Journal of Biomolecular NMR (2005) 31: 21–34

Influence of chemical shift tolerances on NMR structure calculations using ARIA protocols for assigning NOE data

Michele Fossi[a], Jens Linge[b], Dirk Labudde[a], Dietmar Leitner[a], Michael Nilges[b] & Hartmut Oschkinat[a]

[a]*Forschungsinstitut für Molekulare Pharmakologie, Robert-Rössle-Str. 10, 13125 Berlin, Germany;* [b]*Institut Pasteur, 25-28 Rue du Docteur Roux, 75015 Paris, France*

Received 16 July 2004; Accepted 8 October 2004

Key words: ambiguous distance restraint (ADR), ARIA, chemical shift tolerances, NMR, NOESY, structure calculation

Abstract

Large-scale protein structure determination by NMR via automatic assignment of NOESY spectra requires the adjustment of several parameters for optimal performance. Among those are the chemical shift tolerance windows (Δ), which allow for the compensation of badly matching chemical shifts in the assignment-list and peak-lists, and the maximum number of assignment possibilities allowed per peak (n_{max}). Here, we test the influence of different values for Δ and n_{max} on the performance of automated assignment of NOESY spectra by ARIA. Using Cesta.py (a Python script available from http://pasteur.fr/binfs/), we analyse the number of rejected peaks and the average number of assignments as a function of Δ and derive criteria for optimising Δ and n_{max} prior to structure calculation. The analysis also makes it possible to detect inconsistencies in the dataset, e.g., badly matching frequencies in the NOESY peak-lists and in the provided assignment-list. We show that ARIA can deal with a large number of assignment possibilities for each peak, provided the correct option is present, and that consequently narrow tolerances should be avoided.

Abbreviations: NMR – nuclear magnetic resonance; NOE – nuclear overhauser effect; rmsd – root mean square deviation; 2D – two-dimensional; 3D – three-dimensional; ARIA – ambiguous restraints for iterative assignment; ADR – ambiguous distance restraint; Δ – vector of chemical shift tolerances; Δ_{max} – values of Δ for which the number of accepted peaks is maximal; $n(C_j)$ – number of assignment options for the peak C_j; n_{max} – maximum number of assignment possibilities allowed per peak; n_{av} – average number of assignment possibilities per peak in the first iteration; ArgR – arginine repressor N-terminal domain; – HRDC – helicase and RNaseD C-terminal domain; EVH1 – Ena/VASP Homology 1 domain; VASP – vasodilator stimulated phosphoprotein.

Introduction

The use of NMR structure determination in structural genomics projects is still severely hampered by incomplete automation and standardisation of the data analysis step (resonance assignment, peak picking, NOE assignment, etc.). Despite the development of strategies for including powerful new experimental parameters such as residual dipolar couplings (Tjandra et al., 1997), high resolution NMR structure determination still mostly depends

*To whom correspondence should be addressed. E-mail: oschkinat@fmp-berlin.de

on distance restraints from NOESY cross-peaks. The manual identification of NOESY cross-peaks (NOEs) is an error-prone and tedious process, demanding automation by robust computer algorithms. Several routines for the assignment of NOEs have thus been published (ARIA – Kharrat et al., 1995; Nilges et al., 1997; Nilges and O'Donoghue, 1998; Linge et al., 2001, 2003, CANDID – Herrmann et al., 2002, DYANA – Güntert et al., 1997, KNOWNOE – Gronwald et al., 2002, NOAH – Mumenthaler and Braun, 1995; Mumenthaler et al., 1997, AUTOSTRUCTURE – Moseley and Montelione, 1999; Moseley et al., 2001) and have been recently reviewed (Güntert, 2003). All these programs require the sequence-specific chemical shift assignment (assignment-list) and lists with cross-peaks from NOESY-type spectra (peak-lists) as input.

In these programs, several parameters can be adjusted, e.g., the set of chemical shift tolerances (Δ) associated with each dimension of each spectrum, in order to account for the unavoidable experimental uncertainties in determining peak positions. Often, in the case of very complete and redundant datasets (including a set of additional distance restraints like residual dipolar coupling, hydrogen bond and dihedral angle restraints), the influence of Δ on the calculations is not dramatic, hence the use of default values for this parameter is common (Herrmann et al., 2002). In contrast, in the more challenging case of structure calculations based on unassigned NOE data alone and especially without hydrogen bond restraints, the choice of Δ may play a crucial role. However, when choosing values for Δ, the user has limited criteria to make a rational choice. The optimal Δ is not known *a priori* and experience suggests that digital resolution alone is an insufficient guide for choosing correct values. In fact, other factors (line-width, resonance dispersion, presence of noise or artefacts, sample instability, varying measurement conditions, etc.) all contribute to the actual uncertainty in chemical shift.

A second important parameter that can be adjusted prior to calculations is the maximal number of assignment possibilities allowed per peak ($n_{\max}$). To restrict the computational effort, cross-peaks displaying more than $n_{\max}$ alternative assignment options are not used for the structure calculation. The choice of Δ should not be made independently of that of $n_{\max}$, since both parameters together determine the number of accepted peaks and the number of assignment possibilities.

Here, we derive a procedure for choosing optimal values for Δ and $n_{\max}$ via an analysis of the NOE assignment in the first iteration of ARIA, prior to structure calculations. This analysis provides diagnostic information regarding the consistency of resonance assignment-list and peak-lists and about the degree of spectral overlap affecting the spectra. We present a script (Cesta.py*) to automatically perform this pre-calculation analysis: the output of this analysis is the evaluation of four diagnostic functions (defined hereafter), which provide insight into the peculiarity of each protein dataset.

We then performed structure calculations for five different proteins, using several different combinations of Δ and $n_{\max}$, in order to understand their influence on the quality of the structures calculated by ARIA. This allowed us to develop a strategy for predicting optimal values of these two parameters by a simple pre-calculation analysis.

Theory and computational methods

Peak annotation, distance restraint evaluation and structure calculation

The co-ordinates of a 2D NOESY cross-peak are the chemical shifts of two interacting protons. In 3D or 4D NOESY spectra, the 2D peaks are dispersed along one or two more orthogonal axes using the chemical shifts of one or two bonded heteronuclei. The peak co-ordinates in 3D or 4D NOESY spectra are therefore the chemical shifts of the two protons and those of the heteronuclei. Peak-picking is the procedure generating the peak-list of the spectrum. Peak-picking associates every signal intensity identified as cross-peak j to a chemical shift vector $C_j = [(c_j^{\text{het1}}), c_j^{\text{prol}}, (c_j^{\text{het2}}), c_j^{\text{pro2}}]$, entry of the peak-list C. The values referring to the heteronuclear dimensions of the spectrum are indicated in parenthesis since they are only present in 3D and 4D spectra.

Formally, chemical shift values are affected by an intrinsic degree of uncertainty represented

* Available from http://pasteur.fr/binfs/

by the vector of digital resolutions $R = [(r^{\text{het1}}), r^{\text{prol}}, (r^{\text{het2}}), r^{\text{pro2}}]$. In practice, changing experimental conditions in acquiring different spectra, peak overlap, heating effects, etc. further contribute to the actual uncertainty that affects NOE cross-peak co-ordinates. These factors may influence every peak individually. For these reasons, it is more appropriate to use the term *actual chemical shift uncertainty*, that we represent with the vector $U_j = [(u_j^{\text{het1}}), u_j^{\text{prol}}, (u_j^{\text{het2}}), u_j^{\text{pro2}}]$. It is a function of each individual peak and not a global parameter of the spectrum like the digital resolution. In order to account for the limited precision in chemical shift measurements and for the systematic experimental errors, ARIA, like other automatic methods for the assignment of NOESY spectra (NOAH, CANDID, KNOWNOE…), globally applies a vector of chemical shift tolerances $\Delta = [(\delta^{\text{het1}}), \delta^{\text{pro1}}, (\delta^{\text{het2}}), \delta^{\text{pro2}}]$. Sufficiently large values for Δ should be chosen to compensate for all sources of inconsistencies between resonance assignment-list and peak-lists.

During the *NOE annotation*, the program generates a list of assignment options for each peak. This procedure depends intimately on Δ. Every set of spins represented by the two couples of frequencies $A_m = [(a_m^{\text{het1}}), a_m^{\text{pro1}}]$ and, $A_n = [(a_n^{\text{het2}}), a_n^{\text{pro2}}]$, contained in the resonance assignment-list A, which fulfils the conditions:

$$\begin{cases} C_J^{\alpha} - \delta^{\alpha} \le a_m^{\alpha} \le c_J^{\alpha} + \delta^{\alpha} & \alpha = (\text{het1}), \text{pro1}, \\ C_J^{\beta} - \delta^{\beta} \le a_n^{\beta} \le c_J^{\beta} + \delta^{\beta} & \beta = (\text{het2}), \text{pro2}, \end{cases} \tag{1}$$

is accepted as a possible assignment of the peak C_j. NOEs with no assignment options or with a number of assignment options which exceeds $n_{\max}$ are rejected. Depending on the number of assignment options according to Equation 1, the accepted NOEs are divided into *unambiguous* (only one assignment possibility) and *ambiguous* (several assignment possibilities).

In case of complete resonance assignment, if

$$\begin{cases} u_j^{\alpha} < |\delta^{\alpha}| & \alpha = (\text{het1}), \text{pro1}, \\ u_j^{\beta} < |\delta^{\beta}| & \beta = (\text{het2}), \text{pro2}, \end{cases} \tag{2}$$

the chemical shift tolerances account for the actual uncertainty affecting the peak C_j and the correct assignment is taken into account: we will refer to this peak as *correctly annotated*. We will refer to a peak as *incorrectly annotated* anytime one or more conditions in Equation 2 are not satisfied.

The annotated NOEs are used to generate structural restraints. ARIA is able to handle ambiguous NOEs by treating them as ambiguous distance restraints (ADR) (Nilges, 1993, 1995). In this approach, every ambiguous cross-peak is treated as a superposition of signals arising from all possible assignments, each of which contributes to the global volume of the peak, proportional to the inverse sixth power of the corresponding interatomic distances d_k. Distance restraints are calculated by

$$\overline{D}_j \equiv \left(\sum_{k=1}^{n(C_j,\Delta,A)} d_k^{-6} \right)^{-1/6}, \tag{3}$$

where $n(C_j, \Delta, A)$ represents the number of assignment possibilities for the peak C_j, which depends on Δ and A according to Equation 1. The distances d_k are calculated from preliminary template structures: during the initial assignment, when no preliminary structure is available, a template with extended main chain conformation is used. The presence of some incorrect options among the assignment possibilities of an ADR does not lead to inconsistencies, as long as the correct assignment is present, since the r^{-6} weighted average distance $\overline{D}_j$ is always shorter than and strongly weighted to the shortest of the distances d_k. In contrast, whenever the correct assignment is not included among the assignment possibilities, the derived distance restraint is likely to be inconsistent with the others and thus potentially able to induce distortions in the structures. After their evaluation, distance restraints are subjected to simulated annealing (Brunger et al., 1998) to generate a set of structures. ARIA then repeats the cycle (iteration) of peak assignment, restraint evaluation and structure calculations several times, until satisfying convergence of the structure bundles is achieved. During the iterations, an increasing number of less-representative assignment options are rejected for each peak in Equation 3 (Linge et al., 2001), resulting in the unambiguous assignment of most ambiguous cross-peaks.

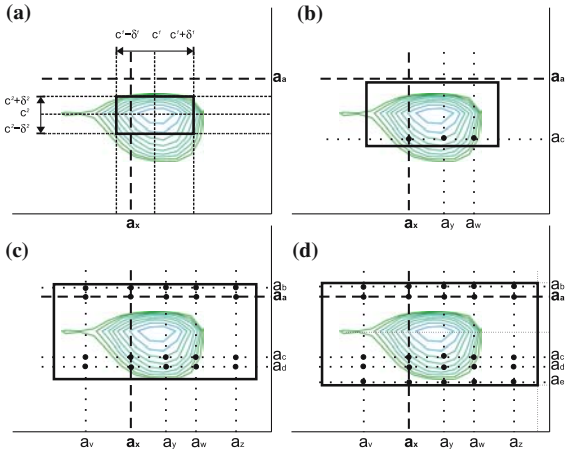

Figure 1. Influence of Δ and n_{max} on the annotation of a generic cross-peak. In this example, n_{max} is set to 20. The black dots indicate the assignment options for the peak, obtained by taking all combinations of resonance assignments falling inside different tolerance windows. The co-ordinates (a_a, a_x) (bold dashed lines) represent the only correct assignment for the peak. (a) The peak is rejected because of a lack of assignment possibilities; (b) the peak is accepted as ambiguous restraint but is incorrectly annotated, because the correct frequency a_a lies outside the tolerance window; (c) the peak is accepted as an ambiguous restraint and is correctly annotated: however, the large number of ambiguities makes it a very loose restraint; (d) the peak is rejected because the number of assignment possibilities exceeds n_{max}.

Influence of Δ and n_{max} on the peak assignment

Δ and n_{max} influence the automated assignment of the NOESY spectrum in various ways. The effects of four different choices for Δ on the assignment of a cross-peak are shown in Figure 1. *C* is a cross-peak of a 2D spectrum with co-ordinates c^1 and c^2. Let a_a, a_b … and a_v, a_w… be examples of resonance assignments close in frequency to c^1 and c^2, respectively. The co-ordinates (a_a, a_x) (bold dashed lines) indicate the only correct assignment for the peak *C*; any other combination of resonance assignments represents an incorrect assignment option. Black dots designate the accepted assignment possibilities for the peak *C*. In this example, $n_{max} = 20$ (the default value) is chosen. In Figure 1a, the choice of very narrow Δ values leads to the rejection of the cross-peak due to the lack of assignment possibilities, as no frequency in the resonance assignment-list matches the range $[c^2-\delta^2, c^2 + \delta^2]$. In Figure 1b, the choice of slightly larger Δ values leads to some assignment possibilities, hence the peak is accepted. However, since the tolerance window is too small to compensate for the actual uncertainty in chemical shift position (i.e., one of the conditions of Equation 2 is not satisfied), the correct frequency a_a lies outside the tolerance window and the assignment possibilities do not contain the correct one. As a result, the peak is accepted but is incorrectly annotated, thus it will be incorrectly assigned at the end of the calculation. In contrast, with much larger Δ values (Figure 1c), the correct assignment is taken into account, although together with many more assignment possibilities. Since the total number of assignment options (20) does not exceed n_{max}, the peak is accepted and correctly annotated. However, due to its high ambiguity, the derived structural restraint will be very loose. A further increase of Δ (Figure 1d) leads to the removal of the peak, since the number of assignment options (25) exceeds n_{max}.

This example shows that, depending on the values assigned to the parameters Δ and n_{max}, a cross-peak can be accepted or rejected, and if accepted, correctly annotated or incorrectly annotated. Therefore, Δ and n_{max} influence globally the calculation by determining the number of accepted NOEs, the percentage of these that are correctly annotated and the average number of assignment possibilities per peak. Thus, we need a strategy to choose Δ large enough to avoid the exclusion of the correct assignments (Figures 1a and b), without increasing excessively the number assignment options (Figures 1c and 1d).

A tool for a pre-calculation analysis of the automated NOE assignment: Cesta.py

The set-up of all calculations plus the collection and analysis of the results presented in this work were performed automatically with the help of the Python script Cesta.py (ChEmical Shift Tolerances Analysis).

This Python script analyses the influence of chemical shift tolerance windows in ARIA v1.2 calculations. The user can run the script after the set-up of an ARIA v1.2 run, when the full ARIA directory tree is already present and the parameter file *run.cns* has already been edited. The script sets up a series of analogous ARIA runs differing only in the values of Δ, which are increased from

small to large values. The script starts each of these ARIA runs and allows the software to annotate the cross-peaks and subsequently merge the various peak-lists in the first iteration. Merging indicates the process of creating a unique peak-list (the *merged list*) from all supplied peak-lists by removing duplicate peaks, arising from the same NOE interactions being present in different spectra. We indicate with ${}^{m}N_{\mathrm{tot}}$ the total number of entries in the merged list. The script interrupts the ARIA calculation just after the annotation and the merging of spectra in the first iteration, prior to any structure calculation. The script then analyses the annotated spectra and the merged list and evaluates for each of them the following four diagnostic functions (i)–(iv):

(i) $N^{\mathrm{rej}}_{\mathrm{noassig}}(\Delta)$ *the number of rejected peaks due to a lack of assignment options as function of* Δ

This function depends on the quality of the alignment of chemical shifts between the assignment-list and the NOESY peak-list and allows for the assessment of differences between both lists. In the ideal case of optimal alignment and complete resonance assignment, no peak is rejected due to a lack of assignment possibilities even for extremely small Δ values, because at least the correct assignment is taken into account for each peak. On the contrary, when dealing with real datasets, the frequencies in the (rarely complete) resonance assignment-list match only within a certain error limit the chemical shift co-ordinates of NOESY cross-peaks. The poorer the consistency between the resonance assignment-list and the peak-lists, the larger the area defined by the function curve and the x-axis (compare the solid and the dotted lines in Figure 2a). Thus, $N^{\mathrm{rej}}_{\mathrm{noassig}}(\Delta)$ is a useful diagnostic function to quantify the agreement between the frequencies in the two lists and, consequently, to identify those datasets which suffer from dramatic frequency inconsistencies and to which larger Δ should be applied. In these cases, the digital resolution alone would be a misleading parameter as a basis for the choice of Δ.

Values of Δ which leave many cross-peaks unassigned are very likely to underestimate the real uncertainty affecting all other cross-peaks; such values should be avoided, as they lead to unnecessary peak rejection (Figure 1a), and, even worse, to the acceptance of many incorrectly annotated peaks (Figure 1b). Thus, the point where $N^{rej}_{\mathrm{noassig}}(\Delta)$ becomes minimal provides a criterion to set a lower limit for Δ.

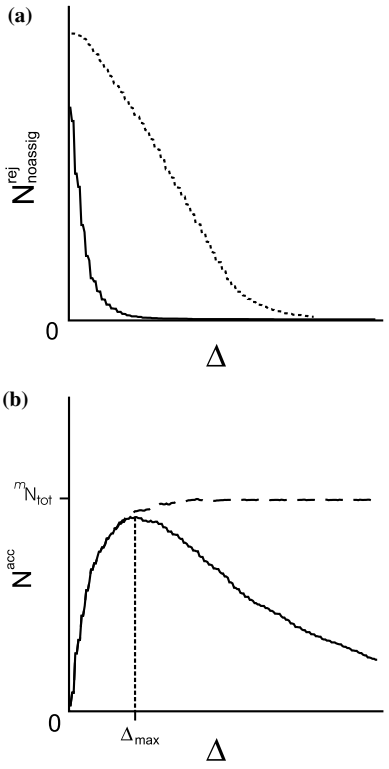

Figure 2. (a) The number of rejected peaks due to a lack of assignment options as a function of Δ ($N^{\mathrm{rej}}_{\mathrm{noassig}}(\Delta)$) for two hypothetical datasets, one featuring a good (solid) and the other a bad (dotted) frequency alignment between the resonance assignment-list and peak-lists.(b) Typical behaviours of the number of accepted peaks $N^{\mathrm{acc}}(\Delta)$ when the removal of the most ambiguous peaks by $n_{\max}$ is on (solid) or off (dashed). In the first case, the function displays a maximum in correspondence to $\Delta = \Delta_{\max}$. In the second case, the function reaches a constant value for large Δ equal to the total number of entries in the merged list ${}^{m}N_{\mathrm{tot}}$.

(ii) $N^{\mathrm{rej}}_{n_{\max}}(\Delta, n_{\max})$ *: the number of rejected peaks due to an excess of assignment options as a function of* Δ

This function represents the number of peaks rejected owing to an excess of assignment options as a function of Δ. In contrast to $N^{\mathrm{rej}}_{\mathrm{noassig}}(\Delta)$, which is independent of $n_{\max}$, $N^{\mathrm{rej}}_{n_{\max}}(\Delta, n_{\max})$ depends intimately on $n_{\max}$. The lower the values chosen for $n_{\max}$ and the larger Δ, the higher the number of peaks rejected by means of $n_{\max}$.

$N^{\mathrm{rej}}_{n_{\max}}(\Delta, n_{\max})$ allows one to quantify the effects of $n_{\max}$ on the rejection of peaks. Hence, the function can be used to detect inadequate choices of the parameter $n_{\max}$, when, for example, large proteins are investigated, avoiding unnecessary removal of cross-peaks.

(iii) $N^{\mathrm{acc}}(\Delta, n_{\max})$: *the number of accepted peaks as a function of* Δ *and* $n_{\max}$

For a peak-list ${}^{k}C$ of spectrum k the total number of accepted peaks in the first iteration (${}^{k}N^{\mathrm{acc}}(\Delta, n_{\max})$) is the total number of cross-peaks (${}^{k}N_{\mathrm{tot}}$), minus the number of peaks rejected because no assignment is possible (${}^{k}N^{\mathrm{rej}}_{\mathrm{noassig}}(\Delta)$) and because of exceeding $n_{\max}$ assignment options (${}^{k}N^{\mathrm{rej}}_{n_{\max}}(\Delta, n_{\max})$). After the exclusion of duplicate restraints ($N^{\mathrm{rej}}_{\mathrm{duplicate}}(\Delta)$), the total number of accepted peaks in the merged list ${}^{m}C$ in the first iteration is represented by

$$\begin{aligned} N^{acc}(\Delta, n_{\max}) &= \sum_k \left[{}^{k}N^{acc}(\Delta, n_{\max})\right] - N^{\mathrm{rej}}_{\mathrm{duplicate}}(\Delta) \\ &= \sum_k \Big[{}^{k}N_{\mathrm{tot}} - {}^{k}N^{\mathrm{rej}}_{\mathrm{noassig}}(\Delta) \\ &\quad - {}^{k}N^{\mathrm{rej}}_{n_{\max}}(\Delta, n_{\max})\Big] - N^{\mathrm{rej}}_{\mathrm{duplicate}}(\Delta) \\ &= \left(\sum_k {}^{k}N_{\mathrm{tot}} - N^{\mathrm{rej}}_{\mathrm{duplicate}}(\Delta)\right) \\ &\quad - \sum_k {}^{k}N^{\mathrm{rej}}_{\mathrm{noassig}}(\Delta) \\ &\quad - \sum_k {}^{k}N^{\mathrm{rej}}_{n_{\max}}(\Delta, n_{\max}) \\ &= {}^{m}N_{\mathrm{tot}} - {}^{m}N^{\mathrm{rej}}_{\mathrm{noassig}}(\Delta) \\ &\quad - {}^{m}N^{\mathrm{rej}}_{n_{\max}}(\Delta, n_{\max}). \end{aligned} \tag{4}$$

To describe the number of accepted peaks in all other iterations, an extra term has to be added to Equation 4, to account for the rejection of systematically inconsistent peaks by means of the noise-removal mechanisms of ARIA (Linge et al., 2001).

At small Δ values, the last term in Equation 4 (${}^{m}N^{\mathrm{rej}}_{n_{\max}}(\Delta, n_{\max})$) is 0. Hence $N^{\mathrm{acc}}(\Delta, n_{\max})$ increases with Δ since a decreasing number of peaks are left without an assignment.

When $n_{\max}$ is assigned a very large value, no peak is rejected by means of $n_{\max}$ (${}^{m}N^{rej}_{n_{\max}}(\Delta, n_{\max}) = 0$) even for very large Δ values. Since with sufficiently large Δ all peaks have at least one assignment option (${}^{m}N^{\mathrm{rej}}_{\mathrm{noassig}}(\Delta) \simeq 0$), the last two terms of Equation 4 vanish. Thus, for large values of $n_{\max}$, the number of accepted peaks increases with increasing Δ until a constant value equal to ${}^{m}N_{\mathrm{tot}}$ is obtained (Figure 2b, dashed line). At intermediate values of $n_{\max}$, peaks are rejected due to an exceeding of $n_{\max}$ assignment options at higher Δ: thus, the function in Equation 4 first increases and then decreases with increasing Δ values (Figure 2b, solid line). We define $\Delta_{\max}$ as the point at which $N^{\mathrm{acc}}(\Delta)$ reaches its maximum. Depending on the value of $n_{\max}$, it can happen that, within an interval of Δ a number of peaks are rejected due to an excess of assignment options while others are excluded because no assignment option can be found (the last two terms in Equation 4 are both different from 0). If this is the case, $\Delta_{\max}$ is obtained at Δ values where a fraction of peaks are left without assignment, thus at smaller values for Δ than the lower limit, determined as discussed in (i), using $N^{\mathrm{rej}}_{\mathrm{noassig}}(\Delta)$ as a criterion. Therefore, whenever $n_{\max}$ is excessively small, $\Delta_{\max}$ becomes a misleading parameter to direct the choice for Δ. Consequently, a good strategy to choose Δ should never rely exclusively on the total number of accepted peaks, but rather on an analysis of the different sources of peak rejection.

However, $\mathrm{N}^{\mathrm{acc}}(\Delta, n_{\max})$ is a useful diagnostic function which allows for an immediate estimation of the overall number of accepted peaks for different settings of Δ and $n_{\max}$ and thus helps to avoid erroneous choices for the two parameters leading to unnecessary removal of peaks.

(iv) $n_{\mathrm{av}}(\Delta, n_{\max})$: *the average number of assignments per peak* n_{av} *as a function of* Δ *and* $n_{\max}$

For each peak-list ${}^{k}C$ *of a spectrum* k with ${}^{k}N_{\mathrm{tot}}$ cross-peak entries, we define the *average number of assignment possibilities per peak in the spectrum* k *in the first iteration* (${}^{k}n_{\mathrm{av}}$) as

$$\begin{aligned} {}^{k}n_{\mathrm{av}} = \frac{1}{{}^{k}N^{\mathrm{acc}}(\Delta, n_{\max})} \sum_{j=1}^{{}^{k}N_{\mathrm{tot}}} \Theta[n_{\max} \\ - n({}^{k}C_j, \Delta, A) + 1] \cdot n({}^{k}C_j, \Delta, A), \end{aligned} \tag{5}$$

where $n({}^{k}C_j, \Delta, A)$ is the function introduced above which associates each entry ${}^{k}C_j$ of the peak-list the spectrum k to its number of assignment possibilities and $\Theta(x)$ is the Heaviside step function, which takes the value of 1 if the argument is larger than 0, otherwise 0. The factor $\Theta[n_{\max} - n({}^{k}C_{\mathrm{j}}, \Delta, A) + 1]$ in Equation 5 accounts

for the fact that peaks with more than n_{max} assignment possibilities are discarded. If more spectra are supplied, the *average number of assignment possibilities per peak in the merged list in the first iteration* (n_{av}) is

$$n_{av} = \frac{1}{N^{acc}(\Delta, n_{max})} \sum_{j=1}^{N^{acc}(\Delta, n_{max})} n({}^{m}C_j, \Delta, A) \qquad (6)$$

for a merged list ${}^{m}C$ containing $N^{acc}(\Delta, n_{max})$ entries ${}^{m}C_j$. For a resonance assignment-list A, ${}^{k}n_{av}$ and n_{av} depend in the first iteration on the values assigned to the two parameters Δ and n_{max} alone. Generally speaking, the average number of assignment possibilities increases with increasing Δ and cannot assume values larger than n_{max}. Due to larger overlap problems in 2D rather than in 3D spectra, it grows much faster with increasing Δ when using 2D data rather than 3D data for the same protein. In general, the effects increase with increasing protein size and are more severe for predominantly α-helical proteins, which notably display low chemical shift dispersion. Finally, Equations 5 and 6 slightly overestimate the real number of assignment possibilities in that, for degenerate protons belonging to the same heavy atom (e.g. methyl groups), each proton is counted as possible assignment.

Therefore, ${}^{k}n_{av}$ (Δ, n_{max}) and n_{av} (Δ, n_{max}) can be used to investigate the overlap problems affecting the spectra and the merged list and help to avoid incorrect choices of Δ and n_{max}, which would lead to an undesirably high average number of assignment possibilities per peak.

Materials and methods

Proteins and datasets used in these calculations

We used five different protein NMR datasets for the ARIA calculations: the C-terminal domain of Lac Repressor (Lac) (Bell and Lewis, 2001); the N-terminal domain of Arginine Repressor (ArgR) (Sunnerhagen et al., 1997; Ni et al., 1999); the C-terminal domain PB1 of yeast CDC24p (PB1) (D. Leitner et al., 2003, personal communication); the HRDC domain of RecQ (HRDC) (Liu et al., 1999) and the EVH1 domain of human VASP (EVH1) (Fedorov et al., 1999; Ball et al., 2002). We used only 2D NOESY data for ArgR and Lac and ^{13}C- and ^{15}N-edited 3D NOESY data for PB1, HRDC and EVH1. In all calculations, we supplied hydrogen bond and dihedral angle restraints. The main characteristics of these proteins and their spectra, relevant to this work, are summarised in Table 1.

Pre-calculation analysis by Cesta.py

The script Cesta.py created 165 different Δ sets, ranging from very small ($\delta^{het1} = 0.00062$, $\delta^{pro1} = 0.00005$, $\delta^{pro2} = 0.000025$) to very large values ($\delta^{het1} = 1.25$, $\delta^{pro1} = 0.1$, $\delta^{pro2} = 0.05$) and evaluated the four diagnostic functions for all five protein datasets. On a 1.8 GHz processor, the pre-calculation analysis required 8–12 h of CPU time, depending on the dataset. The calculation time can be reduced by lowering the number of values to be evaluated. For example, evaluating

Table 1. The five different datasets used for the calculations

Name	PDB entries	Chain length	Secondary structure	Type of data	Number of peaks
Lac	1JWL	56	3α	2D NOESY	2D (H_2O): 2122 2D (D_2O): 2106
PB1	1PQS	77	4β + 2α	3D NOESY	^{13}C-edited 3D: 1909 ^{15}N-edited 3D: 766
ArgR	1AOY	78	3α + 2β	2D NOESY	2D (H_2O): 1403 2D (D_2O): 1245
HRDC	1D8B	91	3α	3D NOESY	^{13}C-edited 3D: 2455 ^{15}N-edited 3D: 824
EVH1	1QC6	115	7β + 1α	3D NOESY	^{13}C-edited 3D: 3506 ^{15}N-edited 3D: 2772

The NOESY peak-lists of Lac, ArgR, HRDC and EVH1 were obtained with manual peak-picking, while those of PB1 were generated automatically with the internal peak-picking algorithm of Sparky v.3.1 (http://www.cgl.ucsf.edu/home/sparky/). Diagonal and obvious noise peaks were removed manually. No manual NOE assignments were included into the peak-list.

the diagnostic functions using only 30 Δ sets required just 90–150 min.

ARIA calculations

The structures were calculated by ARIA v.1.2 on a Dual Athlon M1800 + cluster at the Pasteur Institute. The number of calculated structures was 20 for iterations 0–7 and 100 for the final iteration 8. The calculations were evaluated by computing a pair-wise rmsd (precision) and an rmsd to a reference structure (accuracy) of the 20 lowest-energy structures. As a reference, we used the X-ray structure for Lac and EVH1 and the averaged NMR solution structure for ArgR, HRDC and PB1.

Results and discussion

Analysis of the NOE assignment in the first iteration: evaluation of $N^{\mathrm{rej}}_{\mathrm{noassig}}(\Delta), N^{\mathrm{rej}}_{n_{\max}}(\Delta, n_{\max})$, $N^{\mathrm{acc}}(\Delta, n_{\max})$ *and* n_{av} $(\Delta, n_{\max})$ *for five different protein datasets by means of Cesta.py*

The script Cesta.py was used to analyse the dependence of $N^{\mathrm{rej}}_{\mathrm{noassig}}(\Delta), N^{\mathrm{rej}}_{n_{\max}}(\Delta, n_{\max})$ $N^{\mathrm{acc}}(\Delta, n_{\max})$ and $n_{\mathrm{av}}(\Delta, n_{\max})$ on Δ for all five protein datasets. In this pre-calculation analysis 165 different sets of Δ (see above) were used to evaluate the four diagnostic functions. The same analysis was performed three times using three different values (200, 20 and 5) for $n_{\max}$. We plotted, $N^{\mathrm{rej}}_{\mathrm{noassig}}(\Delta), N^{\mathrm{rej}}_{n_{\max}}(\Delta, n_{\max})$ in Figure 3, $n_{\mathrm{av}}(\Delta, n_{\max})$ in Figure 4 and $N^{\mathrm{acc}}(\Delta, n_{\max})$ in the top sections of each plot in Figure 5 for all five proteins. sixteen out of the 165 Δ sets used for this pre-calculation analysis were chosen to perform structure calculations (their results will be discussed in the following section). The values of these 16 Δ sets are summarised in Table 2: the numbers in parentheses correlate each of these 16 Δ sets to one of the 165 Δ sets used for the pre-calculation analysis. Although all the 165 values of the diagnostic functions evaluated by Cesta.py were employed for the plots in Figures 3–5, we decided to use the 16 Δ sets of Table 2 for labelling the *x*-axis in these figures to facilitate the comparison of the output of the pre-calculation analysis by Cesta.py with the results of the structure calculations.

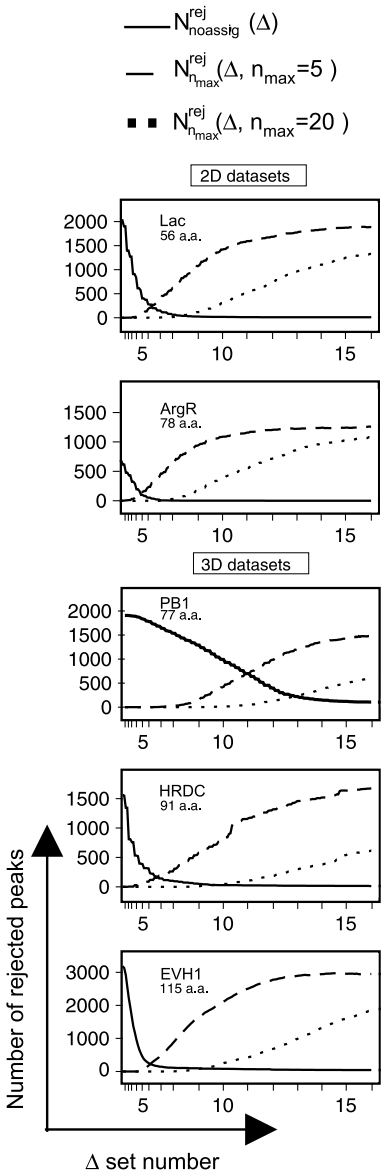

Figure 3. $N^{\mathrm{rej}}_{\mathrm{noassig}}(\Delta)$ and $N^{\mathrm{rej}}_{n_{\max}}(\Delta, n_{\max})$ are plotted as a function of Δ and $n_{\max}$. For each protein, we represent on the *y*-axis the number of peaks rejected during the initial NOE annotation because no assignment option was found ($N^{\mathrm{rej}}_{\mathrm{noassig}}(\Delta)$ solid line) and the number of peaks rejected because of exceeding of $n_{\max}$ ($N^{\mathrm{rej}}_{n_{\max}}(\Delta, n_{\max})$): the latter was evaluated for $n_{\max} = 5$ (dotted line) and $n_{\max} = 20$ (dashed line). The plotted data refer to the H_2O-2D spectra for Lac and ArgR and to the ^{13}C-edited NOESY for PB1, HRDC and EVH1 (Table 1). The labelling of the x-axis refers to the 16 Δ sets of Table 2.

$N^{\mathrm{rej}}_{\mathrm{noassig}}(\Delta)$ (solid line in Figure 3) is independent of $n_{\max}$ and provides insight into the self-consistency of the dataset. The significantly lower slope of the curve for the PB1 domain signals an

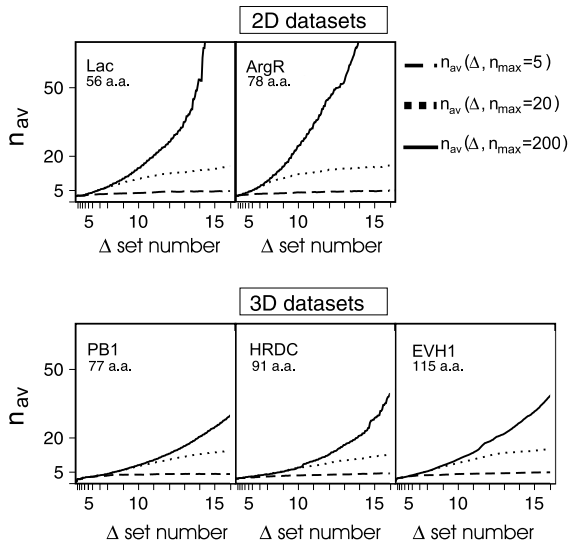

Figure 4. Δ and n_{max} determine the average number of assignment possibilities per peak. On the *y*-axis, n_{av} (Δ, n_{max}) is plotted for different combinations of Δ and n_{max} (dashed line: n_{av} (Δ, $n_{max} = 5$); dotted line: n_{av} (Δ, $n_{max} = 20$); solid line: n_{av} (Δ, $n_{max} = 200$)). The labelling of the *x*-axis refers to the 16 Δ sets of Table 2.

important anomaly in this dataset. Owing to sample decay, frequencies in the peak-lists and resonance assignment-list do not match properly; therefore, many peaks remain unassigned, even when relatively large Δ values are used. In this dataset, the uncertainty with respect to the chemical shift values is much greater than expected from the digital resolution.

Following our discussion above, the evaluation of $N^{rej}_{noassig}(\Delta)$ allowed us to find appropriate lower limits for Δ: the values of set 8 for Lac, set 7 for ArgR, set 10 for HRDC, set 7 for EVH1 and the larger values of set 13 for PB1 (see Table 2).

Figure 3 shows that different values for n_{max} have large effects on $N^{rej}_{n_{max}}(\Delta, n_{max})$. This diagnostic function allows an assessment of which value of n_{max} is more appropriate in avoiding unnecessary peak rejection with respect to the chosen Δ values. For all five proteins, $n_{max} = 5$ led to extensive peak rejection with Δ equal to or larger than the lower limits suggested above. In contrast, the number of rejected peaks was very small with $n_{max} = 20$. This shows that 20 is in general an appropriate value for moderately-sized proteins.

The plots in Figure 4 show that, for very large n_{max} ($n_{max} = 200$), n_{av} can assume very large values when increasing Δ. Generally speaking, the larger the resonance-overlap affecting the spectra, the more dramatic the growth of n_{av} with increasing Δ. The average number of assignments per peak grows much faster with increasing Δ when peak-lists from 2D spectra rather than 3D spectra are used for proteins of comparable size, as can be seen by comparing $n_{av}(\Delta)$ for ArgR and PB1. Furthermore, these effects usually increase with protein size and are more dramatic for dominantly α-helical folds. Hence, the curves representing $n_{av}(\Delta, n_{max} = 200)$ for HRDC (three α-helices) and EVH1 (seven β-strands and only one α-helix) are only marginally different, although EVH1 contains 24 more residues than HRDC.

Influence of Δ and n_{max} on the quality of the structures

An inappropriate choice of Δ and n_{max} can lead to imprecise or inaccurate structure calculations for three different reasons: (i) the number of accepted peaks is too small; hence, the set of restraints is insufficient to define the protein fold; (ii) the average number of assignment possibilities is too high and this hampers calculation convergence; (iii) the percentage of incorrectly annotated peaks is too high and the resulting high number of incorrect distance restraints leads to the calculation of inaccurate folds.

To analyse these situations, we performed one set of calculations for each protein dataset with a very low and a second with a very high value of n_{max} ($n_{max} = 5$ and $n_{max} = 200$, respectively). For comparison, a third set of calculations was performed with the default value of $n_{max} = 20$ as an example of a more realistic calculation scheme. Each set comprises 16 ARIA calculations performed using the 16 different sets of Δ values of Table 2. The results are summarised in twelve plots (Figures 5a–d: calculations with $n_{max} = 5$; e–h: calculations with $n_{max} = 20$; j–m: calculations with $n_{max} = 200$). The analysis of calculations for the PB1 domain is not included: owing to the problems discussed in the previous section, we did not obtain any *de novo* convergent structure with the standard ARIA protocol. This particularly difficult case will be discussed at the end of this section.

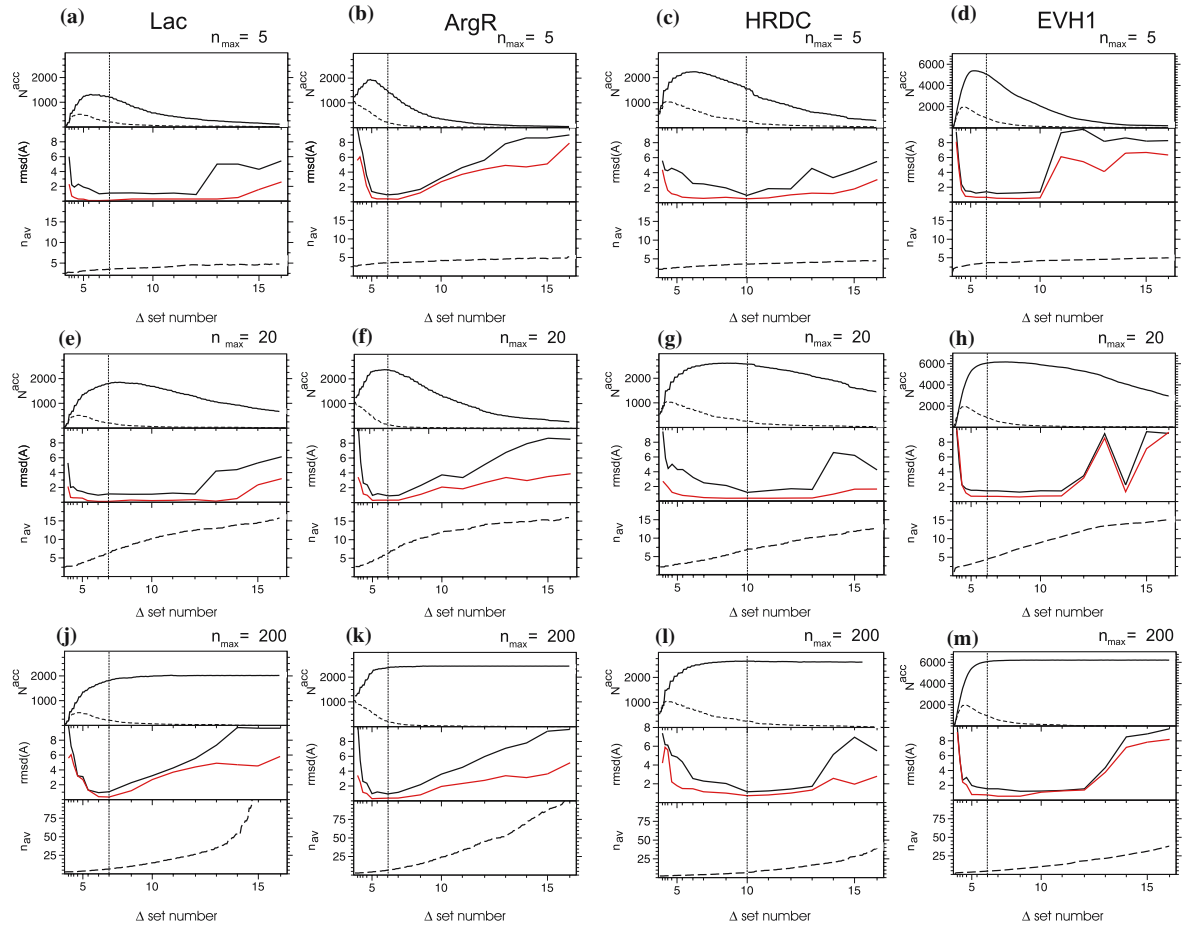

Figure 5. The influence of Δ and n_{max} on the quality of the final structures of Lac, ArgR, HRDC and EVH1. For each protein, three plots are presented, corresponding to different choices for n_{max} (5, 20 and 200): the 12 plots are labelled with small letters a–m. Each plot contains three sections, representing several parameters as a function of Δ: in the top sections, we represented the total number of accepted peaks in the merged list $N^{acc}(\Delta)$ (solid) and the number of unambiguously assigned peaks (dashed) is shown; the accuracy (black) and precision (red) of the calculated structures are shown in the middle sections $n_{av}(\Delta)$ in the bottom sections. A different scale for the *y*-axis is used in plots j–m to allow for the display of the larger values of $n_{av}(\Delta)$ when the cut-off n_{max} is effectively not used. The vertical dashed line in plots a–m indicates the lower limits for Δ as determined by means of $N^{rej}_{noassig}(\Delta)$. The labelling of the *x*-axis refers to the 16 Δ sets of Table 2.

Each plot in Figure 5 is composed of three sections. The top represents $N^{acc}(\Delta)$ vs. Δ (solid line). Additionally, we used a dashed line to indicate the number of accepted peaks with only one assignment possibility, to assess if a fraction of unambiguously assigned peaks in the early iterations is required to obtain correct structures. The middle section shows the accuracy (black) and the precision (red) of the calculated structures after nine ARIA iterations. The curves supply information about the effects of Δ on the calculated structures and allow for assessing the ranges of values yielding the best structures. In the bottom section, the growth of n_{av} with increasing Δ is displayed. In each plot the vertical line indicates the lower limit for Δ as determined by inspection of $N^{rej}_{noassig}(\Delta)$.

(a) Calculations with $n_{max} = 5$

As shown in Figure 3, when n_{max} is too small with respect to the size of the molecule, the number of peaks rejected for excess of assignment options is high even with relatively small Δ; as a result, Δ_{max} corresponds to relatively small Δ values. This can be seen in Figure 5, top sections: the curve representing $N^{acc}(\Delta)$ in plots a–d shows a much

Table 2. The 16 sets of chemical shift tolerances Δ used for the structure calculations

Δ sets		δ^{het1}	δ^{pro1}	δ^{pro2}
Structure calculations	Pre-calculation analysis			
1	(3)	0.0144	0.00115	0.00057
2	(5)	0.0281	0.00225	0.00112
3	(7)	0.0419	0.00335	0.00167
4	(10)	0.0625	0.005	0.0025
5	(14)	0.0930	0.0074	0.0037
6	(18)	0.124	0.01	0.005
7	(26)	0.185	0.015	0.0075
8	(34)	0.247	0.02	0.01
9	(51)	0.377	0.03	0.015
10	(67)	0.500	0.04	0.02
11	(84)	0.624	0.05	0.025
12	(100)	0.754	0.06	0.03
13	(116)	0.877	0.07	0.035
14	(132)	1.00	0.08	0.04
15	(148)	1.12	0.09	0.045
16	(165)	1.25	0.10	0.05

The 16 sets were chosen among the 165 sets used by Cesta.py for the pre-calculation analysis, as indicated by the set number in parentheses. Each set consists of 3 Δ values: the tolerance for the heteronuclear dimension (δ^{het1}), the indirect proton dimension (δ^{pro1}) and the detected proton dimension (δ^{pro2}). The values increase from set 1 to set 16. The increment is smaller for the first five sets to allow thorough sampling of small Δ values. The detected proton dimension of a NOESY spectrum is better resolved than the indirect one, hence smaller tolerance windows for this dimension are used.

narrower shape than in plots e–h and j–m. The curves in the middle sections of plots a–d indicate that the settings of Δ at which we obtained the best structures are centred on Δ values larger than Δ_{max}. This is an interesting result, since we obtained better structures with Δ settings which led to the acceptance of fewer peaks and to a larger average number of ambiguities than with $\Delta = \Delta_{max}$. This is particularly clear for HRDC (Figure 5, plot c). Δ_{max} is not far from Δ set 7 ($\delta^{het1} = 0.185$, $\delta^{pro1} = 0.15$, $\delta^{pro2} = 0.0075$), but only with Δ set 10 ($\delta^{het1} = 0.5$, $\delta^{pro1} = 0.04$, $\delta^{pro2} = 0.02$) we obtained accurate structures within 1 Å rmsd of the reference structure. With Δ set 7, the number of accepted peaks was 2231, of which 518 were unambiguously assigned, and $n_{av} = 2.92$. In contrast, with Δ set 10, the number of accepted peaks was only 1550 (of which just 189 were unambiguous) and $n_{av} = 3.6$. We deduce that in the first case the use of smaller Δ values has led to an incorrect annotation of a number of peaks (as for the peak in Figure 1b).

The structural information contained in NOESY spectra is redundant and this allows for correct structure calculations even when substantial fractions of NOESY cross-peaks are omitted from the peak-lists (Jee and Güntert, 2003). We obtained correct structures with an accuracy of 2 Å with even 65.3% of peaks rejected for EVH1, 67.3% for HRDC, 74.8% for ArgR and 84.6% for Lac, as can be seen in the plots a–d of Figure 5. It is important to note that these percentages do not refer to statistic omissions, but rather to a systematic removal of the most ambiguous peaks by means of n_{max}.

In summary, these results with $n_{max} = 5$ show that the presence of the correct assignment option among the assignment possibilities for annotated cross-peaks is a far more important prerequisite for a proper structure calculation than the completeness of the NOESY peak-list.

(b) Calculations with $n_{max} = 200$

For the proteins studied here, when n_{max} was set to 200, no peaks are rejected for exceeding n_{max} assignment options, allowing larger tolerance windows to be used without loss of restraints. This can be observed in plots j–m of Figure 5: the top curves are characterised by a plateau where $N^{acc}(\Delta)$ is constant and approximately equal to ${}^{m}N_{tot}$ (see Equation 4). The price paid for effectively including all peaks in the calculation is an increase in n_{av} with increasing Δ, as can be seen by comparing the bottom sections in plots a–h with those in plots j–m (taking into account the different scale on the *y*-axis required for the latter). In plots j–m, over a certain value for Δ, calculations lead to inaccurate structures, but for a completely different reason than in plots a–d: the program includes now all peaks in the calculation, but it is not able to handle the number of assignment options when it exceeds a critical value (middle sections, plots j–m). This dramatically affects the calculations with 2D datasets (Lac and ArgR), for which we obtained correct structures only within a narrower range of values for Δ (compare plots j and k with plots l and m). An inspection of the curves in the middle sections of plots j–m allows for estimating the highest tolerated values for n_{av} that still led to

good results for the four proteins. These critical values are approximately 10 for Lac and ArgR (2D NOESY spectra) and 17 for HRDC and EVH1 (3D NOESY spectra), showing that they can be significantly different if exclusively 2D or 3D spectra are used. However, all of them are surprisingly high, indicating that the program is very robust towards high levels of ambiguity in the constraints. This is supported by the observation that optimal performance was obtained with Δ values where most of the peaks were ambiguous (see the dotted line in plots j–m, top sections). This shows that a significant fraction of unambiguously assigned NOEs is not a prerequisite for good performance and that accurate structures can be obtained starting from purely ambiguous data (Nilges, 1995).

(c) Calculations with $n_{\max} = 20$

With the default value of $n_{\max} = 20$, the effects of peak loss and increase of ambiguity are less dramatic than in calculations with $n_{\max} = 5$ and $n_{\max} = 200$, respectively. This results in a more regular, flatter curve for $N^{\text{acc}}(\Delta)$ in the top sections of plots e–h, as compared to plots a–d and j–m in Figure 5. With the exception of HRDC, peaks are rejected due to an excess of assignment options at values for Δ where only a marginal fraction of peaks are left without a single possible assignment: in fact, for Lac, ArgR and Vasp, $\Delta_{\max}$ is obtained for larger values than the lower limits for Δ determined by means of $N^{\text{rej}}_{\text{noassig}}(\Delta)$, as can be seen by comparing the relative position of the maximum of the function $N^{\text{acc}}(\Delta)$ and the vertical dotted line in plots e–h. In contrast, for HRDC we obtained $\Delta_{\max}$ for Δ values smaller than the lower limit. If now we observe the quality of the calculated structures, we see that for Lac, ArgR and Vasp, calculations with $\Delta = \Delta_{\max}$ led in fact to good-quality structures, whereby for the HRDC domain Δ values larger than $\Delta_{\max}$ were necessary to obtain correct structures.

This result tells us that, provided that Δ is not smaller than the lower limit assessed with $N^{\text{rej}}_{\text{noassig}}(\Delta)$, the best structures are obtained by choosing the parameters such that $N^{\text{acc}}(\Delta)$, the number of accepted peaks, is maximised and n_{av}, the average number of ambiguities, is minimised. The case of the HRDC domain, analogously to calculations with $n_{\max} = 5$, shows that whenever $\Delta_{\max}$ is lower than the lower limit, the total number of accepted peaks represents a misleading parameter to choose Δ.

Calculations of the PB1 domain

Despite serious attempts, no calculation for the PB1 domain led to satisfactory results, due to the poor agreement between resonance assignment-list and peak-lists, as discussed above. Compensating for such frequency discrepancies can be achieved only by applying very large Δ values. The analysis by Cesta.py led to the conclusion that the high values of Δ set 13 ($\delta^{\text{het1}} = 0.88$, $\delta^{\text{pro1}} = 0.07$, $\delta^{\text{pro2}} = 0.035$) should be chosen as a lower limit. However, the price to pay for this choice was a high value of n_{av} (> 16.4), preventing convergence. We tried to rescue the calculation by slowing down the cooling phase of the simulated annealing protocol, as recently suggested (Lougheed et al., 2002): this enabled the program to handle this high number of ambiguities and led to accurate structures within an rmsd of 1.5 Å of the reference (Fossi et al., 2004; unpublished). Interestingly, the same modified protocol used in conjunction with Δ values smaller than those of set 13 in Table 2 did not lead to satisfactory results, showing that the diagnostic function $N^{\text{rej}}_{\text{noassig}}(\Delta)$ did indicate a suitable lower limit for Δ.

A strategy for choosing most suitable values for Δ *and* n_{max}

The observations made in the analysis above can be summarised as follows: (i) choosing excessively small values for Δ may exclude the correct assignment from the assignment possibilities for an accepted peak; (ii) ARIA is robust towards high numbers of assignment possibilities per peak; (iii) the automatic removal of a large number of ambiguous peaks by ARIA due to exceeding $n_{\max}$ has little influence on the quality of the structures.

Keeping this in mind, the analysis in terms of the diagnostic functions $N^{\text{rej}}_{\text{noassig}}(\Delta), N^{\text{rej}}_{n_{\max}}(\Delta, n_{\max})$ (Figure 3), $N^{\text{acc}}(\Delta)$ and n_{av} (Figures 3 and 5) suggests a strategy for determining optimal values for Δ and $n_{\max}$. The point where $N^{\text{rej}}_{\text{noassig}}(\Delta)$ becomes minimal, i.e. when it is close to 0, such that most of the peaks contain at least one possible assignment, provides a starting point to set Δ. We recommend using values for Δ which are

slightly (approx. 30%) larger than the lower limit, to ensure the presence of the correct assignment option among the assignment possibilities for annotated peaks. However, values much larger than the lower limit should be avoided, as they lead to an unnecessary increase of n_{av}. The extreme case of PB1 shows that with the help of $N^{rej}_{noassig}(\Delta)$ we can detect such particularly inconsistent datasets which require larger values for Δ.

n_{max} should be adjusted after the choice of Δ. As shown above, the default value of 20 should usually work for moderately-sized proteins. Erroneous choices for this parameter can be detected by inspecting $N^{rej}_{n_{max}}(\Delta)$. In these cases, n_{max} should be adjusted such that few peaks are rejected for excess of assignment options in correspondence to the chosen Δ. By imposing that Δ_{max} assumes similar values to the chosen Δ, we obtain a criterion to optimise n_{max}.

If the chosen Δ and n_{max} lead to an excessively large average number of ambiguities per peak, n_{av} should be reduced by using a smaller n_{max} rather than a smaller Δ. Our results have shown that it is preferable to lose some highly ambiguous cross-peaks (which result in loose structural restraints) rather than to include a large number of incorrectly annotated peaks in the calculation. With a standard ARIA protocol, we recommend avoiding $n_{av} > 8$ for 2D spectra and $n_{av} > 15$ for 3D spectra; these values correspond to the largest tolerated n_{av} values (see calculations with $n_{max} = 200$), reduced by two units for precaution. As an alternative, large values for n_{av} may be handled by conveniently slowing the cooling phase of the simulated annealing protocol in CNS, as shown by Lougheed et al. with the melanoma inhibitory protein.

Conclusions

We performed ARIA structure calculations for five different proteins, applying systematically different combinations of Δ and n_{max}. The results showed how these parameters influence the performance of the program and the quality of the obtained structures. We achieved a quantitative assessment of the software's robustness in terms of assignment ambiguity and peak-list incompleteness: calculations tolerate high levels of peak losses and assignment ambiguity, and thus larger values for Δ; conversely, choosing excessively small values for Δ may lead to misassignments caused by the exclusion of the correct assignment from the assignment possibilities for an accepted peak. Furthermore, we have shown that a fraction of unambiguously assigned peaks in the early iterations is not a prerequisite for correct performance and that convergence can be achieved even without unambiguous peaks. Hence, it is important to avoid the use of excessively small Δ values. On the other hand, the use of overly large Δ values may lead to structure calculation failures resulting either from the rejection of too many peaks for having too many assignment possibilities, or from an excessive average number of assignment options per peak.

We have shown that this can be avoided by performing an analysis of the influence of Δ and n_{max} on the initial NOE assignment prior to structure calculation. Based on the output of this pre-calculation analysis by the Cesta.py script, we developed a strategy for choosing optimal values for Δ and n_{max} which takes into account the peculiarity of each dataset. In particular, this analysis allows the recognition of datasets with poor agreement between the chemical shifts in the assignment-list and NOE cross-peak coordinates. Furthermore, the proposed method is computationally efficient, as it does not involve time-consuming structure calculations.

Acknowledgements

Support from Protein Struktur Fabrik (grant FKZ 01GG9812/4) is gratefully acknowledged. We thank Dr Linda Ball for helpful discussion and for kindly providing the dataset of the EVH1 domain of VASP. We thank Dr Katja Heuer, Layton J. Culter and Richard Walker for carefully reading the manuscript.

References

Bell, C.E. and Lewis, M. (2001) *J. Mol. Biol.*, **312**, 921–926.

Ball, L.J., Jarchau, T., Oschkinat, H. and Walter, U. (2002) *FEBS Lett.*, **513**, 45–52.

Brunger, A.T., Adams, P.D., Clore, G.M., Delano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T. and Warren, G.L. (1998) *Acta Crystallogr. D – Biol. Crystallogr.*, **54**, 905–921.

Fedorov, A.A., Fedorov, E., Gertler, F. and Almo, S.C. (1999) *Nat. Struct. Biol.*, **6**, 661–665.
Gronwald, W., Moussa, S., Elsner, R., Jung, A., Ganslmeier, B., Trenner, J., Kremer, W., Neidig, K.P. and Kalbitzer, H.R. (2002) *J. Biomol. NMR*, **23**, 271–287.
Güntert, P. (2003) *Prog. Nucl. Magn. Reson. Spectrosc.*, **43**, 105–125.
Güntert, P., Mumenthaler, C. and Wüthrich, K. (1997) *J. Mol. Biol.*, **273**, 283–298.
Herrmann, T., Güntert, P. and Wüthrich, K. (2002) *J. Mol. Biol.*, **319**, 209–227.
Jee, J. and Güntert, P. (2003) *J. Struct. Funct. Genom.*, **4**, 179–189.
Kharrat, A., Macias, M.J., Gibson, T.J., Nilges, M. and Pastore, A. (1995) *EMBO J.*, **14**, 3572–3584.
Linge, J.P., Habeck, M., Rieping, W. and Nilges, M. (2003) *Bioinformatics*, **19**, 315–316.
Linge, J.P., O'Donoghue, S.I. and Nilges, M. (2001) *Nucl. Magn. Reson. Biol. Macromol., Pt B*, **339**, 71–90.
Liu, Z., Macias, M.J., Bottomley, M.J., Stier, G., Linge, J.P., Nilges, M., Bork, P. and Sattler, M. (1999) *Struct. Fold. Des.*, **7**, 1557–1566.
Lougheed, J.C., Domaille, P.J. and Handel, T.M. (2002) *J. Biomol. NMR*, **22**, 211–223.
Moseley, H.N. and Montelione, G.T. (1999) *Curr. Opin. Struct. Biol.*, **9**, 635–642.
Moseley, H.N., Monleon, D. and Montelione, G.T. (2001) *Meth. Enzymol.*, **339**, 91–108.
Mumenthaler, C. and Braun, W. (1995) *J. Mol. Biol.*, **254**, 465–480.
Mumenthaler, C., Güntert, P., Braun, W. and Wüthrich, K. (1997) *J. Biomol. NMR*, **10**, 351–362.
Ni, J.P., Sakanyan, V., Charlier, D., Glansdorff, N. and Van Duyne, G.D. (1999) *Nat. Struct. Biol.*, **6**, 427–432.
Nilges, M. (1993) *Proteins*, **17**, 297–309.
Nilges, M. (1995) *J. Mol. Biol.*, **245**, 645–660.
Nilges, M. and O'Donoghue, S.I. (1998) *Prog. Nucl. Magn. Reson. Spectros.*, **32**, 107–139.
Nilges, M., Macias, M.J., O'Donoghue, S.I. and Oschkinat, H. (1997) *J. Mol. Biol.*, **269**, 408–422.
Sunnerhagen, M., Nilges, M., Otting, G. and Carey, J. (1997) *Nat. Struct. Biol.*, **4**, 819–826.
Tjandra, N., Garrett, D.S., Gronenborn, A.M., Bax, A. and Clore, G.M. (1997) *Nat. Struct. Biol.*, **4**, 443–449.