Chapter 3 Influence of chemical shift tolerances on NMR structure calculations using ARIA protocols for assigning NOE data.

3.1 Introduction

In ARIA and analogous programmes, 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 couplings, 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[37]. 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 affecting chemical shifts.

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.

In this part of the work, a large number of ARIA structure calculations for five different proteins, using several different combinations of Δ and n_{max}, were performed in order to understand the influence of these two parameters on the quality of the calculated structures.

Moreover, a strategy is presented for choosing optimal values for Δ and n_{max}, prior to structure calculations, *via* an analysis of the peak annotation (§ 1.4.2.1) in the first iteration of ARIA. This analysis provides diagnostic information regarding the consistency of the list of resonance assignments and the peak lists and about the degree of spectral overlap affecting the spectra. A Python script, Cesta.py (available from http://pasteur.fr/binfs) was developed 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.

3.2 Cesta.py: a pre-calculation analysis of the influence of Δ and n_{max} on peak annotation

3.2.1 A description of 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 is a tool for a pre-calculation analysis of the automated NOE assignment, which provides information about Δ and n_{max} during the peak annotation (§ 1.4.2.1). 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 to merge the various peak lists into a unique 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 peak lists and the merged list and evaluates for each of them the four diagnostic functions described hereafter.

3.2.2 The output of Cesta.py: four diagnostic functions

The output of Cesta.py consists of the evaluation of four functions of Δ and n_{max}. These functions are described in the following sections, where formulas are derived to describe their theoretical dependence on the two parameters of interest.

3.2.2.1 $N^{rej}_{noassig}(\Delta)$: the number of rejected peaks due to a lack of assignment options as a function of Δ

This function depends on the quality of the alignment of chemical shifts between the list of resonance assignments 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) list of resonance assignments match only within a certain error limit the chemical shift co-ordinates of NOESY cross-peaks. The poorer the consistency between the list of resonance assignments

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 3.1a). Thus, $N^{rej}_{noassig}(\Delta)$ 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 1.15a), and, even worse, to the acceptance of many incorrectly annotated peaks (Figure 1.15b). Thus, the point where $N^{rej}_{noassig}(\Delta)$ becomes minimal provides a criterion to set a lower limit for Δ.

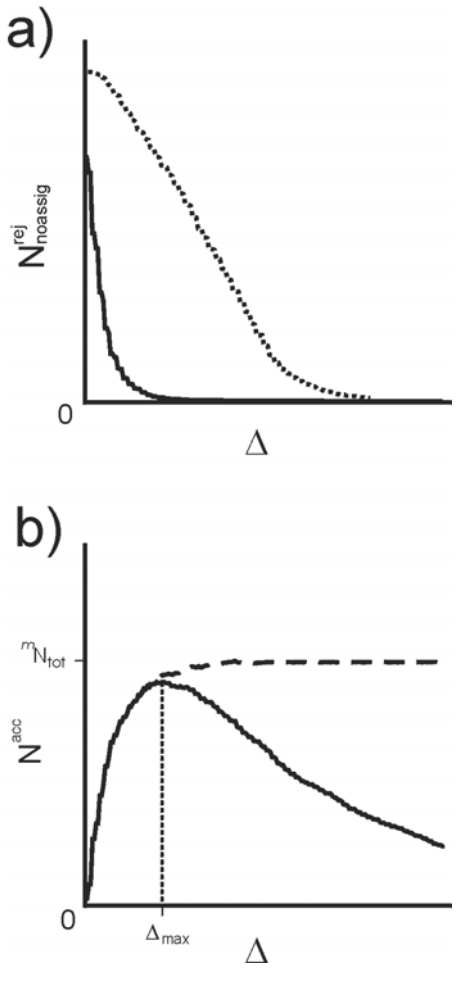

Figure 3.1 a) The number of rejected peaks due to a lack of assignment options as a function of Δ ($N^{rej}_{noassig}(\Delta)$) for two hypothetical datasets, one featuring a good (solid) and the other a bad (dotted) frequency alignment between the list of resonance assignments and the peak lists. b) Typical behaviours of the number of accepted peaks $N^{acc}(\Delta)$ 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 $^{m}N_{tot}$, the maximum number of entries that can be present in the merged list when no peak is rejected due to a lack or to an excess of assignment options..

3.2.2.2 $N^{rej}_{n_{max}}(\Delta, n_{max})$: 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 $N^{rej}_{noassig}(\Delta)$, which is independent of n_{max}, $N^{rej}_{n_{max}}(\Delta, n_{max})$ 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}. $N^{rej}_{n_{max}}(\Delta, n_{max})$ 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.

3.2.2.3 $N^{acc}(\Delta, n_{max})$: the number of accepted peaks as a function of Δ and n_{max}

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

$$
\begin{aligned}
N^{acc}(\Delta, n_{max}) &= \sum_{k} \left[{}^{k}N^{acc}(\Delta, n_{max}) \right] - N^{rej}_{duplicate}(\Delta) \\
&= \sum_{k} \left[{}^{k}N_{tot} - {}^{k}N^{rej}_{noassig}(\Delta) - {}^{k}N^{rej}_{n_{max}}(\Delta, n_{max}) \right] - N^{rej}_{duplicate}(\Delta) \\
&= \left(\sum_{k} {}^{k}N_{tot} - N^{rej}_{duplicate}(\Delta) \right) - \sum_{k} {}^{k}N^{rej}_{noassig}(\Delta) - \sum_{k} {}^{k}N^{rej}_{n_{max}}(\Delta, n_{max}) \\
&= {}^{m}N_{tot} - {}^{m}N^{rej}_{noassig}(\Delta) - {}^{m}N^{rej}_{n_{max}}(\Delta, n_{max})
\end{aligned}
\tag{3.1}
$$

where ${}^{m}N_{tot}$ represents the maximum number of entries that can be present in the merged list when no peak is removed due to a lack or to an excess of assignment options.

To describe the number of accepted peaks in all other iterations, an extra term has to be added to Equation 3.1, to account for the rejection of systematically inconsistent peaks by means of the noise-removal mechanisms of ARIA[33] (§ 1.4.5). At small Δ values, the last term in Equation 3.1 (${}^{m}N^{rej}_{n_{max}}(\Delta, n_{max})$) is 0. Hence $N^{acc}(\Delta, n_{max})$ increases with Δ as a consequence of the fact that 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 (${}^{m}N^{rej}_{noassig}(\Delta) \cong 0$), the last two terms of Equation 3.1 vanish. Thus, for large values of n_{max}, the number of accepted peaks increases with increasing Δ until a constant value equal to ${}^{m}N_{tot}$ is obtained (Figure 3.1b, 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 3.1 first increases and then decreases with increasing Δ values (Figure 3.1b, solid line). Δ_{max} is defined as the point at which $N^{acc}(\Delta)$ 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 3.1 are both different of 0). If this is the case, Δ_{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 § 3.2.2.1, using $N^{rej}_{noassig}(\Delta)$ as a criterion. Therefore, whenever n_{max} is excessively small, Δ_{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, $N^{acc}(\Delta, n_{max})$ 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.

3.2.2.4 $n_{av}(\Delta, n_{max})$: the average number of assignments per peak n_{av} as a function of Δ and n_{max}

For each spectrum *k* with peak list ${}^{k}C$ containing ${}^{k}N_{tot}$ cross-peak entries, the *average number of assignment possibilities per peak in the spectrum k in the first iteration* (${}^{k}n_{av}$) is defined as:

$$ {}^{k}n_{av} = \frac{1}{{}^{k}N^{acc}(\Delta, n_{max})} \sum_{j=1}^{{}^{k}N_{tot}} \Theta[n_{max} - n({}^{k}C_j, \Delta, A) + 1] \cdot n({}^{k}C_j, \Delta, A), \qquad (3.2) $$

where $n({}^{k}C_j, \Delta, A)$ is the function introduced above which associates each entry ${}^{k}C_j$ of the peak list of 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 $\Theta[n_{max} - n({}^{k}C_j, \Delta, A) + 1]$ in Equation 3.2 accounts for the fact that entries 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:

$$ n_{av} = \frac{1}{N^{acc}(\Delta, n_{max})} \sum_{j=1}^{N^{acc}(\Delta, n_{max})} n({}^{m}C_j, \Delta, A) \qquad (3.3) $$

for a merged list ^{m}C containing $N^{acc}(\Delta, n_{max})$ entries $^{m}C_j$. For a list of resonance assignments A, $^{k}n_{av}$ and n_{av} in the first iteration depend exclusively on the two parameters Δ and n_{max}. 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 3.2 and 3.3 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, $^{k}n_{av}(\Delta, n_{max})$ and $n_{av}(\Delta, n_{max})$ 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.

3.3 Results and discussion

3.3.1 Analysis of the NOE assignment in the first iteration: evaluation of $N^{rej}_{noassig}(\Delta)$, $N^{rej}_{n_{max}}(\Delta, n_{max})$, $N^{acc}(\Delta, n_{max})$ 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 $N^{rej}_{noassig}(\Delta)$, $N^{rej}_{n_{max}}(\Delta, n_{max})$, $N^{acc}(\Delta, n_{max})$ and $n_{av}(\Delta, n_{max})$ on Δ for all five protein datasets. In this pre-calculation analysis 165 different sets of Δ (§ 2.1.2) were used to evaluate the four diagnostic functions. The same analysis was performed three times using three increasingly restrictive values (200,

20 and 5) for n_{max}. $N^{rej}_{noassig}(\Delta)$, $N^{rej}_{n_{max}}(\Delta, n_{max})$ were plotted in Figure 3.2, $n_{av}(\Delta, n_{max})$ in Figure 3.3 and $N^{acc}(\Delta, n_{max})$ in the top sections of each plot in Figure 3.4 for all five proteins. 16 out of the 165 Δ-sets used for this pre-calculation analysis were chosen to perform structure calculations (their results will be discussed in § 3.3.2). The values of these 16 Δ-sets are summarised in Table 3.1: the numbers in parentheses correlate each of these 16 Δ-sets to one of the 165 Δ-sets used for the pre-calculation analysis.

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

Table 3.1 The 16 sets of chemical shift tolerances used for the structure calculations. The 16 sets were chosen among the 165 sets used by Cesta.py for the pre-calculations 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 5 sets to allow for 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 are used for this dimension.

Although all the 165 values of the diagnostic functions evaluated by Cesta.py were employed for the plots in Figures 3.4, the 16 Δ-sets of Table 3.1 were used 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.

$N^{rej}_{noassig}(\Delta)$ (solid line in Figure 3.2) 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 important anomaly in this dataset. Owing to sample decay, frequencies in the peak lists and the list of resonance assignments 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 the discussion above, the evaluation of $N^{rej}_{noassig}(\Delta)$ provides a way to set appropriate lower limits for Δ: the values of Δ-set 8 for Lac, Δ-set 7 for ArgR, Δ-set 10 for HRDC, Δ-set 7 for EVH1 and the larger values of Δ-set 13 for PB1 (Table 3.1). Figure 3.2 shows that different values for n_{max} have large effects on $N^{rej}_{n_{max}}(\Delta, n_{max})$. This diagnostic function allows for 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 3.3 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.

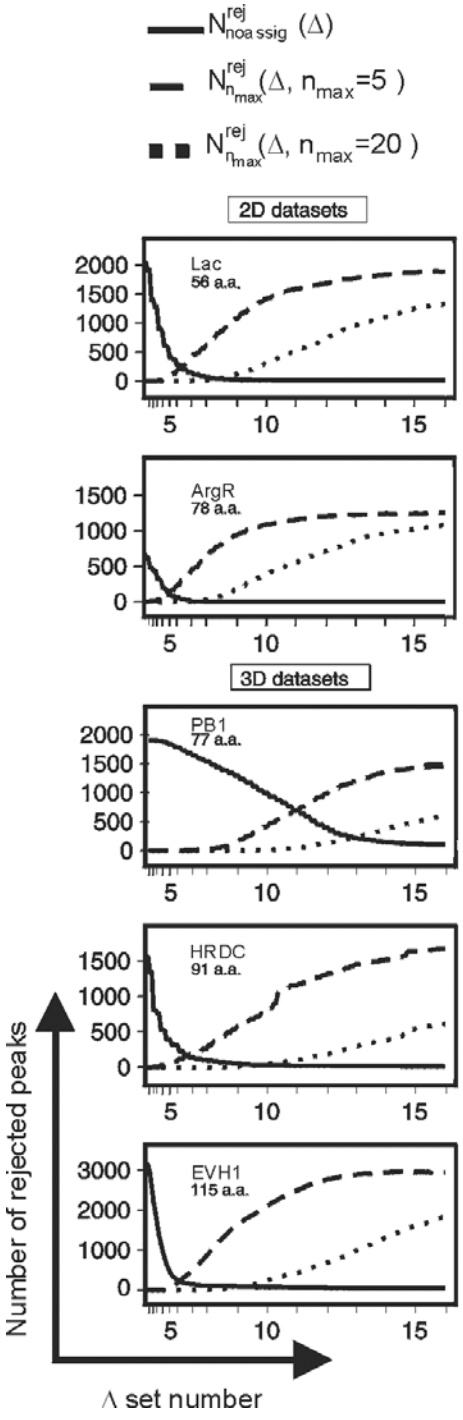

Figure 3.2 For each protein, the number of peaks rejected during the initial NOE annotation owing to a lack of assignment options ($N^{rej}_{noassig}(\Delta)$, solid line) and the number of peaks rejected because of exceeding of n_{max} ($N^{rej}_{n_{max}}(\Delta, n_{max})$) are represented on the y-axis: the latter function 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 2.1). The labelling of the x-axis refers to the 16 Δ-sets of Table 3.1.

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.

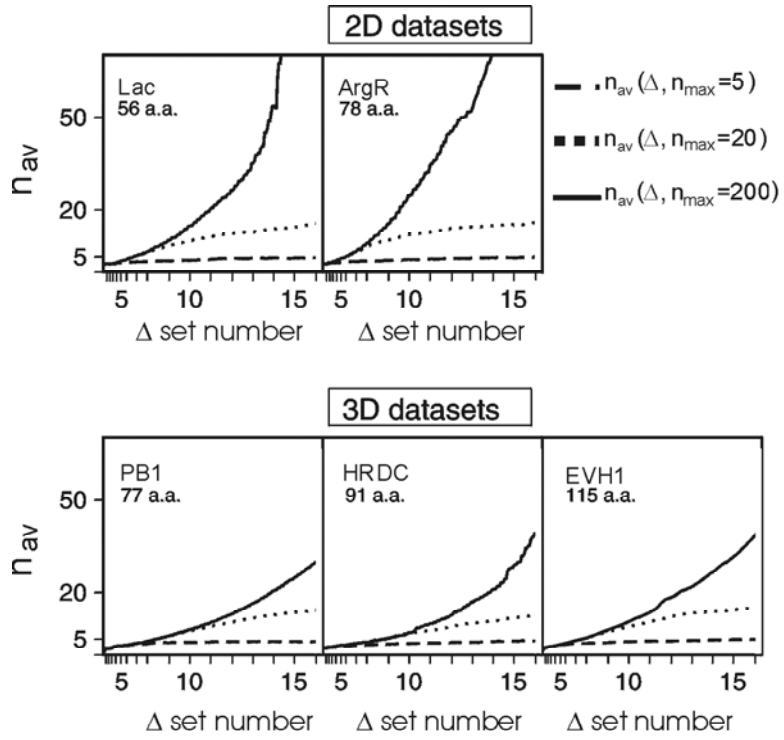

Figure 3.3 Δ and n_{max} determine the average number of assignment possibilities per peak. On the y-axis, $n_{av}(\Delta, n_{max})$ is plotted for different combinations of Δ and n_{max} (dashed line: $n_{av}(\Delta, n_{max}=5)$; dotted line: $n_{av}(\Delta, n_{max}=20)$; solid line: $n_{av}(\Delta, n_{max}=200)$). The labelling of the *x*-axis refers to the 16 Δ-sets of Table 3.1.

3.3.2 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, 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) were performed. 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 3.1. The results are summarised in twelve plots (Figure 3.4, A-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, it was not possible to obtain any *de novo* convergent structure with the standard ARIA protocol. This particularly difficult case will be discussed in § 3.3.3.

Each plot in Figure 3.4 is composed of three sections. The top section represents $N^{acc}(\Delta)$ *versus* Δ (solid line). Additionally, a dashed line is included to indicate the number of accepted peaks with only one assignment possibility, in order to assess whether a fraction of unambiguously assigned peaks in the early iterations is a prerequisite 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 quality of 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 black dashed vertical line indicates the lower limit for Δ as determined by inspection of $N^{rej}_{noassig}(\Delta)$. Δ_{max} is indicated by a red dashed vertical line.

3.3.2.1 Calculations with n_{max}=5

As shown in Figure 3.2, 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

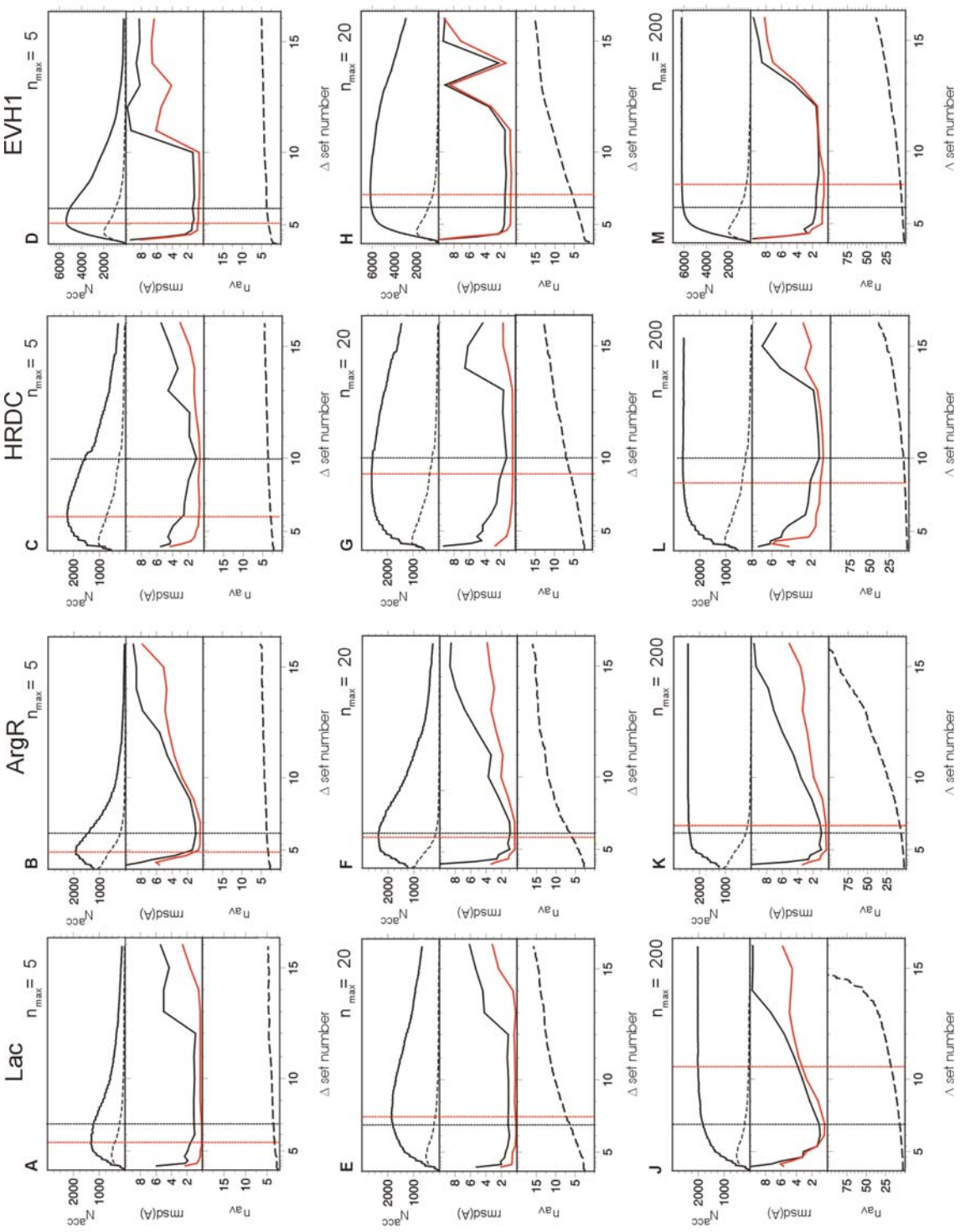

Figure 3.4 Influence of Δ and n_{max} on the quality of the final structures of Lac, ArgR, HRDC and EVH1. For each protein, three plots are presented, corresponding to different choices for n_{max} (5, 20 and 200): the 12 plots are labelled with capital letters A-M. Each plot contains three sections, representing several parameters as a function of Δ: in the top sections, the total number of accepted peaks in the merged list $N^{acc}(\Delta)$ (solid) and the

number of unambiguously assigned peaks (dashed) are shown; the accuracy (black) and precision (red) of the calculated structures are shown in the middle sections; $n_{av}(\Delta)$ is shown 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 lower limits for Δ as determined by means of $N^{rej}_{noassig}(\Delta)$ and are indicated in plots A-M with a black dashed vertical line. The red dashed vertical line indicates Δ_{max}. The labelling of the x-axis refers to the 16 Δ-sets of Table 3.1.

Δ; as a result, Δ_{max} corresponds to relatively small Δ values. This can be seen in Figure 3.4, top sections: the curve representing $N^{acc}(\Delta)$ in plots A-D shows a much 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 the best structures were obtained are centred on Δ values larger than Δ_{max}.

This is an interesting result, since better structures were obtained 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 3.4, plot C). Δ_{max} is not far from Δ-set 7 (δ^{het1}=0.185, δ^{pro1}=0.15, δ^{pro2}=0.0075), but only with Δ-set 10 (δ^{het1}=0.5, δ^{pro1}=0.04, δ^{pro2}=0.02) accurate structures within 1 Å rmsd of the reference structure were obtained. 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. Clearly, 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 1.15b).

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[57]. Correct structures were obtained 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 3.4. 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.

3.3.2.2 Calculations with n_{max}= 200

For the proteins studied here, when n_{max} was set to 200, no peaks were 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 3.4: the top curves are characterised by a plateau where $N^{acc}(\Delta)$ is constant and approximately equal to ${}^{m}N_{tot}$ (Equation 3.1). The price paid for effectively including all peaks in the calculation is a substantial 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 programme 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 correct structures were obtained only when using 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 programme is quite 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[46].

3.3.2.3 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 $N^{acc}(\Delta)$ in the top sections of plots E-H, as compared to plots A-D and J-M in Figure 3.4. 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 EVH1, Δ_{max} is obtained for larger values than the lower limits for Δ determined by means of $N^{rej}_{noassig}(\Delta)$, as can be seen by comparing the relative position of the red and the black vertical lines in plots E-H. In contrast, for HRDC Δ_{max} was obtained 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 EVH1, calculations with $\Delta=\Delta_{max}$ led in fact to good-quality structures, whereby for the HRDC domain Δ values larger than Δ_{max} were necessary to obtain correct structures.

This result tells us that, provided that Δ is not smaller than the lower limit assessed with $N^{rej}_{noassig}(\Delta)$, the best structures are obtained by choosing the parameters such that $N^{acc}(\Delta)$, 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 Δ_{max} is lower than the lower limit, the total number of accepted peaks represents a misleading parameter to choose Δ.

3.3.3 Calculations of the PB1 domain

Despite serious attempts, no calculation for the PB1 domain led to satisfactory results, due to the poor agreement between the list of resonance assignments and the peak lists, as discussed in § 3.3.1. 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 (δ^{het1}=0.88, δ^{pro1}=0.07, δ^{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. In § 4.2.2, the calculation is rescued by slowing down the cooling phase of the simulated annealing protocol, as recently suggested[58]: this enabled the programme to handle this high number of ambiguities and led to accurate structures within an rmsd of 1.5 Å of the reference. Interestingly, the same modified protocol used in conjunction with Δ values smaller than those of Δ-set 13 in Table 3.1 did not lead to satisfactory results, showing that the diagnostic function $N^{rej}_{noassig}(\Delta)$ did indicate a suitable lower limit for Δ.

3.3.4 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 $N^{rej}_{noassig}(\Delta)$, $N^{rej}_{n_{max}}(\Delta, n_{max})$ (Figure 3.2), n_{av} (Figure 3.3) and $N^{acc}(\Delta)$ (top sections of the plots in Figure 3.4) suggests a strategy for determining optimal values for Δ and n_{max}. The point where $N^{rej}_{noassig}(\Delta)$ 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. Values for Δ which are slightly (approximately 30%) larger than the lower limit are recommended, 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 $N^{rej}_{noassig}(\Delta)$ such particularly inconsistent datasets which require larger values for Δ can be detected.

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. Alternatively, it should be chosen such that few peaks are rejected for an excess of assignment options in correspondence to the chosen values for Δ. By imposing that Δ_{max} assumes similar values to the chosen Δ we obtain a criterion to optimise n_{max}. Furthermore, the observation of $N^{rej}_{n_{max}}(\Delta)$ provides a direct way to measure the effects of this parameter on the calculation.

The results in § 3.3.2.1 have shown that it is preferable to remove many highly ambiguous cross-peaks (which result in loose structural restraints) by means of the cut-off n_{max} rather than to include in the calculation even a small fraction of incorrectly annotated peaks. Hence, if the chosen values for Δ and n_{max} lead to an excessively large average number of ambiguities per peak n_{av}, the latter should be reduced by using a smaller n_{max} rather than a smaller Δ. With a standard ARIA protocol, it is recommended 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, much larger values for

n_{av} may be handled by conveniently slowing the cooling phase of the simulated annealing protocol in CNS (see Chapter 4).

3.4 Conclusions

In this work, ARIA structure calculations for five different proteins, applying systematically different combinations of Δ and n_{max}, were performed. The results showed how these parameters influence the performance of the programme and the quality of the obtained structures. A quantitative assessment of the software's robustness in terms of assignment ambiguity and peak list incompleteness was achieved: 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 erroneous assignments caused by the exclusion of the correct assignment from the assignment possibilities for an accepted peak. Furthermore, the results show 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.

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, a strategy was developed 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 list of resonance assignments and NOE cross-peak co-ordinates. The proposed

method is computationally efficient, as it does not involve time-consuming structure calculations.