



Deliverable 3.2

Overview of software for nucleic acids

SEVENTH FRAMEWORK PROGRAMME Research Infrastructures

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: 31-01-2009

Date released: 31-01-2009

Overview of software for DNA and RNA structure validation:

Assessment of existing software, strategy and project planning proposal

Chris Spronk and Geerten Vuister

Existing nucleic acids structure validation software

In contrast to protein structure validation software, which is now highly developed and widely used, "true" nucleic acids structure validation software is currently virtually non-existent. The main reasons for the absence of structure validation methods and software for RNA and DNA are the following:

- The nature of experimental methods used to determine 3D-structures of nucleic acids:
 - NMR: intrinsic lack of experimental information due to the low density of protons in nucleic acids.
 - X-ray: Accurate determination of sugar pucker and backbone structure requires very high resolution structures, which are largely unavailable currently. Base pairing is not a problem, but the high degree of freedom caused by the large number of torsion angles in the phosphate backbone makes it extremely difficult to model the backbone at atomic accuracy in the electron density. The local structure in these underdetermined regions are usually modelled with restraints, which in many cases have been shown to be inaccurate (Robbie Joosten, CMBI, personal communication)
- Limited availability of accurate structures in data bases. A sufficient number of accurate structures is essential for the construction of a high quality reference data base for structure validation that is based on comparison with data base derived distributions.

Some data bases have been constructed from a large set of structures, and can be used as a starting point. However a critical review of the criteria used to construct these data bases, and thus their usability is needed. An example is the RNA DB [1].

Given this lack of good reference data bases and software, structure validation of nucleic acids is therefore largely limited to methods that derive descriptive parameters for the structure of the molecule. These programs may help in identifying abnormalities or suspicious regions, but are however not true validation methods. An example of a program that provides structure descriptions is the recent X3DNA [2], which includes the functionality and improvements over the older software packages CURVES [3] and SCHNAAP [4] (both obsolete), and provides parameters that describe base-pairs, step and local helical features. For an overview of existing software that is considered useful for nucleic acid structure analysis and validation, see Table 1.

Validation based on structure characteristics

One of the most useful and promising programs available is MolProbity [5]. The main functionality of MolProbity is an all-atom contact analysis for nucleic acids and their complexes. It analyses, adds and optimises the hydrogen positions. The all-atom contact analyses are used to identify clashes and thus problems in the structures. MolProbity analyses can be used to improve the phosphate backbone, and a similar analysis for proteins has been shown to improve measures of fit in X-ray structures.

Validation against experimental restraints

Validation of nucleic acid structures against experimentally derived restraints can be performed using the traditional methods implemented in the structure calculation programs that are used to calculate the structures, such as XPLOR [6], CNS [7,8] and CYANA [9]. Alternatively, these analyses can be performed in the CING package, a new software suite that integrates many different structure validation and NMR restraint analysis tools (in development). The usefulness of such analyses is very limited, particularly in the case of nucleic acids with its low experimental information content. Usually it is no more than a measure of fit to the data, and will not provide a lot of information on the quality of the final structure after curation of the input experimental restraint in successive rounds of structure optimisation. The most useful method for structure validation using experimental restraints is the application of complete cross-validation or independent validation. Given the low information content in nucleic acid NMR data, it is highly unlikely that researchers will apply these methods in practice.

It should be mentioned here that the program QUEEN [10] can in principle be used for nucleic acids NMR restraint validation based on information content. However, applying the method will be non trivial, due to the different nature of nucleic acid structures as compared to protein structures (Sander Nabuurs, CMBI, personal communication). The application of QUEEN to nucleic acids is also scheduled in the CING development.

Validation against chemical shifts

Prediction of chemical shifts for both DNA (DSHIFT [11]) and RNA (NUCHEMICS [12]) have been shown to perform rather well, although at this point it is not clear whether the existing software and underlying methods and data allow application to all structural variants, in particular in RNA. Where applicable, the predictability of nucleic acids NMR chemical shifts may prove to be one of the most useful structure validation tools available. Implementation of these methods is considered relatively easy. As a specific example, AA-mismatch chemical shift predictions for DNA have been developed as well, although no software was found for it yet. This should be low priority initially, given its highly specific nature.

Validation using secondary structure prediction methods

Secondary structure prediction methods in proteins have proven to be a very useful tool for interpretation of NMR data, automatic resonance assignments, and in validating NMR derived structure models. The accuracy of protein secondary structure prediction is sufficient to identify potentially wrong folds. For nucleic acids, a number of online services are available for predicting secondary structures and nucleic acid folding. These methods should be considered as potential candidates for use and further development in the field of NMR structure validation. Automation of such methods however may pose a problem, since parameters need to be set by specialists. The main application of these methods in structure validation probably will come initially from parsing output files provided by users. Several

online servers are available, of which the main have been developed by Zuker and co-workers [13,14], see also: <http://www.bioinfo.rpi.edu/applications/mfold/>.

Validation using database potentials

A potentially interesting means for nucleic acid structure validation could be evaluation of structures using database potentials as developed by Clore and Kuszewski [15,16]. The XPLOR implementation of these database potentials may provide powerful means to do a per-residue evaluation of the energy of the structure in the data base potential, even if the underlying data bases would be generally of limited use for structure validation. High energies could be used as indicators for inspection of the structure locally.

Planning and strategy

Based on the overview above, current developments in structure validation software and the available resources, we have chosen to use the latest structure validation software package CING as the basis for implementation of interfaces to existing and development of new methods. CING contains a large number of tools for analysis and graphical representation of validation results. Examples are the calculation of the equivalents of the Janin Chi1-Chi2 plots [17] for all dihedral angles in nucleic acids. Therefore, integration and rendering of output from external programs are considered relatively trivial procedures.

We see 2 main challenges in the project:

1. The selection of methods that could be useful for structure validation (Table 1), and assessing in which cases selected methods can be applied sensibly. Given the limited time of the project, focus should be on a selection structure based validation methods and chemical shift predictions:
 - a. MolProbity
 - b. X3DNA
 - c. Suitename
 - d. DSHIFT
 - e. NUCHEMICS
2. The construction and or selection of reliable reference data bases and test cases for structure based validation. As good starting points for this are considered:
 - a. the RNA DB from Richardson (<http://kinemage.biochem.duke.edu/databases/rnadb.php>)
 - b. The Ribosomal Database Project (RDP) (<http://rdp.cme.msu.edu>)
 - c. A DNA data base could be constructed from the PDB or the Nucleic Acid Data Base (NDB) (<http://ndbserver.rutgers.edu/index.html>)

The project should focus in the beginning on creating or completing where needed the infrastructure in CING that is required for the integration of the software packages mentioned under 1, starting with X3DNA and MolProbity. Further, at an early stage criteria should be defined for construction of a reference data base for nucleic acids, in particular for DNA, and the creation of tests sets. For RNA, the existing RNA DB can be used to create initial distributions for Janin plots for validation of RNA structures. We consider the following work to be feasible within the time frame available for the project, and which would lay the basis for further development after the project end:

1. Creation of:
 - a. 10 data sets, used for validation and cross validation
 - b. RNA data base
 - c. DNA data base

2. Implementation of CING interface to:
 - a. X3DNA
 - b. MolProbit
 - c. DSHIFT
 - d. NUCHEMICS

The feasibility of the application of secondary structure predictions and other methods mentioned should be evaluated at a later stage in the project. Table 2 lists the intended task division among the partners.

Note: Some issues that need addressing before the start of the project are:

Will we limit ourselves to A-, B-, Z- forms initially? What about Pseudoknots, hammerheads etcetera?

Table 1. Overview of nucleic acid structure validation software

Structure based validation software	
WHAT IF [18]	<ul style="list-style-type: none">• Bond lengths and angles• Hydrogen bonds (see below)
MolProbity	<ul style="list-style-type: none">• All atom contact analysis• Identification of problematic regions• Hydrogen bonds (see below)
XPLOR	<ul style="list-style-type: none">• Use of database potentials to identify abnormal conformations• Feasibility study is needed
Structure description software	
X3DNA	<ul style="list-style-type: none">• Base pairing• Step parameters• Local helical parameters
Suitename (MolProbity)	RNA backbone type classification in 46 discrete sugar-sugar "suite" conformers
CURVES	see X3DNA, obsolete
SCHNAAP	see X3DNA, obsolete
Experimental restraint validation	
Structure calculation programs	All types of restraints
<ul style="list-style-type: none">• XPLOR• CNS• CYANA• CING	
RDC specific	All types of restraints (RDCs in development) in development in CING
<ul style="list-style-type: none">• MODULE2 [19]• PALES [20]• REDCAT [21]• iDC [22]	
Chemical shifts	
DSHIFT	DNA
NUCHEMICS	RNA
Hydrogen bonding	
WHAT IF	<ul style="list-style-type: none">• Unsatisfied buried hydrogen donors and acceptors• Needs extensive testing for nucleic acids
MolProbity	Adds, optimizes hydrogens and calculates hydrogen-bonds
Secondary structure prediction	
UNAFold [23]	<ul style="list-style-type: none">• Simulates folding, single stranded RNA or DNA hybridization, and melting pathways for one or two single-stranded nucleic acid sequences• Command line tool
RNAfold server [24]	<ul style="list-style-type: none">• Predicts secondary structures of single stranded RNA or DNA sequences• http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi
DINAMelt server [25]	<ul style="list-style-type: none">• Prediction of Melting Profiles for Nucleic Acids• http://dinamelt.bioinfo.rpi.edu/twostate.php
Pseudoknot prediction	
PLMM_DPSS [26]	http://bioinformatics.ist.unomaha.edu:8080/x/PLMM_DPSS.html
vsfold5 server [27]	http://www.rna.it-chiba.ac.jp/~vsfold/vsfold5/
ILM [28]	http://cic.cs.wustl.edu/RNA/

Table 2. Task division

Tasks	Participant
Selection of external software	eNMR/Nijmegen
Reference data bases	
Criteria	eNMR/Nijmegen
Construction (structure selection)	eNMR
Test data sets	eNMR
Infrastructure in CING	Nijmegen (eNMR?)
Creation graphical representations of reference data	Nijmegen

eNMR: Chris Spronk; Henry Jonker

References

- [1] Murray LJ, Arendall WB 3rd, Richardson DC & Richardson JS (2003). *Proc. Natl. Acad. Sci. U S A.* **100** (24), 13904-13909.
- [2] Lu XJ & Olson WK (2003). *Nucleic Acids Res.* **31** (17), 5108-5121.
- [3] Lavery R & Sklenar H (1988). *J. Biomol. Struct. Dyn.* **6** (1), 63-91.
- [4] Lu XJ, El Hassan MA & Hunter CA (1997). *J. Mol. Biol.* **273** (3), 668-680.
- [5] Davis IW et al. (2007). *Nucleic Acids Res.* **35**, W375-W383.
- [6] Brünger AT. X-PLOR (1992). A system for X-ray Crystallography and NMR, Yale University Press, New Haven, CT.
- [7] Brünger AT et al. (1998). *Acta Cryst.* **D54**, 905-921.
- [8] Brünger AT (2007). *Nature Protocols* **2**, 2728-2733.
- [9] Güntert P (2004). *Methods Mol. Biol.* **278**, 353-378.
- [10] Nabuurs SB, Spronk CAEM, Krieger E, Maassen H, Vriend G & Vuister GW (2003). *J. Am. Chem. Soc.* **125**, 12026-12034.
- [11] Lam SL (2007). *Nucleic Acids Res.* **35**, W713-W717.
- [12] Wijmenga SS, Kruithof M & Hilbers CW (1997). *J. Biomol. NMR* **10**, 337-350.
- [13] Zuker M (2003). *Nucleic Acids Res.* **31** (13), 3406-3415.
- [14] Mathews DH, Sabina J, Zuker M & Turner DH (1999). *J. Mol. Biol.* **288**, 911-940.
- [15] Clore GM & Kuszewski J (2003). *J. Am. Chem. Soc.* **125** (6), 1518-1525.
- [16] Kuszewski J, Schwieters C & Clore GM (2001). *J. Am. Chem. Soc.* **123** (17), 3903-3918.
- [17] Janin J & Wodak S (1978). *J. Mol. Biol.* **125** (3), 357-386.
- [18] Vriend G (1990). *J. Mol. Graph.* **8**, 52-56.
- [19] Dosset et al. (2001). *J. Biomol. NMR* **20**, 223-231.
- [20] Zweckstetter M & Bax A (2000). *J. Am. Chem. Soc.* **122**, 3791-3792.
- [21] Valafar H & Prestegard JH (2004). *J. Magn. Reson.* **167** (2), 228-241.
- [22] Wei Y & Werner MH (2006). *J. Biomol. NMR.* **35** (1), 17-25.
- [23] Markham NR & Zuker M (2008). *Methods Mol. Biol.* **453**, 3-31.

- [24] Hofacker IL. (2003). *Nucleic Acids Res.* **31** (13), 3429-3431.
- [25] Markham NR & Zuker M (2005). *Nucleic Acids Res.* **33** (Web Server issue):W577-W581.
- [26] Huang X & Ali H (2007). *Nucleic Acids Res.* **35** (2), 656-663.
- [27] Dawson W, Fujiwara K, Kawai G, Futamura Y & Yamamoto K (2006). *Nucleosides, Nucleotides & Nucleic Acids* **25** (2), 171-189 (19).
- [28] Ruan J, Stormo GD & Zhang W. "ILM: A Web Server for Predicting RNA secondary structures with Pseudoknots", Submitted.