ABSTRACT

Summary: The conventional approach to calculating biomolecular structures from nuclear magnetic resonance (NMR) data is often viewed as subjective due to its dependence on rules of thumb for deriving geometric constraints and suitable values for theory parameters from noisy experimental data. As a result, it can be difficult to judge the precision of an NMR structure in an objective manner. The inferential structure determination (ISD) framework, which has been introduced recently, addresses this problem by using Bayesian inference to derive a probability distribution that represents both the unknown structure and its uncertainty. It also determines additional unknowns, such as theory parameters, that normally need to be chosen empirically. Here we give an overview of the ISD software package, which implements this methodology.

Availability: http://www.bioc.cam.ac.uk/isd
Contact: wolfgang.rieping@bioc.cam.ac.uk, michael.habeck@tuebingen.mpg.de

1 INTRODUCTION

High-resolution nuclear magnetic resonance (NMR) spectroscopy has become, along with X-ray crystallography, a routine method for determining the 3D structure of biological macromolecules. However, adherent to the method is a certain element of subjectivity, as calculating an NMR structure is rather indirect and involves a number of ad hoc rules, many of which are difficult to derive from rigorous principles. In particular, translating experimental data into geometrical constraints that, combined with physical background information, define the target structure requires some heuristics in order to deal with the errors in the data. The same holds for the estimation of theory parameters (such as the alignment tensor in the case of residue dipolar coupling (RDC) data), for which several empirical rules can be found in the literature (e.g. see Habeck et al., 2008). Both aspects add to the ‘subjectiveness’ of NMR structures and the associated difficulty of deriving the precision of the calculated coordinates in an objective manner.

We have argued that the aforementioned problems can be avoided if structure determination is viewed as an inference problem. The recently introduced inferential structure determination (ISD) framework follows this avenue and employs probability calculus to infer the 3D structure of a molecule, including its uncertainty, from incomplete and uncertain information (Rieping et al., 2005a). Unlike conventional techniques, ISD uses the experimental data to rank all possible conformations of a molecule, instead of converting them into geometrical constraints. Quantitatively, such a ranking requires us to assign a probability to every conformation. To solve an inferential structure determination, all we need to do is explore the distribution of these probabilities.

The probabilities can readily be set up: An error model accounts for deviations calculated from measured data, and a force field that expresses our prior knowledge about biomolecular structures (such as values for bond lengths and bond angles) in the absence of any data. The probability assignment, and hence the inferred structure, is objective because it only depends on the data and required background information. Other unknown parameters, such as the error of the data and theory parameters are estimated during the calculation, along with the 3D coordinates. Hence, ISD has no free parameter.

2 ALGORITHM

One way of exploring the probability distribution arising in ISD is by generating structures with a frequency proportional to the probability of the respective conformation. These samples can then be used, for example, to obtain the most probable structure of a molecule and to calculate the uncertainty of the 3D coordinates. ISD follows a replica-exchange Monte Carlo (MC) sampling strategy to generate probable conformations and parameter values (Habeck et al., 2005). Basically, the algorithm works with ‘replicas’ of the target distribution at different temperatures, and periodically exchanges conformations between the replicas in order to overcome conformational energy barriers. Compared to conventional structure calculation techniques, ISD is computationally more challenging, because a whole distribution of structures needs to be explored rather than locating a number of low energy conformers. Therefore, the algorithm has been parallelized to better support computer clusters. For an average sized protein, and depending on the amount and quality of the data, exploring conformational space takes a few days (assuming 10,000 samples on a PC cluster with 50 nodes).
Most of the calculation time is spent on the convergence of a simulation. The final ensemble typically contains about 1000 representative structures.

3 SOFTWARE

Our program ISD implements the methodology outlined above. An object-oriented software library forms the heart of the program. It provides functionality required to set up a project, to perform the structure calculation, and to analyse the calculation results. A Project file controls each of these steps. It also holds information on the type and location of the experimental data. Supported are NOE intensities (ambiguous and unambiguous), RDCs and scalar couplings (all assigned), as well as distance and dihedral angle information. Associated with each of these parameters is an error model, such as the Lognormal distribution for distances and NOE intensities (Rieping et al., 2005b), to account for deviations calculated from measured data. In order to use the available data in the most optimal way possible, each dataset is described by an individual error model. That way, data of a lower quality are down-weighted automatically (Habeck et al., 2006). The program is particularly suited for inferring structures from sparse data. Theory parameters are estimated automatically during a calculation, and it is therefore not necessary, for instance, to provide RDC alignment tensor parameters or Karplus coefficients etc. However, the user can turn off this feature and provide own estimates.

4 STRUCTURE CALCULATION REPORT

The Report feature of the program provides a convenient way for summarizing the calculation results in the form of a PDF document. A report can be generated during the run-time of a calculation. It provides statistics on the performance of the replica-exchange algorithm to assist the user in assessing the convergence of a calculation. It also contains analyses specific to each dataset, such as its error and estimates of associated theory parameters. The uncertainty of a structure is quantified both directly via confidence intervals for the backbone torsion angles and indirectly, based on a Gaussian approximation of the conformational ensemble (Rieping, 2005). The median CA coordinate uncertainty serves as an approximate estimate of the global structural precision. Other measures can be readily implemented by using the functionality provided by the ISD program library. In addition, typical validation scores, such as Ramachandran statistics, the molecular packing and the normality of the backbone conformation, are calculated with the programs WhatIf and Procheck. To validate a dataset in detail, violation probabilities are calculated to assess the reliability of individual measurements in the light of the entire dataset. Violation probabilities constitute the probabilistic generalization of a violation analysis for NOE data.

5 DATA HANDLING AND CCPN SUPPORT

Experimental data are represented in an XML-based format (following the IUPAC naming system) and can also be provided in X-PLOR/CNS or TALOS format. In addition, the data can be read from a CCPN project. The CCPN data model (Fogh et al., 2005) defines a common storage model that integrates all information emerging in a structure determination project. Supporting the CCPN data model provides direct access to information (e.g. restraint lists) generated by other programs that support CCPN, such as ARIA (Nilges et al., 1997; Rieping et al., 2007) or CcpNmr Analysis (Vranken et al., 2005). Inter-conversion of data formats is not required. Furthermore, the data can be imported from >20 proprietary formats by using the FormatConverter (Vranken et al., 2005). ISD exports the results of a calculation, such as the probabilistic structure ensemble, back to CCPN.

6 SOFTWARE LIBRARY & IMPLEMENTATION

The software library is written in Python and is freely available. Time critical routines, such as energy gradients, are implemented in C and can be accessed from Python. The library provides functionality to access the complete information generated during a calculation and thus enables the user to perform in-depth analyses of the MC samples. Novel NMR parameters that are not yet supported by ISD can be readily incorporated into a calculation by adding user-defined theories and error models. Non-bonded interactions are modelled as in ARIA (purely repulsive; atom radii taken from version 5.3 of the PARALLHDG force field parameters). The program has been tested on different Linux environments. A manual, example calculations, and supplemental material are available from the web site. The ISD discussion group is accessible at http://groups.google.com/group/isd-discuss.

ACKNOWLEDGEMENTS

This work was supported by EU grant LSHG-CT-2005-018988. W.R. thanks E.D. Laue for providing infrastructure, and EMBO for financial support.

Conflict of Interest: none declared.

REFERENCES