Journal of Biomolecular NMR (2005) 33: 123–134
DOI 10.1007/s10858-005-2826-5

Article

Definition of a new information-based per-residue quality parameter

Sander B. Nabuurs[a], Elmar Krieger[a], Chris A. E. M. Spronk[a], Aart J. Nederveen[b], Gert Vriend[a] & Geerten W. Vuister[c,*]

[a]*Center for Molecular and Biomolecular Informatics, NCMLS, Radboud University Nijmegen, Toernooiveld 1, 6525 ED, Nijmegen, The Netherlands;* [b]*Department of NMR spectroscopy, University of Utrecht, Padualaan 8, 3584 CH, Utrecht, The Netherlands;* [c]*Department of Biophysical Chemistry, IMM, Radboud University Nijmegen, Toernooiveld 1, 6525 ED, Nijmegen, The Netherlands*

Received 18 July 2005; Accepted 29 August 2005

Key words: molecular dynamics, NMR, protein structure, structure validation, structure refinement

Abstract

For biomolecular NMR structures typically only a poor correspondence is observed between statistics derived from the experimental input data and structural quality indicators obtained from the structure ensembles. Here, we investigate the relationship between the amount of available NMR data and structure quality. By generating datasets with a predetermined information content and evaluating the quality of the resulting structure ensembles we show that there is, in contrast to previous findings, a linear relation between the information contained in experimental data and structural quality. From this relation, a new quality parameter is derived that provides direct insight, on a per-residue basis, into the extent to which structural quality is governed by the experimental input data.

Introduction

Within the field of biomolecular structure determination by NMR spectroscopy, the relation between the amount and quality of the experimental input data and the precision and accuracy of the resulting protein structures has been extensively investigated (Oshiro et al., 1991; Liu et al., 1992; Clore et al., 1993; Zhao and Jardetzky, 1994). In addition to the nuclear Overhauser effect (NOE) derived distance restraints, the primary source of structural information, other types of experimental data have been evaluated and implemented in NMR structure calculations. Several studies have shown the beneficiary effects of *J*-couplings (Garrett et al., 1994; Kim and Prestegard, 1990; Mierke et al., 1994), chemical shifts (Kuszewski et al., 1995a, 1995b), residual dipolar couplings (Clore et al., 1999; Tjandra et al., 2000, 1997b), T_1/T_2-ratios (Tjandra et al., 1997a) and paramagnetic shifts (Banci et al., 1997, 2004) on both the precision and accuracy of biomolecular structures determined by NMR.

Recently, an interest has arisen in not only the precision and accuracy, but also the quality of protein structures determined by NMR spectroscopy (Doreleijers et al., 1998; Linge and Nilges, 1999; Spronk et al., 2002; Linge et al., 2003; Snyder et al., 2005). The quality of biomolecular structure models is commonly evaluated by indicators describing the packing of core residues and the normality of backbone and side-chain conformations (as reviewed in Spronk et al., 2004). Such

*To whom correspondence should be addressed. E-mail: g.vuister@science.ru.nl

quality indicators are often expressed as a Z-score (Hooft et al., 1997; Spronk et al., 2004), defined as the deviation from the average value for this indicator observed in a database of high resolution crystal structures, expressed in units of the standard deviation of this database derived average.

From X-ray crystallography, it is known that good experimental data are a prerequisite, but not a guarantee, for a high-quality structure and that refinement techniques play a crucial role (Kleywegt and Jones, 1995). This notion also holds true for the NMR structure determination process (Linge et al., 2003; Spronk et al., 2002). A recent re-refinement of 100 NMR derived protein structures in explicit solvent demonstrates a significant improvement in all structural quality scores combined with an equally good or better fit to the experimental data (Nabuurs et al., 2004). Similar results were obtained when a large set of NMR structures was recalculated from scratch using several well-established protocols (Nederveen et al., 2005). These recalculated structures initially showed marginal improvements over the original ones, and only after a final refinement step using molecular dynamics in explicit water did the structures exhibit a similar improvement in quality as observed previously (Nabuurs et al., 2004).

For crystal structures it has been observed that quality indicators, such as those for the Ramachandran plot, tend to improve as the accuracy of the structural model, as judged by resolution and R_{free} factor, improves (Kleywegt and Jones, 1996, 2002). Studies on NMR structures, however, have only shown a poor correlation between structural quality indicators and indicators representing the available experimental NMR data (Doreleijers et al., 1998; Nederveen et al., 2005). Figure 1 shows that classical descriptors of the amount of available NMR data, such as the number of restraints per residue (hereafter referred to as data density) and the NOE completeness (Doreleijers et al., 1999), are only weakly correlated with structural quality indicators (Doreleijers et al., 1999; Nederveen et al., 2005). The structural uncertainty (see Equation 1), calculated from the experimental input data by the QUEEN method (Nabuurs et al., 2003), shows a reasonable correlation with the backbone root mean square deviation (RMSD) of the final structure ensembles, but nevertheless also correlates weakly with overall structural quality indicators. At first glance, these findings seem to point in the direction that the quality of NOE derived NMR structures is foremost determined by the final refinement step and not, as one would expect, by the amount of available experimental data.

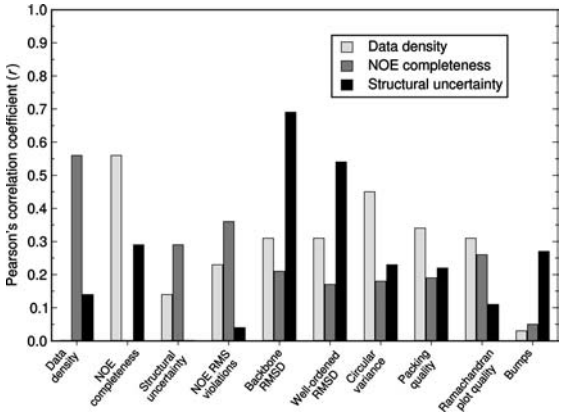

Figure 1. Analysis of experimental data and recalculated structures in the RECOORD database (Nederveen et al., 2005). Three data derived quality scores and seven other common structural quality indicators were correlated. The absolute Pearson's correlation coefficients (r) for data density, NOE completeness and QUEEN structural uncertainty ($H_{structure}$) are displayed on the vertical axis in white, grey and black bars, respectively, as function of the ten parameters along the horizontal axis.

Here, we investigate the relation between the amount of experimental NMR data and the structural quality of the resulting structure ensembles. Using the QUEEN method to determine the data information content, we find, in contrast to previous studies using classical methods, a clear relation between the amount of data and structural quality. This relation is revealed by monitoring structural quality as more experimental information is gradually introduced into the structure calculation process. For this purpose, information measures determined by QUEEN are used to construct subsets of experimental data with a predetermined information content. We show that the information content of these subsets is directly correlated to the accuracy of the resulting structure ensembles, as assessed by cross-validation. Subsequently, the validation results obtained from these structures are used to demonstrate for the first time that the structural quality of NMR structures does indeed directly relate to the amount of experimentally obtained information, albeit that this relation can only be assessed on a per-residue basis. Finally, we then

use these results to define, apply and discuss a new information based per-residue quality parameter.

Material and methods

Datasets

We use the experimental datasets of the B1 immunoglobulin binding domain of protein G (GB1) (Kuszewski et al., 1999), the protein ubiquitin (UBI) (Cornilescu et al., 1998), the alternatively spliced form of the second PDZ domain of PTP-BL (PDZ) (Walma et al., 2004), the cold-shock domain of the Y-box protein YB-1 (YB1) (Kloks et al., 2002) and the presequence peptide of the protein 5-aminolevulinate synthase (PSA) (Goodfellow et al., 2001) as obtained from the BioMagResBank (BMRB) (Doreleijers et al., 2003). For simplicity, only unambiguously assigned NOE distance restraints are taken into account in this study, ambiguous and other types of restraints were removed from all datasets. The UBI, GB1 and PDZ datasets present examples of well-determined, folded proteins. In contrast, the YB1 dataset is less well determined, mostly due to the high flexibility of this protein in solution. The PSA dataset describes a partially unfolded protein, with only ~25% of its residues involved in regular secondary structure elements, and is used to demonstrate that the presented method can also be applied to these types of systems.

For each experimental dataset subsets of restraints were generated, containing a pre-determined fraction of the total available structural information, using the QUEEN program (Nabuurs et al., 2003). Restraints were randomly selected from the complete dataset and sequentially added to the new, initially empty, dataset until the desired fraction of the total information was obtained. Datasets with less than 25 restraints, or whose information content deviated more than 1% from the target information content, were rejected. All subsets were generated in five-fold to assess the variance for each individual target value.

Structural uncertainty

The structural uncertainty resulting from the different datasets was calculated using the QUEEN program (Nabuurs et al., 2003). This method is based on a representation of the structure in distance space and concepts derived from information theory. As most experimental NMR data is readily represented in distance space, it is possible to construct a distance matrix representing all available distance information. The structural uncertainty of an individual residue is defined in QUEEN as

$$H_{\mathrm{res}} = \frac{1}{N_r(N_s-1)} \sum_{r=1}^{N_r} \sum_{s \neq r}^{N_s} \log(d_{rs}^{\mathrm{upper}} - d_{rs}^{\mathrm{lower}}) \qquad \text{(Eq. 1)}$$

with N_r the number of atoms in the residue, N_s the number of atoms in the structure, d_{rs}^{upper} the upper bound for the distance between atoms r and s and d_{rs}^{lower} the lower bound for that same distance. The structural uncertainty of the complete structure ($H_{\mathrm{structure}}$) can be calculated by extending the first sum over all atoms in the structure ($N_r = N_s$) (Nabuurs et al., 2003).

With this definition for structural uncertainty, the information contained in a set of experimental restraints (I_{set}) is defined as the difference in structural uncertainty of the structure before ($H_{\mathrm{structure}|}$ 0) and after ($H_{\mathrm{structure}|}$ set) addition of the experimental dataset:

$$I_{\mathrm{set}} = H_{\mathrm{structure}|0} - H_{\mathrm{structure}|\mathrm{set}} \qquad \text{(Eq. 2)}$$

The relative information content of a subset of restraints is defined as $I_{\mathrm{subset}}/I_{\mathrm{set}}$ and expressed as a percentage:

$$I_{\mathrm{rel}} = \frac{I_{\mathrm{subset}}}{I_{\mathrm{set}}} \times 100\%. \qquad \text{(Eq. 3)}$$

Structure calculations

Structure calculations were performed in torsion angle space using the default simulated annealing algorithm implemented in the program CNS (Brünger et al., 1998). For each dataset 20 accepted structures were calculated with an NOE violation threshold of 0.5 Å. Subsequently, all structures were refined in explicit solvent, which was previously shown to significantly improve structural quality (Linge et al., 2003; Nabuurs et al., 2004). Structures were validated using the WHAT IF program (Vriend, 1990) and visualized using YASARA (http://www.yasara.org). The computation

time for the determination of the proposed *U*-factor (see below) requires the calculation and refinement of structure ensembles containing typically 20 structures for 4×5 datasets, which for a medium-sized protein can take up to a day on a single processor PC. However, the algorithm is highly parallelizable and can be efficiently run a dual processor machine or Linux cluster.

Results and discussion

Evaluating quality on a per residue basis

It has been shown that overall quality descriptors derived from experimental NOE data correlate poorly with structural quality indicators derived from the resulting ensembles (Doreleijers et al., 1999; Nederveen et al., 2005) (cf. Figure 1). As structural variability is known to be non-uniform for different regions of a molecule, we first evaluated all quality scores, information measures and structural indicators on a per residue basis, using the experimental data and structure ensembles of the RECOORD database (Nederveen et al., 2005). For each of the 500 experimental NMR datasets, we calculated the per-residue structural uncertainty (Equation 1) from the input data and correlated these scores with several per-residue quality indicators as determined by WHAT CHECK (Hooft et al., 1996). The Pearson's correlation coefficient between the average per-residue quality scores for the 20 members of each RECOORD ensemble and the per-residue structural uncertainty (Equation 1) is shown in Figure 2 for each of the 500 structural ensembles.

Although different correlation coefficients are obtained for the individual RECOORD entries as function of the different WHAT CHECK quality scores, the plots clearly show the overall trends. For example, Figure 2a shows that the per-residue structural uncertainty (H_{res}) typically correlates reasonably well ($\bar{r} = 0.6$) with the per-residue accessibility. This illustrates that the conformation of more exposed residues often cannot be determined very well by NMR spectroscopy, resulting in a higher structural uncertainty for these residues. In a similar fashion, residues with lower uncertainty values tend to have higher packing quality scores ($\bar{r} = -0.6$, cf. Figure 2b). In general, however, only a weak correlation can be observed

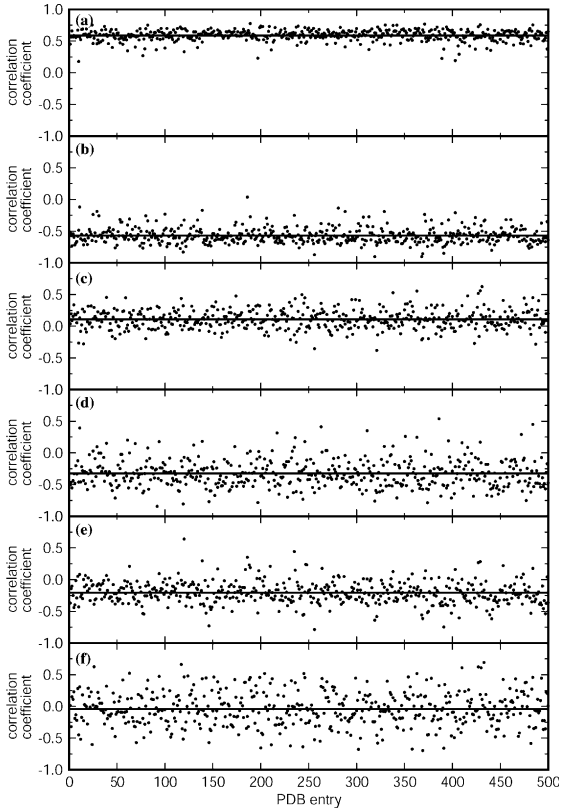

Figure 2. Per-residue Pearson's correlation coefficients (r) of the predicted structural uncertainty versus different ensemble averaged structural properties and quality indicators as determined by WHAT CHECK (Hooft et al., 1996). Shown are the correlation coefficients of structural uncertainty versus (a) accessibility ($\bar{r} = 0.6$), (b) packing quality ($\bar{r} = -0.6$), (c) side-chain rotamer normality ($\bar{r} = 0.1$), (d) backbone normality ($\bar{r} = -0.3$), (e) Ramachandran plot quality ($\bar{r} = -0.2$) and (f) the number of bumps ($\bar{r} = 0.0$) for each of the 500 entries from the RECOORD database (Nederveen et al., 2005). The average value of the correlation coefficient over the 500 entries is indicated with a solid line for each indicator.

between structural uncertainty and other quality indicators of a local nature, e.g., the Ramachandran plot Z-score ($\bar{r} = -0.2$, cf. Figure 2e) (Hooft et al., 1997) or the backbone normality score ($\bar{r} = -0.3$, cf. Figure 2d).

These findings also seem to support the counter-intuitive observation that there is no strong relation between the final per-residue structural uncertainty (H_{res}), and the quality of the corresponding residues in the resulting structure ensembles. From practice and automated structure determination approaches (Güntert, 2003) it is known, however, that structural quality does quite clearly improve as more

experimental information is added to the system and structural uncertainty decreases, which seems to contradict the findings as displayed in Figure 2. Therefore, we decided to monitor how structural quality evolves as more structural information is gradually introduced into a calculation, by calculating structure ensembles from subsets of the experimental data with a predetermined amount of experimental information.

Creating subsets of experimental data

We constructed multiple subsets of experimental data with an information content varying between 60 and 95%. The lower limit of 60% was chosen as datasets with lower information content typically did not yield native like structures (data not shown). Figure 3 shows for different subsets the number of included restraints as function of the information content using two different selection methods. For GB1, the five subsets with an identical information content constructed for each target value, consist of varying numbers of experimental restraints (see Figure 3a). These findings are in-line with our previous results that showed greatly varying information content for individual restraints (Nabuurs et al., 2003). The effect is even more pronounced for the smaller and less well-defined PSA dataset (cf. Figure 3b). Conversely, constructing subsets based on the number of included restraints, results in large variations in their information content (see Figures 3c and 3d). These findings again illustrate that the number of restraints is a poor indicator for the amount of available experimental information and hence subsets constructed on the basis of information content were used.

Validating subset quality

To assess how different structural properties depend on the amount of information contained in the different subsets, an ensemble of 20 structures was calculated for each subset. The structural variance, as judged by the heavy atom RMSD of the resulting structure ensembles, is shown in Figures 4a–b and e–f for the GB1 and PSA datasets, respectively. For comparison, results are shown as function of the number of included restraints (Figure 4a and e) and as function of the relative information content (Figure 4b and f). Figure 4 shows for both the GB1 and the PSA datasets a clear relation between the relative experimental information content (cf. Equation 3) and the RMSD of the resulting structure ensembles (see Figure 4b and 4f), in line with similar results obtained previously (Nabuurs et al., 2003). For the PSA dataset this relation becomes less clear if structural variability is analyzed as function of the number of restraints (cf. Figure 4e).

The structural variance of the resulting ensembles however, does not provide a clear and unbiased measure for the quality of the applied dataset (Spronk et al., 2003). To assess the quality of the resulting structures more rigorously, we determined the agreement of each structure ensemble with the restraints not included in the subset, a procedure commonly referred to as cross-validation. This is analogous to the complete cross-validation procedure described by Brünger et al. (1993). In this technique, however, the available NMR data is partitioned into multiple test sets with an equal number of restraints and cross-validation is performed with each of the test sets. Statistical quantities, e.g., the number and size of NOE violations, are then averaged over the different test sets. Thus, the differences in information content between the test sets is expected to average out, resulting in more meaningful cross-validated measures of fit for NMR datasets. In our present method however, all sets are constructed based on their information content instead of the number of restraints, rendering the use of many test sets in principle unnecessary.

The relation between the information content of the different subsets and the cross-validated RMS violations (cvRMS) of the NOEs not included in these subsets is shown in Figures 4c–d and g–h. Figure 4c and g show that the relation between subset size and the cvRMS of the NOE is similar to that observed for the structural variance. The information content, however, shows a near perfect linear relation with the cross-validated RMS of the NOE for the GB1 dataset (see Figure 4d; $r = -0.99$). For the PSA dataset, despite its largely unfolded nature, we also observe a strong correlation between information content and cvRMS, albeit with more scatter in the data points than for the GB1 dataset (Figure 4h; $r = -0.95$). These findings clearly demonstrate that the relative information

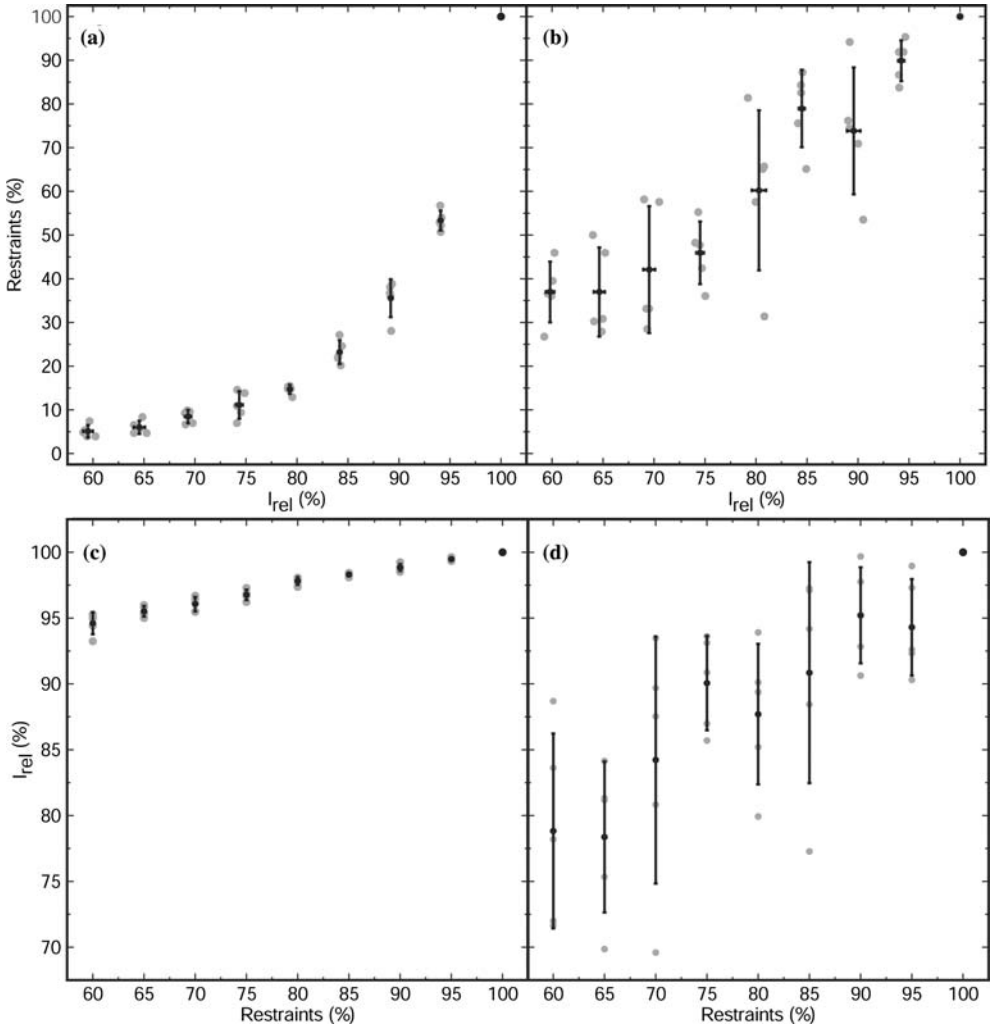

Figure 3. Information content of generated subsets versus the number of restraints included in these datasets. Subsets were constructed based on information content for both the GB1 (a) and PSA (b) datasets and based on the number of restraints for both the GB1 (c) and PSA (d) datasets. The scores for all individual sets are shown in grey; the average scores for each target value are indicated in black.

content within an experimental dataset is directly and linearly related to the accuracy of the resulting structure ensemble as judged by the cross-validated RMS of the NOEs.

Assessing the relation between data and structural quality

After establishing that the overall accuracy of the different structure ensembles correlates linearly with the information content of the data from which they were calculated, we subsequently evaluated the relation between information content and structural quality on a per-residue basis. The structural uncertainty of each individual residue was determined and converted to a relative per-residue information content score using Equations 1 and 2. Structural quality was assessed by the average per-residue Ramachandran plot quality *Z*-score for all twenty members of each ensemble. The Ramachandran plot quality was chosen over other quality indicators as it is a well-known, simple and sensitive parameter for assessing the quality of a protein model (Kleywegt and Jones, 1996; Hooft et al., 1997). Additionally, the per-residue Ramachandran quality score is, as it is directly related to the ϕ and ψ backbone torsion angles, readily translated into structural terms.

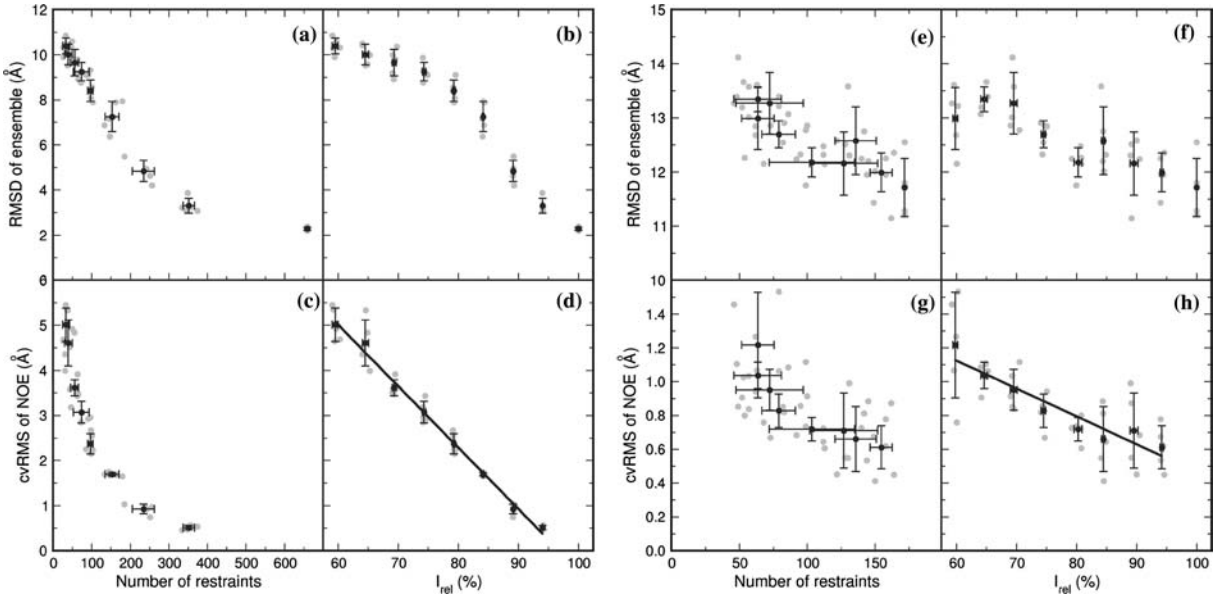

Figure 4. Relationship between dataset properties and structural properties of the resulting structure ensembles. The all heavy atom RMSD versus the number of restraints for GB1 (a) and PSA (e). The all heavy atom RMSD versus the relative information content (I_{rel}) for GB1 (b) and PSA (f). The cvRMS of the NOE versus the number of restraints for GB1 (c) and PSA (g). The cvRMS versus the relative information content (I_{rel}) for GB1 (d) and PSA (h). The scores for all individual sets are shown in grey; the average scores for each target value are indicated in black.

Figure 5 shows the relation between the per-residue relative information content and the resulting Ramachandran plot quality score for four different regions of the GB1 domain. From the different panels in Figure 5 it is evident that there is a clear correlation between the amount of experimental input data and the quality score when evaluated on a per-residue basis. The nature of this relation however, varies widely between residues and between different regions of the GB1 protein. The quality of the residues involved in regular secondary structure elements (as shown in Figures 5b and 5d) tends to improve as more information is introduced into the structure calculations, as expressed in the high average correlation coefficient of a linear fit to the data ($\bar{r} = 0.90$). Other residues, for example Asn 37 (cf. Figure 5c) or Asn 8 (cf. Figure 5a), do not exhibit any improvement or even decrease as more data is added. This very different behavior of individual residues in response to an increase in structural information provides a likely explanation to why we observed only very poor overall correlations between structural uncertainty and local quality indicators, such as the Ramachandran plot appearance (cf. Figure 2). More importantly, correlation plots like those presented in Figure 5 provide a means to establish the relationship between the structural quality of individual residues and the experimental input data. For example, for a Ramachandran plot outlier it can readily be assessed if it is consistently supported by the experimental data, and thus a genuinely interesting feature of the structure, or if it is likely a result of the low information content of the data.

Defining a new quality parameter

To make the information contained in Figure 5 more easily accessible we define a per-residue quality indicator that describes the correlation of the quality of that particular residue with the amount of experimental information known for that residue. For all per-residue correlation plots both the correlation coefficient of the information content with the Ramachandran plot appearance and the slope of the best linear fit to the average data points was determined. As the value of the correlation coefficient (r) expresses the degree to

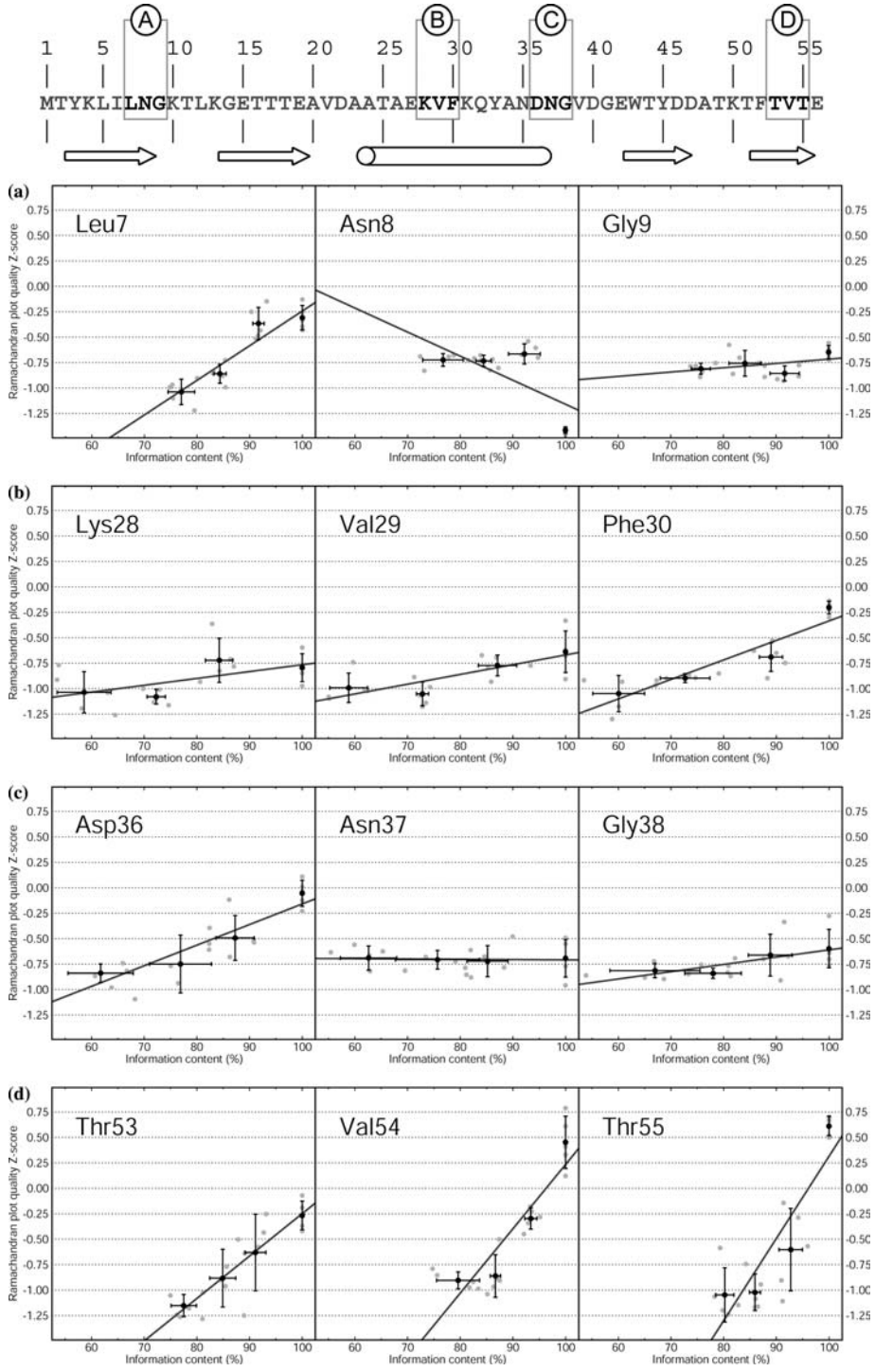

Figure 5. Relative per-residue information (I_{rel}) content versus Ramachandran plot quality *Z*-score, shown for four different regions of the GB1 domain: (a) residues 7–9, (b) residues 28–30, (c) residues 36–38 and (d) residues 53–55. Scores for the individual subsets are shown in grey, averages and standard deviations of the five subsets for each target value are shown in black. The best fit after linear regression of the four average target value scores is indicated in all panels.

which structural quality relates to the amount of information, and the slope (s) provides insight into the nature of this relation, we propose to combine these two measures into one quality indicator U:

$$U_i = C \cdot |r_i| \cdot s_i, \qquad \text{(Eq. 4)}$$

with U_i equal to the U-score for residue i, C an arbitrary scaling factor ($C = 100$) and r_i and s_i, the aforementioned correlation coefficient and slope for residue i, respectively. To limit computational requirements, only four subsets containing 70%, 80%, 90% and 100% of the total information content are used in the determination of the U-factor. Test results show that the inclusion of additional subsets does not lead to significant changes in the obtained values of the U-factor (data not shown).

Figure 6 shows U_i values as derived for five different experimental datasets. Figure 6a shows that the highest U_i values for the GB1 dataset are observed for the N and C-terminal β-strands. These two strands, located in the core of the GB1 domain, run in an anti-parallel fashion and NOEs linking the residues involved were previously found to be among the on average most important restraints in the GB1 dataset (Nabuurs et al., 2003). Of the residues in this dataset only Asn 8 exhibits a very negative U score. Although the correlation with the information content is not very high (see Figure 5a), its relatively low Ramachandran plot Z-score is confirmed by the X-ray structure of the GB1 domain (Gallagher et al., 1994).

As shown in Figure 6b, there is a striking difference in the UBI dataset between the U-scores for the N-terminal part of the α-helix (residues 23–29) and the C-terminal region (residues 30–36). Close examination of both NMR (Cornilescu et al., 1998) and X-ray (Vijay-Kumar et al., 1987) derived ubiquitin models reveals that Lys33 and Glu34 are likely to be responsible for this. The side-chains of both residues contact residues in an adjacent β-strand, with Lys33 hydrogen bonding to Thr14 and Glu34 forming a salt-bridge with Lys11. Formation of these favorable interactions requires both residues to be in a less favored region of the Ramachandran plot. As electrostatic interactions and side-chain hydrogen bond formation are usually not very well defined by the available experimental NMR data, but typically result from refinement in explicit solvent, this likely explains the weaker dependence on the amount of experimental data observed for this region.

In the graph for the PSA dataset (see Figure 6d), the folded part of the peptide is readily recognized by the single large region with positive U-scores. The sections with near-zero U_i values coincide with the unfolded regions of the PSA protein. In a similar fashion, the large flexible loops in the PDZ (Figure 6c, residues 34–46 (Walma et al., 2004)) and YB1 (Figure 6e, residues 42–54 (Kloks et al., 2002)) proteins can be recognized by subsequent residues with near-zero U-values. Additionally, the YB1 dataset has lower U-scores compared to the other datasets describing folded domains (GB1, UBI and PDZ). This reflects the relatively small amount of structural information available to determine the fold of this protein (Kloks et al., 2002; Nabuurs et al., 2003), and thus clearly demonstrates that the structural quality of this domain is to lesser extent determined by the experimental input data.

A legitimate question to ask at this point is what additional information the proposed U-factor conveys with respect to underlying Ramachandran plot quality score. To answer this question, the average per-residue Ramachandran plot quality scores of a structure ensemble calculated using the complete dataset are also shown in Figure 6. Comparison of the two quality indicators reveals that the U-factor does indeed convey additional information, not available from the Ramachandran plot quality score alone. As enclosed in its definition, the U-factor is closely related to the Ramachandran plot quality score. This is especially evident for the well-determined GB1, UBI and PDZ datasets (cf. Figure 6a–c). This relation is less clear for the less well-defined PSA and YBOX datasets (cf. Figure 6d–e), indicating that the quality of these structures is indeed less governed by the experimental data. This is most evident in the flexible regions of these proteins were near-zero U-factors clearly indicate that the sometimes observed high Ramachandran quality scores are actually more determined by the applied force field rather than by the experimental input data. However, also in the set of well-defined structures several residues can be identified, both in loops as in secondary structure elements, where the U-factor indicates that structural quality is to a

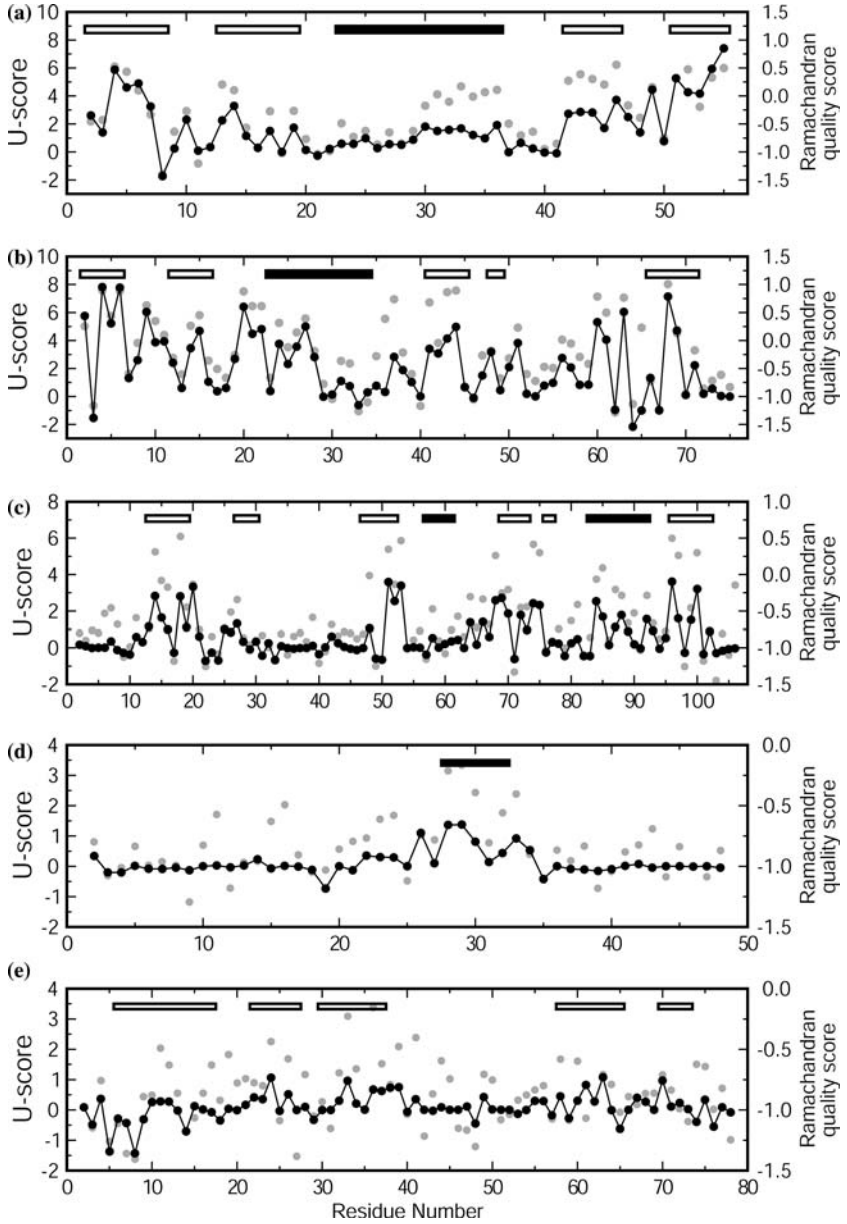

Figure 6. Per-residue *U*-scores (black filled circles and left *y*-axis) and Ramachandran quality Z-score (grey filled circles and right *y*-axis) for the (a) GB1, (b) UBI, (c) PDZ, (d) PSA and (e) YB1 dataset. Secondary structure is indicated by open (β-sheet) and filled (α-helix) boxes at the top of each panel.

lesser extent determined by the experimental data as to what might be expected from the Ramachandran plot quality score alone (e.g., residues 32–36 and 42–46 in the GB1 dataset).

In summary, the different vertical U_i scales in Figure 6 reflect the varying information content and quality of these experimental structures. Large U_i values correspond to a significant response to experimental data resulting from both a positive slope and a high correlation with the information content. Likewise, near zero values indicate that for these residues there is no apparent relation between structural quality and the amount of available experimental data. Residues with negative U_i values are certainly interesting, as these indicate a negative response to the available structural data, which

could either indicate the presence of structurally wrong information (e.g., misassigned NOEs) or structurally correct, but unusual, conformations.

Conclusions

Our results show that global quality parameters provide only limited insight into the quality of NMR structures. Therefore, validation of biomolecular NMR structures should ideally be performed on a per-residue basis, in agreement with the notion that most experimentally accessible NMR parameters are of a local nature. By constructing datasets with a pre-determined amount of experimental information, and deriving per-residue structural quality scores from the corresponding structure ensembles, we showed that there is a clear relation between these two parameters. However, the response of individual residues to an increase in structural information varies widely, both within a single dataset and between different datasets. Based on these results, we have defined a new information-based per-residue quality parameter: the *U*-factor. This indicator provides clear insight into the extent to which the structural quality of individual residues is governed by the experimental input data and provides a useful tool to evaluate and validate possible outliers identified by structure validation software. The software developed for calculating the described *U*-factors will be made available within the QUEEN software package, which can be downloaded at http://www.cmbi.ru.nl/software/queen/.

Acknowledgements

Financial support from the European community (5th Framework program NMRQUAL Contract Number QLG2-CT-2000-01313) to S.N., A.N. and E.K. and from BioRange to C.S. is gratefully acknowledged.

References

Banci, L., Bertini, I., Cavallaro, G., Giachetti, A., Luchinat, C. and Parigi, G. (2004) *J. Biomol. NMR*, **28**, 249–261.

Banci, L., Bertini, I., Savellini, G.G., Romagnoli, A., Turano, P., Cremonini, M.A., Luchinat, C. and Gray, H.B. (1997) *Proteins*, **29**, 68–76.

Brünger, 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 Cryst.*, **D54**, 905–921.

Brünger, A.T., Clore, G.M., Gronenborn, A.M., Saffrich, R. and Nilges, M. (1993) *Science*, **261**, 328–331.

Clore, G.M., Robien, M.A. and Gronenborn, A.M. (1993) *J. Mol. Biol.*, **231**, 82–102.

Clore, G.M., Starich, M.R., Bewley, C.A., Cai, M.L. and Kuszewski, J. (1999) *J. Am. Chem. Soc.*, **121**, 6513–6514.

Cornilescu, G., Marquardt, J.L., Ottiger, M. and Bax, A. (1998) *J. Am. Chem. Soc.*, **120**, 6836–6837.

Doreleijers, J.F., Mading, S., Maziuk, D., Sojourner, K., Yin, L., Zhu, J., Markley, J.L. and Ulrich, E.L. (2003) *J. Biomol. NMR*, **26**, 139–146.

Doreleijers, J.F., Raves, M.L., Rullmann, T. and Kaptein, R. (1999) *J. Biomol. NMR*, **14**, 123–132.

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

Gallagher, T., Alexander, P., Bryan, P. and Gilliland, G.L. (1994) *Biochemistry*, **33**, 4721–4729.

Garrett, D.S., Kuszewski, J., Hancock, T.J., Lodi, P.J., Vuister, G.W., Gronenborn, A.M. and Clore, G.M. (1994) *J. Magn. Reson. B*, **104**, 99–103.

Goodfellow, B.J., Dias, J.S., Ferreira, G.C., Henklein, P., Wray, V. and Macedo, A.L. (2001) *FEBS Lett.*, **505**, 325–331.

Güntert, P. (2003) *Prog. NMR Spec.*, **43**, 105–125.

Hooft, R.W., Sander, C. and Vriend, G. (1997) *Comp. Appl. Biosci.*, **13**, 425–330.

Hooft, R.W., Vriend, G., Sander, C. and Abola, E.E. (1996) *Nature*, **381**, 272 .

Kim, Y. and Prestegard, J.H. (1990) *Proteins*, **8**, 377–385.

Kleywegt, G.J. and Jones, T.A. (1995) *Structure*, **3**, 535–540.

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

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

Kloks, C.P., Spronk, C.A., Lasonder, E., Hoffmann, A., Vuister, G.W., Grzesiek, S. and Hilbers, C.W. (2002) *J. Mol. Biol.*, **316**, 317–326.

Kuszewski, J., Gronenborn, A.M. and Clore, G.M. (1995a) *J. Magn. Reson. B*, **107**, 293–297.

Kuszewski, J., Gronenborn, A.M. and Clore, G.M. (1999) *J. Am. Chem. Soc.*, **121**, 2337–2338.

Kuszewski, J., Qin, J., Gronenborn, A.M. and Clore, G.M. (1995b) *J. Magn. Reson. B*, **106**, 92–96.

Linge, J.P. and Nilges, M. (1999) *J. Biomol. NMR*, **13**, 51–59.

Linge, J.P., Williams, M.A., Spronk, C.A., Bonvin, A.M. and Nilges, M. (2003) *Proteins*, **50**, 496–506.

Liu, Y., Zhao, D., Altman, R. and Jardetzky, O. (1992) *J. Biomol. NMR*, **2**, 373–388.

Mierke, D.F., Huber, T. and Kessler, H. (1994) *J. Comput. Aided Mol. Des.*, **8**, 29–40.

Nabuurs, S.B., Nederveen, A.J., Vranken, W., Doreleijers, J.F., Bonvin, A.M., Vuister, G.W., Vriend, G. and Spronk, C.A. (2004) *Proteins*, **55**, 483–486.

Nabuurs, S.B., Spronk, C.A., Krieger, E., Maassen, H., Vriend, G. and Vuister, G.W. (2003) *J. Am. Chem. Soc.*, **125**, 12026–12034.

Nederveen, A.J., Doreleijers, J.F., Vranken, W., Miller, Z., Spronk, C.A., Nabuurs, S.B., Guntert, P., Livny, M., Markley, J.L., Nilges, M., Ulrich, E.L., Kaptein, R. and Bonvin, A.M. (2005) *Proteins*, **59**, 662–672.

Oshiro, C.M., Thomason, J. and Kuntz, I.D. (1991) *Biopolymers*, **31**, 1049–1064.

Snyder, D.A., Bhattacharya, A., Huang, Y.J. and Montelione, G.T. (2005) *Proteins*, **59**, 655–661.
Spronk, C.A., Linge, J.P., Hilbers, C.W. and Vuister, G.W. (2002) *J. Biomol. NMR*, **22**, 281–289.
Spronk, C.A., Nabuurs, S.B., Bonvin, A.M., Krieger, E., Vuister, G.W. and Vriend, G. (2003) *J. Biomol. NMR*, **25**, 225–234.
Spronk, C.A., Nabuurs, S.B., Krieger, E., Vriend, G. and Vuister, G.W. (2004) *Prog. Nucl. Magn. Reson. Spectrosc*, **45**, 315–337.
Tjandra, N., Garrett, D.S., Gronenborn, A.M., Bax, A. and Clore, G.M. (1997a) *Nat. Struct. Biol.*, **4**, 443–449.
Tjandra, N., Marquardt, J. and Clore, G.M. (2000) *J. Magn. Reson.*, **142**, 393–396.
Tjandra, N., Omichinski, J.G., Gronenborn, A.M., Clore, G.M. and Bax, A. (1997b) *Nat. Struct. Biol.*, **4**, 732–738.
Vijay-Kumar, S., Bugg, C.E. and Cook, W.J. (1987) *J. Mol. Biol.*, **194**, 531–544.
Vriend, G. (1990) *J. Mol. Graph*, **8**52–56.
Walma, T., Aelen, J., Nabuurs, S.B., Oostendorp, M., van den Berk, L., Hendriks, W. and Vuister, G.W. (2004) *Structure (Camb)*, **12**, 11–20.
Zhao, D. and Jardetzky, O. (1994) *J. Mol. Biol.*, **239**, 601–607.