



Deliverable 1.4

Second assessment of the state-of-the art in computational aspects of bio-NMR

INFRA-2007-1.2.2 - Deployment of eInfrastructures for scientific communities

**Grant agreement for: Combination of Collaborative projects &
 Coordination and support actions**

Proposal/Contract no.: 213010 – **e-nmr**

Project full title: Deploying and unifying the NMR e-Infrastructure in System Biology

Project coordinator: Prof. Dr. Harald Schwalbe

Project website: <http://www.enmr.eu/>

Date due: 30-04-2010

Date released: 07-05-2010

Preparation of the 2010 workshop

The present activity represented a follow-up of the assessment carried out in the second project year, which developed into a community-wide initiative aimed at assessing the feasibility of obtaining in an unsupervised manner solution structures of proteins with a quality suitable for direct deposition into the Protein Data Bank (as described in the CASD-NMR manifesto published in *Nature Methods* 6: 325-626, 2009). For the present workshop, only data sets for protein structures not already available to the general public (so-called “blind” data sets) were used. A data set consisted of the assignment of chemical shift values for the protein nuclei and of the integrated NOESY peak lists without assignments; all the lists were the same as those used for calculation of the reference structure, apart from deletion of the assignments in the NOESY peak lists. Ten data sets had been made available over the period June 2009 – February 2010 (to the entire world, not only to CASD-NMR participants). All the data sets had been provided from laboratories involved in the North East Structural Genomics (NESG) consortium (www.nesg.org), which is part of the NIH-funded Protein Structure Initiative (PSI). Seven teams participated in this assessment, of which one managed an automated web server for calculations that is already available via the e-NMR portal. Details on the targets, access to the data and the list of participants are available through the web site of the initiative, which is hosted by e-NMR at <http://www.enmr.eu/CASD-NMR>. Because some structures were calculated more than once using different computational schemes, the total of structures produced and analyzed for this second assessment was 81. The entries were directly deposited by the participants into an online database that is hosted locally on Florence servers.

The workflow of the preparation of the assessment was as follows:

1. blind data sets were released on the CASD-NMR website (and simultaneously on the PSI knowledge base web site). The manually solved (reference) structure was simultaneously deposited in the PDB, with coordinates on hold
2. within eight weeks the participants had to deposit their automatically generated structure(s) in the CASD-NMR database. This db is password-protected so that the participants could not see each other's results before the workshop

3. one month before the workshop, two teams of validators, of which one not involved in CASD-NMR structure calculations, were granted full access to all the data
4. one week before the workshop, full access to all the data was granted also to all participants (but not the general public)
5. all the results were discussed in a workshop held in Florence on May 5-7, 2010

It is worth mentioning that several of the participants had introduced with respect to the 2009 workshop improvements of various kinds to their programs, which lead, as detailed below, to a very high quality of the automatically generated results. In this sense, it is fair to state that the CASD-NMR initiative that e-NMR has launched is already demonstrating a significant impact in pushing the development of programs for automated NMR structure determination. Indeed, some of the participants observed that CASD-NMR provides an unique venue for developers in the field to compare performance, discuss calculation strategies and exchange experiences on the basis of a representative common ensemble of data.

Results

The results of the assessment were evaluated in two manners:

- i - by looking at the stereochemical quality of each structure
- ii - by comparison to the manually solved structures (reference structures).

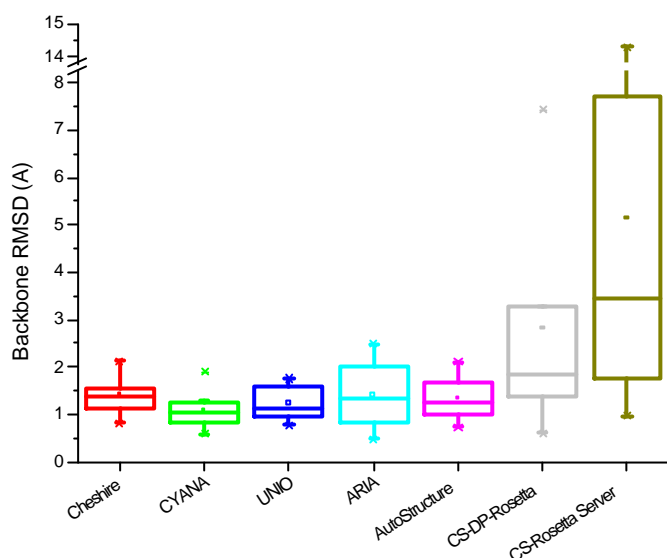
For point i above, two different structure validation tools were used: the PSVS server and the CiNG suite. All analyses with these tools were carried out by their respective developers.

For point ii, a potentially important aspect is which residues are used for structure superimposition. This was extensively discussed by the participants via e-mail, as there are different approaches available in the literature (some of which are also implemented in the aforementioned validation programs or in the structure calculation program themselves) for the automated selection of these residues. The residue ranges proposed by a few different approaches were thus compared and resulted only marginally different.

A consensus was thus established for all participants to enable the comparison of each participant's individual analyses.

Target	Residue range (CYANA)	Residue range (PSVS)	Residue range (CiNG)	Consensus
AR3436A	13-95	13-96	13-95	
AtT13	5-57,70-118	6-119	7-57,62-67,71-117	7-57,71-117
CGR26A	15-129	16-20,23-53,58-71,75-128	15-21,24-71,74-127	16-20,24-53,58-71,75-128
CtR69A	4-53	6-51	5-53	6-51
HR5537A	32-135	37-109,112-134	39-104,117-134	39-104,117-134
NeR103A	18-98	23-82	21-87,90-96	23-82
ET109Aox	91-189	91-136,138-189	91-137,140-154,158-189	91-136,140-154,158-189
PGR122A	418-479	418-444,447-478	418-425,429-478	418-425,429-444,447-478
ET109Ared	91-190	91-189	91-137,140-154,158-189	91-137,140-154,158-189
VpR247	2-100	1-101	2-43,48,52-99	2-43,48,52-99

The results in comparison to the reference structures, with the consensus ranges for RMSD calculations, can be summarized by the boxplot below (the boundaries of the box correspond to the first and third quartile; the line in the box corresponds to the median; the small square in the box corresponds to the mean; the whiskers correspond to 25% - 1.5 times the box size and 75% + 1.5 times the box size; stars indicate outliers):

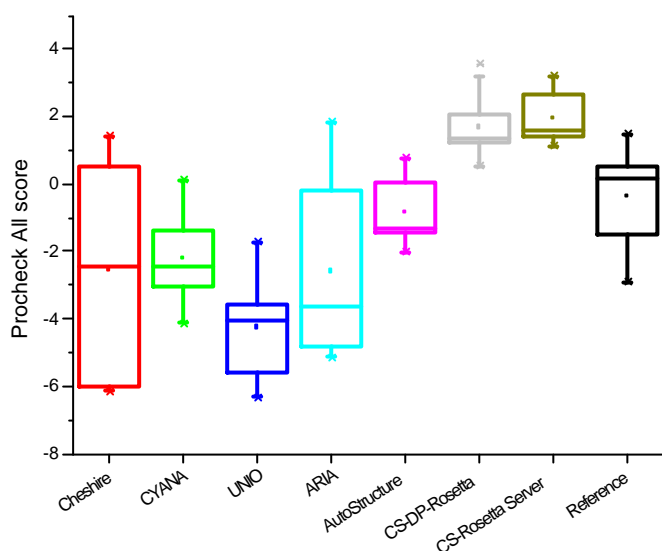


Cheshire and the two cs-Rosetta variants are calculations based on chemical shifts, which can be filtered by agreement with NOESY data (CS-CP-Rosetta and some of the Cheshire entries). Note that some calculations did not converge with Cheshire and therefore there are no corresponding entries.

The above graph shows that using manually refined peak lists, i.e. from which spurious peaks e.g. due to experimental artifacts had been manually removed, calculation programs based on NOESY data (CYANA, UNIO, ARIA, AutoStructure) would typically automatically generate structures within 2 Å from the reference, with median values around 1 Å, i.e. comparable to the indetermination of the reference structures. Chemical-shift based calculations instead has a significantly higher variability in their performance; filtering of the results by comparison to the NOESY data improved the performance. For the NOESY-based programs it is important to note that in some cases the calculation protocols have been improved during the course of this assessment, resulting in more recent targets having been better calculated than the earlier ones.

If the automatically generated structures are evaluated with respect to the references using other measures of structure similarity that do not depend on the definition of residues ranges, such as done in the CASP initiative, the overall conclusions remain unchanged.

For parameters linked to stereochemical quality, such as PROCHECK (see Figure below), the conclusion is that they do not allow users to discriminate between wrong or correct structures



Indeed, it can be seen that the best performances are given by the two cs-Rosetta variants, which perform even better than the reference structures, in spite of their poorer performance in terms of consistently obtaining structures close to the manually solved ones. It is also important to notice that actually most programs tend to have a relatively high variability, which can be in part ascribed to the energy refinement strategy employed after the structure determination itself.

The correlation between pairs of indicators of stereochemical quality was always poor or very poor, with the exception of indicators of global fold correctness such as Prosa II and Verify3D, suggesting that the aspects measured by the different indicators are somewhat independent (i.e. one needs to perform all checks).

A detailed structure-by-structure analysis at the residue level has been performed with CiNG. As in the previous workshop, this analysis has provided a wealth of indications on various errors and inconsistencies of both sporadic and systematic nature. In the present assessment, it was remarked that systematic errors, if any, tended to occur in segments lacking a defined secondary structure content, which can be considered as an improvement with respect to the previous year when the occurrence of wrong packing of secondary

structure elements had been observed. Structural deviations possibly related to the existence of local dynamic processes has also been identified.

Finally it is worth mentioning that CiNG identified a few instances in which the reference structure could have some local inaccuracies that instead were not present in the automatically generated structures.

Detailed tables of results are available on-line and will be reorganized in a new CASD-NMR web site.

Conclusions and future directions

The outcomes of the 2010 assessment indicate that the currently available software tools are well capable of producing in an unsupervised manner structures that are satisfactorily close to the reference structure. The agreement with the data is one of the best indicators of structure correctness, whereas stereochemical quality indicators *per se* are not particularly informative in this respect. On the other hand, these indicators can be very useful at the local level to identify potentially erroneous conformations induced by a wrong interpretation of the experimental restraints, especially in the case of manual analysis. Public dissemination of the present results will take place at the EUROMAR/ISMAR meeting of July 2010.

Because the present calculations started from manually cleaned NOESY peak lists in nearly all cases, all participants agreed that the most logical step forward will be in the use of unrefined (i.e. automatically generated from the spectra) peak lists. This puts more work on the data providers as they will have to go back to the original spectra to generate such peak lists; it appears reasonable that the CASD-NMR organization provide macros with pre-defined parameters to help the data providers. A suitable mechanism to reward the data providers should be defined and put in operation. It is expected that a consistent flow of data will continue to be provided by the NESG, as it is in a good position to receive funds from the NIH to continue its operations. If possible, preference will be given to targets that can be used for analysis of the kind that users could perform, such as comparison of different ligand-bound forms.

It is agreed that the next workshop will take place after completion of the analysis of approximately ten new targets. The current mechanism for target distribution and deposition will be maintained.

Note:

The 2009 CASD-NMR manifesto has been described in the following newsletters/highlights:

<http://www.genomeweb.com/informatics/new-critical-assessment-project-targets-software-nmr-based-protein-structure-det?page=show>

[http://news.eu-egee.org/index.php?id=193&tx_ttnews\[swords\]=nmr&tx_ttnews\[tt_news\]=120&tx_ttnews\[backPid\]=194&cHash=d275d4dd2e](http://news.eu-egee.org/index.php?id=193&tx_ttnews[swords]=nmr&tx_ttnews[tt_news]=120&tx_ttnews[backPid]=194&cHash=d275d4dd2e)

http://kb.psi-structuralgenomics.org/update/2010/06/full/th_psisgkb.2010.25.html

List of Attendees

A. Bagaria, University of Frankfurt, D
A. Bonvin, Utrecht University, NL
A. Cavalli, University of Cambridge, UK
A. Giachetti, University of Florence
P. Guentert, University of Frankfurt, D
P. Guerry, ENS Lyon, F
T. Herrmann, ENS Lyon, F
Y. Huang, Rutgers University, USA
C. Luchinat, University of Florence
T. Malliavin, Institut Pasteur, F
G. Montelione, Rutgers University, USA
M. Nilges, Institut Pasteur, F
A. Rosato, University of Florence
G. van der Schot, Utrecht University, NL
G. Vuister, University of Nijmegen, NL