ARIA2: Automated NOE assignment and data integration in NMR structure calculation

Wolfgang Rieping1,†, Michael Habeck1,†, Benjamin Bardiaux1,2, Aymeric Bernard1, Thérèse E. Malliavin1 and Michael Nilges1,*

1Unité de Bioinformatique structurale, CNRS URA 2185, Institut Pasteur, 25-28 rue du docteur Roux, 75015 Paris, France and 2Laboratoire de Biochimie Théorique, CNRS UPR 9080, Institut de Biologie Physico-Chimique, 13 rue P. et M. Curie, 75005, Paris, France

Received on September 11, 2006; revised and accepted on November 16, 2006
Advance Access publication November 22, 2006
Associate Editor: Dmitrij Frishman

ABSTRACT
Summary: Modern structural genomics projects demand for integrated methods for the interpretation and storage of nuclear magnetic resonance (NMR) data. Here we present version 2.1 of our program ARIA (Ambiguous Restraints for Iterative Assignment) for automated assignment of nuclear Overhauser enhancement (NOE) data and NMR structure calculation. We report on recent developments, most notably a graphical user interface, and the incorporation of the object-oriented data model of the Collaborative Computing Project for NMR (CCPN). The CCPN data model defines a storage model for NMR data, which greatly facilitates the transfer of data between different NMR software packages.
Availability: A distribution with the source code of ARIA 2.1 is freely available at http://www.pasteur.fr/recherche/unites/Binfs/aria2
Contact: nilges@pasteur.fr

INTRODUCTION
The assignment of nuclear Overhauser enhancement (NOE) peaks is the most time-consuming step in the analysis of nuclear magnetic resonance (NMR) data and structure calculation. Though several programs exist that facilitate a manual analysis of spectra, the NOE assignment is tedious due to the large number of assignment possibilities, peak overlap and potential artifacts in the spectra. Therefore, a widely employed approach to NMR structure determination is to calculate a structural model from the experimental data by using programs for automated assignment, such as CANDID/CYANA (Herrmann et al., 2002), AUTOSTRUCTURE (Montelione et al., 2000) or ARIA (Ambiguous Restraints for Iterative Assignment; Nilges et al., 1997). Subsequently, the model is validated against the original data and, if necessary, refined using additional assignments derived with the aid of the model structure. ARIA uses an iterative protocol and the concept of ambiguous distance restraints (ADR) (Nilges, 1995) to automatically assign NOE cross-peaks.

Most software packages use proprietary formats for data storage, which need to be inter-converted for transferring data between different applications. This usually requires manual intervention and can lead to a loss of information since the data conversion is often incomplete. The Collaborative Computing Project for NMR (CCPN) data model (Fogh et al., 2005) alleviates these problems by defining a storage model that integrates all information emerging in a structure determination project in a common framework. This includes details on the molecular system, experimental data such as chemical shifts, NOEs, or residual dipolar couplings, as well as the results of a calculation, most importantly the assigned spectra and the 3D coordinates of the structures. We have incorporated the CCPN data model into ARIA to enable spectroscopists to use existing NMR computer programs in a very efficient way. Version 2.1 of ARIA has several other new features, which we summarize in this article.

ITERATIVE NOE ASSIGNMENT
ARIA assigns NOE cross-peaks by first deriving all possible assignments for each peak by matching a list of chemical shifts with frequency ‘windows’ centered around the position of a peak. Peak volumes are converted into distance restraints by using the isolated spin pair approximation, which relates the volume to the inverse sixth power of the distance between the two interacting spins. Ambiguous assignments are converted into ADRs, so that all assignment possibilities contribute to the target distance. However, most of the assignments are inconsistent, and thus cannot be fulfilled simultaneously by a single structure. ARIA performs an iterative protocol to identify wrong assignments and noise peaks: an iteration begins with correcting the restraint list by filtering out unlikely assignments and noise peaks (cf. Nilges, 1997, for details). Based on the filtered restraint list, a new structure ensemble is calculated which is analyzed in the next iteration.

The simplified treatment of non-bonded forces and missing solvent contacts during structure calculation can result in artifacts, such as unrealistic side-chain packing and unsatisfied hydrogen bond donors or acceptors. To reduce such artifacts, we refine the structures in explicit water (Linge et al., 2003), which also leads to a considerable improvement of structural quality (Nederveen et al.,...
2005). Finally, the water-refined ensemble is validated using several computer programs. We employ WHATIF to compare each conformer with a typical structure found in a database of high-resolution X-ray structures. The program PROCHECK analyzes the local fold in terms of a Ramachandran statistics, and the software PROSA can detect errors in the global fold of a protein. Several reports summarize the validation results and provide information on the assignments and restraints used for the calculation.

DATA INTEGRATION
Experience with earlier versions of the program showed that many problems occurring at later stages of a calculation are due to misformatted or inconsistent input files. To facilitate data validation, we have developed a new data format based on the extensible markup language (XML). ARIA defines XML formats to describe molecular systems (we follow the IUPAC recommendation), chemical shifts, and NOE cross-peaks, and uses the CcpNmr FormatConverter (Vranken et al., 2005) to convert >20 proprietary data formats into ARIA XML.

The program also offers the option to retrieve this information, along with other restraints, directly from a CCPN project. A CCPN project can be created manually by using, for example, the FormatConverter. The recommended approach, however, is to employ NMR analysis programs that support the data model directly, such as the software CcpNmr Analysis (Vranken et al., 2005). That way, the user can seamlessly launch an ARIA calculation, without prior data conversion (Fig. 1). Furthermore, ARIA automatically exports the result of a calculation, namely the assigned peak lists, restraint lists, and the structure ensembles, to a CCPN project. This simplifies the submission of the calculation results to the databases PDB or BMRB, and enables the user, for example, to access the assignments directly from within CcpNmr Analysis, or to validate the restraints by using programs such as QUEEN (Nabuurs et al., 2003).

GRAPHICAL USER INTERFACE
A new graphical user interface (GUI) replaces the HTML web-form used in previous versions of ARIA to simplify and streamline the setup of a calculation (Fig. 1). The GUI enables one to modify all relevant program parameters, such as the iterative protocol and the simulated annealing schedule, the shape of the restraining potentials, as well as names and locations of the data files. Loading data from a CCPN project is straightforward by specifying the location of the project and the internal name used to identify the respective data set within a CCPN project.

IMPLEMENTATION
ARIA comes as a software library written in the object-oriented programming language Python. The modular design makes it easy for the user to extend and modify the program. The GUI is based on the graphics libraries Tcl/Tk and Tix, interfaced by Python. We use the program CNS (Bruenger et al., 1998) and a simulated annealing (SA) strategy (Nilges et al., 1997) to perform the structure calculation. In principal, the open design of ARIA facilitates the use of other structure calculation engines than CNS. Force field parameters and topology files (version 5.3 of the PARALLHDG parameters), as well as the SA protocol are part of the distribution. ARIA has been tested extensively on different Linux environments, and also runs on SGI machines and Mac OS X. At the time we write this article, more than 500 users worldwide exchange their know-how on a mailing list accessible at http://groups.yahoo.com/group/aria-discuss

ACKNOWLEDGEMENTS
The authors thank W. Boucher, R. Fogh, T. Stevens, W. Vranken, and E. D. Laue for their support in incorporating the CCPN data model. This work was supported by EU grants QLG2-CT-2000-01313 and QLG2-CT-2002-00988. W.R. thanks the European Molecular Biology Organization for financial support.

Conflict of Interest: none declared.

REFERENCES