[5] Automated Assignment of Ambiguous Nuclear Overhauser Effects with ARIA

By J. P. LINGE, S. I. O'DONOGHUE, and MICHAEL NILGES

Introduction

ARIA (Ambiguous Restraints for Iterative Assignment) is a software protocol that integrates automated nuclear overhauser effect (NOE) assignments into structure calculations. The user provides a list of assigned chemical shifts and uninterpreted or partly assigned multidimensional homonuclear or heteronuclear-resolved NOE cross-peak lists. Additionally, torsion angle, *J* coupling, residual dipolar coupling, H-bond, disulfide bridge, and planarity restraints can be specified. ARIA converts NOE peak lists from several formats to generate calibrated ambiguous distance restraints. The calibration method includes a CPU-efficient spin diffusion correction in order to improve the accuracy of the distance restraints. Putative artifacts on the peak lists are recognized by violation analysis and can be treated in several ways. ARIA then merges the distance restraint lists and sets up all restraints for automated structure calculation. Explicit assignments are obtained iteratively from chemical shift assignments and successive generations of calculated structures.

A browser-driven user-friendly interface facilitates editing of parameters, protocols for spectra calibration, and Cartesian or torsion angle simulated annealing calculation with CNS.[1] It also provides an interface to interactive assignment programs, which makes it possible to inspect the assignments together with the original data. Scripts for the analysis of the peak tables and the structure ensembles are an integral part of the program and facilitate the control of the automated assignment process. Refinement of the final structure ensemble in explicit water with the CSDX/OPLS hybrid force field is fully integrated. ARIA 1.0 is freely available from www.pasteur.fr/recherche/unites/Binfs/. Running ARIA requires the installation of CNS[1] and Python (www.python.org).

One of the major bottlenecks in the determination of solution nuclear magnetic resonance (NMR) structures of proteins or nucleic acids is the assignment of ambiguous NOEs. More often than not, in the NMR spectra of biological macromolecules, several protons will have the same chemical shift. Therefore, most NOESY cross peaks are ambiguous. That is, in the absence of additional information, they cannot be attributed to a single interaction between two protons.

[1] A. T. Brünger, P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J.-S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T. Simonson, and G. L. Warren, *Acta Cryst. D* **54,** 905 (1998).

0076-6879/00 $35.00

Furthermore, because of limited spectral dispersion, a NOESY cross peak may in fact arise as a sum of two or more distinct NOEs. Ambiguous NOEs that can be reasonably assigned based on the proximity of covalent bonds or secondary structure are often not very useful for determining the tertiary fold of the protein or nucleic acid. Thus, many critical ambiguous long-range NOE interactions can only be interpreted on the basis of an initial three-dimensional model. Structure calculations are therefore usually performed in an iterative ("bootstrapping") way, using preliminary three-dimensional structures based on a few unambiguous NOEs to further assign additional ambiguous NOEs.

Iterative assignment strategies have been in use for some time.[2–6] The main difficulties with complete automation lie in defining rules for explicit assignment based on an ensemble of structures (possibly with incorrect features), in providing mechanisms for correcting wrong NOE assignments, and for treating cross peaks that are genuinely sums of several NOEs. A major step toward a fully automated solution of this problem was the development of computational methods involving ambiguous distance restraints (ADRs).[7,8] Two fully automated iterative assignment methods have been proposed, one based on the use of ADRs (ARIA, Ambiguous Restraints for Automated Assignment)[9,10] the other on self-correcting distance geometry (NOAH)[11].

Prior to the introduction of ADRs, ambiguous data were generally not used in NMR structure calculation for the simple reason that there was no easy way to specify their direct use in the calculation. However, it is simple to see that ambiguous information can be combined to give unambiguous results. As an example, suppose we have the information that ARIA was developed in a German-speaking country. This is ambiguous since it may mean Germany, Austria, or Switzerland. If we are also told that it was developed in a country that borders the sea, this is again very ambiguous, since many countries do. However, combining both these ambiguous statements narrows the possibilities down to only one country—Germany — that satisfied both ambiguous "constraints." In a similar way, unambiguously defined structures can be obtained by combining ambiguous distance data derived from NOESY spectra.

[2] W. Braun, *Quart. Rev. BioPhys.* **19,** 115 (1987).
[3] P. L. Weber, R. Morrison, and D. Hare, *J. Mol. Biol.* **204,** 483 (1988).
[4] P. K. Kraulis, G. M. Clore, M. Nilges, T. A. Jones, G. Pettersson, J. Knowles, and A. M. Gronenborn, *Biochemistry* **28,** 7241 (1989).
[5] R. P. Meadows, E. T. Olejniczak, and S. W. Fesik, *J. Biomol. NMR* **4,** 79 (1994).
[6] P. Güntert, K. D. Berndt, and K. Wüthrich, *J. Biomol. NMR* **3,** 601 (1993).
[7] M. Nilges, *Proteins* **17,** 297 (1993).
[8] M. Nilges, *J. Mol. Biol.* **245,** 645 (1995).
[9] M. Nilges, M. J. Macias, S. I. O'Donoghue, and H. Oschkinat, *J. Mol. Biol.* **269,** 408 (1997).
[10] A. Kharrat, M. J. Macias, T. Gibson, M. Nilges, and A. Pastore, *EMBO J.* **14,** 3572 (1995).
[11] C. Mumenthaler and W. Braun, *J. Mol. Biol.* **254,** 465 (1995).

In this review, we outline the different tasks performed by the program ARIA, describe the theory behind the automatic assignment protocols, including some recent developments, and discuss practical experiences with using ARIA.

Program Flow

An overview of the program flow in ARIA is given in Fig. 1. The principal task of ARIA is to select and assign NOE peaks from peak lists, given the chemical shift assignment of protons and heteronuclei. To facilitate the data exchange between ARIA and the spectrum assignment software, the user can directly start with peak list files from most common spectral analysis programs. Documentation on several possible data formats is included in the program distribution. ARIA provides an HTML interface to edit all the important parameters for setting up and running the iterative NOE assignment and structure calculation. The interface is based on cgi

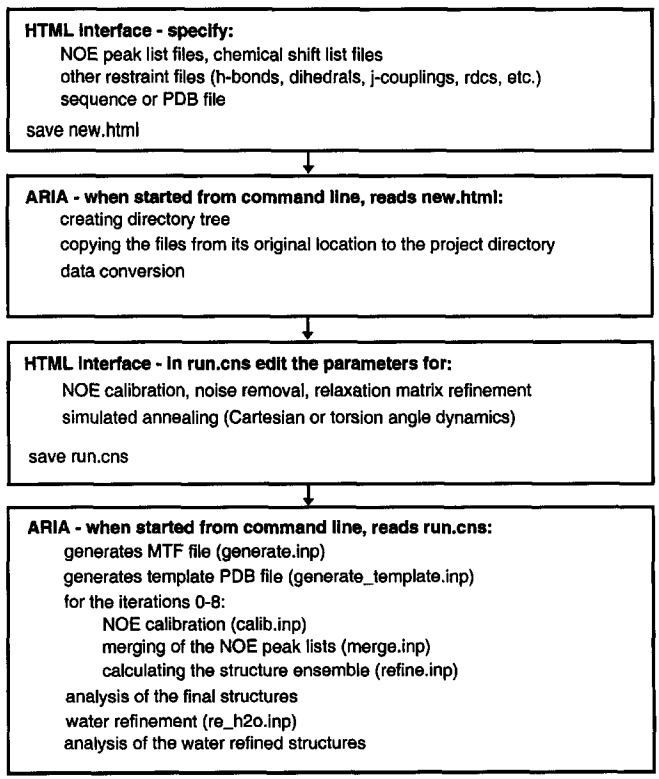

FIG. 1. A flow chart of ARIA.

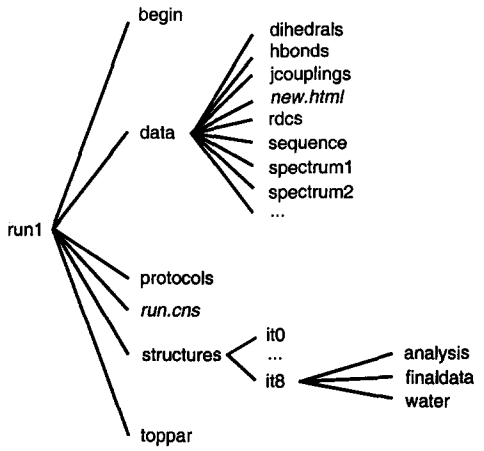

FIG. 2. Directory tree of an ARIA project. Apart from the files *new.html* and *run.cns*, only directories are shown.

scripts developed for the program CNS[1] to read the information from the browser window and to save it to a file. In the ideal case, the only operation necessary to set up a new ARIA calculation is the editing of two files, *new.html* and *run.cns*. ARIA is then called from the command line in the directory where *new.html* or *run.cns* was saved. In the following we describe briefly how a typical project proceeds.

Setting Up a New Project

The file *new.html* contains all information about the location and names of the data files (amino acid or nucleic acid sequence, chemical shift assignments, NOE peaks lists, other experimental data). The user specifies the filenames of the input files and the directory to which the data are written.

Creation of a New Project

ARIA reads the parameters from the file *new.html* and creates a directory tree for the new project (see Fig. 2). All primary data files are copied to the *data* directory. A separate subdirectory of *data* is created for the sequence, dihedral, *J*-coupling, residual dipolar coupling, disulfide bridge, and H-bond restraints, as well as one subdirectory for each NOE spectrum. Files from common spectra assignment programs [ANSIG,[12] NMRView (B. Johnson, Merck), PIPP (D. S. Garett,

[12] P. Kraulis, P. J. Domaille, S. L. Campbell-Burk, T. van Aken, and E. D. Laue, *Biochemistry* **33,** 3515 (1994).

NIH), and XEASY[13]) are converted automatically to ARIA/CNS format and written to the spectrum directory. The user can also convert the data to a CNS readable format manually and start directly from these files. The Python modules NoeList.py and PpmList.py can easily be adapted to convert NMR data in formats not supported by ARIA (examples are provided in the documentation).

Furthermore, there are directories containing the topology and parameter files (*toppar*) and the CNS protocols (*protocols*), which are copied from a central location. The *begin* directory contains the automatically generated molecular topology file (MTF) and template coordinate files, which are necessary for structure calculations with CNS, and lists of equivalent protons and prochiral groups for floating assignments. The *structures* directory is divided into subdirectories for the iterations 0–8, each of which will contain calculated structures for that iteration, along with an energy sorted list of the filenames ("file.list"), restraint lists, and peak tables. The directory *it8* contains subdirectories for the water refined structures and the output of the analysis protocols.

Parameters for Assignment and Structure Calculation

Once the directory tree has been created, the user can use the HTML interface (or a conventional text editor) to edit the file *run.cns*, which contains all the parameters for the assignment and structure calculation process. Of course, the user can also change the CNS protocols themselves in the *protocols* directory, e.g., in order to modify calibration or simulated annealing schemes.

The CNS "module" *run.cns* simply stores all the parameters necessary for ARIA, the simulated annealing protocol, and the analysis as CNS "compound parameters." This module is read by every CNS input file. Based on the iteration number or the spectrum name currently worked on, which are passed to the module, it returns the relevant parameters. This way, only one file needs to be edited to specify parameters for all operations (assignment, structure calculation, analysis). The file also contains some parameters used by the Python script, e.g., the number of processors used in parallel, or the location of the CNS executable.

Running ARIA

Typing "ARIA" on the command line in the directory with the file *run.cns* executes a Python script to perform all operations necessary to assign spectra and calculate structures. The first operations are the creation of the MTF file, a template structure with an extended chain conformation, and a table with the methyl and methylene protons for the floating chirality assignment. The results are written to the *begin* directory.

[13] T. Xia and C. Bartels, "XEASY." Institute of Molecular Biology and Biophysics, Eidgenössische Technische Hochschule, Zürich, Switzerland, 1994.

In every iteration, ARIA calibrates and assigns the NOE spectra, merges the restraint lists, and calculates an ensemble of structures with CNS. The final operation in each iteration is the generation of an energy-ordered list of structures in the file "file.list." By default, the template structure is used for calibration and assignment in iteration 0, with appropriate parameters (see below for possible choices and parameters for iteration 0). In the following iterations, the user specifies how many structures are calculated (typically 20), how many of the best structures are used for calibration, assignment, and violation analysis (typically 7), and how many structures are kept for refinement in the subsequent iteration without randomization (typically 5 to 10). The final structure ensemble can be further refined by a short simulated annealing run in an explicit water shell.[14] ARIA also starts automated analysis protocols for the final structure ensemble and the water refined structures.

When ARIA is restarted (e.g., after a system crash, or when the user wishes to introduce or remove restraints for the subsequent iterations), it checks which files have already been created (i.e., whether "file.list" or the restraint lists are present, and how many calculated structures exist). The program then continues at exactly the same point where it had stopped. The program is parallelized in a straightforward manner and can calculate the structures of one iteration in parallel, either on a multiprocessor computer or on several computers (using a queuing system). This is achieved by creating on the fly a separate input file for each structure in a temporary location. Since CNS log files are rather verbose, they are best written to a temporary disk.

NOE Peak Lists in ARIA

The result of the calibration, assignment, and violation analysis are written to restraint files (with extension ".tbl") and to peak lists (with extension ".list") for each NOE spectrum. The merged restraint lists are then written to two files (ambig.tbl and unambig.tbl), and a merged peak list (called merged.list). The peak lists contain the information present in the experimental NOE peak lists, and all information derived from these data by ARIA: the chemical shifts, chemical shift errors, volumes, volume errors, peak numbers, weights, calibrated distances, upper and lower bounds, assignments, and types (ambiguous/unambiguous, accepted/rejected) for each peak. A complete description of the data format is provided in the ARIA distribution.

Analysis Protocols

In each iteration, the results of the violation analysis (against the uninterpreted peak lists) are written to the calibration/assignment output file ("calib.out"). They

[14] J. P. Linge and M. Nilges, *J. Biomol. NMR* **13,** 51 (1999).

are also documented in the peak lists through the specification of peak types. At the end of the generation of each individual structure, the energies, root mean square (RMS) differences, and violations for different classes of experimental restraints and ideal geometry are calculated and written to the headers of the coordinate files.

For the last iteration and after the water refinement, a more extensive analysis is performed for each class of experimental restraints and the deviations from ideal covalent geometry, and stored in the subdirectory *analysis*. In addition, the circular order parameters are calculated,[15] and a list of ϕ, ψ pairs is generated for a Ramachandran diagram. An average structure is defined based on ordered regions within the final ensemble,[16] the lowest energy structures are superimposed onto this average structure, and atomic RMS differences from the average structure are calculated.

To inspect the restraints and violations with the structures on a graphics display, the restraints are converted to upper and lower bound files in MOLMOL[17] format.

Iterative Assignment and Structure Calculation

Distance Calibration and Relaxation Matrix Calculation

In each iteration after iteration 0, NOE assignment and calibration are based on distances d_{ij} calculated from the ensemble of S lowest energy structures of the previous iteration (typically 7 out of 20). Iteration 0 plays a special role since usually no previous structures are available (see below). For each proton pair, a distance characteristic for the ensemble, $\hat{d}_{ij}$, is calculated as the arithmetic average

$$\hat{d}_{ij} = \frac{1}{S} \sum_{s=1}^{S} d_{ij,s} \tag{1}$$

If the ensemble of structures had physical reality and represented slowly interconverting conformers, the correct average for calculating the relaxation matrix would be the $\langle r^{-6} \rangle^{-1/6}$ average. However, in particular in the early iterations, the ensemble may be very disordered, and this average would weight too strongly to the shortest distance in the ensemble.

In the simplest case, the distances $\hat{d}_{ij}$ are used directly to calibrate experimental peaks and extract distance restraints for iteration $i + 1$, by setting the calibration factor C to

$$C = \sum_{\text{NOEs}} \frac{\hat{d}^{-6}}{V} \tag{2}$$

[15] S. G. Hyberts, M. S. Goldberg, T. F. Havel, and G. Wagner, *Protein Sci.* **1,** 736 (1992).
[16] M. Nilges, G. M. Clore, and A. M. Gronenborn, *FEBS Lett.* **219,** 11 (1987).
[17] R. Koradi, M. Billeter, and K. Wüthrich, *J. Mol. Graph.* **14,** 51 (1996).

where V is the observed NOE volume or intensity, and the sum runs over all experimentally measured NOEs for which the corresponding $\hat{d}$ is smaller than a cutoff (6 Å). The observed distance d^{obs} is calculated as

$$d^{\text{obs}} = (CV)^{-1/6} \tag{3}$$

The calibration will obviously improve from iteration to iteration with the improving structures, but is sufficiently robust that reasonable values are obtained even when using the template structure. Alternatively, reference NOEs with known distance can be used to derive the calibration factor. If the same calibration factor is used for all NOEs, this calibration scheme is also valid for ambiguous NOEs. In this case, $\hat{d}$ is calculated as a "summed distance" over different assignment possibilities [see Eq. (8)].

A better calibration can be achieved by simulating an NOE spectrum A, rather than using the distances $\hat{d}$ directly:

$$A_{ij} = (\exp[-R\tau_{\text{m}}])_{ij} \tag{4}$$

with τ_{m} being the mixing time, and R the relaxation matrix calculated from the $\hat{d}$:

$$R_{ij} = n_i \left(\frac{1}{\hat{d}_{ij}}\right)^6 \frac{\pi}{5} \gamma^4 \hbar^2 \left(\frac{6}{1+4\omega^2\tau^2} - 1\right) \tag{5}$$

Equation (5) can be solved numerically either by diagonalization of the relaxation matrix[18] or by numerical integration of the differential equation.[19] In ARIA, this is performed with an efficient matrix squaring scheme.[20] A strong advantage of the integration scheme is that a distance cutoff d_{cut} in the setup of the relaxation matrix leads to a substantial saving in CPU time through sparse matrix multiplication [from $\mathcal{O}(N_{\text{spin}}^3)$ to $\mathcal{O}(N_{\text{spin}}n^2)$, where N_{spin} is the total number of spins, and n is the typical number of spins within the cutoff d_{cut}].

Details of the inclusion of the relaxation matrix calculation in ARIA will be published elsewhere (J. P. Linge and M. Nilges, to be published). Initial experiences show that severe cases of spin diffusion are reliably detected and corrected. Because of the operations performed in ARIA, in particular the violation analysis and automated restraint removal, the influence on the calculated structures is small.

Target Distances and Error Bounds

Without relaxation matrix calculation, Eq. (3) gives directly the observed distance corresponding to each NOE cross peak using a single calibration factor. If a relaxation matrix calculation is performed, the deviation of the calculated NOE

[18] T. L. James, *Curr. Opin. Struct. Biol.* **1,** 1042 (1991).
[19] M. Madrid and O. Jardetzky, *Biochim. Biophys. Acta* **953,** 61 (1988).
[20] M. W. Kallnik, P. F. Yip, and S. Szalma, *J. Cell. Biochem.* **21B D2,** 421 (1995).

from the ISPA is used as a correction factor for the target distance, essentially as described in ref. 21.

The calibration step used in ARIA does not produce upper limits but gives an estimate of the measured distance itself. Lower and upper bounds are derived from this empirically by estimation of the error from a simple polynomial in d^{obs}:

$$\begin{aligned} L &= d^{obs} - \Delta^- \\ U &= d^{obs} + \Delta^+ \\ \Delta^+ = \Delta^- &= \epsilon_0 + \epsilon_1 d^{obs} + \epsilon_2 (d^{obs})^2 + \epsilon_3 (d^{obs})^3 \end{aligned} \tag{6}$$

By default, $\epsilon_0 = \epsilon_1 = \epsilon_3 = 0$ and $\epsilon_2 = 0.125$, such that the estimated error rises with 12.5% of the square of the target distance.

Noise Removal by Violation Analysis

Structural consistency is often taken as the final criterion to evaluate distance restraints. This is the central idea in the original distance geometry algorithms,[22–24] and it is intimately related to the way distance data are usually specified. That is, the error bounds are set wide enough that all experimental and covalent data are geometrically consistent, and the calculation attempts to find structures that do not violate any of the bounds (i.e., the final value of the energy or target function is zero[25]).

In building three-dimensional structures for NOE and covalent data, most noise peaks will be inconsistent with each other and real peaks. On the other hand, if structures are calculated with restraints from both real and noise peaks, the latter will preferentially be violated in the calculated structures. To facilitate the appearance of violations of incorrect restraints and prevent violations of real restraints, it is necessary to choose an error-tolerant target function. For example, by putting a high penalty on large violations of erroneous restraints, a standard harmonic potential would introduce structural distortions and lead to violations in correct restraints.

Therefore, we expect violations due to incorrect restraints to be present systematically (in the majority of structures) rather than randomly. As proposed in the self-correcting distance-geometry algorithm,[11,26] a violation analysis is performed as follows: calculate the fraction R_{vio} of structures in which a particular restraint

[21] G. Lancelot, J.-L. Guesnet, and F. Vovelle, *Biochemistry* **28,** 7871 (1989).

[22] L. M. Blumenthal, "Theory and Application of Distance Geometry." Chelsea, New York, 1970.

[23] G. M. Crippen, *J. Comp. Phys.* **24,** 96 (1977).

[24] G. M. Crippen and T. F. Havel, "Distance Geometry and Molecular Conformation." Research Studies Press, Taunton, U.K., 1988.

[25] T. F. Havel, *Prog. Biophys. Mol. Biol.* **56,** 43 (1991).

[26] G. Haenggi and W. Braun, *FEBS Lett.* **344,** 147 (1994).

is violated by more than a threshold v_{tol}:

$$R_{vio} = \frac{1}{S_{conv}} \sum_{s}^{S_{conv}} \Theta(D - U - v_{tol}) \tag{7}$$

$\Theta(x)$ is the Heaviside step function (which takes the value of 1 if the argument is larger than 0, otherwise 0), S is the number of converged structures, and U is the upper distance bound. If R_{vio} exceeds the threshold R_{tol} (typically 0.5) for a particular restraint, three options are possible. The peak is either only identified and listed, the bounds can be moved to [0.0 . . . 6.0], or the restraint can be removed from the calculation of the current iteration. The automated removal of each individual restraint can be avoided by marking it directly in the experimental peak list.

Ambiguous Distance Restraints and Partial Assignment

The most important concept to make direct use of ambiguous NOEs in structure calculations is the ADR. This can be derived by defining an effective distance (the summed distance) $\overline{D}$, which contains contributions from distances between all pairs of protons that are possible assignments of the NOE:

$$\overline{D} \equiv \left(\sum_{a=1}^{N_\delta} d_a^{-6} \right)^{-1/6} \tag{8}$$

where the index a runs over all N_δ possible assignments of the NOE within a chemical shift tolerance δ. Within the isolated spin pair approximation (ISPA), $\overline{D}$ has the same dependency on an ambiguous NOE as a single interproton distance on an unambiguous NOE, i.e., the ambiguous NOE depends on the inverse sixth power of $\overline{D}$:

$$\text{NOE} \propto \overline{D}^{-6} = \left(\sum_{a=1}^{N_\delta} d_a^{-6} \right) \tag{9}$$

An ADR is then generated by demanding that $\overline{D}$ stay between limits derived from the size of the NOE given by Eq. (6). For this, the number of possible assignments N_δ, or the number of protons involved, is irrelevant.

Equivalent protons (in aromatic amino acids and methyl groups) are treated as ambiguous NOEs, thus removing the need for any additional corrections.[7,27] This is the same as dividing the experimental NOE volume by the number of equivalent atoms, and using the $\langle r^{-6} \rangle^{-1/6}$ average. It should be noted that, at van der Waals contact, the summed distance between a methyl group and another proton is smaller than 1.8 Å. Therefore, the lower bound should not be set to 1.8 Å, but either reduced to 0.0 Å, or derived from the estimated distances as in Eq. (6).

[27] C. M. Fletcher, D. N. M. Jones, R. Diamond, and D. Neuhaus, *J. Biomol. NMR* **8**, 292 (1996).

To partially assign an NOE peak, we can estimate the contribution of each possibile proton–proton pair a from the inverse sixth power of the ensemble averaged distance $\hat{d}_a$ or the back-calculated NOEs A_a;

$$C_a \propto A_a \tag{10}$$

where the C_a are normalized such that

$$\sum_{a=1}^{N_\delta} C_a = 1 \tag{11}$$

A partial assignment is then achieved by ordering the contributions according to size, and removing the smallest contributions such that

$$\sum_{a=1}^{N_p} C_a > p \tag{12}$$

where p is the assignment cutoff (usually a value between 0.8 and 1.0), and N_p is the number of contributions to the peak necessary to exceed p.

Structure Calculation

The structure calculation protocols have been described in detail elsewhere.[28] The user has the choice between torsion angle[29,30] and Cartesian dynamics.[31] An important difference between the protocols used in ARIA and previous ones[30,31] is that they have been optimized for ambiguous distance restraints and violation analysis. Force constants are therefore relatively low and the potentials used are error tolerant (see ref. 28 for a discussion).

Unassigned prochiral groups are treated with a floating assignment approach.[32] By default, all prochiral groups are automatically specified in the *begin* directory. In cases where stereospecific assignments were made, these can be specified. In this approach, a different assignment for the prochiral groups may result in each calculated structure. In the present version of ARIA, floating is not performed by swapping atom positions, but by swapping the chemical shift assignments during the calculation. Although this approach has the advantage that the final coordinates follow standard atom nomenclature, a problem arises in that the coordinates may not be consistent with the input restraints, since the chemical shift assignments

[28] M. Nilges and S. I. O'Donoghue, *Progr. NMR Spectr.* **32,** 107 (1998).

[29] L. M. Rice and A. T. Brünger, *Proteins* **19,** 277 (1994).

[30] E. G. Stein, L. M. Rice, and A. T. Brünger, *J. Magn. Reson.* **124,** 154 (1997).

[31] M. Nilges, J. Kuszewski, and A. T. Brünger, *in* "Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy" (J. C. Hoch, F. M. Poulsen, and C. Redfield, eds.), Vol. 225 of NATO ASI Series, pp. 451–455. Plenum, New York, 1991.

[32] R. H. A. Folmer, M. Nilges, R. N. H. Konings, and C. W. Hilbers, *EMBO J.* **14,** 4132 (1995).

may change during the calculation. How the results of floating assignment are best documented and submitted to databases is still an open question.

Use of Additional Information

Other Experimental Data

It is clear that the use of additional experimental information will simplify the assignment task since it reduces the conformational space that is searched. A case in point is α-helical regions, which can readily be defined based on secondary chemical shift or coupling constant information yet, because of low dispersion, exhibit many NOE ambiguities. The effect of using torsion angle restraints in combination with H-bond distance restraints for the helices in one example is discussed below.

When setting up a new structure calculation in ARIA, the user can specify restraint files for residual dipolar couplings, torsion angles, *J*-couplings, disulfide bridges, and H-bonds. The energy term in CNS for the residual dipolar couplings evaluates either the angles between the bond vectors and an external axis[33] or, in an extended version, intervector projection angles.[34] The use of intervector projection angles has the advantage that the alignment tensor does not have to be oriented during the simulated annealing protocol. In CNS, *J*-couplings and cross-correlated relaxation data can be either used as torsion angle restraints or directly refined against Karplus curves.[35,36] H-bonds and disulfide bridges are defined as simple (ambiguous or unambiguous) distance restraints that can be switched on and off separately during the simulated annealing protocol. In the present version of ARIA, systematic violations of these restraints are only analyzed in the final iteration.

Apart from refining directly against chemical shifts in CNS,[37,38] these data can be used to derive restraints for hydrogen bonds or ϕ and ψ angles. ARIA includes conversion scripts to create these restraints from the output of the program CSI.[39] Alternatively, the user can create these restraint files manually, using the output of other software packages, such as TALOS.[40] These restraints can, for example, be used to generate starting structures for the iteration 0.

[33] N. Tjandra, D. S. Garrett, A. M. Gronenborn, A. Bax, and G. M. Clore, *Nature Struct. Biol.* **4,** 443 (1997).

[34] J. Meiler, N. Blomberg, M. Nilges, and C. Griesinger, *J. Biomol. NMR* **16,** 245 (2000).

[35] D. S. Garrett, J. Kuszewski, T. J. Hancock, P. J. Lodi, G. W. Vuister, A. M. Gronenborn, and G. M. Clore, *J. Magn. Reson. B* **104,** 99 (1994).

[36] R. Sprangers, M. J. Bottomley, J. P. Linge, J. Schultz, M. Nilges, and M. Sattler, *J. Biomol. NMR* **16,** 47 (2000).

[37] J. Kuszewski, A. M. Gronenborn, and G. M. Clore, *J. Magn. Reson. B* **107,** 293 (1995).

[38] J. Kuszewski, J. Qin, A. M. Gronenborn, and G. M. Clore, *J. Magn. Reson. B* **106,** 92 (1995).

[39] D. S. Wishart and B. D. Sykes, *J. Biomol. NMR* **4,** 171 (1994).

[40] G. Cornilescu, F. Delaglio, and A. Bax, *J. Biomol. NMR* **13,** 289 (1999).

Using Additional Information in Iteration 0

A structure with an extended chain conformation (the template structure) is automatically generated as a starting point for the structure calculation process. There are several options to start the NOE assignment in ARIA, by generating different starting structures in iteration 0.

1. By default, iteration 0 is a full structure calculation, using the template structure for calibration of NOE restraints, but bypassing violation analysis and assignment. That is, the parameters are set such that no restraints are excluded and partial assignments are based on chemical shifts only (the assignment parameter p to 1.0, and the violation tolerance to a very large value, 1000.0 Å). In iteration 1, the violation tolerance is then reduced to an intermediate value (5 Å).
2. Generate an ensemble of random structures, and use them for calibration and NOE assignment in iteration 1. The assignment parameter p is set to a value smaller than but still close to 1 (e.g., 0.9999), and the violation tolerance to a very large value, 1000.0 Å. In this way, NOEs are preferentially assigned for cases in which the interproton distance in the random structure ensemble is consistently short, i.e., predominantly intraresidual and sequential NOEs.
3. Generate a set of structures with correct secondary structure, based on secondary chemical shifts, J couplings, and manual inspection of NOEs diagnostic of secondary structure. The assignment parameter p is again set to a value smaller than but close to 1 (e.g., 0.9999), and the violation tolerance to a very large value, 1000.0 Å. This will help in the partial assignment of NOEs within α helices (and β sheets if the pairing of strands can be defined) already in the early iterations.
4. Use structures precalculated with a set of manually selected and assigned restraints. This is a useful option if the project has started with manual NOE assignment, or if preliminary three-dimensioal structures have been determined, for example, using few NOEs and residual dipolar couplings (e.g., see D. Baker's Web site http://depts.washington.edu/bakerpg/ROSETTA). The assignment parameter p is set to a value smaller than but close to 1 (e.g., 0.999), and the violation tolerance to an intermediate value, e.g., 5.0 Å. This set of parameters will search the peak lists for additional NOEs around the fold found in the precalculated structures.
5. Use models built on the basis of sequence homologies. The assignment parameter p is set to a value smaller than but close to 1 (e.g., 0.999), and the violation tolerance to an intermediate value, e.g., 5.0 Å.

The last option will be increasingly important since in more and more cases homologous proteins with known three-dimensional structures can be identified based on sequence or secondary structure analysis. Currently, homologous structures are generated automatically for each new sequence where a match to a

previously solved structure can be detected (e.g., the HSSP database[41] and the SwissModel server[42]). A future goal could be to link these databases to ARIA, so that, where possible, homology models are generated automatically once the squence is specified. At the moment, the homology models have to be explicitly copied into iteration 0, and the file list ("file.list") has to be created.

One of the effects of using homology models is a reduction of the NOE ambiguity. Experience shows that the principal effect is the exclusion of NOEs inconsistent with the model. This is helpful if very noisy peak lists are used. It should be kept in mind, however, that the initial model introduces a strong bias for subsequent iterations and can potentially lead to precisely determined yet incorrectly folded structures.

Clearly, the use of homology models for NOE assignment in ARIA is only a first step toward a true molecular replacement method for NMR, the challenge being to use this information also to simplify chemical shift assignment. In this spirit, the chemical shift assignment has been extended from preliminary structures, using ARIA, by specifying expected frequencies for nuclei with unassigned chemical shifts.[43]

Practical Experiences with ARIA

Previous reviews have discussed aspects of the application of ARIA, in particular the use of ADRs in the structure determination of symmetric oligomers.[28,44] Here we highlight a few additional points.

Effect of Assignment Schedule

The two parameters with predominant influence on the results of automated assignment with ARIA are the assignment parameter p and the violation tolerance v_{tol}. Values of p smaller than 1.0 eliminate assignment possibilities corresponding to the largest interproton distances in the previous iterations, whereas restraints as a whole are removed if they are systematically violated by more than v_{tol} Å. We have tested the influence of the assignment parameter p on the performance of the program, using the data for the *Escherichia coli* arginine repressor N-terminal domain. The averaged NMR solution structure[45] of this domain served as the reference structure. This structure was largely determined with the help of ARIA

[41] C. Sander and R. Schneider, *Proteins* **9,** 56 (1991).

[42] M. C. Peitsch, *Biochem. Soc. Trans.* **24,** 274 (1996).

[43] B. J. Hare and G. Wagner, *J. Biomol. NMR* **15,** 103 (1999).

[44] S. I. O'Donoghue and M. Nilges, *in* "Structure Computation and Dynamics in Protein NMR" (R. Krishna and J. L. Berliner, eds.), Vol. 17 of Biological Magnetic Resonance, pp. 131–161, Kluver Academic/Plenum, New York, 1999.

[45] M. Sunnerhagen, M. Nilges, G. Otting, and J. Carey, *Nature Struct. Biol.* **4,** 819 (1997).

TABLE I
PARAMETERS USED FOR NOISE EXCLUSION (v_{TOL}) AND PEAK ASSIGNMENT (p)[a]

Iteration (it)	Run1		Run2		Run3	
	p	v_{tol}	p	v_{tol}	p	v_{tol}
1	0.9999	1000.0	0.9999	1000.0	0.90	1000.0
2	0.999	1.0	0.999	1.0	0.90	1.0
3	0.99	0.5	0.99	0.5	0.90	0.5
4	0.98	0.1	0.95	0.1	0.90	0.1
5	0.96	2.5	0.90	2.5	0.90	2.5
6	0.93	0.1	0.90	0.1	0.90	0.1
7	0.90	0.1	0.90	0.1	0.90	0.1
8	0.80	0.1	0.90	0.1	0.90	0.1

[a] For the two spectra of the Arg repressor N-terminal domain (ArgRN).

(in particular, no NOEs were manually assigned for the initial structure calculation), the assignments were extensively checked manually, and the structures were refined with heteronuclear data and refined in explicit solvent. An X-ray crystal structure of a remote sequence homolog[46] shows a virtually identical conformation of the N-terminal domain. In contrast to the solution of the experimental structure, only two homonuclear data sets (spectra recorded in H_2O and D_2O) with a larger frequency tolerance δ_{ppm} (0.03 ppm in both frequency dimensions) were used for the present calculations. We also included the H-bond restraints for the helices, and torsion angle restraints for the backbone angle ϕ where $^3J_{NH-H\alpha}$ coupling constant data were available.

Three different schemes were used (see Table I). In all cases, the structures in iteration 0 were calculated from H-bond restraints and coupling constants alone. The scheme for v_{tol} was the same in all three calculations, with a very large value for iteration 1 (no restraint exclusions), and values between 0.1 and 2.5 Å in the subsequent iterations. By using a larger violation tolerance in iteration 5, ARIA tries to reinclude restraints that are almost consistent with the current structures. The simulated annealing protocol consisted of one high-temperature torsion angle annealing phase, one torsion angle cooling phase, and two Cartesian dynamics cooling phases (similar to that used in ref. 47).

The scheme run1 corresponds to the scheme used in most of the structure determinations so far, run2 uses a constant value for p of 0.9 from iteration 4 on,

[46] J. Ni, V. Sakanyan, D. Charlier, N. Glansdorff, and G. D. van Duyne, *Nature Struct. Biol.* **6,** 427 (1999).

[47] Z. Liu, M. J. Macias, M. J. Bottomley, G. Stier, J. Linge, M. Nilges, P. Bork, and M. Sattler, *Structure* **7,** 1557 (1999).

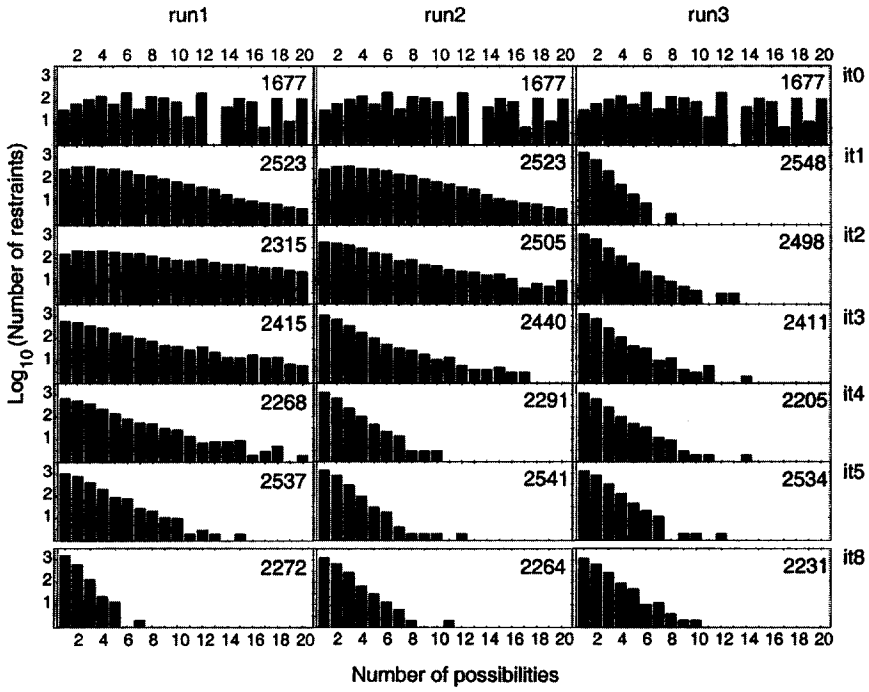

FIG. 3. Ambiguity histograms for three different assignment schemes. The number in the top right-hand corner shows the total number of restraints with fewer than 20 assignment possibilities.

and run3 uses 0.9 as value for p for all iterations, including iteration 1, with random structures as a basis. Clearly, this will result in assignment mistakes in iteration 1.

Figure 3 shows the assignment statistics for the different schemes. There is a substantial amount of ambiguity in the data if the assignment is based on chemical shifts alone ("it0"). Interestingly, there is more ambiguity in the data in iteration 2 than in iteration 1, in spite of a smaller p value. This is because, as a result of increased compactness, more assignment possibilities come within the effective cutoff defined by p.

Figure 4 shows the rms difference from the reference structure (accuracy) and from the average structure (precision) for residues 7 to 70 (the ordered core of the molecule) and the total conformational energy. The first two assignment schemes give almost identical results. The third scheme also converges to the same fold, but is significantly further away from the reference structure than the first two schemes, becuase of overassignment in the early iterations. However, the structures still seem to converge toward the reference structure in iterations 6, 7, and 8; thus, the assignment method is self-correcting. For the first two schemes there is little structural change after iteration 5.

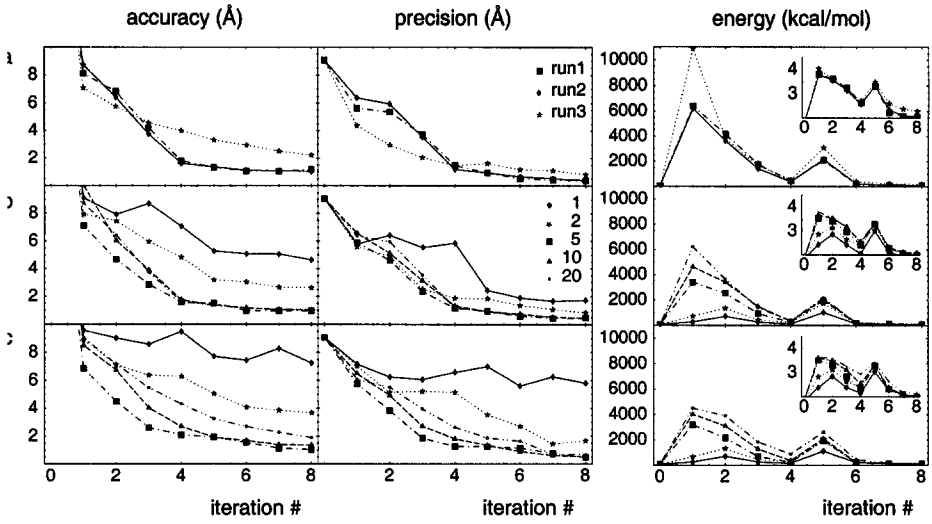

FIG. 4. Accuracy, precision, and total energy for three sets of calculations. (a) Different sets of p values (see Table I). (b) For the set used in run2, different maximum numbers of possibilities per restraint. (c) Same as (b), but the calculations were performed without any of the non-NOE restraints (H-bond restraints for slowly exchanging amide protons in helices, experimental torsion angle restraints). *Insets*: $\log_{10}$ of the energy values.

We have also tested the effect of limiting the maximum number of possibilities a restraint may have to be included in the calculation (parameter n_{max}). Figure 4b shows the result for the n_{max} parameter set to 1, 2, 5, 10, and 20 possibilities. There is a considerable difference in precision and accuracy between 1 and 2 possibilities and the other calculations. The optimal compromise between completeness of the restraints used in the calculation (high n_{max}) and simplicity of the energy surface (low n_{max}) is clearly at $n_{max} = 5$. Additional data (torsion angle restraints and helical H-bonds) improve the convergence, in particular for $n_{max} = 1$ and $n_{max} = 2$ (Fig. 4c).

Caveats

Since restraints inconsistent with the structure can be removed by ARIA, it is difficult to judge the correctness of a final structure by the total energy or violations of the restraints in the final data set. This is evident from Fig. 4, where all final structures have comparable energies, although some are as much as 8 Å away from the reference structure. It is therefore important to check excluded restraints manually.

Model calculations with ADRs[8] and practical experience with ARIA have shown that a number of assignment errors can occur. Similar experiences have been reported with the automatic NOE assignment program NOAH.[48] Careful

[48] C. Mumenthaler, P. Güntert, W. Braun, and K. Wüthrich, *J. Biomol. NMR* **10,** 351 (1997).

TABLE II
PDB ENTRIES USING ARIA AS REFINEMENT MODEL

Structure	PDB access code	Ref.[a]
Pleckstrin homology domain from β-spectrin	1MPH	[1,2]
dsRBD domain	—	[3]
Protein disulfide-isomerase A domain	1MEK	[4]
KH domain from vigillin	1VIG	[5]
Chromo domain	1AP0	[6]
142-residue recombinant prion protein	1B10	[7]
Cyclin-dependent kinase inhibitor p19Ink 4d	1AP7	[8]
First KH domain of FMR1	2FMR	[9]
Spectrin repeat	1AJ3	[10]
Arginine repressor N-terminal domain	1AOY	[11]
Dbl homology domain	1BY1	[12]
Titin fnIII domain	1BPV	[13]
Numb PTB domain–peptide complex	2NMB	[14]
Cdc42–peptide complex	1CEE	[15]
C-terminal domain of p73	1COK	[16]
HAT bromo domain	1B91	[17]
Hydrophobic core variant of ubiquitin	1UD7	[18]
Hydrophobic core variant of ubiquitin	1C3T	[19]
Protein disulfide isomerase B domain	1BJX	[20]
HRDC domain of RecQ	1D8B	[21]
Phl p 2 from timothy grass pollen	1BMW	[22]
CX3C chemokine domain of fractalkine	1B2T	[23]
Frataxin	1DLX	[24]
Shadow chromo domain dimer	1DZ1	[25]
Domain 1 of yeast TFIIS	1EO0	[26]
WASP GTPase binding domain	1EJ5	[27]
Cytokine-binding domain	1D4Q	[28]
β-Lactoglobulin dimer	1DV9	[29]
Numb PTB domain–peptide complex	1DDM	[30]
RNA polymerase subunit 10	1EFY	[31]

[a] References for Table II: [1] M. J. Macias, A. Musacchio, H. Ponstingl, M. Nilges, M. Saraste, and H. Oschkinat, *Nature* **369,** 675 (1994); [2] M. Nilges, M. J. Macias, S. I. O'Donoghue, and H. Oschkinat, *J. Mol. Biol.* **269,** 408 (1997); [3] A. Kharrat, M. J. Macias, T. Gibson, M. Nilges, and A. Pastore, *EMBO J.* **14,** 3572 (1995); [4] J. Kemmink, N. J. Darby, K. Dijkstra, M. Nilges, and T. E. Creighton, *Biochemistry* **35,** 7684 (1996); [5] G. Musco, G. Stier, C. Joseph, M. A. Castiglione Morelli, M. Nilges, T. J. Gibson, and A. Pastore, *Cell* **85,** 237 (1996); [6] L. J. Ball, N. V. Murzina, R. W. Broadhurst, A. R. Raine, S. J. Archer, F. J. Stott, A. G. Murzin, P. B. Singh, P. J. Domaille, and E. D. Laue, *EMBO J.* **16,** 2473 (1997); [7] T. L. James, H. Liu, N. B. Ulyanov, S. Farr-Jones, H. Zhang, D. G. Donne, K. Kaneko, D. Groth, I. Mehlhorn, S. B. Prusiner, and F. E. Cohen, *Proc. Natl. Acad. Sci. USA* **94,** 10086 (1997); [8] F. Y. Luh, S. J. Archer, P. J. Domaille, B. O. Smith, D. Owen, D. H. Brotherton, A. R. Raine, X. Xu, L. Brizuela, S. L. Brenner, and E. D. Laue, *Nature* **389,** 999 (1997); [9] G. Musco, A. Kharrat, G. Stier, F. Fraternali, T. J. Gibson, M. Nilges, and A. Pastore. *Nature Struct. Biol.* **4,** 712 (1997); [10] J. Pascual, M. Pfuhl, D. Walther, M. Saraste, and M. Nilges, *J. Mol. Biol* **272,** 740 (1997); [11] M. Sunnerhagen, M. Nilges, G. Otting, and J. Carey, *Nature Struct. Biol.* **4,** 819 (1997);

manual checking of automated assignments or comparison with a previous manual assignment revealed that, for example, peaks may be assigned differently within the frequency tolerance δ_{ppm}, usually with few structural consequences. A more detailed discussion of possible assignment errors can be found in Refs. 28 and 48.

Errors may appear because of the use of structural consistency to identify noise peaks: good peaks are removed from the data list because of incorrect assignment, and obvious noise peaks may be used as real data. Although ARIA is robust enough to tolerate a certain level of noise,[9] it is clearly advisable to strive for maximal quality in the NOE peak lists used with ARIA.

Structures Solved with ARIA

At EMBL, ARIA was used as integral part in the course of a series of structure determinations[10,49,50,9,51] with homo- and heteronuclear data. In some of these,[10,50,9] manual assignment of the NOEs had proven very difficult, and ARIA was necessary to determine the structure. ARIA is also being used increasingly in a number of other laboratories, as evidenced by a list of Protein Data Bank (PDB) entries that reference ARIA as a refinement method (Table II).

---

[12] B. Aghazadeh, K. Zhu, T. J. Kubiseski, G. A. Liu, T. Pawson, Y. Zheng, and M. K. Rosen., *Nat. Struct. Biol.* **5,** 1098 (1998); [13] C. M. Goll, A. Pastore, and M. Nilges, *Structure* **6,** 1291 (1998); [14] S. C. Li, C. Zwahlen, S. J. Vincent, C. J. McGlade, L. E. Kay, T. Pawson, and J. D. Forman-Kay, *Nat. Struct. Biol.* **5,** 1075 (1998); [15] N. Abdul-Manan, B. Aghazadeh, G. A. Liu, A. Majumdar, O. Ouerfelli, K. A. Siminovitch, and M. K. Rosen. *Nature* **399,** 379 (1999); [16] S. W. Chi, A. Ayed, and C. H. Arrowsmith, *EMBO J.* **18,** 4438 (1999); [17] C. Dhalluin, J. E. Carlson, L. Zeng, C. He, A. K. Aggarwal, and M. M. Zhou, *Nature* **399,** 491 (1999); [18] E. C. Johnson, G. A. Lazar, J. R. Desjarlais, and T. M. Handel. *Structure Fold. Des.* **7,** 967 (1999); [19] G. A. Lazar, E. C. Johnson, J. R. Desjarlais, and T. M. Handel, *Protein Sci.* **8,** 2598 (1999); [20] J. Kemmink, K. Dijkstra, M. Mariani, R. M. Scheek, E. Penka, M. Nilges, and N. J. Darby, *J. Biomol. NMR* **13,** 357 (1999); [21] Z. Liu, M. J. Macias, M. J. Bottomley, G. Stier, J. Linge, M. Nilges, P. Bork, and M. Sattler, *Structure* **7,** 1557 (1999); [22] De S. Marino, M. A. Morelli, F. Fraternali, E. Tamborini, G. Musco, S. Vrtala, C. Dolecek, P. Arosio, R. Valenta, and A. Pastore, *Structure Fold. Des.* **7,** 943 (1999); [23] L. S. Mizoue, J. F. Bazan, E. C. Johnson, and T. M. Handel, *Biochemistry* **38,** 1402 (1999); [24] G. Musco, de T. Tommasi, G. Stier, B. Kolmerer, M. Bottomley, S. Adinolfi, F. W. Muskett, T. J. Gibson, T. A. Frenkiel, and A. Pastore, *J. Biomol. NMR* **15,** 87 (1999); [25] S. V. Brasher, B. O. Smith, R. H. Fogh, D. Nietlispach, A. Thiru, P. R. Nielsen, R. W. Broadhurst, L. J. Ball, N. V. Murzina, and E. D. Laue, *EMBO J.* **19,** 1587 (2000); [26] V. Booth, C. Koth, A. M. Edwards, and C. H. Arrowsmith, *J. Biol. Chem.* (2000); [27] A. S. Kim, L. T. Kakalis, N. Abdul-Manan, G. A. Liu, and M. K. Rosen, *Nature* **404,** 151 (2000); [28] T. D. Mulhern, A. F. Lopez, R. J. D'Andrea, C. Gaunt, L. Vandeleur, M. A. Vadas, G. W. Booker, and C. J. Bagley, *J. Mol. Biol.* **297,** 989 (2000); [29] S. Uhrinova, M. H. Smith, G. B. Jameson, D. Uhrin, L. Sawyer, and P. N. Barlow, *Biochemistry* **39,** 3565 (2000); [30] C. Zwahlen, S. C. Li, L. E. Kay, T. Pawson, and J. D. Forman-Kay, *EMBO J.* **19,** 1505 (2000); [31] C. D. Mackereth, C. H. Arrowsmith, A. M. Edwards, and L. P. McIntosh, *Proc. Natl. Acad. Science USA* **97,** 6316 (2000).

Advantages of Automated Methods

The most important advantage is the substantial speedup of structure calculation that can be achieved by automation of one of the most time-consuming steps. Compared with a manual approach where initial structures are calculated based on a small fraction of the NOEs, the automated approach uses many more data to direct the calculation from the start. The restraints that need to be rejected during the calculation can be viewed as an integral part of the result, can be analyzed in a systematic way, and might even be submitted with the structures and the active restraints. In addition to the more complete description of the result, all parameters used in assigning and selecting restraints are documented, and hence the entire calculation can be reproduced. For manual methods, experience has already shown that such a level of documentation is usually not feasible.

When an automated assignment method is used, the assignment of every single peak will not be checked manually, since this task is about as complex as doing the assignment itself. Although this loss of control over the assignment process may seem dangerous, it is in the spirit of the way the generation of NMR structures has been perceived from the beginning: models should not be built manually but rather calculated with little human intervention. ARIA and similar methods extend this principle to the NOE assignment, and the uniqueness and reproducibility of the assignment can be tested by repeating it with different initial conditions.

Acknowledgments

An initial implementation of the ARIA graphic user interface based on Open-Step and Perl was written by Dinu Gherman and François-Regis Chalaoux. We thank the growing user base of ARIA for fruitful discussions and suggestions (especially Michael Sattler, Remco Sprangers, Alexandre Bonvin, Niklas Blomberg, Helena Berglund, Johan Kemmink, Giovanna Musco, Helen Mott, and Maria Macias), and Lawrence McIntosh for a careful reading of the manuscript. J.P.L. thanks the Boehringer-Ingelheim Fond for a Ph.D. fellowship.

[49] J. Kemmink, N. J. Darby, K. Dijkstra, M. Nilges, and T. E. Creighton, *Biochemistry* **35,** 7684 (1996).
[50] G. Musco, G. Stier, C. Joseph, M. A. Castiglione Morelli, M. Nilges, T. J. Gibson, and A. Pastore, *Cell* **85,** 237 (1996).
[51] J. Kemmink, N. J. Darby, K. Dijkstra, M. Nilges, and T. E. Creighton, *Curr. Biol.* **7,** 239 (1997).