WeNMR: Structural Biology on the Grid

Tsjerk A. Wassenaar[1,14], Marc van Dijk[1], Nuno Loureiro-Ferreira[1,15], Gijs van der Schot[1], Sjoerd J. de Vries[1], Christophe Schmitz[1], Johan van der Zwan[1], Rolf Boelens[1], Andrea Giachetti[2], Lucio Ferella[2], Antonio Rosato[2], Ivano Bertini[2], Torsten Herrmann[3], Hendrik R. A. Jonker[4], Anurag Bagaria[5], Victor Jaravine[5], Peter Güntert[5], Harald Schwalbe[4], Wim F. Vranken[6,16], Jurgen F. Doreleijers[7,8], Gert Vriend[8], Geerten W. Vuister[9,7], Daniel Franke[10], Alexey Kikhney[10], Dmitri I. Svergun[10], Rasmus Fogh[11], John Ionides[11], Ernest D. Laue[11], Chris Spronk[12] , Marco Verlato[13], Simone Badoer[13], Stefano Dal Pra[13,17], Mirco Mazzucato[13], Eric Frizziero[13], Alexandre M.J.J. Bonvin[1,*]

[1] Bijvoet Center for Biomolecular Research, Faculty of Science, Utrecht University, Padualaan 8, 3584 CH, Utrecht, The Netherlands.

[2] Magnetic Resonance Center, University of Florence, 50019 Sesto Fiorentino, Italy.

[3] Centre de RMN à très Hauts Champs, Institut des Sciences Analytiques, Université de Lyon, UMR-5280 CNRS, ENS Lyon, UCB Lyon 1, 5 rue de la Doua, 69100 Villeurbanne, France.

[4] Institute of Organic Chemistry and Chemical Biology and Biomolecular Magnetic Resonance Center, Goethe University Frankfurt, 60438 Frankfurt am Main, Germany.

[5] Institute of Biophysical Chemistry and Biomolecular Magnetic Resonance Center, Goethe University Frankfurt, 60438 Frankfurt am Main, Germany.

[6] European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK.

[7] Protein Biophysics/IMM, Radboud University Nijmegen, Geert Grooteplein 26-28, Nijmegen, The Netherlands.

[8] CMBI, Radboud University Nijmegen Medical Centre, Geert Grooteplein 26-28, Nijmegen, The Netherlands.

[9] Department of Biochemistry, Henry Wellcome Building, University of Leicester, Lancaster Road, Leicester LE1 9HN, U.K.

[10] European Molecular Biology Laboratory, Hamburg Outstation, Notkestrasse 85, D22603 Hamburg, Germany.

[11] Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK.

[12] UAB "Spronk NMR Consultancy" Palangos gatvė 4 LT-01402, Vilnius, Lithuania.

[13] Istituto Nazionale di Fisica Nucleare, Sez. di Padova, 35131 Padova, Italy.

[14] Current address: Groningen Biomolecular Sciences and Biotechnology Institute, Rijksuniversiteit Groningen, Nijenborgh 7, 9747AG, The Netherlands.

[15] Current address: Stichting European Grid Initiative (EGI), 140 Science Park, 1098 XG Amsterdam, The Netherlands.

[16] Current address: Department of Structural Biology, VIB, and Structural Biology Brussels, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium.

[17] Current address: Istituto Nazionale di Fisica Nucleare, CNAF, 40127 Bologna, Italy.

ABSTRACT

The WeNMR (http://www.wenmr.eu) project is an EU-funded international effort to streamline and automate structure determination from Nuclear Magnetic Resonance (NMR) data. Conventionally calculation of structure requires the use of various softwares, considerable user expertise and ample computational resources. To facilitate the use of NMR spectroscopy in life sciences the eNMR/WeNMR consortium has set out to provide protocolized services through easy-to-use web interfaces, while still retaining sufficient flexibility to handle more specific requests. Thus far, a number of programs often used in Structural Biology have been made available through portals, including HADDOCK, XPLOR-NIH, CYANA and CS-ROSETTA, MARS, MDDNMR. The implementation of these services, in particular the distribution of calculations to the Grid, involves a novel mechanism for submission and handling of jobs that is independent of the type of job being run. With over 280 registered users (April 2011), eNMR/WeNMR is currently one of the largest Virtual Organization (VO) in life sciences. With its large and worldwide user community, WeNMR has become the first Virtual Research Community officially recognized by the European Grid Infrastructure (EGI).

*To whom correspondence should be addressed.

1 INTRODUCTION

NMR Spectroscopy is one of two techniques that allow determining three dimensional (3D) structures of biomacromolecules, such as proteins, RNA, DNA, and their complexes, at atomic resolution. Knowledge of their 3D structures is vital for understanding functions and mechanisms of action of macromolecules, and for rationalizing the effect of mutations. 3D structures are also important as guides for the design of new experimental studies and as starting point for rational drug design. An advantage of NMR over X-ray crystallography is that it also allows investigation of time-dependent chemical and conformational phenomena, including reaction and folding kinetics and intramolecular dynamics. For these reasons, NMR plays an important role within the life sciences.

The principles underlying NMR are modulation of the natural magnetic moment of atomic nuclei, and measurements of how the system relaxes back to the initial state (Bloch, 1946; Purcell, et al., 1946). The signal thus obtained is a fading wave consisting of many individual frequency contributions: the Free Induction Decay, FID. Typically, up to 27000 different frequencies can be resolved at the highest magnetic fields that are nowadays available. To investigate the frequency contributions and their decays, such measurements have to be repeated many times, due to the low signal-to-noise ratio. To obtain structural information from NMR data, many more, but also more complex measurements have to be run, yielding substantial amounts of data that need processing.

Processing data from NMR to obtain a 3D structure typically involves the following steps, summarized graphically in Figure 1. First the raw data have to be processed, more specifically Fourier-transformed, to obtain spectra revealing the different frequency contributions and their relations. These frequencies are the resonances of the atoms measured, but to infer structural information from them, these resonances subsequently have to be assigned to individual contributors (atoms/residues). If the assignment is sufficiently complete, structural restraints can be determined from the spectra, including inter-atomic distance restraints, dihedral angle restraints, and orientation restraints. These structural restraints are then used to calculate a number of structures using a variety of molecular modeling approaches, after which structure validation checks are performed to assert the quality of the results.

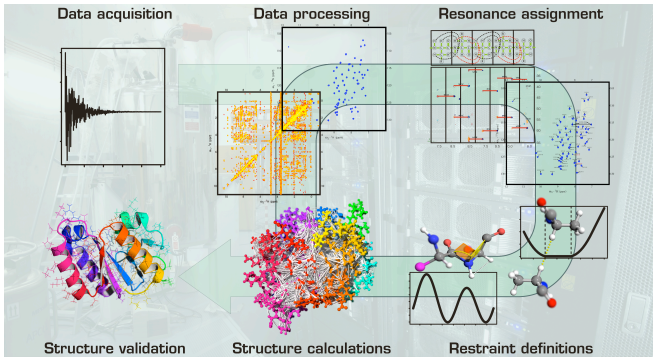

Fig. 1 NMR data processing from signal to 3D structure After acquisition of the primary NMR data, these are Fourier transformed to obtain spectra in which the individual frequency contributions or resonances of spin systems, and their relations, are revealed. The resonances subsequently have to be assigned to individual atoms. If sufficient resonances have been assigned, restraints can be inferred from the data, pertaining to distances between atoms, dihedral angles, domain orientations, etc. When an adequate number of restraints is available, these can be used to calculate a set of three-dimensional structures optimally satisfying these restraints. The resulting structures represent the structure of the protein in solution, which is validated against the available experimental data. Although the process is here depicted linearly, intermediate stages may involve iterative cycles of refinement.

For each of the steps involved, specialized computer programs are available, each with its own characteristics and often with its own data format. Processing of NMR data has thus become a task for specialists, who can understand the data and their formats, as well as the programs, with installation requirements and usage details. Furthermore, NMR data processing requires considerable data storage and computational resources. These factors together currently represent a barrier for groups in life sciences to employ the full power of NMR. Against this background, the eNMR project was ran as a European initiative funded under the Framework 7 e-Infrastructure programme to considerably facilitate this process. It is now carried on by the WeNMR project since November 2010. It aims at allowing groups lacking the resources to add NMR to their toolbox, as well as to allow dedicated NMR groups to improve their standard from basic practice towards cutting-edge research.

The main objectives of the WeNMR project are:

- to provide integrated protocols for NMR data processing
- to provide access to end users through user-friendly web interfaces
- to exploit Grid technology for computationally demanding tasks in structural biology
- to lower the barriers for access to Grid resources in life sciences, notably in structural biology
- to build a virtual research community around a web portal
- to initiate SAXS (Small-angle X-ray scattering) integration into the WeNMR project

Considering the background sketched, these objectives set the challenges to be met within the project. The first of these has been the implementation of a new NMR Grid infrastructure. Historically, due to the requirements for processing of large amounts of data, NMR spectroscopy has always been intimately linked with high performance computing. Therefore, sites with high-end facilities for performing NMR measurements commonly also have considerable computational resources. For the WeNMR partners it thus came as a natural first step to integrate the existing resources into a Grid, offering a single standard for deployment and use of applications across the contributing sites, as well as a natural mechanism to share resources. Currently, the WeNMR project involves an operational Grid, running gLite 3.1 and 3.2 middleware, and the individual sites are being part of the EGI provided by National Grid Initiatives (NGIs) and their infrastructures from Europe and elsewhere.

Having an operational Grid, the programs involved in the different steps, which often require direct user interaction, have to be interfaced in such a way that they can be run automatically. Focus has been initially placed on the CPU intensive programs, which have to be operated remotely as Grid enabled applications. This has to be done in such a way that they can be combined in

automated workflows for protocolized processing of data, raising the issue of interoperability. In addition, web interfaces should be set up to be easy to use, yet sufficiently flexible for expert users. At the same time a mechanism is required to handle job traffic to and from the Grid. In the following paragraphs, these different aspects are discussed in more detail, providing an account of the state of the project thus far. But before discussing the more technical details regarding the implementation, the portals that are available are discussed in more detail.

2 METHODS

2.1 The WeNMR web Portals for Structural Biology

The web portals developed within WeNMR are among the most important elements of the project, as these form the points of entry for the end users. The ultimate goal is to offer to users registered with the eNMR/WeNMR Virtual Organization (VO) complete online protocols for processing NMR data, including all the steps depicted in Figure 1. In addition, each of these steps, and every program involved has value by itself as web based service. For this reason, a piece-wise implementation has been adopted and programs that are ported to the Grid are simultaneously being made available as a web portal. Currently, 12 portals are operational and can be accessed through http://www.wenmr.eu/wenmr/nmr-services. These provide access, among other services, to HADDOCK (De Vries, et al., 2007; Dominguez, et al., 2003) for the prediction of biomolecular complexes, XPLOR-NIH (Schwieters, et al., 2003), CYANA (Guntert, et al., 1997; Herrmann, et al., 2002) and CS-ROSETTA (Shen, et al., 2008; Shen, et al., 2009) for calculating structures from NMR data, AMBER (Case, et al., 2005) for structure refinement and molecular dynamics (MD) simulations CcpNmr (Vranken, et al., 2005) for data conversion, MARS (Jung and Zweckstetter, 2004) for backbone assignment, TALOS+ (Shen, et al., 2009) for torsion angle prediction, and MDDNMR (Jaravine, et al., 2008) for NUS (Non-Uniform Sampling) spectral processing. Next to these available portals, several new ones are in development for various NMR applications, including the UNIO program (Fiorito, et al., 2008; Volk, et al., 2008) that provides computational routines for each individual step depicted in Figure 1. The main WeNMR portal is shown in Figure 2.

2.2 HADDOCK

HADDOCK (De Vries, et al., 2007; Dominguez, et al., 2003) is an acronym for High Ambiguity Driven DOCKing and is a program to predict structures of biomolecular complexes from individual components. As the full name indicates, this approach in docking of biomolecules distinguishes itself from other methods by using external information to guide the docking process. Such information can be empirical, theoretical or both, pertaining to the residues or atoms involved in the binding interface. From this information ambiguous restraints are derived that are used to drive the docking. HADDOCK is particularly useful in predicting complexes from known experimental structures of the partners using NMR data, such as chemical shift perturbations and residual dipolar couplings (RDCs). Chemical shift perturbations and RDCs can be obtained relatively easily and also for macromolecules of increasing size, making the large applicability of HADDOCK as a tool for cutting edge Structural Biology apparent. HADDOCK has proven its value within the CAPRI (Critical Assessment of PRediction of Interactions) experiment, a blind evaluation of the performance of current docking methods (De Vries, et al., 2007; Lensink, et al., 2007; Mendez, et al., 2005; van Dijk, et al., 2005).

The docking process starts with random placement of the individual components with a given separation and random orientations. Subsequently, a large number of complex structures, typically in the order of thousands, is generated by rigid-body docking, driven by the ambiguous restraints. From these a number of structures, typically several hundred, are selected for further refinement, using a scoring function. These structures are first subjected to a further cycle of simulated annealing, introducing flexibility to allow optimization of contacts. After this, a final cycle of refinement follows, in which the complex is solvated. The results are then scored, analyzed and returned to the user. The structure calculations are the CPU intensive part of the process and involve a combination of energy minimization and MD (in torsion angle or Cartesian space) simulations.

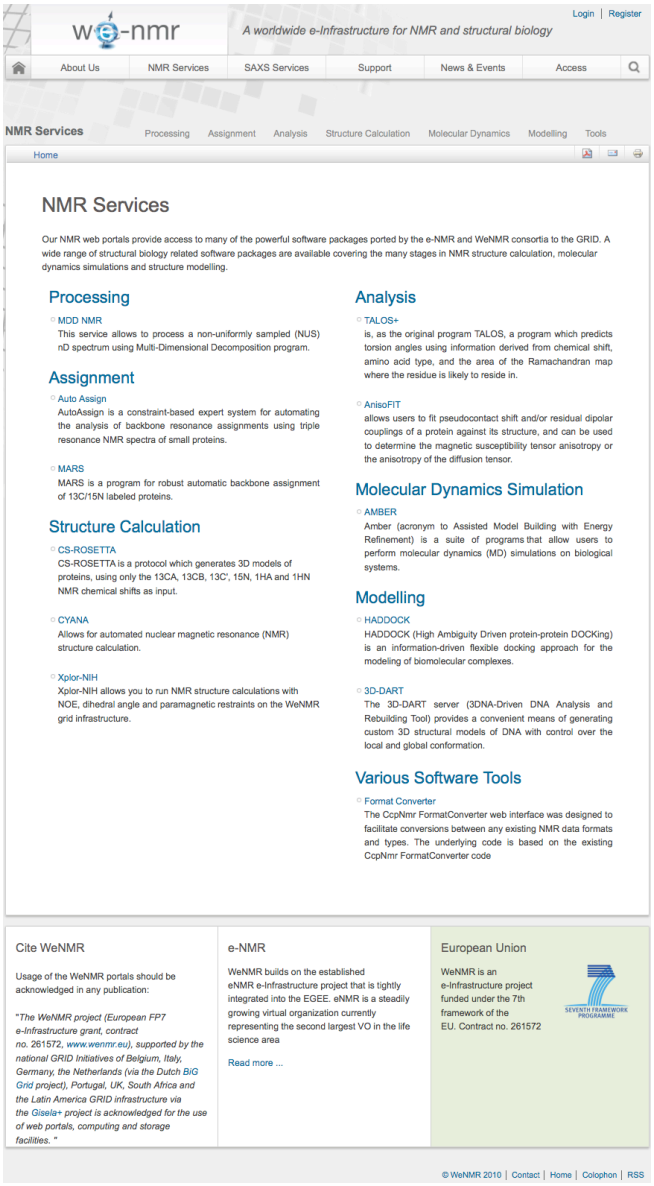

Fig. 2 The WeNMR web portal (http://www.wenmr.eu/wenmr/nmr-services)

HADDOCK offers almost full control of the many parameters involved in the docking process. To offer the full functionality of HADDOCK through a web portal thus requires putting forth a complicated form, contrasting with the objective of having a simple interface. To avoid compromises regarding user friendliness and functionality, two innovations were introduced in the design of the portal. First of all, the portal is divided in four interfaces, corresponding to different levels of control and user experience:

- The **Easy Interface** requires no more than providing the two components of a complex and the residues of each that are involved in the interaction.
- The **Expert Interface** allows the user to provide his own customized restraints to be included in the docking process and to specify certain aspects of the sampling and analysis. In addition, using this interface the user can set protonation states of histidine residues, and define regions of the interacting molecules to be kept flexible during the docking. This allows a certain degree of conformational change to take place during docking.
- The **Guru Interface** offers almost full control of parameters, allowing e.g. specification of symmetry and relaxation anisotropy restraints and RDCs as well as of parameters pertaining to the energy, the scoring and the analysis of results.
- Finally, for complete control a **File Upload Interface** is available, where a HADDOCK run parameter file can be provided. This is particularly useful for those who have their own standard protocol or who want to replicate a previous run with minor modifications. This option also offers a simple way to build pipelines from other applications.

The Expert and Guru interface offer control of the docking process at the expense of making the forms to be filled in more complex. Thus, to facilitate the user's task and keep the forms manageable, foldable menus were introduced that group related parameters under a single header. In this way, users only need to unfold groups of options that should be changed from their default values.

Except for the File Upload Interface, the HADDOCK portals share the data structure, albeit that part of the variables is fixed to predefined values for the Easy and Expert interfaces. This has the advantage that they can all couple to a single back end CGI (Common Gateway Interface) script to handle the request, as will be discussed in more detail in the implementation details.

After issuing a request, the user is presented a link to a site where the progress can be followed. After the run is finished, the results can be viewed online and selected complexes or the complete output data of the run can be downloaded to a local machine.

The use of the HADDOCK portal requires registration with a valid Grid certificate, giving a username and password. These are thereafter used to sign service requests. The requests themselves are handled using an eToken-based robot certificate, as is explained in more detail in the implementation details.

2.3 XPLOR-NIH

XPLOR-NIH (Schwieters, et al., 2006; Schwieters, et al., 2003) is one of the programs for structure calculations that have been ported to the Grid and are available through the WeNMR web portal. It is a versatile program that can be operated through a command line interface or with scripts in the specific XPLOR language.

Performing structure calculations using NMR data commonly starts with the generation of an extended conformation from a topological description of the macromolecule. For standard components, such as protein, DNA and RNA, this topological description can be easily inferred from the sequence of the building blocks. Distance, orientation and other restraints derived from NMR data can then be added to the topological description and used to drive the system to a folded state using simulated annealing. This annealing step is repeated many times to obtain sufficient statistics regarding the goodness of fit of the structures determined against the experimental data. Since the different annealing runs are independent of each other, they can be easily distributed over multiple CPUs. After the annealing runs have finished, the best structures are usually selected for further refinement, including solvent in the calculations.

The portal uses a design which is different from the HADDOCK portal, and is aimed at more direct user interaction during the process. Users log in with their Grid certificate loaded in the web browser, gaining access to an environment where projects can be started, stored and managed. Structure calculation projects are initiated by filling in a form and providing files for the structures and topological descriptions of the molecules, as well as for the different restraints to be included in the calculations.

When the structure calculations have finished, the user can view and download the results. In addition, it is possible to select a number of structures for further refinement and characterization using the AMBER package (Case, et al., 2005) for MD simulations.

2.4 CYANA

Another widely used program for calculating structures from conformational restraints is CYANA (Combined Assignment and dYnamics Algorithm for NMR Applications) (Guntert, et al., 1997; Herrmann, et al., 2002). Its main characteristics are the ability for iterative assignment of NOE peaks, and structure calculations through simulated annealing in torsion angle space. Like with XPLOR-NIH, the structure calculations involve many simulated annealing runs, divided over several iterative cycles.

The design of the web portal for CYANA is similar to that of HADDOCK. Foldable menus are used to hide optional sets of parameters, by default presenting an intuitive menu offering a standard structure calculation protocol. The portal allows three modes of invocation of the service. Users can request structure calculation using a set of upper distance bound restraints, providing a list of assigned peaks, or providing a list of unassigned peaks, in which case the automated peak assignment will be performed.

Use of the service requires having a license for CYANA and registering for use of the portal, presenting a valid Grid certificate. This will give a username and password that can be used to sign service requests.

2.5 CS-ROSETTA

Chemical-Shift ROSETTA or CS-ROSETTA (Shen, et al., 2008; Shen, et al., 2009) is the third program for structure calculations that has been ported to the Grid and made available through the WeNMR web portal. CS-ROSETTA, unlike XPLOR-NIH and CYANA, allows structure determination of proteins, based on chemical shift information alone. It thus bypasses the need for NOE based distance restraints, which usually require considerable time to obtain. Further advantages of using chemical shifts are that these are among the most reliable parameters that can be obtained from NMR spectroscopy and that they can potentially be obtained for larger macromolecules for which NOEs become impractical. On the other hand, direct chemical shift based structure determination is computationally much more expensive than structure calculations using distance restraints. However, the most time consuming part of a CS-ROSETTA run consists of a large number of independent calculations that can be easily distributed over the Grid.

Structure determination using CS-ROSETTA requires as only input the amino acid sequence and a list of chemical shifts and a number of parameters to control the process that can be changed from the default values. Backbone chemical shifts for $^{13}C^{a}$, $^{13}C^{b}$, $^{13}C'$, $^{1}H^{a}$, $^{1}H^{N}$, and ^{15}N that are provided by the user, are validated and stored as the target shifts. These chemical shifts are first used to select a set of protein fragments from a structure database, e.g. the Protein Data Bank (PDB) (Berman, et al., 2000), based on the list of chemical shifts as predicted with SPARTA. Then the regular ROSETTA protocol (Rohl, et al., 2004) for Monte Carlo assembly and relaxation is used to reassemble the protein from the fragments. For the resulting models the chemical shifts are again predicted using SPARTA (Shen and Bax, 2007) and the deviations between the predicted and target values are used as a pseudo-energy term in the scoring of the models, yielding a ranking based on both overall structural quality as well as on the match with the experimental data.

The computationally most expensive step in the process is the construction of a model using Monte Carlo assembly and relaxation. To obtain a reliable prediction, a set of 10 000 to 50 000 models has to be built, each starting from the same fragment library. Using different seeds for generation of random numbers ensures independence of the results from different runs. For the WeNMR implementation of CS-ROSETTA only the Monte Carlo search is performed on the Grid.
The computational cost involved in chemical shift based structure determination makes CS-ROSETTA a typical example of a program that is beyond the capacity of most local sites. Here, the access to Grid resources through a web-portal, combining computational power and ease of use, clearly demonstrates its added value.
To use the service, users have to register with a valid Grid certificate to obtain a user name and password that can subsequently be used to sign service requests. The web interface for CS-ROSETTA itself is straightforward and only requires uploading a file containing chemical shifts in TALOS (Cornilescu, et al., 1999) format, and modifying the parameters to control the calculations. When the job has finished, the user receives an e-mail containing a link to the result page that gives an overview of the run, including some statistics and images to assess the overall quality of the results. The user can then select a number of structures to view in more detail or to download, or choose to download the whole set of results as an archive.

2.6 MARS

The fifth portal, MARS performs automatic backbone assignment of $^{13}C/^{15}N$ labeled proteins and is applicable to a wide variety of NMR data, including RDCs (Jung and Zweckstetter, 2004).
Its advanced features compare favorably with other assignment tools and include:

- simultaneous optimization of the local and global quality of assignment to minimize propagation of initial assignment errors and thus providing robustness against missing chemical shift information; applicable to proteins above 15 kDa using only C_a and C_b chemical shift information with connectivity thresholds as high as 0.5 ppm;
- applicable to proteins with very high degeneracy such as partially or fully unfolded proteins;
- combination of the secondary structure prediction program PSIPRED (McGuffin, et al., 2000) with statistical chemical shift distributions, which were corrected for neighboring residue effects (Wang and Jardetzky, 2002), to improve identification of likely positions in the primary sequence;
- assessment of the reliability of fragment mapping by performing multiple assignment runs with noise-disturbed chemical shifts.

Registration and a valid Grid certificate are needed to use the service. Its interface is similar to Cyana: after uploading peak lists a file containing assigned chemical shifts is downloaded after job has finished. A use-case example is available for download on the WIKI pages (http://www.wenmr.org/wenmr/mars-use-case-example).

2.7 TALOS+

TALOS+ (Shen, et al., 2009), like its predecessor TALOS (Cornilescu, et al., 1999), is a program to predict torsion angles for amino acids given information regarding chemical shift and the probable regions in the Ramachandran plot for each type of amino acid. TALOS+ distinguishes itself by the inclusion of a neural network component, the output of which is added as an empirical term in the conventional TALOS data base search. To prevent assignment of torsion angles to the backbone of flexible regions, TALOS+ first identifies such regions using the flexibility prediction program RCI developed by Berjanskii and Wishart (Berjanskii and Wishart, 2005).
The TALOS+ portal offers a simple interface where an input file in either TALOS or BMRB format can be uploaded. In addition a number of PDB ID codes can be given, indicating structures that should be excluded from the calculations. Unlike most of the other portals, the TALOS+ portal can be used without a Grid certificate, as the calculations are run on a local server.

2.8 MDDNMR

The MDDNMR (Jaravine, et al., 2008) portal can process individual NUS multidimensional NMR spectra. The interface supports the NUS data recorded by two major spectrometer brands: Varian and Bruker.
The main advantage of usage of NUS data is the substantially higher resolution in the indirect spectral dimensions. The NUS acquisition mode of both Vnmr and TopSpin makes use of standard NMR experiments, except that only a fraction of a full data set is recorded. This means that it can be used with virtually any pulse sequence available. After acquisition, MDDnmr replenishes the missing data points in the full matrix. The resulting regular spectra are then processed conventionally with FFT (fast Fourier transform), LP (Linear Programming), window functions etc. The current portal allows this to be repeated for each experiment in the dataset; several such high-resolution experiments are processed sequentially as single matrices, and the resulting high-resolution FT-domain spectra, after peak-picking, are amenable for automatic backbone assignment using MARS. Most of the experimental data types supported by MDDNMR can be processed via the portal, including constant-time acquisition (CT), J-coupling splitting etc. The five use-case examples of 2D, 3Ds, 4D spectra are available for download on the WIKI pages (http://www.wenmr.org/wenmr/mddnmr-use-case-examples); the examples have extensive documentation on algorithms and adequate tutorial on usage. The design of the web portal for MDDNMR is similar to that of Cyana. Use of the service requires registering for use of the portal, presenting a valid Grid certificate.

2.9 CcpNmr

The CcpNmr portal, for CcpNmr (Vranken, et al., 2005) based data conversions, is not directly related to NMR data processing, but is an important element within the WeNMR project. The reason for this is the fact that the programs already ported to the Grid often have their own data formats, making it impossible to combine these as steps in a direct pipeline. Establishing interoperability of such programs requires automated conversion of output from one step to match the input of a next step. This is exactly what CcpNmr was designed for. It has a comprehensive internal data model that can contain all different types of NMR related data. These data can be imported from files in a large number of formats. Likewise, the data stored in a CcpNmr project can be exported in any of the file format required, provided that the data are present in the model. In this way, CcpNmr provides a straightforward approach in meeting one of the challenges of the WeNMR project, namely establishing interoperability of the programs involved in NMR data processing and building automated protocols for complex tasks.
The portal for CcpNmr was developed as the program was ported to the Grid, to offer the WeNMR members an easy solution for matching program output and program input during the steps involved in processing of their data. At present, the portal allows conversion between several different file formats, aimed at facilitating the use of the other portals. The file conversions for the portal are performed locally, as these are not computationally intensive. Use of the CcpNmr portal thus does not require a Grid certificate.

2.10 AMBER

AMBER (Case, et al., 2005) is a collective name for a suite of programs that allow users to carry out MD simulations on biological systems. The web-portal permits the creation and management of MD calculations from a web browser. The portal takes care also of all the grid accounting. A new

user with her/his personal certificate installed in the browser can access to the portal from WeNMR web page and straightforwardly create a new user login and password. The site permits the creation of a new single calculation or the creation of a project where it will be possible to save a number of different calculations that belong to a common project (e.g. MD refinements of protein structures generated with different ensembles of restraints).

The user can select a pre-set MD refinement protocol to energy optimize NMR protein structures. The currently proposed protocol comprises four steps:

- In the first step the protein structure to be optimized is uploaded. Here it is possible to upload one pdb file, whose format will be automatically validated and converted into the format recognized by amber. During this step the user can add explicit water molecules to solvate the protein, add counter ions, insert new bonds for selected atoms (e.g. protein to metal). For NMR structures, which are typically represented by a bundle of 20-40 different conformers, this first step is carried out only for the first conformer and then automatically applied to all the structures in the bundle. For each structure in the bundle, an individual job is sent to the grid.
- The second step manages the NMR restraints. Four types of restraint are allowed: NOE, dihedral angles, RDCs and pseudocontact shifts (PCS). The last restraints are the so-called paramagnetic restraints. For all restraints it is possible to upload Xplor, Dyana, or Cyana files. For paramagnetic restraints, it is possible to fit the anisotropy tensor directly in the web site.
- The third step manages the setting of MD calculations. The page can be visualized in a so-called basic mode and in an extended mode, which allows users to view all the details of the amber settings.
- The fourth and final step allows the user to give a name to the calculation started and submit it to the grid. After the results of a job have been downloaded, the user can browse them and download the various files.

2.11 UNIO

In addition to the portals already available, a UNIO portal has now been tested on local infrastructure and is expected to be finalized and available mid 2011. UNIO comprises elements for all major tasks involved in protein structure determination by NMR (Figure 1). The UNIO portal will allow the user to obtain backbone resonance assignment based on projection NMR spectroscopy of high-dimensional spectra. Such spectra have recently received much interest in the NMR community and presumably represent a substantial and more reliable addition to data analysis programs commonly used for backbone NMR resonance assignment (Volk, et al., 2008). The UNIO portal will be designed as a multiple component web portal, similar to the HADDOCK portal. It will offer expert systems for the subsequent computationally demanding tasks of NMR signal identification, side-chain resonance assignment (Fiorito, et al., 2008) and comprehensive collection of distance restraints (Herrmann, et al., 2002), with the latter task focusing on the primary source of NMR-based protein modeling. UNIO is compatible with powerful NMR structure calculation programs, such as CYANA and CNS, which are already operational on the WeNMR grid infrastructure, and will equip the structural biology community with all computational processing tools necessary for a complete protein structure determination by NMR.

3 STRUCTURAL BIOLOGY ON THE GRID: DESIGN STRATEGIES AND IMPLEMENTATION

Successfully running web portals requires a proper machinery to handle requests. This machinery involves various steps that can be categorized in three layers of operation: The server level involves handling of service requests, either by direct human interaction or through requests from another machine. This stage includes input type checking. The next level involves preparation, trafficking and monitoring of jobs between the server and the Grid. The third layer is the core layer involving the process(es) to be run on a worker node. The tasks associated with these different levels are conceptually unrelated and allow for a component based development approach, in which distinct tasks are programmed in a most generic form. This has the advantage that such building blocks can be easily maintained, adapted and reused.

To facilitate the component-based implementation, a single, simple model, illustrated in Figure 3, was designed within the WeNMR consortium for the representation of processes. This model characterizes any process as a block with four connectors: input, output, dependencies and logging. The input and output connectors allow building larger sequences or complex workflows. The dependencies are considered static to a process and the logging is for messaging and provenance. Logging information can also be used to check the status and react to errors. Processes are all shaped to adhere to this simple model, which can be achieved by rewriting programs or by wrapping them inside a script. Doing so, a set of process modules is obtained that can easily be combined in a pipeline.

Process pipelines are often built imperatively, using one of the standard scripting languages. But an imperative approach has the drawback that it is inflexible: e.g. a failure will cause the whole pipeline to fail and a process has to be started anew. For this reason, a partial declarative approach was designed, in which direct communication between processes from the different layers of operation are eliminated. Rather, output from one level that forms the input for another is 'pooled' on disk. Processes from the next layer that depend on these data are run periodically, scanning the pool for data matching the input requirements. This has the advantage that the state of all processes is naturally check pointed and that the use of computational resources can be better controlled. How this approach is used to connect the web portals to Grid calculations is illustrated in Figure 4 and explained in more detail below. Note that this description is rather general and some of the portals do not yet adhere to this strict separation of the layers.

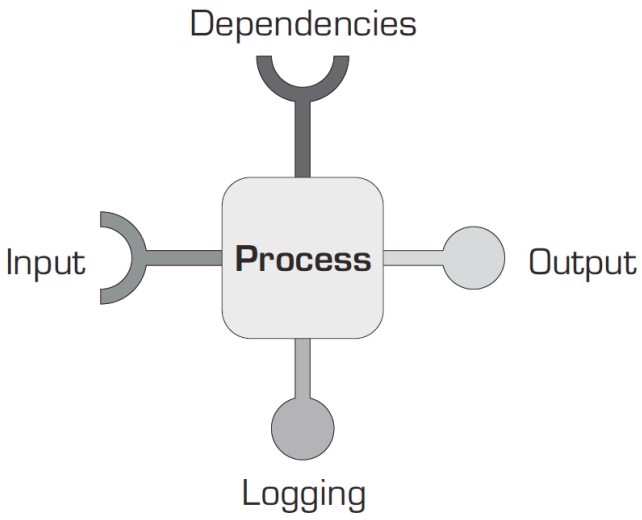

Fig. 3 Process model To facilitate implementation and management, processes are represented and, if needed, rewritten in a manner adhering to a simple five node model, with the process itself as the central block that has four connectors: an input and an output connector, a connector for dependencies and one for logging. The output of one process can be connected to the input of another to build larger sequences or workflows. Obviously, each connection can involve several components, e.g. a process' input can consist of several files and/or option settings. The process itself may be regarded a black box, as long as the connections are well-defined.

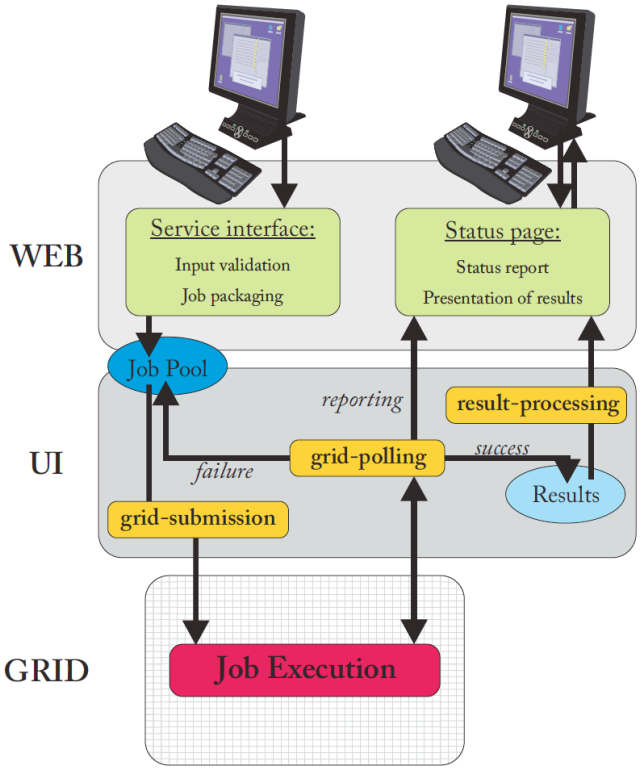

Fig. 4 Grid job submission management using job pooling The figure shows a general scheme for managing job trafficking to and from the Grid, using server side job pooling. This scheme is characterized by a separation of three layers of operation, between which there is no direct communication. Green boxes indicate user interaction, whereas yellow boxes indicate jobs that are running periodically as daemon jobs and that use an eToken-based robot certificate for generating a Grid proxy. The blue ellipses represent 'pools', which are used for storage of job or result packages. User service requests are processed on the server, up to the point of generating a job package that is stored on disk. On the Grid UI (User Interface) a daemon job (grid-submission) is running on a scheduled base scanning the 'job pool' for job packages and submitting these to the Grid when found. Another daemon job (grid-polling) is periodically checking running jobs for their status, retrieving the results when ready and placing these in a result pool. Finally, results are presented back to the user, possibly after post-processing (results-processing). Currently the HADDOCK, CS-ROSETTA, UNIO, CYANA, MARS and MDD-NMR portals, which all send jobs to the Grid, are implemented following this model.

The first level: Request and data handling, invocation of the service, reporting

The first level of operation involves the interaction with the user, both processing the request, as well as presenting the results. A service request is made by filling in a form that is parsed by a CGI script. This script also performs type checking and validation of the user input and presents the user a unique ID with which the results can be retrieved.

Both the web form and the CGI script depend primarily on the data to be provided, and it is possible to generate these automatically from a description of these data. To this purpose, the Spyder framework (http://www.spyderware.nl, S.J. de Vries, unpublished) was designed, initially to facilitate setting up and managing the portals for HADDOCK. Next to the generation of web forms, Spyder natively supports data validation for known types and can convert between data types if all intermediary conversions are defined. Thus, Spyder offers a single framework for managing most of the elements involved in the first stage of processing of requests.

The data parsed and validated are then processed by a request specific script that combines all data and control parameters required for further processing into a self-contained job package. This package is subsequently placed into a job pool directory, which ends processing of the request at the first level.

The second level: Grid job preparation, trafficking and monitoring

At the second level, a daemon job is running periodically, scanning the job pool for jobs that are ready to be run on the Grid and submitting these when found. In principle, this daemon job does not require information regarding the nature of the job, although in practice different instances are run, each linked to one type of job to better control the work load associated with the different tasks.

A separate daemon job is running, also periodically, checking the status of the jobs running on the Grid and retrieving the results when finished. Alternatively, this process can resubmit the job when it has failed. The results are put back, after validation, in a place where they can be accessed through a web page. Like the submission process, the polling and retrieval process is in principle independent, since all information regarding the job, such as the directory to place the results in, are contained in the job package.
Submission, polling and retrieval of output are handled using a standard toolbox for Grid operation, which, in the case of WeNMR, is the gLite 3.1 / gLite 3.2 suite. Accordingly, the jobs that operate at this level require the use of a valid proxy. To facilitate proxy management, all of the processes at the second level of operation are running using an eToken-based robot certificate, in accordance with the security requirements for data portals formulated by the Joint Security Policy Group (https://www.jspg.org/wiki/VO_Portal_Policy).

The third level: Primary tasks

The third level of operation involves the tasks running on the Grid. This requires programs to be ported to the Grid, but that process is relatively straightforward. The only aspect that is different from more common strategies is that the processes have to adhere to the process model discussed previously, facilitating automation and provenance.

4 RESULTS

4.1 Status and Statistics

Since the start of the eNMR project in fall 2007, considerable progress has been made, both in the deployment and the utilization of an infrastructure for structural biology. Currently, the infrastructure is distributed over three partner sites, which together provide a body of 272 dedicated CPUs, and 2.87 Tb of storage. Resources are shared with 16 other sites, giving access to about 10000 CPU cores and 37 Tb of storage.

Over the last year, more than 1 million jobs have been run on the Grid, corresponding to about 500 years of normalized CPU time. The overall CPU efficiency, the total CPU time divided by the total wall time, of all jobs was 99.0% (statistics taken from the EGI accounting portal).

Including the twelve applications that have already been made available through a web portal, twenty programs have been ported to the Grid, several of which will be made available as web portals in the near future.

Currently there are more than 280 users registered, several of which use the portals on a regular basis. The most active portals are the ones for HADDOCK and for CS-ROSETTA, which have processed over 1109 and 545 requests thus far. Together these requests account for over 90% of the jobs that have been run on the Grid, as a result of the task farming approach involved.

4.2 Conclusions

Since the beginning of the eNMR project, the eNMR/WeNMR consortium has managed to set up an operational Grid (http://www.wenmr.eu/wenmr/wenmr-grid-statistics), to port twenty applications and bring up twelve web portals (http://www.wenmr.eu/wenmr/nmr-services), with several others being finalized. At the time of writing, WeNMR has already grown to be nearly the largest virtual organization within the life sciences. This successful start has been underlined by the award for the best demonstration of an application, received at the EGEE (Enabling Grids for E-sciencE) 2009 User Forum. With the present-day momentum, the WeNMR project is rapidly evolving into a factor of importance within structural biology, and life sciences in general. As such it has been the first Virtual Research Organization officially recognized by the EGI. Currently, efforts include writing WSDL definitions for the portals that will allow calling services remotely, e.g. from a workflow-manager. At the next stage of the project the different elements will be combined, providing comprehensive, yet easy-to-use tools for integrated analysis of NMR data. Furthermore, a number of SAXS services are being added to support a wider user community in structural biology.
Up-to-date information, regarding the state of the project, the available services, and how to join the WeNMR virtual organization, can be found on the project web page at http://www.wenmr.eu.

ACKNOWLEDGEMENTS

The WeNMR project is funded by the European Commission under an FP7 e-Infrastructure grant, contract no. 261572 and builds on the previous FP7 e-Infrastructure project e-NMR, contract no. 213010. Support from the former EGEE and the current EGI in terms of expertise and recognition is also acknowledged. The national Grid Initiatives of Belgium, Italy, Germany, the Netherlands (via the Dutch Big Grid project), Portugal, UK, South Africa and the Latin America Grid infrastructure via the Gisela+ project is acknowledged for the use of web portals, computing and storage facilities. Finally, the authors like to thank those that have expressed their interest in and support to the project.

REFERENCES

Berjanskii, M.V. and Wishart, D.S. (2005) A simple method to predict protein flexibility using secondary chemical shifts, *J Am Chem Soc*, **127**, 14970-14971.

Berman, H.M., *et al.* (2000) The Protein Data Bank, *Nucleic Acids Research*, **28**, 235-242.

Bloch, F. (1946) Nuclear Induction, *Physical Review*, **70**, 460.

Case, D.A., *et al.* (2005) The Amber biomolecular simulation programs, *J Comput Chem*, **26**, 1668–1688.

Cornilescu, G., Delaglio, F. and Bax, A. (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology, *J Biomol Nmr*, **13**, 289-302.

De Vries, S.J., *et al.* (2007) HADDOCK versus HADDOCK: New features and performance of HADDOCK2.0 on the CAPRI targets, *Proteins-Structure Function and Bioinformatics*, **69**, 726-733.

Dominguez, C., Boelens, R. and Bonvin, A.M.J.J. (2003) HADDOCK: A protein-protein docking approach based on biochemical or biophysical information, *J Am Chem Soc*, **125**, 1731-1737.

Fiorito, F., *et al.* (2008) Automated amino acid side-chain NMR assignment of proteins using (13)C- and (15)N-resolved 3D [(1)H, (1)H]-NOESY, *J Biomol Nmr*, **42**, 23-33.

Guntert, P., Mumenthaler, C. and Wuthrich, K. (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA, *Journal of Molecular Biology*, **273**, 283-298.

Herrmann, T., Guntert, P. and Wuthrich, K. (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA, *Journal of Molecular Biology*, **319**, 209-227.

Jaravine, V.A., *et al.* (2008) Hyperdimensional NMR spectroscopy with nonlinear sampling, *J Am Chem Soc*, **130**, 3927-3936.

Jung, Y.S. and Zweckstetter, M. (2004) Mars - robust automatic backbone assignment of proteins, *J Biomol Nmr*, **30**, 11-23.

Lensink, M.F., Mendez, R. and Wodak, S.J. (2007) Docking and scoring protein complexes: CAPRI 3rd edition, *Proteins-Structure Function and Bioinformatics*, **69**, 704-718.

McGuffin, L.J., Bryson, K. and Jones, D.T. (2000) The PSIPRED protein structure prediction server, *Bioinformatics*, **16**, 404-405.

Mendez, R., *et al.* (2005) Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures, *Proteins*, **60**, 150-169.

Purcell, E.M., Torrey, H.C. and Pound, R.V. (1946) Resonance Absorption by Nuclear Magnetic Moments in a Solid, *Physical Review*, **69**, 37.

Rohl, C.A., *et al.* (2004) Protein structure prediction using Rosetta, *Methods in Enzymology*, **383**, 66-93.

Schwieters, C.D., Kuszewski, J.J. and Clore, G.M. (2006) Using Xplor-NIH for NMR molecular structure determination, *Progress in Nuclear Magnetic Resonance Spectroscopy*, **48**, 47-62.

Schwieters, C.D., *et al.* (2003) The Xplor-NIH NMR molecular structure determination package, *Journal of Magnetic Resonance*, **160**, 65-73.

Shen, Y. and Bax, A. (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology, *J Biomol Nmr*, **38**, 289-302.

Shen, Y., *et al.* (2009) TALOS plus : a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts, *J Biomol Nmr*, **44**, 213-223.

Shen, Y., *et al.* (2008) Consistent blind protein structure generation from NMR chemical shift data, *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 4685-4690.

Shen, Y., *et al.* (2009) De novo protein structure generation from incomplete chemical shift assignments, *J Biomol Nmr*, **43**, 63-78.

van Dijk, A.D., *et al.* (2005) Data-driven docking: HADDOCK's adventures in CAPRI, *Proteins*, **60**, 232-238.

Volk, J., Herrmann, T. and Wuthrich, K. (2008) Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH, *J Biomol Nmr*, **41**, 127-138.

Vranken, W.F., *et al.* (2005) The CCPN data model for NMR spectroscopy: Development of a software pipeline, *Proteins-Structure Function and Bioinformatics*, **59**, 687-696.

Wang, Y.J. and Jardetzky, O. (2002) Investigation of the neighboring residue effects on protein chemical shifts, *J Am Chem Soc*, **124**, 14075-14084.