Journal of Biomolecular NMR, **26:** 355–366, 2003.
KLUWER/ESCOM

A use of Ramachandran potentials in protein solution structure determinations

Ivano Bertini[a,*], Gabriele Cavallaro[a], Claudio Luchinat[b] & Irene Poli[a]
[a]*Magnetic Resonance Center and Department of Chemistry and* [b]*Magnetic Resonance Center and Department of Agricultural Biotechnology, University of Florence, Via Luigi Sacconi 6, 50019, Sesto Fiorentino, Italy*

Received 10 February 2003; Accepted 14 April 2003

Key words: backbone conformation, protein structure, Ramachandran plot, structure calculation, structure databases

Abstract

A strategy is developed to use database-derived ϕ-ψ constraints during simulated annealing procedures for protein solution structure determination in order to improve the Ramachandran plot statistics, while maintaining the agreement with the experimental constraints as the sole criterion for the selection of the family. The procedure, fully automated, consists of two consecutive simulated annealing runs. In the first run, the database-derived ϕ-ψ constraints are enforced for all aminoacids (but prolines and glycines). A family of structures is then selected on the ground of the lowest violations of the experimental constraints only, and the ϕ-ψ values for each residue are examined. In the second and final run, the database-derived ϕ-ψ constraints are enforced only for those residues which in the first run have ended in one and the same favored ϕ-ψ region. For residues which are either spread over different favored regions or concentrated in disallowed regions, the constraints are not enforced. The final family is then selected, after the second run, again only based on the agreement with the experimental constraints. This automated approach was implemented in DYANA and was tested on as many as 12 proteins, including some containing paramagnetic metals, whose structures had been previously solved in our laboratory. The quality of the structures, and of Ramachandran plot statistics in particular, was notably improved while preserving the agreement with the experimental constraints.

Introduction

The knowledge of the backbone dihedral angles ϕ and ψ is a major issue in protein solution structure determination by NMR, as they are especially important in order to define the secondary structure (Figure 1). Given a new protein whose structure is unknown and not deducible by homology with other similar proteins, these angles are of course not known *a priori*. However, on statistical grounds, ϕ and ψ show a very well defined tendency to cluster within preferred ranges of the so-called ϕ-ψ, or Ramachandran, plot. This tendency, noted and rationalized long ago (Ramachandran et al., 1963), is confirmed day by day by the availability of a continuously increasing number of high resolution/atomic resolution crystal structures. Statistically, newly determined protein structures by X-ray show more than 95% (Kleywegt and Jones, 2002) of their ϕ-ψ pairs within the expected regions, and the remaining pairs not far from them. Exceptions do occur, but they are always related to peculiar structural features. On the other hand, protein structures solved by NMR show a significantly lower percentage of ϕ-ψ pairs in the expected regions, and strong outliers are much more common than in protein structures solved by X-ray (Spronk et al., 2002). It is now widely accepted that this different behavior of solution structures – *in most cases* – does not imply any intrinsic difference with respect to the solid state structures, but rather originates from less accurate, or lack of, experimental information on the ϕ-ψ angle values.

*To whom correspondence should be addressed. E-mail: bertini@cerm.unifi.it

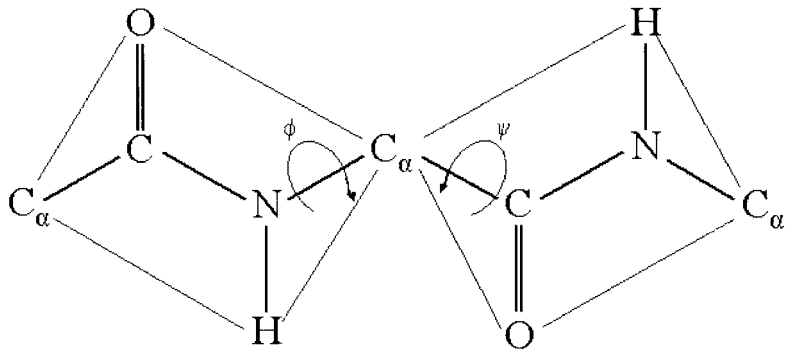

Figure 1. Scheme of the dihedral angles ϕ and ψ. ϕ is defined as the C-N-C_α-C angle, whereas ψ is defined as the N-C_α-C-N angle. The atoms bound to C_α (H_α and side-chain) are not shown for clarity.

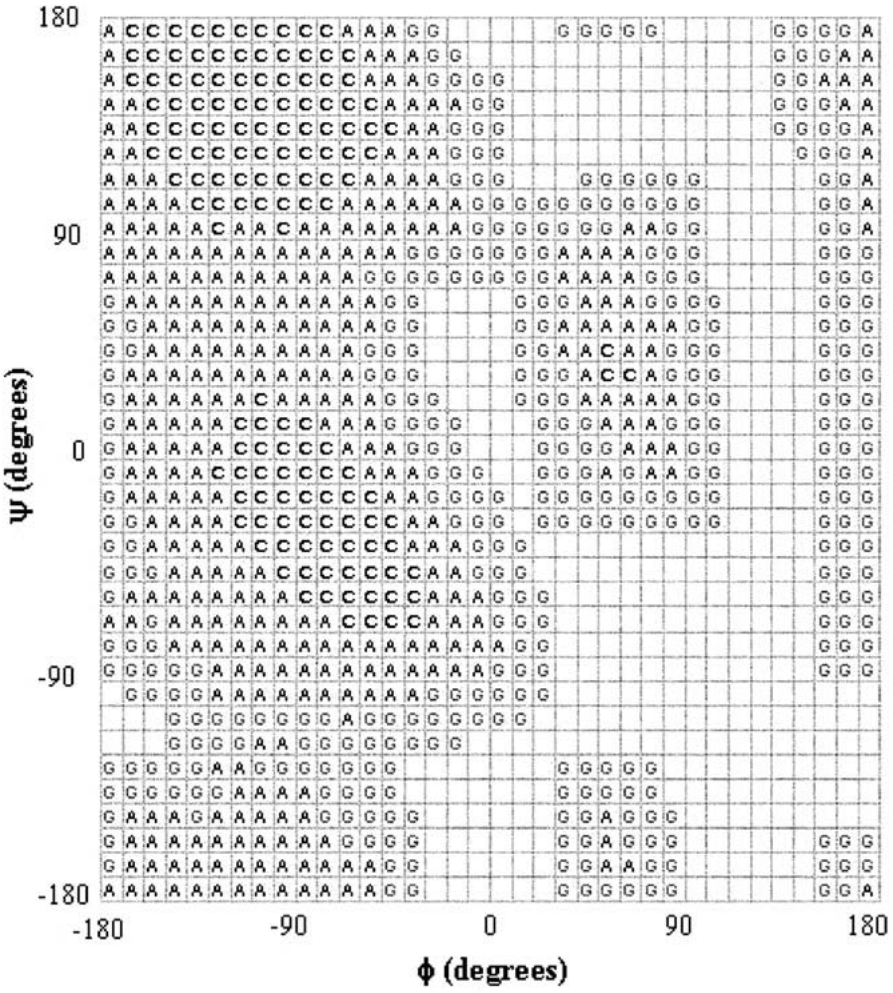

Figure 2. Ramachandran type plot derived from the PROCHECK processed database of high-resolution X-ray structures. The 360° × 360° conformational space of ϕ, ψ was divided in 1296 10° × 10°< pixels. Pixels are labeled according to the region they belong to (C = Core, A = Allowed, G = Generous).

The agreement between the observed and expected distributions of ϕ and ψ is indeed a well-established, significant check of the validity of experimentally determined protein structures. More generally, a so-called stereochemical quality of a protein, which involves checks on the geometrical properties (e.g., bond lengths, bond angles, dihedral angles, etc.), is defined and used to assess the quality of a structure (Markley et al., 1998; Doreleijers et al., 1998). The evaluation of the stereochemical quality is based on comparisons with what is known about standard protein structure and geometry from the wealth of high-resolution X-ray structures already deposited in the PDB. A widely used suite of computer programs performing a large number of validation checks on a given NMR ensemble of protein structures is PROCHECK-NMR (Laskowski et al., 1996), which is an extension of the PROCHECK programs (Laskowski et al., 1993) used for assessing the stereochemical quality of X-ray structures. Using a database of 163 high-resolution non-homologous X-ray structures, these programs can assess how 'normal' or 'abnormal' a given model is, compared with the standard values (Morris et al., 1992; MacArthur et al., 1994). In particular, they examine the dihedral angles ϕ and ψ in a protein to find any angles in the model that are uncommon, and therefore suspect.

With these premises, it would seem natural to include database-derived ϕ-ψ constraints in solution structure determinations. This has been suggested several times, and ϕ-ψ statistical constraints (as well as other dihedral angle constraints) are an option already available (Kuszewski et al., 1996) in XPLOR-NIH, the current freely available version of XPLOR (Clore and Gronenborn, 1998; Schwieters et al., 2003). However, the use of constraints based on database statistics, and in general on previous knowledge, is not universally accepted. The argument is that, while in the majority of cases these constraints will improve the accuracy of the structure, in the few but *a priori* unpredictable cases of true deviations from the favored ϕ-ψ regions, the structure will be forced to adopt a wrong local conformation if the experimental NMR constraints are not enough, and the accuracy will be lowered. A similar argument holds for energy terms based on empirical inter-atomic potentials (Sprangers et al., 2000). Another popular program, DYANA (Güntert et al., 1997) and the more recent CYANA (Herrmann et al., 2002), purposely avoid the use of empirical or database-derived potentials for this reason.

Here we suggest a strategy to use database-derived ϕ-ψ constraints simply as 'filters' during the simulated annealing procedure, to favor those structures that, by still fulfilling the experimental constraints, are in better agreement with ϕ-ψ statistics. In the end, the best structures are selected only on the ground of the agreement with the experimental constraints and not with the database-derived constraints. The procedure, fully automated, was implemented in DYANA and was tested on as many as 12 proteins whose structures had been previously solved in our laboratory. The quality of the structures, and of Ramachandran plot statistics in particular, was notably improved. Importantly, these improvements were not achieved at the expenses of the agreement with the experimental constraints.

Materials and methods

The conformational database potential was derived from the PROCHECK processed database of high-resolution X-ray structures (Morris et al., 1992; Laskowski et al., 1993). In that processing, the distribution of ϕ, ψ angles in the protein structures was analyzed, starting from the well-known observation by Ramachandran et al. (Ramachandran et al., 1963) that ϕ, ψ space for a dipeptide is very restricted for all residues except glycine, due to steric clashes. The $360^\circ \times 360^\circ$ conformational space of ϕ, ψ was divided in 1296 $10^\circ \times 10^\circ$ pixels, and the number of residues in each pixel was calculated. Proline and glycine residues were excluded from the study due to their atypical distributions (Morris et al., 1992; Kuszewski et al., 1996). On the basis of the obtained population density, three sets of allowed ϕ, ψ angles were defined: (i) The CORE region, including all pixels with more than 100 residues, (ii) the ALLOWED region, including all pixels with eight or more residues per pixel, and (iii) the GENEROUS region, defined by extending out by 20° (two pixels) all round the ALLOWED area. The space left after these regions had been defined was designated OUTSIDE. The Ramachandran type plot derived by this procedure, showing the individual pixel assignments, is shown in Figure 2. This plot was straightforwardly transformed into a 36×36 matrix of energy values at evenly spaced points along the ϕ, ψ axes, by arbitrarily assigning the following values of target function to

the four different sets:

$$\begin{aligned} TF_{CORE} &= 0\ \text{Å}^2 \\ TF_{ALLOWED} &= 1\ \text{Å}^2 \\ TF_{GENEROUS} &= 4\ \text{Å}^2 \\ TF_{OUTSIDE} &= 9\ \text{Å}^2 \end{aligned}$$

Other values could have been used as well: This choice was intended to follow the quadratic form of the potential terms used for the other constraints (Güntert et al., 1997).

The contribution of the new potential term (TF_{rama}) to the global DYANA target function can be thus expressed as:

$$TF_{rama} = W\Sigma_i w_i (TF_{set})_i, \quad (1)$$

where the sum is over all the residues on which the constraint is applied, and TF_{set} is equal to one of the four above defined values of target function, depending on the grid pixel of ϕ, ψ space that encompasses the i-th pair of ϕ, ψ angles. W and w_i are global and relative weighting factors, respectively; the former is used to rescale TF_{rama} with respect to the other constraints.

For the integration of the equations of motion in DYANA, explicit expressions for the torques about the rotatable bonds are required. In the module RAMADYANA developed in the present work, the gradients of the new potential energy term with respect to the torsion angles ϕ and ψ are approximated by the local slope of the target function, defined by:

$$\begin{aligned} \partial TF_{rama}(\theta_k)/\partial\theta &= Ww_i[TF_{rama}(\theta_{k+1}) - \\ &\quad TF_{rama}(\theta_{k-1})]/2d, \end{aligned} \quad (2)$$

where $\partial TF_{rama}(\theta_k)/\partial\theta$ is the partial derivative of the target function of pixel k with respect to the rotatable bond θ (i.e. ϕ or ψ), $TF_{rama}(\theta_{k+1})$ and $TF_{rama}(\theta_{k-1})$ are the target function values of the pixels that precede and follow the pixel k along the θ dimension in the grid, and d is the width of a pixel (i.e., 10°).

It is worth to notice that the gradients (2) can be calculated once for all from the target function matrix, yielding two other matrices whose elements are the partial derivatives of TF_{rama} with respect to ϕ and ψ, respectively. Because the gradients (2) do not need to be evaluated at every step of dynamics, the additional computational burden required by the new potential term is thus considerably lightened.

Results

The module RAMADYANA was tested on 12 proteins, which are listed in Table 1. These proteins cover a wide range of folding types (see Table 1), varying from the EF-hand motif of Calbindin D_{9k} (12 in Table 1) to the β-barrel of Cu-free SOD (4 in Table 1), and comprise diamagnetic metalloproteins (2, 4, 6 and 8 in Table 1), as well as paramagnetic metalloproteins (1, 3, 10 and 12 in Table 1).

Successive structure calculations by DYANA were performed on each protein, using the available experimental constraints (Assfalg et al., 1999, 2002; Bartalesi et al., 2002; Banci et al., 1994, 2001, 2002; Arnesano et al., 2001, 2002; Bertini et al., 2001). These constraints are summarized in Table 2. In all the calculations, 200 conformers with initial random values of the torsion angles were subjected to 12000 steps of the simulated annealing procedure, and the 20 structures with the lowest values of the target function were sorted out. In no case, the contribution of TF_{rama} was taken into account to select the models of the final ensemble: the rationale for this choice is described in the Discussion.

In the reference calculation, the new potential term, TF_{rama}, was excluded from the global DYANA target function, by setting the weighting factor of this term (W in Equation 1) to zero. Statistics on the resulting ensembles of structures are collected in Table 3 under the heading NO.

In the first calculation with the RAMADYANA module, the new potential term, TF_{rama}, was included in the global target function, setting the weighting factor W to 1. The new constraint was applied to all the residues of the protein chain, except for proline and glycine residues. The distribution of the values of ϕ and ψ over the selected ensemble of 20 models was then analyzed on a residue-by-residue basis, in order to evaluate whether the application of the new constraint on every single residue was allowed. In particular, the residues were selected for the second run according to their value of circular variance (Allen and Johnson, 1991; MacArthur and Thornton, 1993), which quantifies the degree of spread of ϕ and ψ values across the structures of the ensemble. The circular variance for a given dihedral angle θ is defined as:

$$Var(\theta) = 1 - R/n, \quad (3)$$

the parameter R being given by the expression:

$$R^2 = (\Sigma_{i=1,n} \cos\theta_i)^2 + (\Sigma_{i=1,n} \sin\theta_i)^2, \quad (4)$$

Table 1. Proteins used for test calculations

	Protein	Source	Residues	PDB code	Fold
1	Fully oxidized K9-10A cyt *c*7	Desulforomonas acetoxidans	68	1L30[a] 1KWJ[b]	βαββαβ
2	Reduced iso-1-cyt*c*	Saccharomyces cerevisiae	108	1YFC	All-α
3	Oxidized cyt*c*	Shewanella putrefaciens	81	1KX7[a] 1KX2[b]	All-α
4	Monomeric Cu-free SOD	Homo sapiens	153	1KMG	All-β
5	Apo-CopA (N-terminal domain)	Bacillus subtilis	80	1JWW	βαββαβ
6	Cu(I)-CopA (N-terminal domain)	Bacillus subtilis	80	1KQK	βαββαβ
7	Apo-Atx1	Saccharomyces cerevisiae	73	1FES	βαββαβ
8	Cu(I)-Atx1	Saccharomyces cerevisiae	73	1FD8	βαββαβ
9	Apo-Ccc2A domain	Saccharomyces cerevisiae	72	1FVQ	βαββαβ
10	Oxidized HiPIP I	Ectothiorhodospira halophila	73	1PIH[a] 1PIJ[b]	All-β
11	Apo-CopC	Pseudomonas syringae	102	1M42	All-β
12	Ca(II)-Ce(III) Calbindin D_{9k}	Bos Taurus	79	1KQV[a] 1KSM[b]	All-α

[a]Ensemble of structures.
[b]Minimized mean structure.

Table 2. Type and number of constraints used in DYANA calculations

Protein	NOE constraints	Dihedral angles constraints	RDC constraints	RAMADYANA constraints	Residues with Var larger than the threshold value	Residues consistently falling in disallowed regions	Threshold value of Var
1	1186	–	–	55	2	1	0.11
2	1473	–	–	56	34	–	0.12
3	1310	70	–	64	5	–	0.13
4	2467	165	–	95	27	–	0.11
5	1278	95	–	61	10	–	0.11
6	1415	87	–	59	12	–	0.05
7	1176	–	–	56	9	–	0.18
8	1148	42	60	59	6	–	0.17
9	1970	35	–	61	6	–	0.13
10	1125	–	–	57	3	–	0.15
11	1437	149	–	71	10	1	0.18
12	1715	–	–	64	5	–	0.13

Table 3. Selected parameters describing the structures obtained with all the available NOE, dihedral angles and RDC constraints, without (NO) or with (YES) RAMA constraints

Protein	RAMA constraints	RMSD backbone (Å)	Target function excluding RAMA constraints	Ramachandran			
				Core (%)	All. (%)	Gen. (%)	Dis. (%)
1	NO	1.11 ± 0.29	0.57 ± 0.02	47.5	43.5	5.9	3.0
	YES	1.05 ± 0.17	0.41 ± 0.06	63.4	33.4	1.6	1.6
2	NO	1.01 ± 0.12	0.44 ± 0.07	55.1	37.3	6.6	1.0
	YES	0.95 ± 0.15	0.37 ± 0.07	67.5	29.1	3.1	0.4
3	NO	0.78 ± 0.15	0.63 ± 0.06	63.5	28.9	5.3	2.3
	YES	0.67 ± 0.11	0.65 ± 0.09	70.4	26.4	2.9	0.4
4	NO	1.20 ± 0.16	1.66 ± 0.18	54.5	39.5	5.1	1.0
	YES	1.00 ± 0.15	1.92 ± 0.36	70.6	25.0	2.9	1.5
5	NO	1.65 ± 0.25	0.69 ± 0.11	51.3	42.8	4.9	1.0
	YES	1.62 ± 0.52	0.73 ± 0.11	70.5	24.7	3.4	1.4
6	NO	0.70 ± 0.14	0.47 ± 0.04	73.2	20.7	5.6	0.5
	YES	0.77 ± 0.14	0.63 ± 0.07	77.6	17.8	3.7	0.9
7	NO	0.94 ± 0.19	0.54 ± 0.15	60.3	34.9	2.6	2.2
	YES	1.01 ± 0.23	0.52 ± 0.13	70.6	25.0	2.9	1.5
8	NO	0.56 ± 0.09	0.79 ± 0.08	71.2	25.8	2.5	0.5
	YES	0.57 ± 0.12	0.76 ± 0.11	77.0	20.6	1.8	0.6
9	NO	0.56 ± 0.09	0.59 ± 0.06	67.5	27.9	4.0	0.6
	YES	0.57 ± 0.08	0.65 ± 0.09	80.2	19.3	0.4	0.0
10	NO	0.65 ± 0.09	0.45 ± 0.60	56.1	39.5	2.7	1.8
	YES	0.66 ± 0.11	0.39 ± 0.11	75.9	23.4	0.5	0.2
11	NO	1.36 ± 0.32	0.37 ± 0.09	69.3	23.4	5.6	1.7
	YES	1.33 ± 0.35	0.30 ± 0.12	75.8	17.6	4.7	2.0
12	NO	0.94 ± 0.21	0.09 ± 0.01	72.8	23.8	3.1	0.2
	YES	0.95 ± 0.15	0.17 ± 0.02	94.8	4.2	0.9	0.0

where n is the number of members in the ensemble. The value of the circular variance varies from 0 to 1, with the lower the value the tighter the clustering of the values about a single mean value. An averaged, residue-specific circular variance is thus straightforwardly defined as:

$$\text{Var(residue)} = [\text{Var}(\phi) + \text{Var}(\psi)]/2 \quad (5)$$

with Var(residue) likewise ranging from 0 to 1. Residues with a value of circular variance Var(residue) larger than a given threshold value were considered unsuitable for the application of the new constraint (see Discussion). Moreover, residues with a small value of Var(residue), but located in unfavorable regions of the Ramachandran plot, were excluded as well (see Discussion). The above mentioned threshold value can be assigned by default (e.g., 0.15), or its choice can be prompted by the program through visual inspection of the Ramachandran plots per residue, which are provided as an output by PROCHECK-NMR (Laskowski et al., 1996), and comparison with the corresponding values of Var(residue).

In the second calculation, the new potential term, TF_{rama}, was again included in the global target function, setting the weighting factor W to 1. However, the new constraint was applied only to the residues selected on the basis of the analysis described above, i.e. residues with a value of circular variance Var(residue) larger than the threshold value, as well residues with small circular variance consistently falling in the disallowed regions, were excluded by setting their relative weighting factors (w_i in Equation 1) to zero. The number of constraints and the threshold value of Var(residue) used for each protein, as well as the number of residues excluded on the basis of these criteria, are shown in Table 2. The structures composing the final family were again sorted only on the basis of the target function for the experimental constraints. Selected parameters of the resulting en-

sembles of structures are collected in Table 3 under the heading YES, for comparison with structures obtained without (heading NO) the RAMADYANA strategy.

The value of the target function, calculated with the exclusion of the TF_{rama} contribution, slightly increases in 6 cases (proteins 3, 4, 5, 6, 9 and 12), whereas it decreases in the other 6 cases (proteins 1, 2, 7, 8, 10 and 11). This shows that the agreement with the available NMR constraints is not compromised by the use of the RAMADYANA strategy during the simulated annealing. The average value of backbone RMSD decreases in 6 cases (proteins 1, 2, 3, 4, 5 and 11), whereas it is essentially unchanged in 4 cases (proteins 8, 9, 10 and 12) and slightly increases in the other 2 cases (proteins 6 and 7).

On the other hand, the improvement in the Ramachandran plot statistics is notable, yet not trivial (see Discussion), with an average increase of 12.7% of residues in the CORE region (from 61.8% to 74.5%), and an average decrease of 2.5% of residues in the DISALLOWED + GENEROUS regions (from 5.8% to 3.3%).

When available (Louie and Brayer, 1990; Breiter et al., 1991; Svensson et al., 1992; Rosenzweig et al., 1999; Czjzek et al., 2001), X-ray structures were compared to the solution structures obtained with (YES) or without (NO) the RAMADYANA strategy, evaluating both the backbone RMSD and the correlation between the values of ϕ and ψ (see Table 4): the improved agreement with the crystal structures in the former case provides a further proof of the enhancement in the accuracy of NMR structures due to the use of this strategy. A representative plot of the calculated ϕ-ψ values versus the corresponding ones measured in the X-ray structure is shown in Figure 3 for protein 2.

It is worth to notice that only one Ramachandran plot outlier was found among the available X-ray structures, i.e., the residue Glu39 of protein 10. Conversely, in the NMR structure it falls very well within the Ramachandran B region, both without and with the RAMADYANA strategy. This particular disagreement is thus not an artifact generated by the new constraint. Indeed, this outlier in the X-ray structure is present in only one of the two molecules per asymmetric unit, and has been interpreted as due to crystal packing forces (Breiter et al., 1991).

Discussion

As discussed in the Introduction, a possible drawback in the use of empirical information for defining the backbone conformation of a protein is that it can bias the calculated structure towards the structures present in the databases from which the conformational potential was derived. Errors can be introduced into the protein model in two cases: (i) The structure actually has unusual features which bring to outliers in the Ramachandran plot (Kleywegt and Jones, 1996); (ii) some regions of the backbone are poorly defined due to the lack or scarcity of experimental data. In case (i), residues that truly exhibit ϕ, ψ angles located in unfavorable regions of the Ramachandran plot will be spotted by specific discrepancies between the model and the experimental constraints. Such violations will therefore be accounted for by special structural features of the protein, provided that there is extensive experimental evidence to account for those unusual ϕ, ψ values (Kleywegt and Jones, 1996). In case (ii), the values of ϕ and ψ for residues lacking experimental information will be mainly determined by the conformational potential, an eventuality which we want to avoid. Since we defined this potential as having three equivalent minima corresponding to the three CORE regions (see Methods and Figure 2), the values of ϕ and ψ across an ensemble of calculated structures for residues poorly defined by experimental data will have a tendency to cluster in the three minima. Such an unwanted distribution is hinted at by a relatively large value of the circular variance (see Equation 5), with respect to residues showing values of ϕ and ψ clustered by the experimental constraints about a single mean value. Thus, the conformational potential can cause errors in the calculated protein models either by conflicting with the experimental constraints (case i), or by replacing them in determining the resulting structures (case ii). The strategy that we suggest here eliminates these possible errors, endowing the constraints derived from the experimental data with absolute prevalence with respect to the constraints derived from structure databases. This is achieved through the execution of a preliminary structure calculation aimed at selecting the residues on which the conformational potential can be safely applied. In that calculation, the new constraint is applied to all the residues, and the distribution of the values of ϕ and ψ over the resulting ensemble of structures is analyzed. In order to eliminate errors due to lack of experimental data (case ii), the residues with a value of circular vari-

Table 4. Comparison between the available X-ray structures and the average NMR structures obtained with all the available experimental constraints, without (NO) or with (YES) RAMA constraints. The RMSD for backbone atoms was evaluated, as well as the squared correlation coefficient (R^2) between the calculated values of ϕ and ψ and the corresponding ones measured in the X-ray structure

Protein	PDB code of X-ray structure	RAMA constraints	ϕ and ψ correlation (R^2 value)	Backbone RMSD (Å)
1	1HH5	NO	0.83	1.30
		YES	0.90	1.22
2	1YCC	NO	0.92	0.88
		YES	0.96	0.86
8	1CC8	NO	0.96	1.13
		YES	0.97	1.11
10	2HIP	NO	0.91	0.89
		YES	0.92	0.89
12	4ICB[a]	NO	0.82	1.43
		YES	0.88	1.34

[a]Ca(II)-Ca(II) Calbindin D_{9k}.

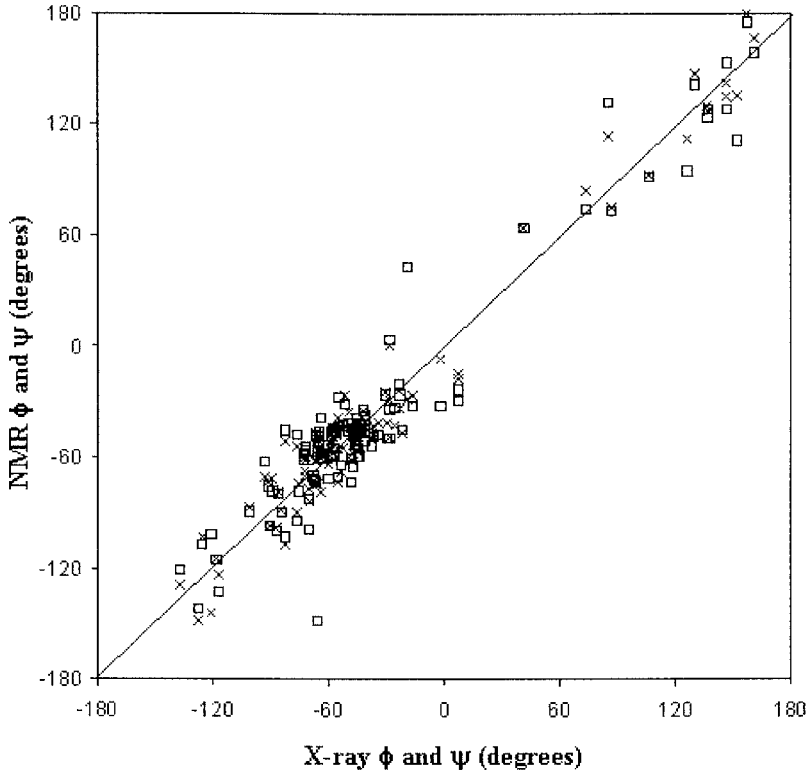

Figure 3. Plot of ϕ and ψ values calculated for protein 2 with (crosses) and without (squares) RAMA constraints versus the corresponding ϕ and ψ values measured in the X-ray structure. The squared correlation coefficient R^2 is 0.92 for the calculation without RAMA constraints and 0.96 for the calculation with RAMA constraints.

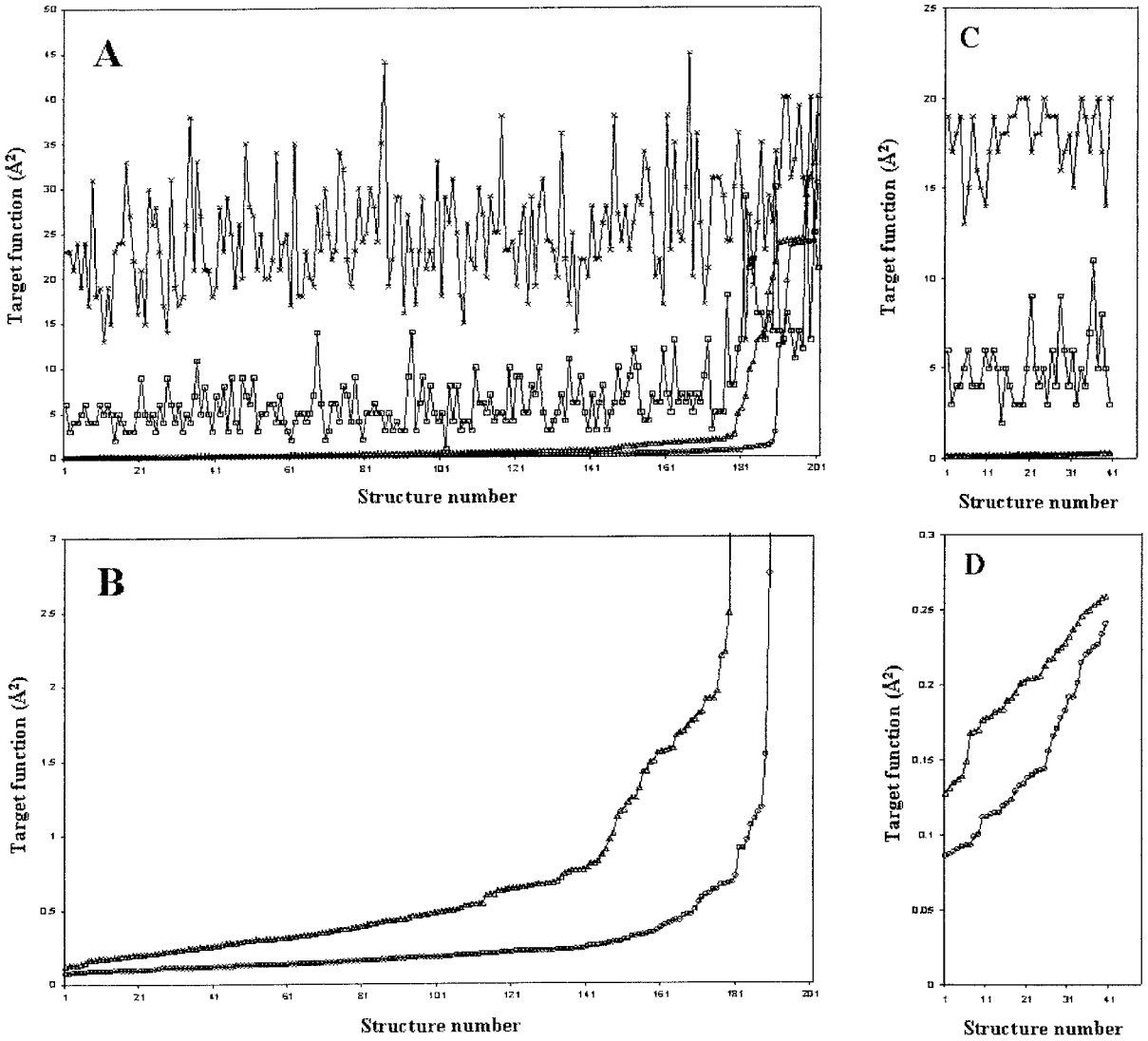

Figure 4. Contributions to the target function value of the structures calculated for protein 12 without and with RAMA constraints. The contribution of the experimental constraints is shown as circles (calculation without RAMA constraints) or triangles (calculation with RAMA constraints), whereas the contribution of RAMA constraints is shown as crosses (calculation without RAMA constraints) or squares (calculation with RAMA constraints). In (A) and (B) all the 200 calculated structures are shown, sorted in order of increasing contribution of the experimental constraints to the target function. In (C) and (D) the 40 structures with the lowest violations of RAMA constraints resulting from the calculation without RAMA constraints are selected and compared to the 40 structures with the lowest violations of experimental constraints resulting from the calculation with RAMA constraints.

ance Var(residue) larger than a given threshold value are considered potentially dangerous, and the new constraint is not applied on them in the successive calculation. Furthermore, in order to eliminate errors due to conflict between the experimental data and the conformational potential (case i), also the residues with a value of Var(residue) under the threshold, yet located in unfavorable regions of the Ramachandran plot, are considered unsuitable for the application of the new constraint in the successive calculation.

A key feature of the RAMADYANA strategy is that the structures are sorted out on the basis of the value of the target function with the exclusion of the Ramachandran constraints contribution. This allows us to use a relatively large weighting factor $W = 1$ (see Equation 1) for the Ramachandran constraints contribution, in such a way that the contribution of TF_{rama}

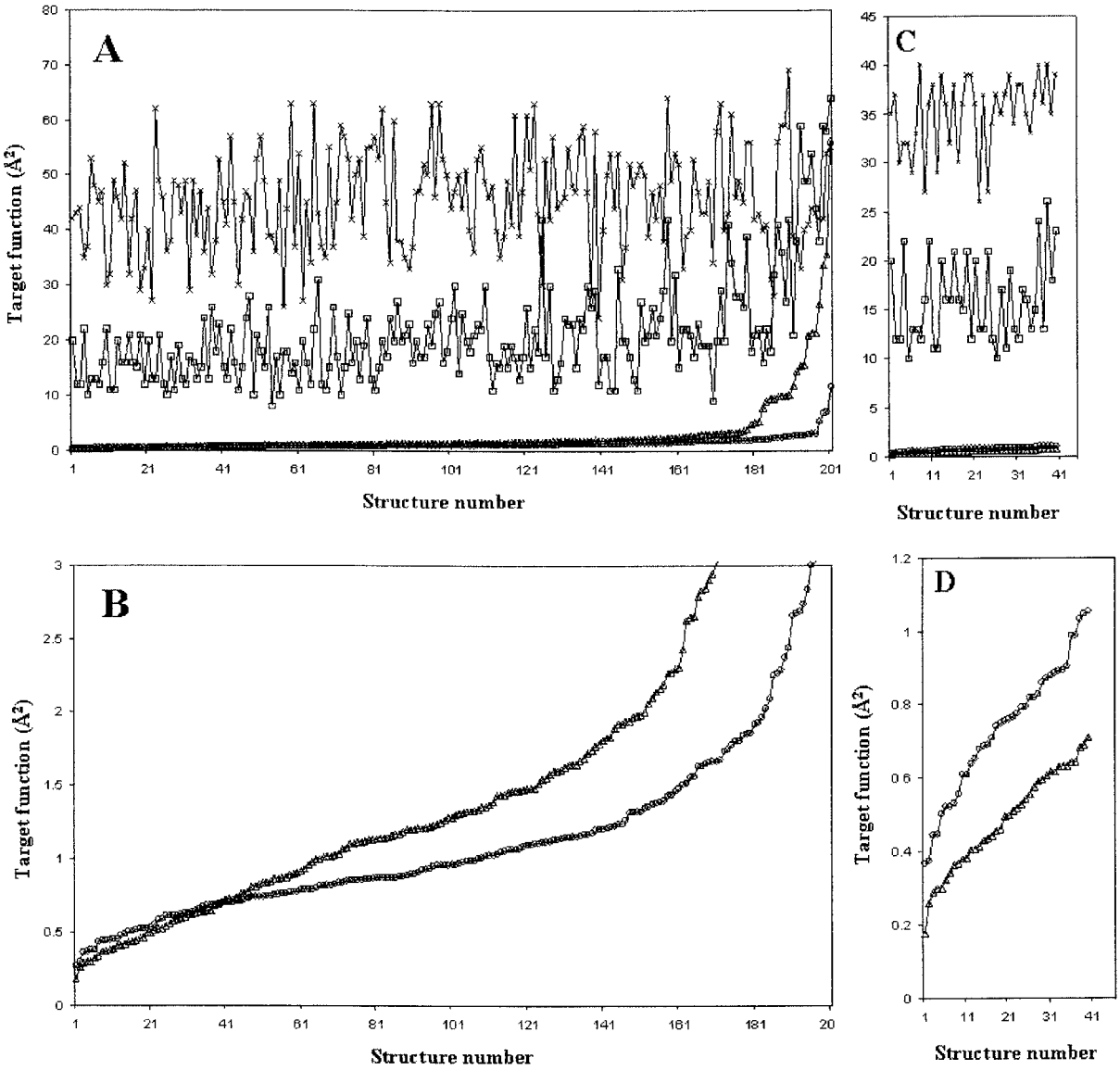

Figure 5. Contributions to the target function value of the structures calculated for protein 2 without and with RAMA constraints. The contribution of the experimental constraints is shown as circles (calculation without RAMA constraints) or triangles (calculation with RAMA constraints), whereas the contribution of RAMA constraints is shown as crosses (calculation without RAMA constraints) or squares (calculation with RAMA constraints). In (A) and (B) all the 200 calculated structures are shown, sorted in order of increasing contribution of the experimental constraints to the target function. In (C) and (D) the 40 structures with the lowest violations of RAMA constraints resulting from the calculation without RAMA constraints are selected and compared to the 40 structures with the lowest violations of experimental constraints resulting from the calculation with RAMA constraints.

to the global target function is of the order of few tens of Å^2, thus largely predominant with respect to the other contributions. The final ensemble of structures however comprises the models which fulfill best the experimental constraints only, although the TF_{rama} contribution was actually taken into account for their determination.

The consequences of this strategy are illustrated in Figures 4 and 5 for proteins 12 and 2, respectively. The Figures illustrate the contributions to the target function due to the experimental and the RAMA constraints for the structures calculated both with and without RAMA constraints. It appears that the addition of RAMA constraints causes a modest increase

of the experimental contribution for protein 12 (Figure 4B), and even its decrease in the best 40 structures of protein 2 (Figure 5B). At the same time, the target function for the Ramachandran violations decreases, on average, from 25 to 5 $Å^2$ for protein 12 (Figure 4A) and from 40 to 15 $Å^2$ for protein 2 (Figure 5A). It can be noted that it would have been impossible to use such a high weight for the Ramachandran constraints from the very beginning of the simulated annealing procedure if their contribution to the target function were to be kept into account for sorting the structures. In that case the Ramachandran constraints would have had an overwhelming role. Moreover, a filtering process done *a posteriori*, i.e., a selection of the structures on the ground of the lowest Ramachandran violations done *after* performing a simulated annealing *without* Ramachandran constraints would not have been as effective as the present strategy (see Figures 4C and 4D for protein 12 and Figures 5C and 5D for protein 2).

The significant improvement in the Ramachandran plot statistics achieved through the use of the conformational potential described in this work (see Results and Table 3) is not deceptive, because the agreement with the experimental constraints was not compromised in any of the calculations: in fact, it even improved in 6 cases (proteins 1, 2, 7, 8, 10 and 11), whereas only a minimal impairment was observed in the other cases. Moreover, even when the impairment is relatively large (proteins 4, 6 and 12), it can be observed (data not shown) that this is not due to the generation of consistent violations in the structures obtained with the RAMADYANA strategy; rather, the value of the target function increases because of the rise of small NOE violations (under 0.05 Å) spread all over the chain, which are not meaningful for the quality of the structures. In any case, it should be kept in mind that in no case the experimental constraints were re-calibrated with respect to the original calibration. Re-calibration, which is possible within the present program, would probably have brought further slight improvements in the structures.

For these reasons, also the decrease in the average value of backbone RMSD observed in 6 cases (proteins 1, 2, 3, 4, 5 and 11) should be seen as a real improvement in the precision of the structure ensembles. The RMSD is instead unchanged for proteins 8, 9, 10 and 12 and slightly increases for proteins 6 and 7. In any case, the increase in precision, where present, should be taken as an additional bonus and not as the goal of this strategy, which is to increase the accuracy of the obtained structures.

Finally, a comment is due on the form chosen for the conformational potential. The potential defined in the present procedure is flat within each type of region, i.e., a force is active only along region boundaries, but not in the interior of forbidden regions. A variety of methods could be used to smooth that function, in order to reduce the number of pixels for which both the local slopes with respect to ϕ and ψ are null. However, test calculations with potentials acting over the whole ϕ, ψ conformational space showed a worsening in the convergence properties of the calculation (data not shown). This was somehow expected, because the original potential fits at best the RAMADYANA strategy, whose innocence is based on using the database-derived constraints simply as filters during the simulated annealing procedure.

Conclusions

The present strategy to use the information from databases of high-resolution protein crystal structures can provide a useful addition to improve the accuracy of NMR structures. In particular, we defined a conformational potential term based on the relative populations of various combinations of ϕ, ψ dihedral angles observed in databases, and we implemented it in the DYANA program through the new module RAMADYANA. The new potential term is aimed at improving the definition of the backbone conformation of a protein through the introduction of a constraint directly acting on ϕ and ψ angles, without influencing the sorting of the resulting structures. The possibility that the use of empirical information bias the conformation of the protein towards the structures existing in the databases was addressed through the definition of a protocol which (i) gives absolute relevance to the available experimental constraints with respect to the conformational constraints, and (ii) automatically eliminates the constraints when they spread the ϕ-ψ values over more than one allowed region or when the ϕ-ψ values concentrate in a disallowed region due to experimental constraints. With this approach, which was tested on as many as 12 proteins, the quality of the structures, and of Ramachandan plot statistics in particular, was notably increased. The improved agreement with the available crystal structures confirms the validity of the strategy. Importantly, these improvements were not achieved at the expenses of the

agreement with the experimental constraints, which was not compromised in any way.

Acknowledgements

Financial support of the EU through contract nos. QLG2-CT-1999-01003 and QLG2-CT-2002-00988 is gratefully acknowledged.

References

Allen, F.H. and Johnson, O. (1991) *Acta Cryst. B*, **47**, 62–67.

Arnesano, F., Banci, L., Bertini, I., Huffman, D.L. and O'Halloran, T.V. (2001) *Biochemistry*, **40**, 1528–1539.

Arnesano, F., Banci, L., Bertini, I. and Thompsett, A.R. (2002) *Structure*, **10**, 1337–1347.

Assfalg, M., Banci, L., Bertini, I., Bruschi, M., Giudici-Orticoni, M.T. and Turano, P. (1999) *Eur. J. Biochem.*, **266**, 634–643.

Assfalg, M., Bertini, I., Turano, P., Bruschi, M., Durand, M.C., Giudici-Orticoni, M.T. and Dolla, A. (2002) *J. Biomol. NMR*, **22**, 107–122.

Banci, L., Bertini, I., Cantini, F., D'Onofrio, M. and Viezzoli, M.S. (2002a) *Protein Sci.*, **11**, 2479–2492.

Banci, L., Bertini, I., Ciofi-Baffoni, S., D'Onofrio, M., Gonnelli, L., Marhuenda-Egea, F.C. and Ruiz-Dueñas, F.J. (2002b) *J. Mol. Biol.*, **317**, 415–429.

Banci, L., Bertini, I., Ciofi-Baffoni, S., Huffman, D.L. and O'Halloran, T.V. (2001) *J. Biol. Chem.*, **276**, 8415–8426.

Banci, L., Bertini, I., Eltis, L.D., Felli, I.C., Kastrau, D.H.W., Luchinat, C., Piccioli, M., Pierattelli, R. and Smith, M. (1994) *Eur. J. Biochem.*, **225**, 715–725.

Bartalesi, I., Bertini, I., Hajieva, P., Rosato, A. and Vasos, P. (2002) *Biochemistry*, **41**, 5112–5119.

Bertini, I., Donaire, A., Jimenez, B., Luchinat, C., Parigi, G., Piccioli, M. and Poggi, L. (2001) *J. Biomol. NMR*, **21**, 85–98.

Breiter, D.R., Meyer, T.E., Rayment, I. and Holden, H.M. (1991) *J. Biol. Chem.*, **266**, 18660–18667.

Clore, G.M. and Gronenborn, A.M. (1998) *Proc. Natl. Acad Sci. USA*, **95**, 5891–5898.

Czjzek, M., Arnoux, P., Haser, R. and Shepard, W. (2001) *Acta Cryst. D*, **57**, 670–678.

Doreleijers, J.F., Rullmann, J.A.C. and Kaptein, R. (1998) *J. Mol. Biol.*, **281**, 149–164.

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.

Kleywegt, G.J. and Jones, T.A. (1996) *Structure*, **4**, 1395–1400.

Kleywegt, G.J. and Jones, T.A. (2002) *Structure*, **10**, 465–472.

Kuszewski, J., Gronenborn, A.M. and Clore, G.M. (1996) *Protein Sci.*, **5**, 1067–1080.

Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) *J. Appl. Crystallogr.*, **26**, 283–291.

Laskowski, R.A., Rullmann, J.A.C., MacArthur, M.W., Kaptein, R. and Thornton, J.M. (1996) *J. Biomol. NMR*, **8**, 477–486.

Louie, G.V. and Brayer, G.D. (1990) *J. Mol. Biol.*, **214**, 527–555.

MacArthur, M.W. and Thornton, J.M. (1993) *Proteins Struct. Funct. Genet.*, **17**, 232–251.

MacArthur, M.W., Laskowski, R.A. and Thornton, J.M. (1994) *Curr. Opin. Struct. Biol.*, **4**, 731–737.

Markley, J.L., Bax, A., Arata, Y., Hilbers, C.W., Kaptein, R., Sykes, B.D., Wright, P.E. and Wüthrich, K. (1998) *J. Biomol. NMR*, **12**, 1–23.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G. and Thornton, J.M. (1992) *Proteins Struct. Funct. Genet.*, **12**, 345–364.

Ramachandran, G.N., Ramakrishnan, C. and Sasisekharan, V. (1963) *J. Mol. Biol.*, **7**, 95–99.

Rosenzweig, A.C., Huffman, D.L., Hou, M.Y., Wernimont, A.K., Pufahl, R.A. and O'Halloran, T.V. (1999) *Struct. Fold Des.*, **7**, 605–617.

Schwieters, C.D., Kuszewski, J., Tjandra, N. and Clore, G.M. (2003) *J. Magn. Reson.*, **160**, 65–73.

Sprangers, R., Bottomley, M.J., Linge, J.P., Schultz, J., Nilges, M. and Sattler, M. (2000) *J. Biomol. NMR*, **16**, 47–58.

Spronk, C.A.E.M., Linge, J.P., Hilbers, C.W. and Vuister, G.W. (2002) *J. Biomol. NMR*, **22**, 281–289.

Svensson, L.A., Thulin, E. and Forsén, S. (1992) *J. Mol. Biol.*, **223**, 601–606.