

NOE Assignment With ARIA 2.0

The Nuts and Bolts

Michael Habeck, Wolfgang Rieping, Jens P. Linge, and Michael Nilges

Summary

The assignment of nuclear Overhauser effect (NOE) resonances is the crucial step in determining the three-dimensional structure of biomolecules from nuclear magnetic resonance (NMR) data. Our program, Ambiguous Restraints for Iterative Assignment (ARIA), treats Noe assignment as an integral part of the structure determination process. This chapter briefly outlines the method and discusses how to carry out a complete structure determination project with the new version 2.0 of ARIA. Two new features greatly streamline the procedure: a new graphical user interface (GUI) and the incorporation of the data model of the Collaborative Computing Project for the NMR community (CCPN). The GUI supports the user in setting up and managing a project. The CCPN data model facilitates data exchange with a great variety of other programs. We give practical guidelines for how to use ARIA and how to analyze results.

Key Words: Automated assignment; ambiguous distance restraints; CCPN; structure calculation; XML; graphical user interface.

1. Introduction

The interpretation of spectra from nuclear magnetic resonance (NMR) experiments is a complex data analysis problem. Despite great progress toward automation, some crucial analysis steps are still only amenable to semiautomatic or manual treatment. The primary observables for structure determination are cross-relaxation rates measured in a nuclear Overhauser effect spectroscopy (NOESY) experiment. Other experiments provide complementary information that can be used in combination with the nuclear Overhauser effect (NOE)-based approach.

The problem of analyzing a NOESY spectrum is twofold. First, cross-peaks must be attributed to pairs of magnetically interacting spins (NOE assignment

problem). Second, structures must be calculated that fulfill conformational restraints derived from assigned NOEs. Regarding the latter aspect, NMR structure determination benefits from advances in simulation techniques, such as molecular dynamics, minimization, and simulated annealing. Robust methods for structure calculation thus exist, whereas methods for NOE assignment are still in the development stage.

The introduction of restraints from ambiguously assigned cross-peaks marked an important conceptual change in the treatment of degenerate NOE assignments. An ambiguous distance restraint (ADR) (**1**) combines alternative assignment possibilities in one restraint. Molecular conformations can be calculated from a list of ADRs and subsequently used to filter the assignment possibilities. The resulting iterative protocol converges to a structure ensemble and NOE assignments that are consistent. The method has been implemented and made publicly available, as in the program package Ambiguous Restraints for Iterative Assignment (ARIA) (**2–4**).

2. Concepts

The standard procedure to determine macromolecular structures from solution NMR data is to minimize an objective function that incorporates experimental data and physical knowledge. Conformational restraints integrate experimental data; a molecular dynamics force field quantifies physical knowledge.

2.1. NOE Models

Traditionally, NOESY measurements serve as the primary source of structural information. ARIA offers two models to analyze cross-peak volumes quantitatively: the isolated spin-pair approximation and relaxation matrix analysis.

2.1.1. Isolated Spin-Pair Approximation

For short mixing times, a good approximation for relating an NOE volume, V_{ij} , to the distance, d_{ij} , of the two contributing spins is

$$V_{ij} = \alpha d_{ij}^{-6} \quad (1)$$

The scale, α , depends on properties of the system under investigation as well as on the experimental setup and cannot be measured directly. Therefore, it has to be estimated during the course of structure calculation (cf. **Subheading 4.2.1.**).

If an NOE is caused by dipolar interaction of two groups of magnetically equivalent spins, such as methyl groups and aromatic rings, **Eq. 1** must be extended to account for averaging effects. Let I and J denote two groups of spins with n_I and n_J members, respectively. We calculate the theoretical cross-peak volume as r^{-6} average over pairwise contributions:

$$V_{IJ} = \alpha n_I n_J \hat{d}_{IJ}^{-6} \quad \text{where} \quad \hat{d}_{IJ}^{-6} = \frac{1}{N_I N_J} \sum_{I'J'} d_{I'J'}^{-6} \quad (2)$$

Introduction of the *effective* distance \hat{d}_{IJ} retains the functional form of **Eq. 1**. **Equation 2** relies on a discrete slow jump model where spins I and J jump between N_I and N_J equilibrium sites, respectively (**5**). Other models exist but quantitative differences are small (**6**). Most commonly, r^{-6} averaging is applied to protons belonging to methyl groups or aromatic rings.

2.1.2. Relaxation Matrix Analysis

Observed NOE intensities involving two spins are affected by the presence of vicinal spins through spin diffusion. The spatial configuration of all proton spins forms a network that establishes alternative pathways for indirect magnetization transfer. Therefore, interproton distances are mostly underestimated when applying the isolated spin-pair approximation. We use relaxation matrix theory to account for indirect magnetization transfer. This formalism allows the calculation of cross-peak volumes at mixing time τ_m given the volumes at $\tau_m = 0$ and the matrix of auto- and cross-relaxation rates, R (**5**):

$$V_{ij}(\tau_m) = \alpha V_{ij}(0) (\exp(-R\tau_m))_{ij} \quad (3)$$

2.2. Ambiguous Distance Restraints

To formulate restraints for structure calculation, distances are derived from cross-peak volumes. By means of **Eq. 1** the relation

$$D = (\alpha^{-1}V)^{-1/6} \quad (4)$$

converts the observed volume V into a target distance D . Isolated spin-pair approximation and relaxation matrix analysis are only approximate models of NOE volumes; they do not consider all mechanisms that affect relaxation rates. To account for this problem, it is common practice to introduce lower and upper *distance bounds*, L and U , by

$$L = \max(0, D - \Delta), U = D + \Delta \quad (5)$$

to restrain the distance to an interval rather than to a unique value. This reduces systematic bias because of potentially erroneous restraint distances. By default, we use a second order polynomial in the target distance to calculate the error margin $\Delta = 0.12D^2$. Other functional forms are supported. The distance bounds enter the objective function via a *flat-bottom harmonic-wall* potential with linear asymptotes (**2**).

For degenerate resonance assignments with n_c pairs of contributing spins or groups of equivalent spins, we sum over *partial volumes* $\{V_i\}$ (Eq. 2), to calculate the effective distance

$$\bar{D} = \left(\alpha^{-1} \sum_{i=1}^{n_c} V_i \right)^{-1/6} \quad (6)$$

An ADR restrains the effective distance to lie between lower and upper distance bounds. Ambiguous distances have proved to provide an efficient way of expressing ambiguities at the logical level. Logical ambiguities occur during initial cross-peak assignment when compiling lists of possible contributions for every cross-peak. The ambiguity is resolved later during the course of structure calculation.

3. The ARIA Package

The program package ARIA implements the outlined ideas for structure determination from unassigned NOESY spectra. Methods for data management and data analysis are bundled in as a software library, thereby avoiding a rigid monolithic program structure and encouraging the user to modify and extend the standard protocols. This is further facilitated by choosing Python (7), a plain and modern scripting language, for programming.

The ARIA core protocols derive distance restraints from a NOESY spectrum. These restraints define molecular conformations that are compatible with the observed cross-peaks. We use CNS (8) employing a simulated annealing strategy (9) to calculate structures. In principle, other structure calculation programs could replace CNS.

ARIA needs a definition of the molecular system and a NOESY spectrum accompanied by a list of chemical shift assignments. Other experimental data, such as scalar and residual dipolar couplings (RDCs), dihedral angles, hydrogen bonds, and distance restraints, can be added easily.

Two new features enhance the functionality and usability of the program: a graphical user interface (GUI) for complete project management and full support of the Collaborative Computing Project (CCPN) data model (10).

3.1. The Project File

A single file contains the locations of the input data as well as all program and protocol parameters. Thus, the user can easily survey and change program settings by directly editing the project file or by using the GUI. The project file can be created during data conversion (cf. Subheading 5.1.), on the command line, or via the user interface. The latter option is the recommended and most convenient way for creating and modifying projects.

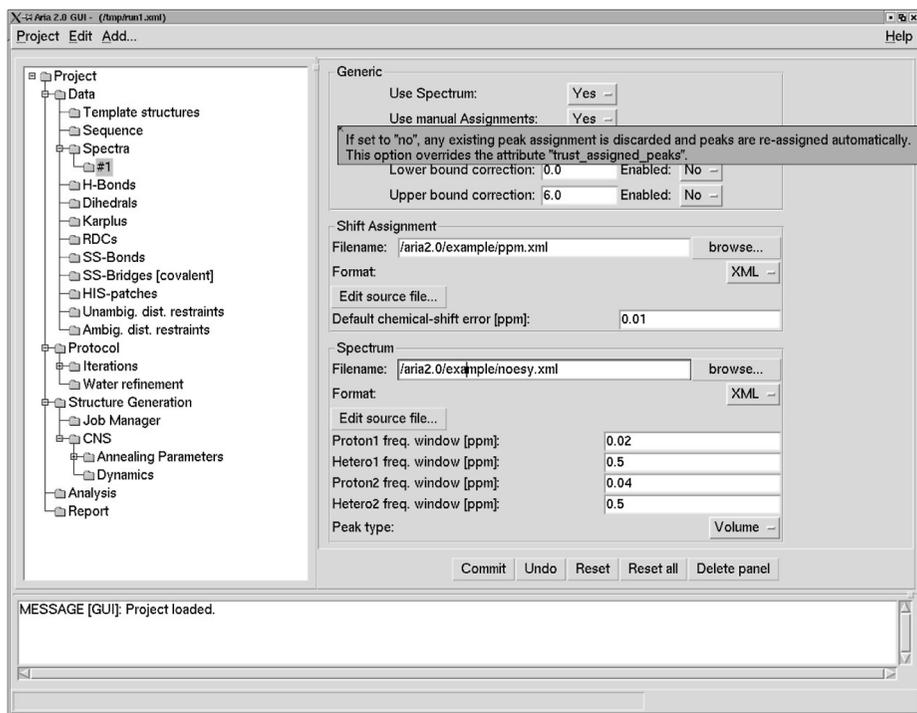


Fig. 1. ARIA 2.0 provides a new user interface for streamlined project management. The complete set of program and protocol settings can now be modified graphically.

3.2. A GUI for Project Management

To rationalize and simplify the setup of an ARIA run, version 2.0 offers a new GUI that replaces the HTML-based interface used in previous versions (*see Fig. 1*). All relevant program and protocol parameters can be adjusted via the GUI. Default values are given if appropriate and a pop-up help menu explains the meaning of each parameter.

3.3. New Formats for Data Representation

NMR data processing packages have introduced several proprietary data formats, each with its own advantages and drawbacks. Experience with earlier versions showed that many problems occurring at later stages of a calculation were the result of misformatted or inconsistent input files. Therefore, version 2.0 runs thorough validation checks on the input data before starting the calculation. We have developed a new data format that is based on the eXtensible Markup Language (XML) (*II*). As of ARIA 2.0, all input data are XML compliant.

XML allows definition of portable and human readable formats for information exchange and supports document validation, thus guaranteeing consistency and integrity of the input data. XML is easy to interpret and can be manually corrected by using either an XML or a text editor. Our data format streamlines the validation process and represents experimental information as explicitly as possible. For instance, we do not support pseudoatoms or wild cards (as, e.g., the methylene group HB#). Instead, the user has to explicitly specify the corresponding group of atoms (i.e., HB2, HB3). Another example is the interpretation of chemical shift assignments of prochiral groups: In the input file, it is explicitly stated whether the assignment is stereospecific or nonstereospecific; this clarifies how an assignment is used throughout the calculation.

3.4. Data Exchange With Other Software Packages

The heterogeneity of NMR data formats hampers integrative analysis approaches. It is often very cumbersome to move data between different software packages, and this process almost always requires manual interference. The CCPN for NMR is an initiative that provides a service to NMR spectroscopists analogous to CCP4 in the X-ray community. One objective is the development of a data model for harvesting and exchange of data between different NMR processing and analysis software packages. ARIA uses the CCPN data model to store all results and analyses of a calculation in a general format. Furthermore, an existing CCPN project can be used to launch an ARIA calculation. However, we also support the more traditional approach of directly converting data generated by other programs. Besides CNS tbl-files, various data formats have become amenable via use of the CCPN program suite: Ansig (*12*), NMRDraw (*13*), NMRView (*14*), Pipp (*15*), Pronto (*16*), CYANA (*17*), XEasy (*18*), Diana (*19*), NMRStar (*20*), and Sparky (Goddard, T.D. and Kneller, D.G. SPARKY3, University of California, San Francisco).

3.5. Complementary Experimental Information

Additional data, such as RDCs, complement NOE measurements and provide valuable information for structure calculation. We allow the integration of various data types.

3.5.1. Hydrogen Bonds

Hydrogen bonds are incorporated by restraining the distance between hydrogen donor and acceptor as well as the distance between acceptor and hydrogen.

3.5.2. J-Couplings

The Karplus curve describes the approximate functional relationship between a three-bond measured J-coupling and the involved dihedral angle. Calculated J-couplings are directly refined against observed J-couplings.

3.5.3. Residual Dipolar Couplings

ARIA offers two approaches to use residual dipolar coupling (RDC) data as restraints: the SANI statement (21) and the VEAN statement (22). For SANI, the user has to specify the rhombicity and magnitude of the alignment tensor. VEAN uses angular restraints that must be precalculated with a separate program (22).

3.5.4. Disulfide Bridges

If the connectivity of a disulfide bridge is known, the respective bond is added to the molecular topology. During torsion angle dynamics (TAD) (23), unambiguous disulfide bridges cannot be treated as fixed; they are modeled as standard harmonic bonds. ARIA automatically deletes the disulfide bonds from the molecular topology during TAD and incorporates them as distance restraints while slowly increasing the weight of the harmonic potential. Unknown connectivities are treated like ambiguous distance restraints from NOE measurements (1).

3.5.5. Dihedral Angle Restraints

ARIA supports the programs TALOS (24) and CSI (25), which predict likely values of phi/psi main-chain dihedral angles given a list of chemical shift assignments. The predictions are incorporated as dihedral angle restraints using a *flat-bottom harmonic-wall* potential.

3.6. IUPAC Atom Name Nomenclature

Markley et al. (26) have proposed a standard for the presentation of NMR structures of proteins and nucleic acids; the standard has been approved by the International Union of Pure and Applied Chemistry (IUPAC). ARIA follows these recommendations. Owing to the inconsistent use of atom name nomenclatures in the NMR community, we want to explicitly list the most common naming problems:

1. The C-terminal carboxyl group is named O' and O''. O'' contains two apostrophes (ASCII 39), not a quotation mark (ASCII 34). The Protein Data Bank (PDB) uses O and OXT or OT1 and OT2 instead.
2. The N-terminus consists of H1, H2, and H3 (not HT1, HT2 and HT3).
3. The protein backbone amide proton is called H (instead of HN).
4. The glycine alpha protons are HA2 and HA3.
5. Pseudoatoms are not supported; r^{-6} averaging is applied to equivalent groups.

4. Program Flow

Figure 2 shows the work flow of an ARIA project. It consists of the following steps: a preparatory stage, iterative structure calculation and NOE assignment, structure refinement in explicit solvent, and final analysis of the results. We explain these steps in more detail.

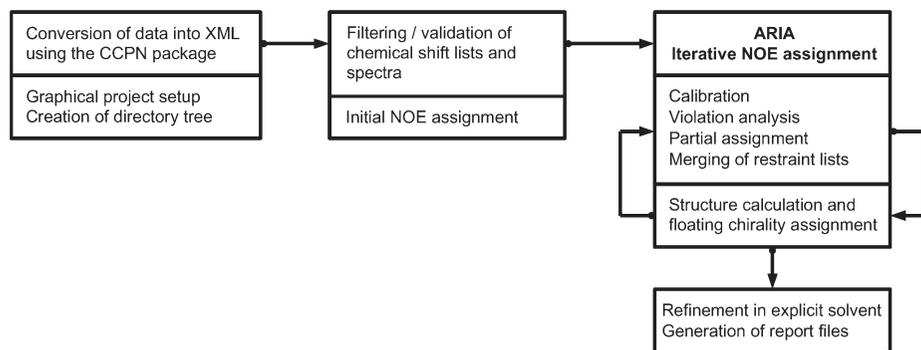


Fig. 2. ARIA integrates the following tasks: conversion of experimental data to XML, iterative structure calculation and NOE assignment, refinement in explicit solvent, and final analysis.

4.1. Preparation Stage

The following steps are performed before structure calculation. First, the data are filtered for errors and inconsistencies. The program then proceeds with creation of the molecular topology. Afterward a seed NOE assignment is derived, based on the chemical-shift lists.

4.1.1. Filtering of Data

ARIA checks the chemical shift assignments for consistency. We consider three cases: First, a unique assignment consisting of one atom and one chemical shift; second, a degenerate chemical shift assignment, where one group of equivalent atoms is assigned to exactly one chemical shift; third, an assignment of the two constituents of a prochiral group, which can have one or two chemical shifts. In the latter case, we use floating chirality assignment (27) to treat the resulting restraints (cf. **Subheading 4.2.2.**). To clean the peak lists, we first delete diagonal peaks that are detected by comparing the proton dimensions within the chemical shift tolerances. Second, peaks with incomplete frequency information or missing peak sizes are removed. We always use absolute values of peak sizes. Files reporting results of data cleaning are generated (cf. **Subheading 5.4.**).

4.1.2. Creating the Molecular Topology

ARIA runs the CNS script `generate.inp` to create a molecular topology file (MTF). The MTF defines the name, chemical type, charge and mass of each atom, as well as the covalent connectivity. The CNS script `geneate_template.inp` constructs the molecule in an extended conformation and writes the coordinates to a PDB file. By default, the extended conformation is randomized and used as the

starting structure for calculation, but it can be replaced with any other structure. In the case of standard biopolymers, the molecular topology is created automatically. The user can define disulfide bridges and histidine protonation states in the GUI. For nonstandard residues or chemical compounds the user has to intervene manually. The script `generate.inp` for setting up the MTF can be edited and then run from ARIA. Alternatively, an MTF can be specified in the project file. Changes must also be made to the CNS topology, linkage, and parameter files.

4.1.3. Initial NOE Assignment

ARIA uses the chemical shift lists to derive possible assignments for every cross-peak. These *seed assignments* solely refer to the peak position in frequency space. A proton is assigned to all cross-peaks whose frequency windows include the proton chemical shift. Either the frequency error or the global frequency window determines the window size. In the case of heteronuclear NOESY experiments, the hetero atom attached to the proton must also match the corresponding chemical shift window. Frequency window sizes are important parameters that ought to be chosen carefully. On the one hand, narrow windows lead to potentially incomplete seed assignments. On the other hand, generous window sizes produce highly degenerate initial assignments that can cause severe convergence problems. The preferred strategy is therefore to choose the window size just large enough that the initial assignments contain the correct assignment, as far as possible. Recommended window sizes are 0.02 ppm for direct, 0.04 ppm for indirect proton dimensions, and 0.5 ppm for heteronuclear dimensions.

Furthermore, the completeness of the chemical shift list affects the accuracy of the initial assignment. Evidently, atoms with missing resonance assignment cannot be assigned to any cross-peak. In this case, automatically generated assignments are almost certainly wrong. As a rule of thumb, the completeness of a chemical shift list should not fall below 90% to achieve reasonable convergence.

ARIA provides two ways for adding manual cross-peak assignments: First, assignments contained in the original peak list are looped through to the spectrum XML file during data conversion (cf. **Subheading 5.1.**). Second, the user may edit the spectrum XML file to add manual assignments (element `assignment`). Manual assignments always enter the seed assignment list and are not modified.

4.2. Iterative Structure Calculation

ARIA detects inconsistent cross-peaks and reduces cross-peak assignment possibilities by iterating through the following scheme:

1. Analysis of the structure ensemble, calibration of spectra, and detection of inconsistent cross-peaks.

2. Creation of a restraint set with fewer assignment possibilities and calculation of a new structure ensemble.

Both stages depend on numerous user-definable parameters. The objective is to achieve mostly unambiguous cross-peak assignments after the last iteration. The parameters are adjusted such that the algorithm converges after a given number of iterations, nine by default. Obviously, the parameter values affect the performance of the algorithm and hence the quality of the produced structures. We have optimized the default parameter settings in terms of consistency and quality of the structure ensemble, not necessarily with respect to speed.

4.2.1. Analysis of the Structure Ensemble

Every cycle begins with an analysis of the previously calculated structures. All analyses relate to ensemble averages. To obtain reliable estimates, it is essential to exclude outliers. The minimization algorithm, however, cannot be expected to converge in every instance. Hence, it is important to identify non-converged structures. We use the total energy to discriminate between converged and nonconverged conformers by selecting S structures with lowest energy, 7 out of 20 by default. In the following, all ensemble averages are calculated with respect to the S lowest energy structures.

4.2.1.1. CALIBRATION

According to **Eq. 4** calculation of distances from observed volumes involves an unknown scale or “calibration” factor, α . We use the ratio of the average of the experimental to the average of the theoretical volumes

$$\alpha = \frac{\sum_i V_i^{\text{exp}}}{\sum_i V_i^{\text{th}}} \quad (7)$$

as an estimator for the calibration constant. Theoretical volumes can be calculated using either the isolated spin-pair approximation or relaxation matrix analysis. For the latter case, test calculations demonstrated a significant improvement of quality indexes, such as Ramachandran coverage, packing quality, and a reduction of NOE restraint violations (**28**). Partial volumes are calculated as ensemble averages. We then introduce a *cutoff distance*, d_c . Cross-peaks with theoretical volume less than $V_c = d_c^{-6}$ are considered as outliers and therefore excluded from calibration. d_c is a user-defined parameter that is set to 6.0 Å by default. We estimate a preliminary calibration factor from the remaining cross-peaks according to **Eq. 7**. Every spectrum is calibrated separately.

For the first iteration, the user may specify an initial structure ensemble for calibration. If omitted, the calibration routine is modified: We assume that the average distance, d_{avg} , of spin pairs causing a NOE is known. Using **Eq. 7**, the initial estimate of the calibration factor, α_0 , is

$$\alpha_0 = n^{-1} \sum_{i=1}^n V_i^{\text{exp}} / d_{avg}^{-6} \quad (8)$$

We use the calibration factor, α (respectively, α_0), to calculate a set of preliminary distance bounds on the basis of **Eq. 6**.

4.2.1.2. DETECTION OF INCONSISTENT RESTRAINTS BY VIOLATION ANALYSIS

To identify wrong assignments and noise peaks, the obtained restraints are subject to a violation analysis. Violation analysis relies on the hypothesis of structural consistency (**2,29**). To assess whether a restraint follows the “general trends” imposed on the structures by the entire data set, we compare its distance bounds with the corresponding distances found in the ensemble. A restraint is “violated” if the distance found in the structure lies outside the bounds by more than a user-defined *violation tolerance*, t . To identify restraints that are systematically violated, each structure in the ensemble is analyzed. Let L_i and U_i denote the lower and upper bound of the i -th restraint. We calculate the fraction, f_i , of structures violating restraint i according to

$$f_i := S^{-1} \sum_{j=1}^S \left[\theta(L_i - d_i^{(j)} + t) + \theta(d_i^{(j)} - U_i - t) \right] \quad (9)$$

where $d_i^{(j)}$ denotes the distance found in the j -th structure; $\theta(\cdot)$ is the Heaviside function. We classify a restraint as violated if f_i exceeds a user-defined *violation threshold*, which is set to 0.5 by default. To calculate the final calibration factor we apply **Eq. 7** to the set of nonviolated restraints.

4.2.2. Reduction of Assignment Possibilities and Calculation of a New Ensemble

Cross-peaks are assigned in an indirect fashion by eliminating unlikely assignment possibilities. Because of the r^{-6} dependence, assignments with large distances contribute only marginally to the NOE intensity. We weight each assignment possibility by its normalized partial volume, $w_i = V_i / \sum_1^{n_c} V_i$. To reduce the number of possibilities, only the m largest contributions satisfying $\sum_1^m w_i \geq w_c$ are kept. w_c denotes a user-defined *ambiguity cutoff* that is set to 1.0 in the first iteration. The cutoff is gradually decreased to 0.8 to obtain almost unambiguous assignments after the last iteration. It may also be advantageous to limit the maximum number of contributions, max_n , to improve the quality of NMR structure ensembles (**3**) (the default value is 20).

4.2.2.1. STRUCTURE CALCULATION

Restraints involving the same set of atoms lead to a biased restraint list. They are the result of symmetry peaks or duplicated peaks from multiple spectra. Equivalent restraints produce artifact in the structures because of overrepresentation of certain distance data. Therefore, we detect nonviolated restraints with equivalent atom content. We keep the restraint with smallest distance and discard the others.

We calculate a new structure ensemble on the basis of the merged restraint set and other restraints (e.g., hydrogen bonds, dihedral angles, RDCs). For every iteration, the number of structures can be controlled with the parameter *n_structures*.

4.2.2.2. FLOATING CHIRALITY ASSIGNMENT

It is often difficult to assign the chemical shifts of the two substituents of a prochiral center stereospecifically. In proteins, these are the two methylene protons or the methyl protons of the isopropyl groups of valine or leucine. A resonance matching one of the chemical shifts in the proton dimensions potentially involves either of the two prochiral substituents. ARIA compensates this lack of information by testing both alternatives during structure calculation. For each calculated structure, the energetically preferred options are written to a file with a `.float` extension.

4.3. Refinement in Explicit Solvent

Because of the simplified treatment of nonbonded forces and missing solvent contacts, calculated structures often show artifacts, such as unrealistic sidechain packing and unsatisfied hydrogen bond donors or acceptors. ARIA provides a protocol for refining macromolecular structures in a shell of water or dimethyl sulfoxide molecules. For each structure of the final ensemble, we calculate a short trajectory with a full molecular dynamics force field, including electrostatic and Lennard–Jones potentials (*see* **ref. 30** for details). The parameters for refinement (`PARALLHDG 5 . 3`) are consistent with the force field used for structure calculation and validation. Hence, there are no systematic differences that could influence validation results.

4.4. Analyses and Output Files

ARIA generates various output files to report analysis results. Calculated structures are stored in PDB format respecting IUPAC atom name nomenclature. For every iteration, ARIA creates the following report files:

1. `noe_restraints.unambig`, `noe_restraints.ambig`
These files tabulate unambiguous and ambiguous restraints, respectively. Restraints discarded by the merging procedure are excluded. For every restraint, information is given on its reference cross-peak, restraint bounds, the average distance found in the ensemble, and the result of violation analysis.

2. `noe_restraints.violations`
Lists all violated restraints sorted with respect to their upper-bound violations and contains the same information as no. 1.
3. `noe_restraints.assignments`
Lists primarily restraint-wise assignments and gives information on whether the assignment(s) stem from fully, partially, or unassigned cross-peaks.
4. `noe_restraints.merged`
Reports all restraints that have been discarded by the merging procedure.
5. `noe_restraints.xml`, `noe_restraints.pickle`
The complete list of NOE-based distance restraints stored in XML and Python binary format for persistent object serialization. Both files are intended to serve as a database for advanced users working with their own analysis scripts.
6. `report`
Summarizes analyses of the restraint lists and the structure ensemble.

4.4.1. MOLMOL

For the last iteration, MOLMOL (**31**) distance restraint files containing lower and upper bounds are written to the subdirectory */molmol/*. These can be used to visualize distance restraints from ARIA and may assist in identification of erroneous restraints.

4.4.2. Quality Checks

ARIA uses the programs WHAT IF (**32**), PROCHECK (**33**), and PROSA II (**34**) to evaluate the quality of both the final set of structures and the solvent-refined ensemble. For every program separate report files, `quality_checks`, are stored in the directories of the respective ensembles.

4.4.3. Miscellaneous Analyses

Several CNS scripts calculate restraint energies, ensemble root-mean-square deviations (RMSDs), and an average structure. Results are stored in */analysis/*.

By default, ARIA creates all analysis and report files except for Python binary data.

5. Working With ARIA

An ARIA calculation is easy to set up and proceeds largely automatically. The standard procedure is to run ARIA repeatedly and use the results to revise the input data (**Fig. 3**). We give some practical guidelines that enable beginners to start their own structure determination project.

5.1. Conversion of Input Data to XML

Prior to structure calculation, all input data must be converted to XML. Besides formats and filenames of the raw data (sequence, chemical shift lists and spectra), data conversion requires some additional information that must be

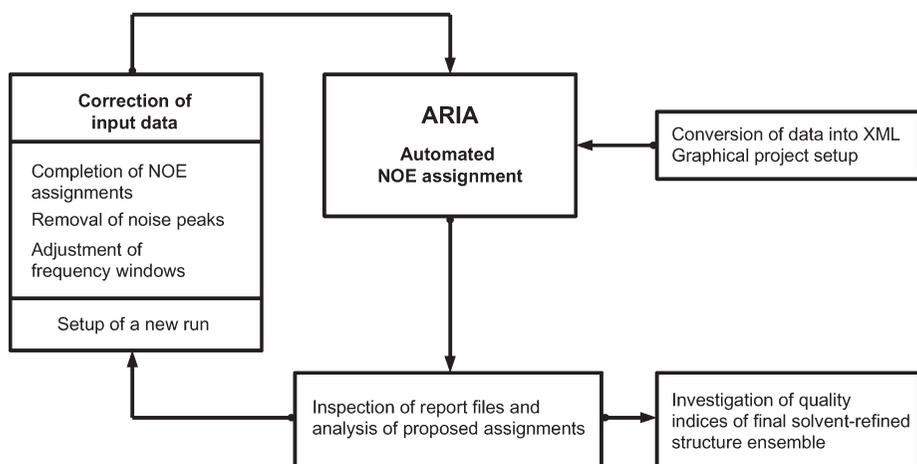


Fig. 3. A typical structure determination project consists of repeated ARIA runs followed by manual correction of the input data.

provided in the form of a simple XML *conversion file*. Most important, the user has to specify the mapping between nuclei and frequency dimensions; for convenience, a preformatted project file is autogenerated. The conversion file should be filled in carefully to avoid the introduction of unnecessary errors at the beginning of a project. A common source for inconsistencies is, for example, mismatches in the residue numbering in the molecule definition and in the other data files. The command

```
aria2 --convert -t conversion.xml
```

creates an empty conversion file template, *conversion.xml*, that must be completed. To start the conversion, invoke the command

```
aria2 --convert conversion.xml
```

A general conversion issue is atom names. Most software packages rely on their own nomenclature and also use wild cards or pseudoatom names. These have to be translated into their IUPAC counterparts. The conversion module is mostly automated. However, in dubious cases, the user is either asked to resolve the ambiguity or is notified on the program's decision. After conversion, the user should nevertheless check whether the data were translated in the intended way. **Figure 4** shows an excerpt of a converted chemical shift list stemming from the HRDC-domain data (35). Further examples as well as a complete conversion file can be found in the directory */examples/werner/ready-to-use* of the distribution. It is important to bear in mind how chemical shift assignments are converted, as this determines the way they are translated into distance restraints.

<pre> 167 2.051 0.001 HB2 10 168 1.970 0.001 HB3 10 169 999.000 0.000 QB 10 </pre>	<pre> <shift_assignment method="FLOATING"> <spin_system averaging_method="NONE"> <atom segid=" " residue="10" name="HB2"/> <chemical_shift value="1.97" error="0.001"/> <chemical_shift value="2.051" error="0.001"/> </spin_system> <spin_system averaging_method="NONE"> <atom segid=" " residue="10" name="HB3"/> <chemical_shift value="1.97" error="0.001"/> <chemical_shift value="2.051" error="0.001"/> </spin_system> </shift_assignment> </pre>
--	---

Fig. 4. Example for the conversion of a chemical shift list. The *left hand side* shows chemical shift assignments in XEasy format. The *right hand side* displays the corresponding assignments in ARIA XML format. Because protons HB2 and HB3 belong to a methylene group, they make up a floating chirality assignment. Pseudo atom names are superfluous (as QB in this case).

5.2. Nonstereospecific Chemical Shift Assignments

ARIA uses r^{-6} averaging (Eq. 2) to treat protons with degenerate chemical shift assignments. Methyl protons usually exhibit degenerate chemical shifts because they are in fast exchange. Consequently, only the whole group can match the proton shift dimension of an NOE peak. The same holds for N-terminal amide protons. Protons in an aromatic ring are also equivalent when assigned to the same chemical shift. In case both frequencies of the two substituents of a prochiral group are identical, the substituents make up a group of equivalent spins handled by r^{-6} averaging. This applies to methylene groups and the isopropyl groups in valine and leucine. Averaging might not always be physically justified. The user should verify whether it is indeed correct. Thymine is the only nucleotide that has a group of fast exchanging methyl protons at the C7 carbon.

Prochiral groups often have ambiguous chemical shift assignments. The generic case is a methylene group giving rise to two distinct resonances that cannot be assigned stereospecifically. Consider, for example, the case where HB2 and HB3 were assigned to different chemical shifts, δ_1 and δ_2 . Floating chirality assignment takes both possibilities, $\text{HB2} \rightarrow \delta_1$, $\text{HB3} \rightarrow \delta_2$ and $\text{HB2} \rightarrow \delta_2$, $\text{HB3} \rightarrow \delta_1$, into account. The same applies when only one shift is given or in the case of an assignment of a wild card corresponding to a methylene group (e.g., HG# in Glu). The prochiral proline N-terminus is handled like a methylene group. The two methyl groups in valine and leucine may have non-stereospecific assignments in which case the floating chirality method is applied. Floating chirality applies also if the HD and HE protons of aromatic rings are assigned to different shifts. In nucleotides methylene groups are located at the 5' carbon and, in the case of DNA, also at C2'.

Parameter	XML element / attribute	Location in GUI	Default value
Project environment	infrastructure	Project	
Project nickname	run		1
File root	file_root		
Working directory	working_directory		
Temporary directory	temp_root		
Data specification	data	Data	
Frequency windows	freq_window_proton1	→ Spectra	0.02
Frequency windows	freq_window_hetero1	→ Spectra	0.5
Trust assignments	trust_assigned_peaks	→ Spectra	no
CNS topology file	topology_definition	→ Sequence	topallhdg5.3.pro
CNS linkage file	linkage_definition	→ Sequence	topallhdg5.3.pep
CNS parameter file	parameter_definition	→ Sequence	parallhdg5.3.pro
Initial ensemble	template_structure	→ Template structures	
Protocol parameters	protocol	Protocol	
Number of structures	n_structures	→ Iterations	20
Distance cutoff d_c	distance_cutoff	→ Iterations	6.0
Violation tolerance t	violation_tolerance	→ Iterations	1000.0 – 0.1
Violation threshold	violation_threshold	→ Iterations	0.5
Ambiguity cutoff w_c	weight_threshold	→ Iterations	1.0 – 0.8
Maximum number of contributions, $max.n$	max_contributions	→ Iterations	20
Number of lowest energy structures S	n_best_structures	→ Iterations	7
Solvent for refinement	solvent	→ Water refinement	water
Structure calculation	job_manager	Structure Generation	
Local CNS executable	default_executable	→ CNS	
Command to start calculation on remote machine	command	→ Job Manager	
CNS executable on remote machine	executable	→ Job Manager	
High temperature steps	steps_high	→ CNS → Dynamics	10000

Fig. 5. Important protocol parameters.

5.3. Specification of Protocol Parameters

Data conversion generates a project file that the user has to complete. Alternatively, he can choose the “New” command in the GUI menu “Project” or execute

```
aria2 --project_template project.xml
```

on the command line to create a new project file. Important program and protocol parameters are compiled in **Fig. 5**. A minimal set of input data consists of the description of the molecule and a NOESY spectrum with a corresponding chemical shift list. To supply further data, such as additional spectra, J couplings, or RDCs, choose the menu “Add” in the user interface. For each spectrum, the default frequency window sizes should be adjusted. Other mandatory parameters concern the project’s infrastructure and shall be explicitly listed here:

1. *Working directory*: directory for storing results of an ARIA run.
2. *File root*: used throughout the project as code for the molecular system, for example, when writing PDB files.
3. *Temporary path*: temporary directory used by CNS during structure calculation.
4. *Local cns executable*: path of the CNS executable on the same machine where ARIA is running.

For all other parameters default settings are provided.

5.4. Project Setup

Before calculations can start, the project must be set up:

```
aria2 --setup project.xml
```

This command reads and validates the project file `project.xml`. If both steps are successful, ARIA will create the directory tree shown in **Fig. 6**. Structure calculation results are deposited in `structures/` where each iteration has its own subdirectory `structures/it*`. All data needed by CNS reside in a separate subdirectory `cns/`. These include protocols (`cns/protocols/`), topology, parameter and linkage files (`cns/toppar/`), and data files in CNS format (`cns/data/*`). CNS specific files usually do not need to be modified. Only if the users want to change a protocol or the molecular topology in a nonstandard way, will they need to edit them.

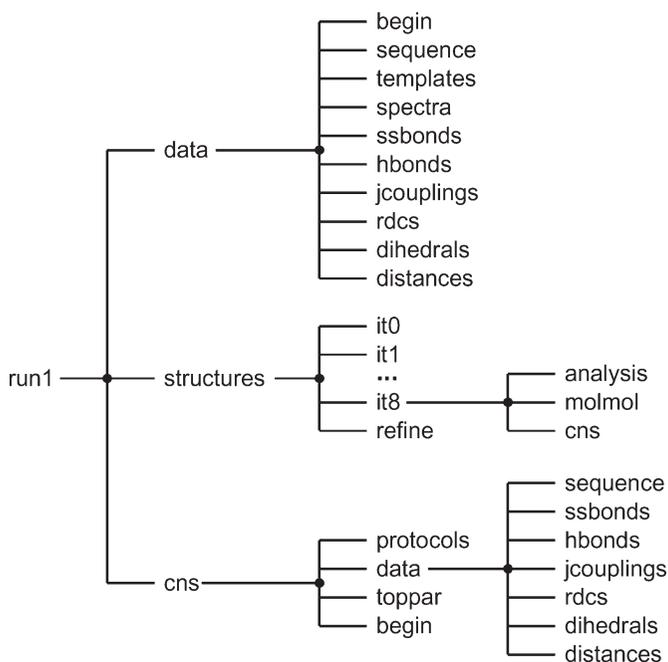


Fig. 6. Directory tree of an ARIA project.

Copies of the experimental data files as well as report files of the data filtering procedures are stored in *data/*. Only local copies are used for structure calculation. Thus, changes in the original files will become active in the next project setup.

5.5. Running ARIA

The setup of the project directory completes the necessary arrangements for starting ARIA. Without further specification, calculations will be performed on the same processor from which ARIA has been started.

5.5.1. Parallel Setup

To accelerate ARIA, we recommend distributing the calculations to multiple processors. We use a temporary shell script to launch the calculations. Hence, the user must provide ARIA with a command (project file, attribute `command`) that first establishes the connection to the remote machine and then starts the script—for example, `ssh machine_name csh`. The parallel setup is easily prepared by using the GUI.

For the first round of calibration, the user can specify an initial ensemble of *template structures* (attribute `template_structure`, see **Fig. 5**). In this case, the default violation tolerance (1000Å) must be reduced to a smaller value (e.g., 5Å).

To start ARIA, use the following command:

```
aria2 project.xml
```

The following sections give some advice on how to inspect the calculation results.

5.6. Convergence Checks

The performance of the minimization protocol depends on both the particular molecular system and on the quality and completeness of the data. Protocol parameters, in particular the number of dynamic steps required for convergence, are primarily determined by the system size. Default values work well for systems up to about 150 residues. However, for larger systems it might become necessary to increase the number of steps. To evaluate whether structure calculation has converged, one first examines the average and the variation of the total energy in the ensemble. As a rule of thumb, well-converged ensembles for systems of about 100 residues show average energies of the order of 1000 kcal mol⁻¹ or less. An energy variation of 10% can be considered normal. The average energy scales approximately linear with the system size. Energy distributions not meeting these requirements can be categorized as follows:

- *Large average energy or large energy variation* indicates that the calculation has not converged. This can be verified by inspecting the conformational variance of the ensemble, for instance in terms of the RMSD (in *it8/analysis/rmsdave.disp*).

In most cases, poor convergence is owing to erroneous protocol settings, in particular the number of dynamics steps or the size of frequency windows, as well as poor data quality.

- *Low average energy, large energy variation.* This case indicates a broad and unpronounced energy minimum that typically occurs when using a small number of restraints. Check the `report` file for information on the number of restraints used for structure calculation. A tight violation tolerance might also cause high restraint rejection rates leading to data too sparse to achieve convergence. In particular, highly ambiguous, incomplete, or noisy data sets require a more generous violation threshold in the first iterations.

5.7. Checking Automatic Resonance Assignments

Figure 7 shows an excerpt of the analysis file `noe_restraints.violations`. It provides information on all restraints rejected by violation analysis. Restraints being violated in the majority of the structures ($\geq 85\%$) in conjunction with large upper-bound violations ($\geq 5\text{\AA}$) are usually inconsistent or stem from wrong assignments and should not be used in a subsequent calculation. To be on the safe side, restraints showing systematic violations of more than 0.1\AA ought to be inspected manually. One should also check frequency window sizes because narrow windows affect the completeness of a cross-peak assignment. It may therefore be advisable to increase the individual window size by 10% in the subsequent run and check the cross-peak again. For automatically generated assignments, one proceeds in the same way. Correct automatic assignments should be treated like manual assignments in a subsequent run.

5.8. Setting Up a New Run

Based on the analyses mentioned in the previous sections, wrong or inconsistent assignments as well as noise peaks have been corrected or marked. The user may also want to tell the program which cross-peak assignments are definitely correct. ARIA therefore distinguishes between NOEs with reliable and

ref_spec ¹	ref_no ¹	lower ²	upper ²	d_avg ³	u_viol	%_viol	a_type ⁴	n_c ⁵
...								
13C NOESY	1568	2.03	5.00	14.73	9.73	100.0	S	2
13C NOESY	1612	1.76	3.30	12.93	9.63	100.0	M	1
13C NOESY	2996	1.89	3.90	13.38	9.49	100.0	A	3
...								

Fig. 7. An excerpt of the file `noe_restraints.violations`. The restraints have been violated in all structures (`%_viol`) and show upper-bound violations (`u_viol`) of more than 5\AA . ¹Reference cross-peak; ²lower and upper distance bounds; ³ensemble average distance; ⁴assignment type: (A)automatic, (S)emi-automatic, (M)anual; ⁵number of contributions.

NOEs with questionable assignments. Reliable assignments always enter structure calculation, that is, some care must be taken not to declare erroneous peak assignments as reliable. To declare a cross-peak as reliable, set its attribute `reliable` (spectrum file) to “yes”. To declare all fully assigned cross-peaks of a spectrum as reliable, set the spectrum-specific attribute `trust_assigned_peaks` to “yes” (see **Fig. 5**).

Changes in protocol parameters, such as those controlling simulated annealing or the initial NOE assignment, will become operative when rerunning the program and do not require further intervention. Previous results, however, in particular PDB files and analysis files, are overwritten. For better bookkeeping, we recommend to use the same project but to create a new run (cf. **Subheading. 5.4.**). Of course, the new run must be set up. To exploit the information generated in the previous calculation the input data must be modified:

- Cross-peak assignments can be augmented by automatically created assignments.
- Erroneous cross-peak assignments can be removed from the respective spectrum.
- Low-energy structures could serve as templates or starting structures.
- Reliable distance restraints can serve to improve convergence.
- Noise peaks should be deleted.

After the new run has successfully been set up, the user proceeds as normal by running ARIA, analyzing the results, and so forth.

5.9. Inspecting the Final Result

5.9.1. Software

When all cross-peaks have been assigned, the quality of the final structure ensemble must be evaluated. However, no cogent definition of a “good” NMR structure exists. Various programs attempt to assess the quality of structures with respect to general properties of folded proteins, such as packing, dihedral angle, or hydrogen-bond configuration. Most methods compare the target structure on the basis of statistics gathered from high-resolution X-ray structures. ARIA uses the programs WHAT IF, PROCHECK, and PROSA II to calculate several quality indices; a summary can be found in the report file. Most indexes depend on both the system under investigation and on data quality, hence no exact values can be given here. However, the following scores should be investigated further:

1. *PROCHECK Ramachandran coverage.* For typical X-ray structures, about 90% of all dihedral angles lie within the preferred region of the Ramachandran plot. For high-quality NMR structures a comparable value is desirable. Unstructured loop regions, however, tend to decrease that value. Typical NMR structures deposited in the PDB show Ramachandran coverages of about 80% (cf. **Fig. 3** in **ref. 36**).

2. *PROSA II Z-score*. Prosa uses a knowledge-based potential, derived from high-resolution crystal structures, to assess how well a structure fits a sequence. The Prosa Z-score quantifies the match between target structure and target sequence compared to alternative folds compiled from the PDB. Z-scores of native proteins are sequence dependent; native structures with 150 residues typically have Z-scores less than -5 .
3. *WHAT IF packing quality and bump-score*. The packing quality measures the agreement between the distribution of local atom configurations and the equivalent distributions derived from a set of high-resolution X-ray structures (37). The bump-score reports van der Waals clashes (32).

5.9.2. Completeness

To further judge the credibility of the structures, it is advisable to reexamine the data. The *completeness* (38) of the restraint set, in particular, provides an insight into the local reliability of each structure. An atom-wise completeness is defined by the ratio of numbers of observed to expected restraints involving the atom. The completeness is zero, if the atom is not restrained at all, and 1 if it is included in all restraints that are structurally possible. Atoms with completeness factors >1 are involved in too many restraints; this might be owing to a wrong resonance assignment. A completeness factor significantly less than 1 indicates an unreliable atom position. The most likely reasons for low completeness values are:

1. Cross-peaks are missing in the data set. They might have either not been picked or could not be assigned because of undersized frequency windows.
2. If the chemical shift list is incomplete some resonances can never be correctly assigned at all.
3. The corresponding cross-peaks were rejected by violation analysis and have thus not been used during calculation (cf. **Subheading 5.7**).
4. If neighboring atoms also show a low completeness, the structure might be locally misfolded.
5. If the local fold of the structure is correct, missing cross-peaks could also provide an indication of local motion or multiple conformations. This hypothesis, however, cannot be verified without further experimental investigation.

6. Conclusions

ARIA is a software package for automated structure determination from NOESY spectra. In principle, the user only has to supply data (chemical shift assignments and NOESY peak lists) stored in one of the standard NMR formats. ARIA then proceeds with data conversion, derivation of ADR lists, and structure calculation. An automated approach has several advantages. First, the data are interpreted in an objective way thus reducing human bias. The method does not require manual intervention, and the results can be judged by external quality criteria. Second, although the program cannot always be expected to

find cross-peak assignments, it is of great use in assisting the experimenter in analyzing his or her data. An otherwise tedious and time-consuming examination of the data can be highly facilitated.

At the time we write this chapter, more than 60 structures have been deposited in the PDB referring to ARIA.

Acknowledgments

M. Habeck and W. Rieping were supported by the European Community (5th Framework program NMRQUAL, contract number QLG2-CT-2000-01313). J. P. Linge thanks the Pasteur Institute and the SPINE network (EU 5th Framework program, contract number QLG2-CT-2002-00988) for financial support.

References

1. Nilges, M. (1995) Calculation of protein structures with ambiguous distance restraints. Automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. *J. Mol. Biol.* **245**, 645–660.
2. Nilges, M. and O'Donoghue, S. I. (1998) Ambiguous NOEs and automated NOESY assignment. *Prog. NMR Spec.* **32**, 107–139.
3. Linge, J. P., O'Donoghue, S. I., and Nilges, M. (2001) Automated assignment of ambiguous nuclear Overhauser effects with ARIA. *Methods Enzymol.* **339**, 71–90.
4. Linge, J. P., Habeck, M., Rieping, W., and Nilges, M. (2003) ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* **19**, 315, 316.
5. Görler, A. and Kalbitzer, H. R. (1997) Relax, a flexible program for the back calculation of NOESY spectra based on complete relaxation matrix formalism. *J. Magn. Reson.* **124**, 177–188.
6. Yip, P. F. and Case, D. A. (1991) Incorporation of internal motions in NMR refinements based on NOESY data. In *Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy* (Hoch, J. C., Poulsen, F. M., and Redfield, C., eds.). Plenum, New York, pp. 317–330.
7. van Rossum, G. and de Boer, J. (1991) Linking a stub generator (AIL) to a prototyping language (Python). In *EurOpen: UNIX Distributed Open Systems in Perspective: Proceedings of the Spring 1991 EurOpen Conference, Tromsø, Norway, May 20–24, 1991* (EurOpen, ed.), EurOpen, Buntingford, Herts, UK, pp. 229–247.
8. Brünger, A. T., Adams, P. D., Clore, G. M., Delano, W. L., Gros, P., Grosse-Kunstleve, R. W., et al. (1998) Crystallography and NMR system (CNS): a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921.
9. Nilges, M., Macias, M. J., O'Donoghue, S. I., and Oschkinat, H. (1997) Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from spectrin. *J. Mol. Biol.* **269**, 408–422.

10. Fogh, R. H., Ionides, J., Ulrich, E., Boucher, W., Vranken, W., Linge, J. P., et al. (2002) The CCPN project: an interim report on a data model for the NMR community. *Nat. Struct. Biol.* **9**, 416–418.
11. The World Wide Web Consortium (1999) *Extensible Markup Language (XML) 1.0, W3C recommendation*. Available at <http://www.w3.org/TR/REC-xml>. Accessed 03/04/04.
12. Kraulis, P., Domaille, P. J., Campbell-Burk, S. L., van Aken, T., and Laue, E. D. (1994) Solution structure and dynamics of ras p21.GDP determined by heteronuclear three- and four-dimensional NMR spectroscopy. *Biochemistry* **33**, 3515–3531.
13. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293.
14. Johnson, B. A. and Blevins, R. A. (1994) NMRView: a computer program for the visualization and analysis of NMR data. *J. Biomol. NMR* **4**, 603–614.
15. Garrett, D., Powers, R., Gronenborn, A., and Clore, G. (1991) A common sense approach to peak picking two-, three- and four-dimensional spectra using automatic computer analysis of contour diagrams. *J. Magn. Reson.* **95**, 214–220.
16. Kjær, M., Andersen, K. V., and Poulsen, F. M. (1994) Automated and semiautomated analysis of homo- and heteronuclear multidimensional nuclear magnetic resonance spectra of proteins: the program PRONTO. *Meth. Enzymol.* **239**, 288–308.
17. Güntert, P., Mumenthaler, C., and Wüthrich, K. (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. *J. Mol. Biol.* **273**, 283–298.
18. Bartels, C., Xia, T.-H., Billeter, M., Güntert, P., and Wüthrich, K. (1995) The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J. Biomol. NMR* **5**, 1–10.
19. Güntert, P., Braun, W., and Wüthrich, K. (1991) Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. *J. Mol. Biol.* **217**, 517–530.
20. Hall, S. R. and Spadaccini, N. (1994) The STAR file: Detailed specifications. *J. Chem. Inf. Comput. Sci.* **34**, 505–508.
21. Tjandra, N., Garrett, D. S., Gronenborn, A. M., Bax, A., and Clore, G. M. (1997) Defining long range order in NMR structure determination from the dependence of heteronuclear relaxation times on rotational diffusion anisotropy. *Nat. Struct. Biol.* **4**, 443–449.
22. Meiler, J., Blomberg, N., Nilges, M., and Griesinger, C. (2000) A new approach for applying residual dipolar couplings as restraints in structure calculations. *J. Biomol. NMR* **16**, 245–252.
23. Stein, E. G., Rice, L. M., and Brünger, A. T. (1997) Torsion-angle molecular dynamics as a new efficient tool for NMR structure calculation. *J. Magn. Reson.* **124**, 154–164.

24. 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.
25. Wishart, D. S. and Sykes, B. D. (1994) The ^{13}C chemical-shift index: a simple method for the identification of protein secondary structure using ^{13}C chemical-shift data. *J. Biomol. NMR* **4**, 171–180.
26. Markley, J. L., Bax, A., Arata, Y., Hilbers, C. W., Kaptein, R., Sykes, B. D., et al. (1998) Recommendations for the presentation of NMR structures of proteins and nucleic acids. *J. Mol. Biol.* **280**, 933–952.
27. Folmer, R. H., Hilbers, C. W., Konings, R. N., and Nilges, M. (1997) Floating stereospecific assignment revisited: application to an 18 kDa protein and comparison with J-coupling data. *J. Biomol. NMR* **9**, 245–258.
28. Linge, J. P. (2001) *New Methods for Automated NOE Assignment and NMR Structure Calculation*. Books on Demand, Norderstedt, Germany.
29. Mumenthaler, C. and Braun, W. (1995) Automated assignment of simulated and experimental NOESY spectra of proteins by feedback filtering and self-correcting distance geometry. *J. Mol. Biol.* **254**, 465–480.
30. Linge, J. P., Williams, M. A., Spronk, C. A., Bonvin, A. M., and Nilges, M. (2003) Refinement of protein structures in explicit solvent. *Proteins* **20**, 496–506.
31. Koradi, R., Billeter, M., and Wüthrich, K. (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55.
32. Vriend, G. (1990) WHAT IF: a molecular modeling and drug design program. *J. Mol. Graph.* **8**, 52–56.
33. Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283–291.
34. Sippl, M. J. (1993) Recognition of errors in three-dimensional structures of proteins. *Proteins* **17**, 355–362.
35. Liu, Z., Macias, M. J., Bottomley, M. J., Stier, G., Linge, J. P., Nilges, M., et al. (1999) The three-dimensional structure of the HRDC domain and implications for the Werner and Bloom syndrome proteins. *Fold. Des.* **7**, 1557–1566.
36. Doreleijers, J. F., Rullmann, J. A., and Kaptein, R. (1998) Quality assessment of NMR structures: a statistical survey. *J. Mol. Biol.* **281**, 149–164.
37. Vriend, G. and Sander, C. (1993) Quality control of protein models: directional atomic contact analysis. *J. Appl. Cryst.* **26**, 47–60.
38. Doreleijers, J. F., Raves, M. L., Rullmann, T., and Kaptein, R. (1999) Completeness of NOEs in protein structure: a statistical analysis of NMR data. *J. Biomol. NMR* **14**, 123–132.