SEVENTH FRAMEWORK PROGRAMME

Research Infrastructures

Deliverable 7.3

A functional and tested data and structure validation platform

Grant Agreement number:	RI-261572
Project acronym:	WeNMR
Project title:	WeNMR, A worldwide e-Infrastructure for NMR and structural biology
Deliverable type and status	Demonstration – Public
Due Date	November 1st, 2012
Delivery Date	December 24th, 2012
Project coordinator name, title:	Prof. Alexandre M.J.J. Bonvin
Organization:	Utrecht University, The Netherlands
Tel:	+31 30 2533859
Fax:	+31 30 2537623
E-mail:	a.m.j.j.bonvin@uu.nl
Project website address:	www.wenmr.eu

Table of Contents

Section 1: Summary of Deliverable

Background

Experimentally determined biomacromolecular three-dimensional (3D) structures typically are deposited in the Protein Data Bank (wwPDB). As of November 2012, there are more than 90,000 entries in the PDB of which ~9500 have been solved by Nuclear Magnetic Resonance (NMR). The Biological Magnetic Resonance Bank (BMRB) serves as a global repository of experimental NMR data, such as restraints, assigned chemical shifts and dynamic order parameters.

The determination of protein structures by NMR spectroscopy involves an elaborate process with many imperfect experimental and computational steps that often are based on largely empirically determined procedures. Consequently, the resulting structures are prone to errors that inevitably become part of the NMR ensembles deposited in the PDB. Not surprisingly therefore, analyses by us and others have shown that NMR structures are often flawed.

NMR structure validation differs substantially from validation of X-ray structures. The primary differences between X-ray and NMR originate from the nature of the experimental data and the computational protocols employed in the structure calculation process. Whereas the X-ray experimental data (the reflections) are well-defined and uniform in structural information content, experimental NMR data originate from different sources and report on different properties.

Goal

This document constitutes the demonstration of deliverable D7.3, “A functional and tested data and structure validation platform”, due by Month 24. It is intended to be the report on the results of the project in WP7 aimed to develop an integrated approach for the validation of the results produced by the structure generation web portals. The CING validation framework and accompanying iCing server are the results of this effort.

Section 2: The CING software framework for structure validation

The CING software package

The CING (Common Interface for NMR structure Generation; pronounced 'king'; http://nmr.cmbi.ru.nl) software package constitutes a framework for the validation of NMR structures (Fig. 1). CING can accommodate both the diversity of the experimental NMR data, and is also capable of properly handling multi-model NMR ensembles in the analysis routines. CING first assembles a set of experimental and structural data and then produces a report that includes the results of ~25 different, both internal and external, programs and routines. The quality of the 3D structure ensembles is assessed in relation to a database of reference structures using WHAT_IF, PROCHECK_NMR and internal routines. The experimental data are tested for consistency and agreement with the ensemble. Validation of chemical shift values based on structural- and sequence information (VASCO) and SHIFTX is an integral part of the analyses. The ensemble is analyzed for secondary structure (DSSP), solvent accessibility, potential disulfide bridges and salt-bridges.

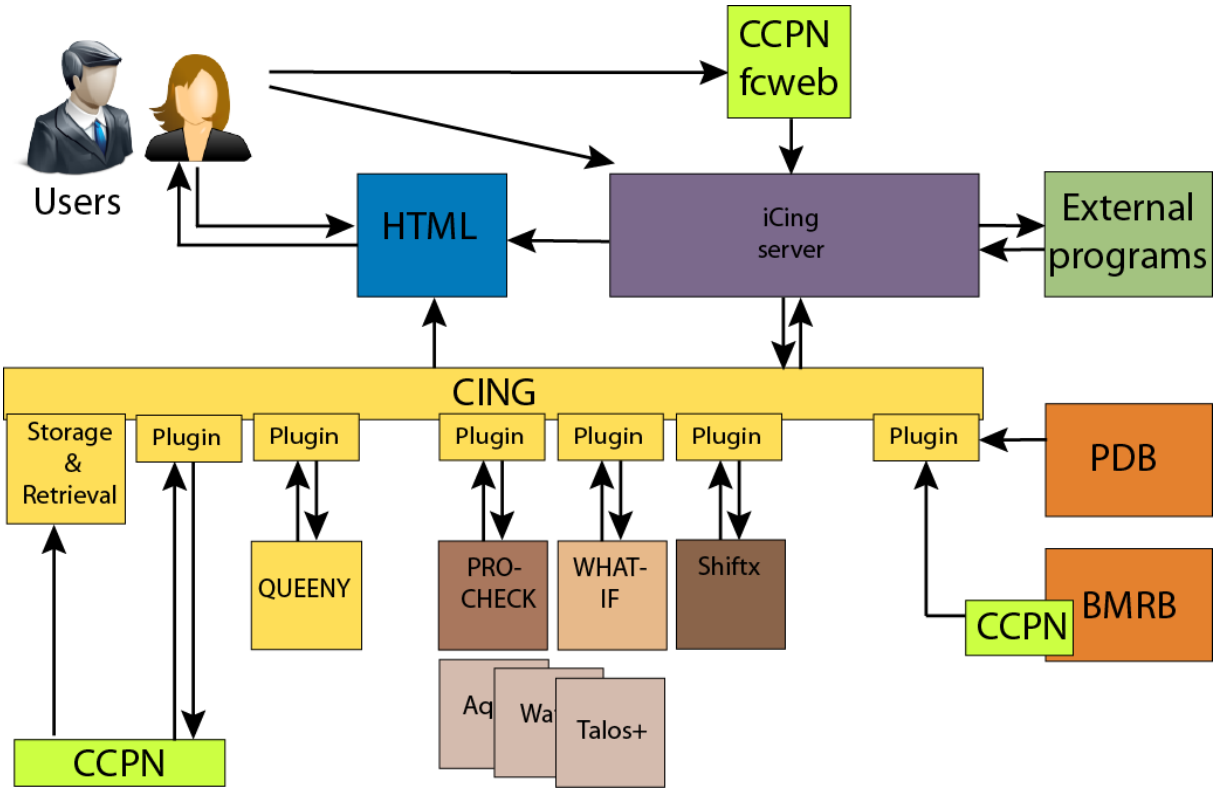

Figure 1. Schematic outline showing the data flow and software components involved in CING (yellow boxes). External programs interfaced to CING, CCPN services/APIs and wwPDB repositories are represented in brown-shades, green and dark-orange, respectively. External programs (olive green) can access the iCing web user interface (purple) through the dedicated icing robot. Figure taken from Doreleijers et al. 2012.

CING generates easily accessible, comprehensive, interactive HTML/Javascript-based validation reports and directs the NMR spectroscopist to troublesome areas. CING uses a simple Red-Orange-Green (ROG) scoring, much like a traffic light at any level of the data, to direct the focus of the author (Fig. 2). The ROG scoring is dependent upon the combined analysis of all results and allows CING to summarize the important issues. A red coloring indicates some potential serious issues, green denotes the absence of any detected issue and orange (amber) is intermediate between these two situations. In addition to this classification mechanism, CING displays the validation results in the direct relation to the experimental data.

Within the CING philosophy, a large emphasis is placed upon the validation of the individual residues. NMR assignment strategies are almost exclusively residue-based, NMR related parameters are residue dependent and the local nature of the NMR-derived restraints also correlates well with a residue-based approach. Structural properties can also be conveniently summarized at the residue level. We previously showed that structurally bad regions are masked when using overall validation parameters, which can also be circumvented by the residue-based approach. Hence, the residue-specific HTML pages and ROG scores presents the ultimate representation of this analysis.
Although the CING program is specialized for NMR-derived biomolecular structures it is not limited to these. For example, the code base is frequently running on ~500 X-ray structures to derive the database frequencies used for shading the Ramachandran and Janin plots.

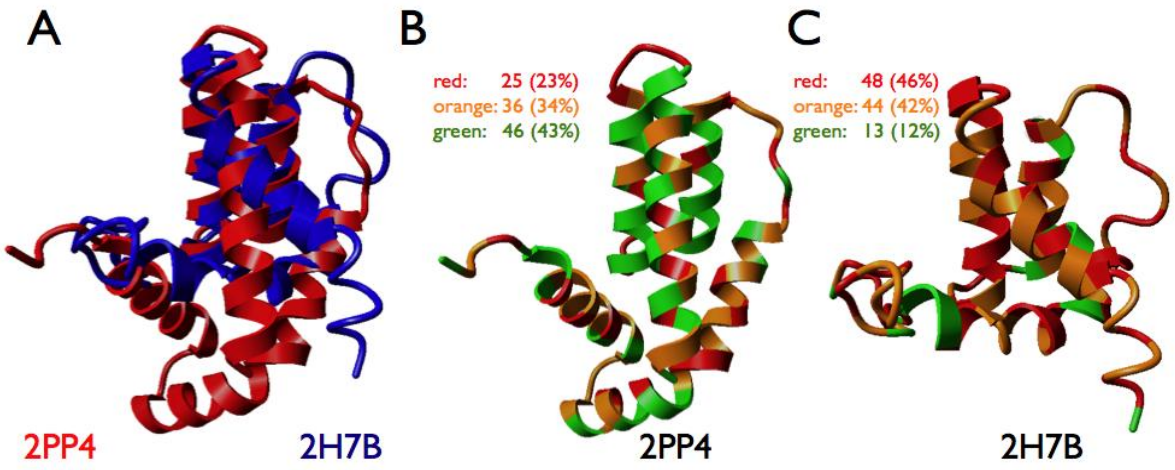

Figure 2. Ribbon representations of the TAF homology domain, PDB entries 2PP4 and 2H7B. (A) Approximate superposition. The current orientation faces the region with the largest structural differences between the two structures. 2PP4 (B) and 2H7B (C) color-coded according to the residue-specific ROG scores; overall residue counts and percentages are listed.

The iCing webserver

The CING multilingual web server and web service together are called iCing. (https://nmr.cmbi.ru.nl/icing) (Fig. 3). The iCing server has been implemented using the Google Web Toolkit (http://code.google.com/p/google-web-toolkit). The technology allows the generation of a Web 2.0 enabled graphical user interface (GUI) that works on any modern version of browsers such as Firefox, Internet Explorer Google Chrome or Safari without actually having to write different Javascript versions for each platform. The resulting CING report should be thought of more as a program than as a collection of static HTML pages. The user can interact with the report in several ways using Web 2.0 Javascript functionality such as provided by JQuery (http://jquery.com) and its plugin Datatables (http://www.datatables.net). CING already has connectivity to four visualization programs, i.e. Jmol, MolMol, PyMol and YASARA.

CING is open as a platform to third-party applications, through web services that interact through the iCing server, employing the so-called iCing robot. XML formatted data of the most important quantities derived by CING, including the per-residue ROG score, are returned in response to specific queries of the third party applications. Users of CCPN Analysis already routinely can use this facility by directly initiating a CING validation run from within the program, combined with automated retrieval and display of the results. Similar usage is currently been developed for UNIO and Xplor_NIH.

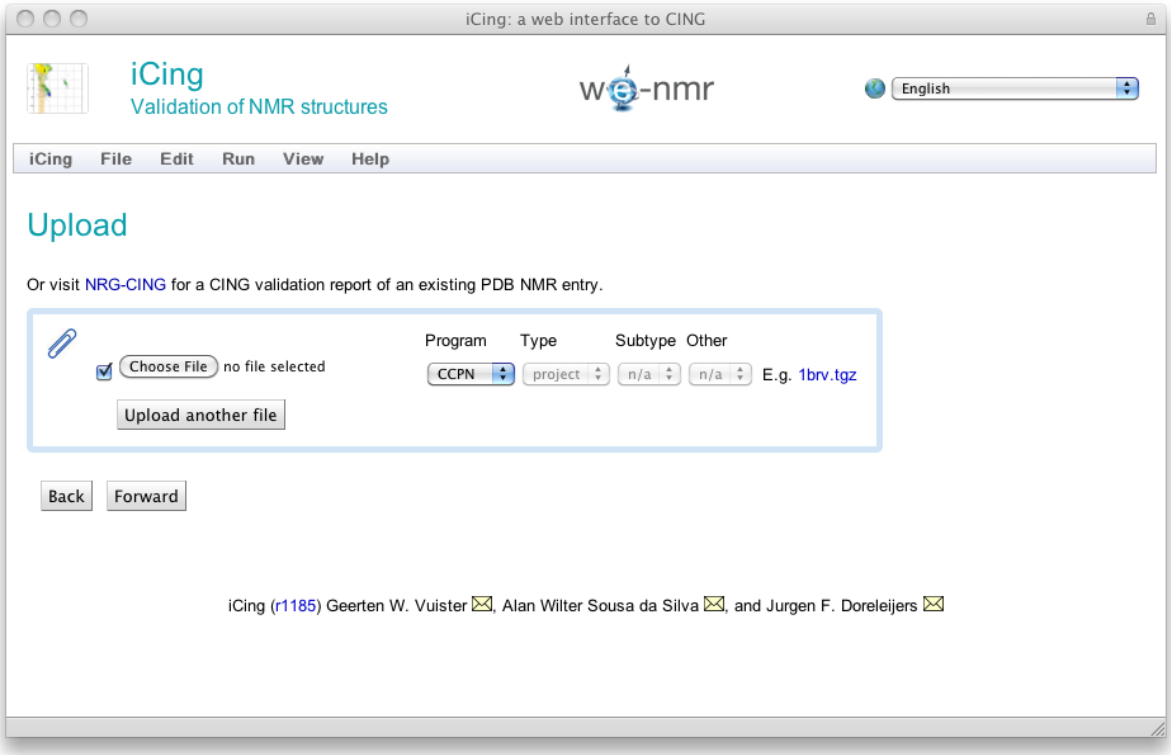

Figure 3. Multilingual iCing server entry page (https://nmr.cmbi.ru.nl/icing).

CING and iCing inputs

CING and the iCing server natively accept PDB, Cyana and CCPN formatted data and have import facilities for other types. Clearly, PDB only data do not contain the experimental input data, and thus prevent all CING analyses with use this kind of information. Both experimental Cyana data (shifts, peaks, restraints) and Cyana structural results can be imported. However, we advocate the usage of CCPN formatted data as this format can accommodate the whole variety of the diverse NMR data and a tested, actively maintained conversion tool is available (http://webapps.ccpn.ac.uk/fcweb/).

CING API

CING also constitutes an 'Application Programming Interface' (API), that allows for the analysis of all data, i.e. experimental, structural and validation data, using high-level functional routines. This setup allows for the easy testing and implementation of new algorithms by both developers and users of the software package.

Section 3: Conclusions

Up to now, it has been common practice to perform only a limited validation of NMR derived biomolecular structures and this analysis was not executed until after release of the wwPDB data to the public. The CING software tools now allow the individual users to perform a complete analysis routinely during the process of structure determination and thus monitor their progress in terms of the quality of their results. As a result, data submitted to the PDB and BMRB databases could be consistent and validated *a priori*, thereby reducing the burden of validation for the two repositories, while simultaneously providing improved structures. Analysis of the (anonymous) iCing server runs indicate that our users do use the server on multiple occasions, presumably to test successive rounds of their structural ensembles *en route* to the final result to be deposited and published.

Reference

Doreleijers, J.F., Sousa da Silva, A.W., Krieger, E., Nabuurs, S.B., Spronk, C.A.E.M., Stevens, T.J., Vranken, W.F., Vriend, G., and Vuister, G.W. (2012). CING: an integrated residue-based structure validation program suite. J Biomol NMR *54*, 267–283.