

CASD-NMR: critical assessment of automated structure determination by NMR

To the Editor: NMR spectroscopy is currently the only technique for determining the solution structure of biological macromolecules. This typically requires both the assignment of resonances and a labor-intensive analysis of multidimensional nuclear Overhauser effect spectroscopy (NOESY) spectra, in which peaks are matched to assigned resonances. Software tools that fully automate the NOESY assignment and the structure calculation steps have the potential to boost the efficiency, reproducibility and reliability of NMR structures.

Within the e-NMR project (<http://www.e-nmr.eu/>), which is funded by the European Commission (project number 213010), we sought to assess whether such automated methods can indeed produce structures that closely match those manually refined by experts using the same experimental data (the 'reference structures'). We just completed the first comparison of automated NMR protein structure calculation methods and now announce its continuation in the form of an ongoing, community-wide experiment, called critical assessment of automated structure determination of proteins by NMR (CASD-NMR). CASD-NMR is open for members of any laboratory to participate and/or to submit targets. The concept closely resembles that of other community-wide experiments, such as the critical assessment of techniques for protein structure prediction (CASP)¹ and the critical assessment of prediction of interactions (CAPRI)². Unlike CASP and CAPRI, CASD-NMR is entirely based on the analysis of experimental data, which presents special issues in assembling, organizing and distributing these data among participants.

In the first year of the CASP-NMR experiment, we provided seven research teams involved in developing fully automated structure assignment tools with ten experimental data sets for various protein systems of known structure and two sets for protein structures not yet publicly available (tests performed in a blinded fashion), courtesy of the Northeast Structural Genomics (NESG) consortium. We then met in Florence, Italy on May 4–6, 2009 to analyze the structures generated (Fig. 1) by comparison to the reference structures and by using independent methods for structure validation. This first experiment indicated that although most of the automated structure

determinations had correct overall folds, for certain targets some programs did not calculate accurate packing and length of secondary structure elements. The root mean square (r.m.s.) deviations of the automatically generated backbone coordinates with respect to the reference structures were typically 1–2 Å but in some cases were as high as 9 Å.

We anticipate that the complete automation of protein solution structure determination from assigned chemical shift lists and unassigned NOESY peak lists may soon reach the point at which 'unsupervised' results can be directly deposited to the Protein Data Bank (PDB). It is therefore meaningful and timely³ to implement CASD-NMR as a community-wide rolling experiment. We invite software developers to participate in CASD-NMR to test their fully automated protocols on masked data sets and produce structures as if they would directly deposit them to the PDB. We will regularly release masked test data sets for proteins whose solution structure will be kept on hold by the PDB for at least eight weeks. We also invite members of any NMR group that is about to deposit a structure in the PDB to contribute a masked test case to CASD-NMR. To guarantee that a sufficient amount of data is available, the NESG consortium of the NIH Protein Structure Initiative (PSI) will also provide one data set per month. A masked data set will include the protein sequence, chemical shift assignments and unassigned integrated NOESY peak lists. Data providers may also include additional biochemical information and raw spectral data. These experimental data will be available from a central database of the e-NMR project (<http://www.e-nmr.eu/CASD-NMR/>) and through the PSI Knowledgebase (<http://kb.psi-structuralgenomics.org/>), also after the release of the reference PDB structure. This will allow

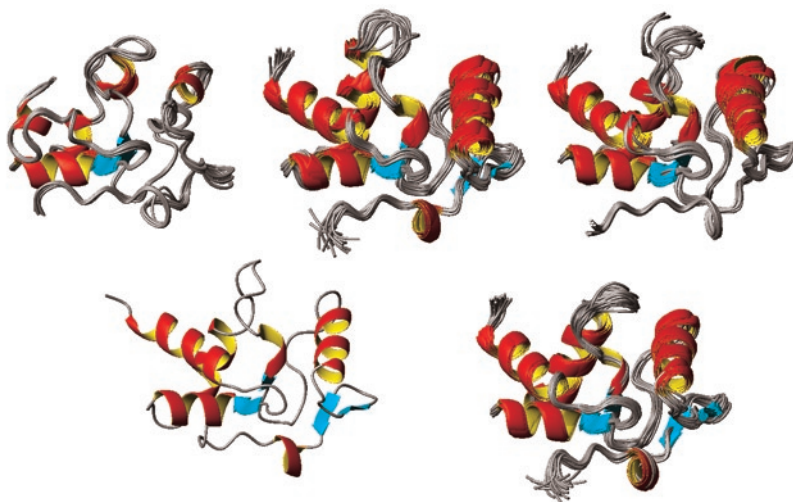


Figure 1 | Performance of various automated structure calculation methods. The results of fully automated calculations by various programs for one of the masked test data sets of the 2009 Florence workshop compared to the reference structure (bottom right) determined by Aramini, J.M. *et al.*, NESG Consortium (unpublished data; PDB: 2kif).

participants to join CASD-NMR at any time. For more information, including file format requirements and details on the submission procedures, see <http://www.e-nmr.eu/CASD-NMR/>.

CASD-NMR participants will have eight weeks to automatically generate structures and deposit their coordinates and the conformational restraints used in calculations to the CASD-NMR database. The participants will be responsible for making their software tools available to the community. Manual intervention on the data other than recalibration of chemical shifts is forbidden. The coordinates generated by participants, their comparisons to the reference structure and their validation scores will be accessible through the CASD-NMR website. An assessment meeting is planned for mid-2010.

We believe that the community-wide CASD-NMR experiment will foster the development of better algorithms and validation tools, and the adoption of state-of-the-art automated structure determination protocols by the wider bio-NMR community. We look forward to a fascinating experiment.

Antonio Rosato^{1,2}, Anurag Bagaria^{3,4}, David Baker⁵, Benjamin Bardiaux⁶, Andrea Cavalli⁷, Jurgen F Doreleijers⁸, Andrea Giachetti¹, Paul Guerry⁹, Peter Güntert^{3,4}, Torsten Herrmann⁹, Yuanpeng J Huang¹⁰, Hendrik R A Jonker^{4,11},

Binchen Mao¹⁰, Thérèse E Malliavin⁶, Gaetano T Montelione¹⁰, Michael Nilges⁶, Srivatsan Raman⁵, Gijs van der Schot¹², Wim F Vranken¹³, Geerten W Vuister⁸ & Alexandre M J J Bonvin¹²

¹Magnetic Resonance Center and ²Department of Chemistry, University of Florence, Sesto Fiorentino, Italy. ³Institute of Biophysical Chemistry and Frankfurt Institute for Advanced Studies and ⁴Center for Biomolecular Magnetic Resonance, Goethe University Frankfurt, Frankfurt am Main, Germany. ⁵Department of Biochemistry, University of Washington, Seattle, Washington, USA. ⁶Unité de Bioinformatique Structurale, Institut Pasteur, Centre National de la Recherche Scientifique Unité de recherche autonome 2185, Paris, France. ⁷Department of Chemistry, University of Cambridge, Cambridge, UK. ⁸Protein Biophysics, Institute of Molecules and Materials, Radboud University Nijmegen, Nijmegen, The Netherlands. ⁹Centre de RMN à très Hauts Champs, Université de Lyon, Centre National de la Recherche Scientifique, Ecole normale supérieure Lyon, Université Claude Bernard Lyon 1, Villeurbanne, France. ¹⁰Center for Advanced Biotechnology and Medicine, Department of Molecular Biology and Biochemistry, and Northeast Structural Genomics Consortium, Rutgers, The State University of New Jersey, Piscataway, New Jersey, USA. ¹¹Institute of Organic Chemistry and Chemical Biology, Goethe University Frankfurt, Frankfurt am Main, Germany. ¹²Bijvoet Center for Biomolecular Research, Faculty of Science, Utrecht University, Utrecht, The Netherlands. ¹³European Bioinformatics Institute, Hinxton, Cambridge, UK.
e-mail: rosato@cerm.unifi.it

1. Moulton, J. *et al. Proteins* **23**, ii–v (1995).
2. Janin, J. *et al. Proteins* **52**, 2–9 (2003).
3. Anonymous. *Nat. Methods* **5**, 659 (2008).