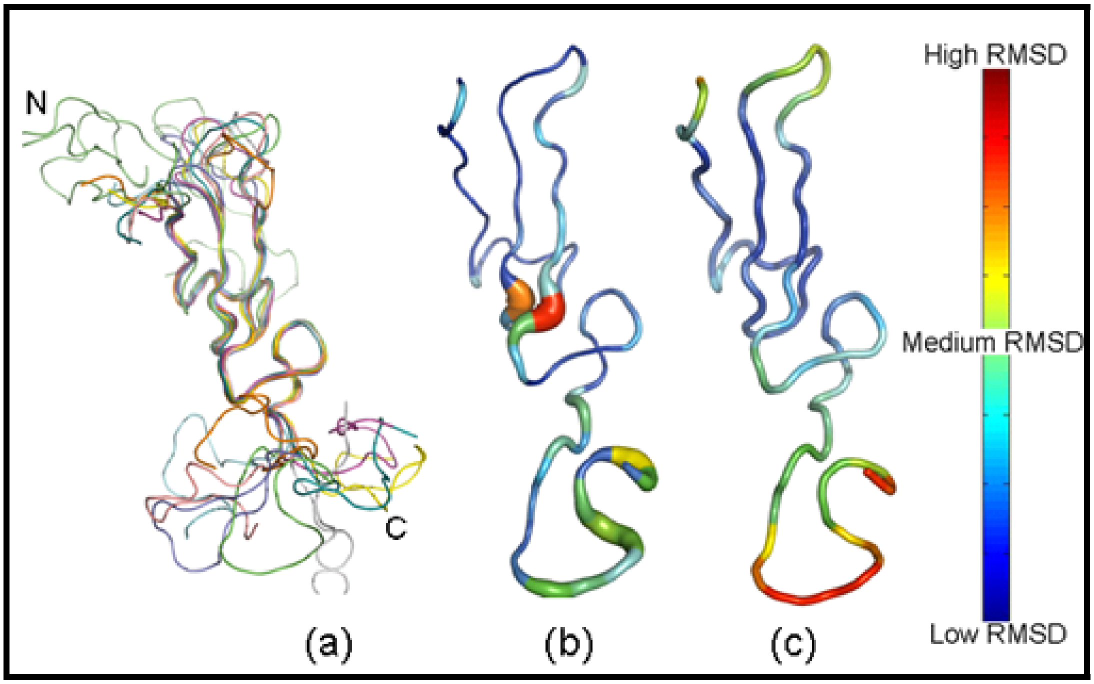

Figure 1: (a) Trace of all 10 models of the 1HRF entry of the Protein Data Bank; (b) First model colored according to the *B-factor putty* script distributed with PyMol; (c) First model colored according to the average residues RMSD computed with CARON. All figures were created with PyMol program (http://www.pymol.org).

Implementation:

CARON is a stand alone program written in C language and can be compiled using any standard ANSI C compiler under Linux or Windows. The input files are PDB files containing coordinates of atoms. The output data are stored in PDB formatted files where the occupancy and the B-factor fields of the ATOM lines report the average RMSD of the atom and of the residue, respectively, computed on the basis of the superposition of all the unique pairs of molecular models. Although the PDB files of NMR structures contains hydrogen atoms and often also non-protein atoms, these are disregarded by CARON and are absent from the PDB formatted output files. The user may decide to superpose all the non-hydrogen protein atoms or only the Cα atoms. At the beginning of the PDB-formatted output file, pertinent information is provided about the overall RMSD values and their distribution. An additional *parse* feature enables user to dissect a PDB file containing N models into N PDB files, each containing one of the models.

Results:

An example of the results obtained with CARON is shown in **Figure 1**. The three-dimensional structures of the human epidermal growth factor-like domain of heregulin-alpha was determined with NMR methods and deposited in the PDB file 1HRF as an ensemble of 10 models **[5]**. It is apparent that the termini are conformationally ill-defined as well as the loop on the top-right corner (**Figure 1a**). This is well shown by coloring the trace of only one model as a function of the RMSD values of the residues (**Figure 1c**). An alternative method, implemented in a script distributed with PyMol, allows one to depict the trace of a single model according to its conformational dispersion - segments with a very variable stereochemistry are large and red - produces on the contrary different results, with a large conformational dispersion in the middle of the molecule, which does not seem to be a genuine structural feature (**Figure 1b**). Eventually, it is necessary to remember that the conformational dispersion observed on the basis of the PDB files does not depend necessarily on the intra-molecular flexibility. The absence of experimental information might also be responsible for the structural divergence of the termini and of some loops. Moreover, work is in progress to use also other superposition tools that might be appropriate to handle multi-domain protein structures and to allow alternative selections of the atoms/residues that must be superposed.

Acknowledgement:

Funding from BIN-II, BIN-III (GEN-AU Austrian research program) and a grant from the Ministry of Science, Republic of Croatia, number 036-0362214-1987 are gratefully acknowledged.

References:

[1] HM Berman *et al.*, *Nucleic Acids Res.* **28**:235 (2000) [PMID: 10592235]
[2] FC Bernstein *et al.*, *J Mol Biol.* **112**:535 (1977) [PMID: PMID: 875032]
[3] L Ying-Hung *et al.*, *International Computer Symposium (ICS'2004)* Taipei, Taiwan (2004) 1000-1005.
[4] W Kabsch *Acta Crystallographica.* **32**:922 (1976).
[5] K. Nagata *et al.*, *EMBO J.* **13**:3517 (1994) [PMID: 8062828]

Edited by P. Kangueane
Citation: Sikic & Carugo, Bioinformation 4(3): 132-133 (2009)