CCPNGrid – A Portal for Biomolecular Structure Calculation from NMR Spectroscopy Data

Alan da Silva[1], Wim Vranken[2], Mark Hayes[3], Michael A. Parker[4], Ernest D. Laue[1]

[1] University of Cambridge, Dept. of Biochemistry
[2] European Molecular Biology Lab, European Bioinformatics Institute, MSD Group
[3] University of Cambridge, DAMTP, Cambridge e-Science Centre
[4] University of Cambridge, Dept. of Physics, Cavendish Laboratory

Abstract

CCPNGrid portal (*http://www.ccpn.ac.uk/ccpn/projects/ccpngrid*) was conceived by Collaborative Computing Project for the NMR community (CCPN) & Cambridge/E-science Centre to be a part of a stable and permanent framework for executing state-of-the-art structure calculations and validation software on a computer grid for biological macromolecules using Nuclear Magnetic Resonance (NMR) spectroscopy data. Among various objectives of this project one can mention: (a) enable researchers to access and execute state-of-the-art structure calculation methods; (b) encourage the use of the CCPN project resources by providing online services based on this framework. This project is founded on CCPN data model framework and intends to be another example of how this model can hugely facilitate the development, integration and use of applications designed for NMR community. To work out the best approach we have implemented established automated NOE assignment and NMR structure calculation software (ARIA/CNS) using as input information stored in the data model framework provided by CCPN over CamGrid (the campus Condor grid at the University of Cambridge). Although the Portal is under constant development and may not fit for every desired case, it is active and ready for production. Project funded by BBSRC grant BB/D006384/1.

Introduction

Nuclear Magnetic Resonance (NMR) has developed into a key experimental tool for determining the 3D atomic structure of biomolecules. The two main steps that determine the speed with which biomolecular NMR data can be processed are the extraction and analysis of information from NMR spectra, and the subsequent 3D structure calculation.

CCPN [1] is part of a consortium which intends to provide the scientific NMR community with the means to process, analyse, execute state-of-the-art 3D-structure calculation and validation software, so that the quality and scientific value of structural coordinates from NMR can be improved (Figure 1).

The application of this project requires large computational resources, because a structure calculation from NMR data is an intensive iterative process that involves cycles of structure calculation and reinterpretation of experimental data in light of the intermediate structures.

Because the NMR data is incomplete, each iteration requires the calculation of a large number of structures (typically 100 or more), so that the landscape of possible solutions is adequately sampled.

Initially, this project was implemented on CamGrid [2], a network of workstations & compute clusters at the University of Cambridge managed by Condor middleware [3] featuring a couple of hundred CPUs. In the longer term the framework we develop will be implemented on the JISC National Grid Service (NGS) [4] for high performance computing, as well as compute clusters available to other NMR groups, e.g. at UCL and at Glasgow (ScotGRID).

The Portal in particular is our first approach to integrating some applications by hiding from users some major chores from the whole NMR structure calculation process and hence job submission to a computer grid. It can be accessed at *http://www.ccpn.ac.uk/ccpn/projects/ccpngrid*.

Aims of CCPNGrid

CCPNGrid was started as a BBSRC eScience pilot project, and a skeletal framework that proves that the basic concept works is now available and being extended with a range of relevant

software in order to have a well defined workflow to really support the scientific community and help them in the complex task of determining and validating biomolecular structures. CCPNGrid will continue with the open source tradition of CCPN and will be freely available. It addresses two key issues that scientists face when confronted with calculations using NMR data. The first is that keeping the relevant software up-to-date, and having the expertise to use it, requires considerable time and effort. The second is that not all laboratories have the computational resources to execute these calculations properly, including computer resources. Our aim is thus to make NMR more accessible to both specialists and non-specialists, within and outside the UK, and in the process provide a tool to educate researchers on the latest structure calculation and validation tools, as well as improve the quality of the structures that finally get deposited at archive databases like the wwPDB [5].

For example, a biologist with an important project may want to study his/her own protein/complex. He/she may obtain suitable NMR spectra from one of the UK, European or International NMR Centres, but will require help to analyse that data and calculate the structure. In the CCPN project we are developing state-of-the-art software for these purposes, but there is a need to make it more accessible to the expert and non-expert alike. CCPNGrid will also give users the opportunity to try out different software and approaches without having to go through a complete local software installation process. We also want to extend the range of molecules that can be handled by NMR, increasing the field of interest to, for example, protein-ligand, protein-drug and protein-carbohydrate complexes.

The way in which we want to make the improved tools we are developing more accessible to the scientific community is by providing a portal to these tools that is simple and straightforward to use. This project will therefore provide a central calculation facility for smaller groups, who either don't have suitable computing resources locally, or the manpower to keep all of the software up-to-date. It will also enable larger NMR laboratories, with their own compute clusters and software specialists, to install a copy of CCPNGrid for internal and/or external use. The importance of such an effort is clear: if state of the art calculation and validation software is made more readily available to scientists, then this will directly affect their ability to use NMR spectroscopy and, it will improve the quality and scientific value of the structures they calculate. An additional advantage is that their data will be stored in one single project using the CCPN data model framework. This not only provides the researcher with an archive of the structure determination project, but it also facilitates exchange of their data with other scientists. The prime public international archiving site for biomolecular structures (the wwPDB) will also, via the MSD group at the EBI, provide a portal so that it will be possible to deposit CCPN project files and extract all the relevant information from them. This deposition process can thus be initiated after a CCPNGrid calculation, with minimal additional effort from the user.

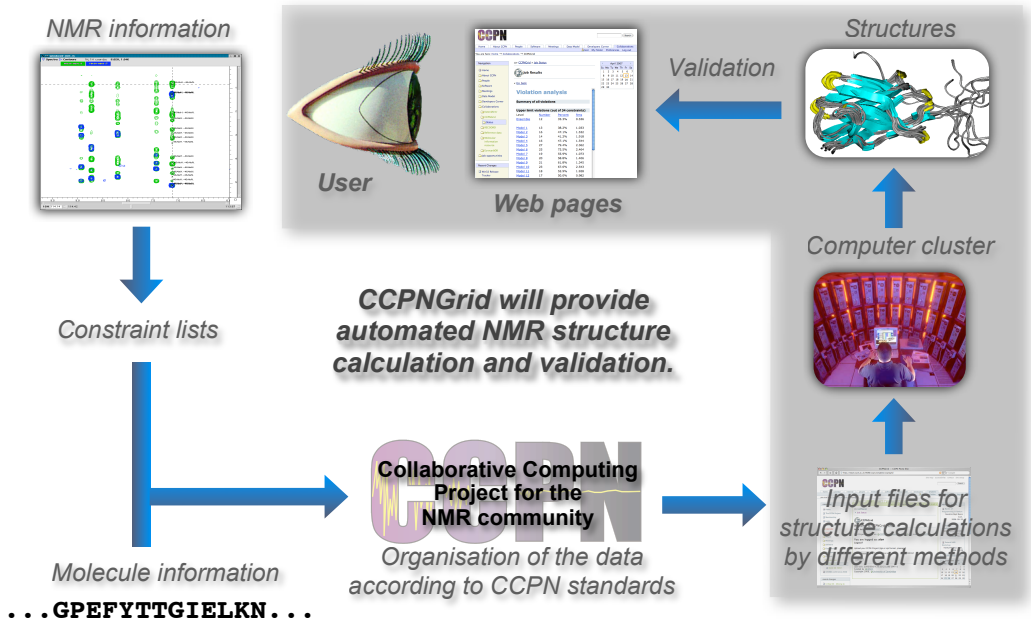

Figure 1 - NMR structure calculation process with the current CCPNGrid setup (detailed in gray).

Portal Description

To work out the best approach we have initially implemented established automated NOEs (Nuclear Overhauser Effect) assignment and NMR structure calculations (ARIA [6] / CNS [7]) using input information stored in the data model framework provided by CCPN (see Figure 1). This procedure runs using a Condor grid compute cluster and it is interfaced by a web application running on Plone/Zope server platform. This platform provides authentication, maintenance, layout framework and productivity tools that simplify the task of implementing, maintaining and developing our interface application. Moreover, the code behind CCPNGrid is written in Python which facilitates the integration between core applications with Plone/Zope, being all built in Python. In addition to the portal functionality, the following products developed for Plone/Zope has to be installed: LocalFS and ZMySQLDA.

The portal may virtually run in any *NIX system with basic shell scripting. So far, CCPNGrid has been implemented in a server with Solaris 5.9 Sun Sparc and tested with Apple Mac OSX 10.4.10. It should also work well with any Linux platform. However, although our intention is to make CCPNGrid portable to other systems and sites, it depends on many variables and some 3rd party software, so that a specific effort will have to be made to make its installation as simple as possible.

Currently CCPNGrid is set up to handle structure calculations from NMR data using the ARIA 2.2 program (under development), which depends on restrained molecular dynamics calculations in the CNS program (version 1.2). ARIA obtains the NMR and molecular system data directly from a CCPN project, and executes a series of calculations to determine an ensemble of 3D structures before storing the results back into the CCPN project. The structures are then validated using the PROCHECK [8] and WHAT_CHECK/WHAT IF [9] programs. There is also a python module called violationStatistics developed by Wim Vraken, originally for RECOORD Project [10], that uses the program R for carrying out some statistical calculations over the structures generated. This module is another great example of how CCPN data model framework enhances the development and integration of 3rd party applications.

The core program here for grid computation is ARIA and CNS is the core engine for structure calculation. ARIA is flexible enough to cope with different types of cluster's architectures and it is already able to run CNS on Condor (and may be on Globus via Condor-G), Mosix/OpenMosix grid/clusters, Beowulf clusters and multi-core/ processors computers. However, to take advantage of idle computing resources usually available in a e.g. university campus a sophisticated middleware is necessary. At University of Cambridge Condor is the option.

The others programs used by the portal are Python 2.3.5, CCPN API and Analysis [1], MySQL, RPy, Ghostscript, Moscript and Dssp.

Further Steps

The main objective henceforth of this project is to maintain, extend and improve the CCPNGrid resource that was developed during the BBSRC eScience pilot project BB/D006384/1, whose support we acknowledge.

This basic setup has been developed as a proof of concept. We now need to add other applications for automated structure calculation from NMR data to this setup, using the same basic framework, to make the whole CCPNGrid resource increasingly valuable (see Figure 2). Not only will different applications be connected together to validate or extend the obtained results, but it will also be possible to run similar applications in parallel, so that the results obtained (from the same data) can be compared and examined by the user. Calculation of structures in this way will provide important validation of the results and help identify which regions of the structure are not well determined by the data, and which therefore require further manual analysis.

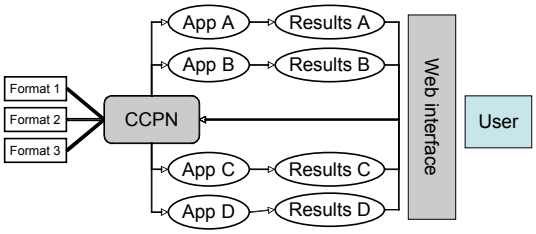

Figure 2 - The workflow behind CCPNGrid. Input data for the applications is taken from the CCPN framework (which can be populated from many external formats). The results are forwarded to the user via the web interface, but are also stored in the CCPN framework. Results from one application can be used as input for another, leading to logical workflow components (e.g. a structure calculation followed by structure validation).

With that in mind a likely application to be added to CCPNGrid resources is the validation tool QUEEN [11], which relies on XPLOR-NIH [12] for calculations and now features CCPN data model integration. It assesses the information content of NMR distance restraints, both on an ensemble and a per-residue basis. It can identify the more important input data (distance restraints) in a structure calculation, and those restraints that might be incorrect. It also has sig-

nificant computational requirements and will promptly benefit from a grid facility.

Further developments for CCPNGrid are rather linked to ARIA development, such as being able to cope with complexed biomolecular systems (e.g. protein-ligand) and which plans to support other programs such as CYANA [13] and XPLOR-NIH. However it also works on the other way, since this project had contributed to ARIA by enabling it, viz., to use Condor and WHAT_CHECK (instead of only WHAT IF).

References

[1] A framework for scientific data modeling and automated software development. R Fogh et al. Bioinformatics (2005), 21(8), 1678-1684. http://www.ccpn.ac.uk/

[2] CamGrid: http://www.escience.cam.ac.uk/projects/camgrid/

[3] Condor: http://www.cs.wisc.edu/condor/

[4] JISC National Grid Service: http://www.ngs.ac.uk/

[5] Announcing the worldwide Protein Data Bank. H Berman et al. Nat. Struct. Biol. (2003), 10, 980.

[6] Assigning Ambiguous NOEs with ARIA. JP Linge, SI O'Donoghue & M Nilges. Methods in Enzymology (2001), 339, 71-90. http://aria.pasteur.fr/

[7] Crystallography & NMR System: A new software suite for macromolecular structure determination. Axel T Brünger et al. Acta Cryst. (1998) D54, 905-921 http://cns.csb.yale.edu/v1.1/

[8] AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. RA Laskowski et al. J. Biomol. NMR (1996), 8, 477-486.

[9] WHAT IF: A molecular modeling and drug design program. G Vriend. J. Mol. Graph. (1990), 8, 52-56.

[10] RECOORD: a REcalculated COORdinates Database of 500+ proteins from the PDB using restraint data from the BioMagResBank. Aart J Nederveen et al. Proteins (2005) 59, 662-672. http://www.ebi.ac.uk/msdsrv/docs/NMR/recoord/main.html/

[11] Quantitative evaluation of experimental NMR restraints. Sander B. Nabuurs et al. J. Am. Chem. Soc. (2003), 125, 12026-12034. http://www.cmbi.kun.nl/software/queen/

[12] The Xplor-NIH NMR Molecular Structure Determination Package. CD Schwieters et al. J. Magn. Res. (2003), 160, 66-74.

[13] Torsion angle dynamics for NMR structure calculation with the new program DYANA. P. Güntert et al. J. Mol. Biol. (1997), 273, 283-298.