Journal of Biomolecular NMR (2007) 37:15–30
DOI 10.1007/s10858-006-9096-8

Article

AUREMOL-RFAC-3D, combination of R-factors and their use for automated quality assessment of protein solution structures

Wolfram Gronwald[a], Konrad Brunner[a], Renate Kirchhöfer[a], Jochen Trenner[a], Klaus-Peter Neidig[b] & Hans Robert Kalbitzer[a,*]

[a]*Department of Biophysics and Physical Biochemistry, University of Regensburg, Universitätsstr.31, D-93040, Regensburg, Federal Republic of Germany;* [b]*Bruker BioSpin GmbH, Software Department, Silberstreifen 4, D-76287, Rheinstetten, Federal Republic of Germany*

Received 18 May 2006; Accepted 20 September 2006

Key words: 3D-NOESY-HSQC, quality assessment, R-factor

Abstract

We present here the computer program AUREMOL-RFAC-3D that is a generalization of the previously published program RFAC for the fully automated estimation of residual indices (R-factors) from 2D NOESY spectra. It is part of the larger AUREMOL software package (www.auremol.de). RFAC-3D calculates R-factors directly from two-dimensional homonuclear NOESY spectra as well as from three-dimensional ^{15}N or ^{13}C edited NOESY-HSQC spectra and thus extends the application range to larger proteins. The fully automated method includes automated peak picking and integration, a Bayesian noise and artifact recognition and the use of the complete relaxation matrix formalism. To enhance the reliability of the calculated R-factors the method is also generalized to calculate combined R-factors from a set of 2D and 3D-spectra. For an optimal combination of the information derived from different sources a plausible formalism had to be derived. In addition, we present a novel direct R-factors based measure that correlates an R-factors as defined in this paper to the root mean square deviation of the actual structure from the optimal structure. The new program has been successfully tested on the histidine containing phosphocarrier protein (HPr) from *Staphylococcus carnosus* and on the Ras-binding domain (RBD) of the Ral guanine-nucleotide dissociation stimulation factor (RalGDS).

Abbreviations: HPr – histidine containing phosphocarrier protein; RalGDS-RBD – Ral guanine-nucleotide dissociation stimulation factor; RBD – Ras-binding domain; HSQC – heteronuclear single quantum coherence; NOE – nuclear Overhauser effect; NOESY – nuclear Overhauser effect spectroscopy; rmsd – root mean square deviation.

Introduction

One of the most important points in any automated or manual structure determination process of proteins in solution is the assessment of the quality of the resulting structures. As a final aim, one wants to know if the solved structure really reflects the true structure present in the natural environment of the molecule of interest. The overall precision of a NMR structure is usually expressed either as an average pair wise root-mean-square deviation (RMSD) of the coordinates of the selected ensemble of structures or as an RMSD of the structures relative to the mean coordinates of the ensemble. However, RMSD

*To whom correspondence should be addressed. E-mail: hans-robert.kalbitzer@biologie.uni-regensburg.de

values are a measure of the precision of the structures in the ensemble but not necessarily for their accuracy. Another mean to analyze a structure is the quality of the geometrical properties of the molecule, e.g. the comparison of bond lengths, bond angles, dihedral angles etc., with standard values obtained for example from a set of high resolution structures (Laskowski et al., 1996). Alternatively, using a set of previously solved three-dimensional structures one can compute a force field consisting of potentials of mean force. In this way, energy potentials for the atomic interactions between the various amino acid pairs are derived as a function of the distance between the involved atoms. Employing such a force field one can compute energy graphs for a given structure to identify problematic regions as is done within PROSA II (Sippl, 1993).

A better measure for the quality of a NMR structure includes information how well the obtained structures agree with the experimental data. Often the overall quality of the experimental data itself is simply judged by the number of restraints per residue. Furthermore, the number and sizes of violated restraints, such as distance, dihedral angle, hydrogen bond, and residual dipolar coupling restraints are analyzed in a qualitative or quantitative way. R-factors (residual factor) are used in crystallography (Brünger et al., 1987) for the quantification of the agreement of the experimental data with the calculated structure. A similar measure can be defined for NMR data where experimental NOESY spectra and NOESY spectra back-calculated from a trial structure are compared. In the literature different definitions for NMR-R-factors can be found (Lefevre et al., 1987; Gupta et al., 1988; Nikonowicz et al., 1990; Lane, 1990; Baleja et al., 1990; Borgias et al., 1990; Borgias and James, 1990; Gonzalez et al., 1991; Nilges et al., 1991; Bonvin et al., 1991; Thomas et al., 1991; Mertz et al., 1991; Brünger et al., 1993; Clore et al., 1993; Xu et al., 1995; Cullinan et al., 1996). Most R-factors are calculated from manually edited distance or peak volume lists, however, they can also derived in an automated way (Gronwald et al., 2000).

Another important point in R-factor calculation is the consideration of unassigned or not assignable experimental cross peaks. We defined earlier (Gronwald et., 2000) that true experimental NOE cross peaks should be assigned to the class of unassigned signals (*U*-list) when in the trial structure the corresponding protons are further apart than an user defined distance cutoff (e.g. 0.6 nm). This class of unassigned signals is then handled differently. The idea of using the presence or absence of simulated and experimental signals was later also employed in the paper by Huang et al. (2005) for the calculation of so called RPF scores.

In a previous publication (Gronwald et al., 2000), we have described a method for automated calculation of NMR-R-factors based on the use of ^{1}H 2D NOESY spectra that is well suited for the investigation of smaller molecules. However, for larger proteins the use of 2D spectra becomes usually prohibitive for NMR R-factor calculations due to the presence of extensive overlap in the experimental spectra. As a consequence we have extended our routines for the automated NMR R-factor calculation to the use of ^{15}N or ^{13}C edited 3D NOESY-HSQC spectra. In principle, the validation of a structure should include the set of all relevant experimental data, e.g. all NOESY-type spectra used in the structure determination. The appropriate weighting of different spectra during the R-factor calculation is a non-trivial problem, which also has to be solved in this context.

One important point in any NMR R-factor calculation that has not been solved satisfactorily yet is the interpretation of the R-factor in geometric terms. As a consequence we present in this paper a clear relationship between R-factors defined by us and rmsd values.

Our approach was tested on two medium size proteins namely the histidine containing protein (HPr) from *Staphylococcus carnosus* and the Ras-binding domain (RBD) of the Ral guanine-nucleotide dissociation stimulation factor (RalGDS) from *human*. The protein HPr is 88 residues in size and its structure consists of three α-helices and a four stranded anti parallel β-sheet (Görler et al., 1999) while RalGDS-RBD is 87 residues in size and displays the ubiquitin super-fold.

Materials and methods

NMR-samples

For HPr the three-dimensional ^{15}N edited NOESY-HSQC spectrum used was recorded from a sample of 3.1 mM ^{15}N-labelled HPr from *Staphylococcus carnosus* in 90% H_2O/10% D_2O (v/v), pH 7.2 while for the homonuclear ^{1}H 2D NOESY spectrum a sample containing 4.3 mM unlabelled HPr in 90% H_2O/10% D_2O (v/v), pH 7.2 was used. For RalGDS-RBD the corresponding spectra were measured from ^{15}N-labeled and unlabeled samples of 1.0 mM RalGDS-RBD in 90% H_2O/10% D_2O (v/v), pH 7.0.

NMR spectroscopy

For HPr the three-dimensional NOESY-HSQC spectrum and the 2D spectrum were recorded at 500 MHz employing a mixing time of 100 ms and at 800 MHz with a mixing time of 150 ms, respectively. The 3D and 2D spectra were acquired using relaxation delay (time between the last $\pi/2$-pulse of the present and the first $\pi/2$-pulse of the following NOESY sequence) of 1.1 and 2.37 s and 224 × 112 × 1024 and 1024 × 8192 time domain data points, respectively. For RalGDS-RBD the three-dimensional NOESY-HSQC spectrum and the 2D spectrum were recorded at 600 MHz employing a mixing time of 100 ms and at 800 MHz with a mixing time of 80 ms, respectively. The 3D and 2D spectra were acquired using relaxation delays of 1.59 and 1.56 s and 128 × 64 × 2048 and 512 × 4096 time domain data points, respectively. All spectra for both proteins were measured at 298 K.

Three-dimensional structure

The corresponding three-dimensional solution structure of HPr from *S. carnosus* (Görler et al., 1999) was taken from the set of structures submitted to the PDB, accession code 1QR5. The overall quality of the structure was further improved by subjecting it to refinement in explicit solvent (Nabuurs et al., 2004) using the same set of structural restraints used before (Linge et al., 2003). The solution structure of *human* RalGDS-RBD (residues 1–97, corresponding to residues 788–884 of the full length protein, Swiss prot accession code: Q12967) has previously been published by Geyer et al. (1997). For the current tests the structure of a shorter construct (amino acid 11 to 97) has been recalculated from the NMR data. Using AUREMOL (Gronwald and Kalbitzer, 2004) the NOESY spectra were automatically peak-picked employing locally adapted thresholds. Separation of artifacts and noise from true signals was accomplished using a Bayesian analysis implemented in AUREMOL (Antz et al., 1995; Schulte et al., 1997). In the next step, signals are integrated using iterative segmentation (Geyer et al., 1995). NOE signals were completely automatically assigned using the KNOWNOE approach (Gronwald et al., 2002) implemented in AUREMOL. KNOWNOE contains as a central part a knowledge driven Bayesian algorithm for solving ambiguities in NOE assignments arising for example from chemical shift degeneracy. In contrast to other known assignment algorithms KNOWNOE uses volume probability distributions obtained from a large number of already solved structures together with the individual cross peak volumes to calculate the most probable assignments. Accurate distance constraints were obtained from the NOE spectra by using the full relaxation matrix approach embedded in the AUREMOL module RELAX (Görler and Kalbitzer, 1997; Görler et al., 1999; Ried et al., 2004)/ REFINE (to be published), taking also experimentally determined order parameters S^2 and correlation times into account. Appropriate individual error bounds were obtained by a local analysis of noise levels and signal overlap and expected order parameter variations. Six cycles of iterative NOE assignments and structure calculations with CNS 1.1 (Brünger et al., 1998) using as input an extended strand starting structure and the sequential resonance assignment were performed to obtain the final solution structures. These structures were also subjected to refinement in explicit solvent. The structure together with the resonance line assignment has been deposited in the protein data bank with the accession number 2B3A.

Software

The NMR data were processed with the program XWINNMR® (Bruker). All other routines required for R-factor calculation are contained in the program AUREMOL (www.auremol.de). AUREMOL is written in ANSI C and a compiled version for PCs running under Microsoft WINDOWS-NT® or higher can be obtained from the above web page. Molecular dynamics simulations were performed with the program CNS (Brünger et al., 1998).

Theoretical considerations and algorithms

In general, the R-factor (residual index) should measure the agreement between the experimental data set and the data back calculated from the structure. The automated NMR R-factor analysis using three-dimensional spectra follows in principle the strategy described previously (Gronwald et al., 2000). Therefore, we will concentrate in the following mainly on new extensions to that strategy. All routines required are included in the newly developed software package AUREMOL (Gronwald and Kalbitzer, 2004). The methods developed for 2D-NOESY spectra have (1) to be extended to 3D-NOESY-HSQC spectra and (2) an appropriate weighting algorithm has to be defined for combining different types of 2D- and 3D-NOESY spectra in a common R-factor.

Calculation of R-factors of 3D-spectra

As in the case of 2D-spectra in the first part, a three-dimensional NOESY spectrum has to be back-calculated from the three-dimensional trial structure using the resonance line assignments. In our implementation we are employing the full relaxation matrix approach using the AUREMOL module RELAX (Görler and Kalbitzer, 1997; Görler et al., 1999; Ried et al., 2004) for back-calculating a 3D NOESY spectrum which gives a list of back calculated peaks (*B*-list) defined by their positions and intensities (volumes). If required, in AUREMOL it is also possible to correct the back-calculated 3D NOESY intensities for losses experimentally observed in the HSQC part of the 3D NOESY-HSQC spectrum. The experimental three-dimensional NOESY spectrum is automatically peak picked and integrated. In addition, the probabilities p_i of the peaks i in the tree-dimensional spectrum to be true NMR signals and not noise or artifact peaks are calculated according to Bayes theorem with a new routine of AUREMOL (to be published), a generalization of the method already published for 2D-spectra (Antz et al., 1995; Schulte et al., 1997). The probability values p_i provide a measure how reliable the peaks i are. They are used as weighting factors during the calculation of the R-factors. The resulting list of unassigned experimental peaks (*U*-list) consists of the peak positions, the volumes V_i and the probability values p_i. Using the backcalculated *B*-signals the yet unassigned experimental signals are automatically assigned with the new AUREMOL module PeakAssign (to be published). In the first step, the program optimally adapts the chemical shift values obtained from the general sequential resonance assignment to the actual experimental data. In the second step, the peak assignment itself is done on local peak clusters. For each back calculated peak a search is performed if a corresponding experimental peak exists in a given search radius. The assigned experimental peaks are assembled in the *A*-list. In PeakAssign, the obtained assignment is dependent on the structure under investigation since only cross peaks are considered where a significant volume is back calculated (i.e. where the contributing atoms are within a given distance limit). In case that more than one signal is back calculated at the same position it can be assumed that the corresponding experimental signal also consists of the sum of several NOEs. Therefore, different to our earlier implementation (Gronwald et al., 2000), in this case the volumes of the overlapping simulated signals will be summed before R-factor calculation.

The remaining *U*-list contains two types of cross peaks, cross peaks where in principle an assignment is possible since its two (2D-NOESY) or three (3D-NOESY) frequency coordinates correspond to known resonance frequencies (U_1-list) and peaks where not all resonance frequencies correspond to known resonance frequencies (U_2-list). For the signals from the U_1-list an experimental peak exists but no corresponding simulated signal was calculated since the distance between the contributing atoms in the trial structure exceeds the threshold value where the corresponding simulated signal volume is in good approximation equal to zero. Therefore, the experimental signals from the U_1-list are strong indicators of structural problems. The U_2-list contains peaks where at least one of the spins contributing to an experimental cross peak has not yet been assigned (incomplete resonance line assignments) and peaks which are either pure spectral artifacts, signals from other compounds (e.g. buffer) or do not represent the main conformation of the protein. The *B*-, *A*-, and U_1-signals are automatically read in by the AUREMOL module RFAC-3D. We have proposed previously a number of different R-factors (Gronwald et al., 2000). Here, two of our previously published R-factors are shown, the PWAR (probability-weighted assigned resonances based

R-factor) and the PWAUR R-factor (probability-weighted assigned and unassigned resonances based R-factor) corresponding to the earlier published R-factors R_3 and R_5, respectively (Gronwald et al., 2000). They are defined as:

$$R_{\mathrm{PWAR}}(\alpha) = \sqrt{\frac{\sum\limits_{i \in A} (sf_\alpha V^{\alpha}_{\mathrm{exp},i} - V^{\alpha}_{\mathrm{calc},i})^2 p^2_{\mathrm{exp},i}}{\sum\limits_{i \in A} sf^2_\alpha V^{2\alpha}_{\mathrm{exp},i} p^2_{\mathrm{exp},i}}} \quad (1)$$

and

$$R_{\mathrm{PWAUR}}(\alpha) = \sqrt{\frac{\sum\limits_{i \in A} (sf_\alpha V^{\alpha}_{\mathrm{exp},i} - V^{\alpha}_{\mathrm{calc},i})^2 p^2_{\mathrm{exp},i} + \sum\limits_{i \in U_1} (sf_\alpha V^{\alpha}_{\mathrm{exp},i} - V^{\alpha}_{\mathrm{noise}})^2 p^2_{\mathrm{exp},i}}{\sum\limits_{i \in A} sf^2_\alpha V^{2\alpha}_{\mathrm{exp},i} p^2_{\mathrm{exp},i} + \sum\limits_{i \in U_1} (sf_\alpha V^{\alpha}_{\mathrm{exp},i} - V^{\alpha}_{\mathrm{noise}})^2 p^2_{\mathrm{exp},i}}} \quad (2)$$

Here, sf_α is a scaling factor, $V_{\mathrm{exp},i}$ is the volume of the cross peak i in the experimental spectrum, $V_{\mathrm{calc},i}$ the sum of back calculated volumes corresponding to cross peak i, $p_{\mathrm{exp},i}$ the signal probability of the experimental peak i and α an exponent (usually set to $-1/6$). In Equation 2, we additionally take the experimental signals of the list U_1 into account. As explained above these are signals where no corresponding simulated signal is available. In Equation 2, all signals of the list U_1 are compared to a standard noise volume V_{noise} that corresponds to the volume of an experimental signal at the detection limit to substitute for the missing corresponding simulated volumes. By also using V_{noise} in the denominator it is ensured that R_{PWAUR} approaches a value of 1 for completely erroneous structures. For this contribution, a new R-factor definition has been developed where additionally the signals of the list U_1 are now assigned. These assignments are based on chemical shifts and in case of ambiguity the assignment corresponding to the shortest distance in the given structure is chosen. Since the corresponding simulated volumes are all approximately zero, the matching distances $d_{\mathrm{PDB},i} = V^{-1/6}_{\mathrm{sim},i}$ are taken instead from the trial structures and are used directly in R_{PWFAR} (probability-weighted full resonance assignment based R-factor) as defined in Equation 3. As a consequence R_{PWFAR} is in some respect similar to R_{PWAR} (Equation 1) using an infinite distance cutoff value for signal simulation. Additionally R_{PWFAR} is not limited to maximum values of 1 but might adopt significantly larger values depending on the size and quality of the used structure. In R_{PWFAR} signals of the U_1-list that correspond to large distances in the given trial structure will therefore have a significant impact on the resulting R-factor. In comparison, in R_{PWAUR} (Equation 2) the influence of these signals is limited by the application of the V_{noise} term

$$R_{\mathrm{PWFAR}}(\alpha) = \sqrt{\frac{\sum\limits_{i \in A} (sf_\alpha V^{\alpha}_{\mathrm{exp},i} - V^{\alpha}_{\mathrm{calc},i})^2 p^2_{\mathrm{exp},i} + \sum\limits_{i \in U_1} (sf_\alpha V^{\alpha}_{\mathrm{exp},i} - d^{-6\alpha}_{\mathrm{PDB},i})^2 p^2_{\mathrm{exp},i}}{\sum\limits_{i \in A} sf^2_\alpha V^{2\alpha}_{\mathrm{exp},i} p^2_{\mathrm{exp},i} + \sum\limits_{i \in U_1} sf^2_\alpha V^{2\alpha}_{\mathrm{exp},i} p^2_{\mathrm{exp},i}}} \quad (3)$$

The R-factor definitions given by Equations 1–3 provide a measure how well the experimental signals are explained by a given test structure. It is also possible to define an R-factor that additionally analyzes how well the simulated signals are explained by the given experimental spectrum (Equation 4). This R-factor definition has the advantage of the large statistical basis used. Therefore, the influence of a single signal with an erroneous volume due to for example base-line roll should be minimal

The last term in R_{PWFASR} (probability-weighted full resonance assignment plus simulated signals based R-factor) takes the simulated signals into account for which no corresponding experimental signals were available (U_{C}-list). Please note that in principle for each simulated signal an experimental signal should be present; however, due to a finite signal to noise ratio in the experimental spectra not

$$R_{\mathrm{PWFASR}}(\alpha) = \sqrt{\frac{\sum_{i\in A} (sf_\alpha V^{\alpha}_{\mathrm{exp},i} - V^{\alpha}_{\mathrm{calc},i})^2 p^2_{\mathrm{exp},i} + \sum_{i\in U_1} (sf_\alpha V^{\alpha}_{\mathrm{exp},i} - d^{-6\alpha}_{\mathrm{PDB},i})^2 p^2_{\mathrm{exp},i} + \sum_{i\in U_c} \left(V^{\alpha}_{\mathrm{cal},i} - V^{\alpha}_{\mathrm{noise\_c}}\right)^2}{\sum_{i\in A} sf^2_\alpha V^{2\alpha}_{\mathrm{exp},i} p^2_{\mathrm{exp},i} + \sum_{i\in U_1} sf^2_\alpha V^{2\alpha}_{\mathrm{exp},i} p^2_{\mathrm{exp},i} + \sum_{i\in U_c} V^{2\alpha}_{\mathrm{cal},i}}} \tag{4}$$

all of these can be detected. As a consequence we only use signals of the U_C-list where the simulated volumes are above the detection limit of the corresponding experimental spectrum. In this context, the term $V^{-1/6}_{\mathrm{noise\_c}}$ specifies the distance limit above which the fraction of the number of experimental signals to the number of simulated signals of the same distance class substantially decreases.

In NMR spectroscopy and X-ray crystallography, one has to normalize the experimental (or calculated) data by a scaling factor *sf*. While in the previous versions of this R-factor the scaling procedure was applied to the calculated signals it is now performed at the experimental signals of class *A* to allow a comparable normalization of different data sets (see below). As a consequence the calculation of the definition of scaling factor sf_α has to be changed as well. We also include the probabilities $p_{\mathrm{exp},i}$ in the calculation of the scale factor (Equation 5). This should help to diminish the influence of artifacts present in the experimental spectra on the obtained scale factor

$$sf_\alpha = \frac{\sum_{i\in A} p_{\mathrm{exp},i}(V_{\mathrm{exp},i} V_{\mathrm{calc},i})^{\alpha}}{\sum_{i\in A} p_{\mathrm{exp},i} V^{2\alpha}_{\mathrm{exp},i}} \tag{5}$$

Definition of a combined R-factor

One important issue that arises when calculating R-factors from a set of spectra, like for example from two 2D spectra measured in H_2O and D_2O, respectively, and a ^{15}N edited NOESY-HSQC spectrum is how to combine the corresponding R-factors into one value that allows to judge the obtained three-dimensional structure. From the theoretical point of view this is not trivial since the scaling of the different spectra and their information content is usually different and has to enter implicitly into the definition of the combined R-factor. In general the combined R-factor R_{comb} should present a reliable measure how well all available experimental data are explained by the proposed structure(s). A general theory for the derivation of a combined R-factor has not been proposed yet in the literature, however, one can find at least some plausible criteria which have to be fulfilled by R_{comb}. It is clear that the simple averaging of the M R-factors calculated selectively from the M spectra available does not provide a meaningful solution of the problem since a spectrum containing only 1 signal would influence the R-factor equally as a complete NOESY spectrum. From that it is evident that in some way the number of events (signals) has to influence the weight of the different R-factors R_j of the spectra *j*. In addition, a different scaling of the experimental spectra (e.g. due to different receiver gains) should not influence the calculation of the combined R-factor. When only NOESY-type spectra are combined, the influence of the experimental conditions can be partly removed when the experimental data are scaled to the back calculated NOESY data as defined in Equation 5. In case of a combination of two- and three-dimensional data one has to define for the back calculation a common intensity basis that is one has to use in the 3D-spectra volumes reduced to the three-dimensional space (the volumes in a 3D-spectrum are calculated in a four-dimensional space since the intensity is the fourth coordinate). This means that in the simplest case only the corresponding two-dimensional spectrum is calculated (as it is done usually). In the more general case where the HSQC-transfer efficiency varies for different cross peaks one can correct the volumes by the transfer efficiencies with full transfer efficiency set to 1.

A plausible criterion for the calculation of a combined R-factor can be derived from the Gedanken experiment that two regions of a normal 2D or 3D spectrum are recorded in two different experiments with different receiver gains. Here, it is evident that the combined R-factor should be equal to the R-factor obtained when the complete spectrum has been recorded and used for the calculation. With the assumption that the scaling factors calculated with Equation 5 provide sufficiently good approximations of true scaling factors (obtained when the number of peaks is very large), the combined R-factor R_{comb} is given by

$$R_{\mathrm{comb}} = \sqrt{\frac{1}{\sum_{j=1}^{M}\sum_{i=1}^{N_j} sf_j^2 p_i^2 V_{\exp,i}^{2\alpha}} \sum_{j=1}^{M} R_j^2 \sum_{i=1}^{N_j} sf_j^2 p_i^2 V_{\exp,i}^{2\alpha}} \tag{6}$$

where each contributing R-factor R_j is weighted according to the experimental volumes $V_{\exp,i}^{\alpha}$ scaled with the scaling factors sf_j together with their peak probability values p_i of the N_j peaks in the spectra j.

Dependence of the R-factor on the accuracy of the 3D-structure

The R-factor should be a measure for the quality of a structure that is how closely the actual structure represents the true protein structure. The "true" structure (or better the set $\mathbf{S}_0$ of true structures S_0^{i}) is generally not known, but only the set $\mathbf{S}_1$ of the structures S_1^{i} optimally fulfilling the given set of experimental restraints $\mathbf{R}_{\mathrm{e}}$ and of restraints from *a priori* knowledge $\mathbf{R}_{\mathrm{p}}$. When we define a measure F (a metric) for the distance of the actual structure $S_{\mathrm{j}}^{\mathrm{i}}$ from the set $\mathbf{S}_0$ than the R-factor ideally should increase when the value of F increases. Under ideal conditions, that is the experimental restraints are error-free and the optimization procedure (in our case the simulated annealing procedure) finds only solutions that optimally fulfill the restraints, the structures of set $\mathbf{S_0}$ should be contained in set $\mathbf{S}_1$. Sets of structures with larger mean values of F should be obtained when the set of experimental restraints $\mathbf{R}_{\mathrm{e}}$ is reduced. An operative way to obtain structures S^{i} which larger $\mathrm{F}(\mathbf{S}_0, S^{\mathrm{i}})$ would be an unrestrained molecular dynamics run *in vacuo* starting with structures of set S_1 and selecting structures with increasing distance $F(\mathbf{S}_1, S^{\mathrm{i}})$. Although this procedure does not grant that the condition $F(\mathbf{S}_0, S^{\mathrm{i}}) > F(\mathbf{S}_0, S_1^{\mathrm{i}})$ is fulfilled it is likely to hold for a single structure S^{i} and should hold for the first moment of the set of structures created in this way.

A possible way to determine the metric F would be the pairwise RMSD value of the Cartesian coordinates of the heavy atoms of the test structures to the set of target structures. Since this metric can be calculated easily, it will be used in this paper. However, it is clearly not the only possible definition of such a metric.

Results

Calculation of R-factors from 2D and 3D-NOESY spectra and their combination

2D-NOESY and 3D-NOESY-HSQC spectra from two small proteins, the Ras-binding domain of RalGDS-RBD and the HPr protein were used to calculate R-factors. Several different R-factor definitions are possible (see Gronwald et al., 2000). Exemplarily, the R-factor R_{pwaur} with $\alpha = -1/6$ defined by Equation 2 was used that is especially well-suited to judge the quality of the three-dimensional structure (see below). In addition a new R-factor R_{PWFAR} defined by Equation 3 is employed where the signals of the U_1-list are assigned based on chemical shifts and the corresponding distances $d_{\mathrm{PDB},i} = V_{\mathrm{calc},i}^{-1/6}$ are taken directly from the trial structure. The R-factor calculation method implemented in AUREMOL works directly on the spectra with a minimum interference of the user and includes automated peak picking, automated integration, automated signal and artifact recognition, and automated back calculation of the NOESY spectra. However, a few default parameters can be selected and modified. An important parameter is the signal probability. Here only peaks with a considerable probability p_i to be true signals with $p_i > 0.8$ were taken, although the inclusion of the probability reduces already automatically the effect of pure artifact peaks in the obtained R-factor. The default value of 0.55 nm was accepted as detection limit that determines the assignment to class A or U_1.

The obtained values are summarized in Table 1. As to be demanded the selection of the type of NOESY spectra used for the calculation does not influence significantly the R-factor calculation, although the number N of peaks contributing to the calculation differs considerably. This is also true for the combined R-factor.

Calculation of the R-factors and the influence of motional models

For each test protein a 3D ^{15}N edited NOESY-HSQC spectrum and a 2D ^{1}H NOESY spectrum were taken for the R-factor calculations. A strip transformation (11.82–5.66 ppm for HPr and 10.00–6.2 ppm for RalGDS-RBD) in the acquisition domain of the 3D spectra was performed to exclude signal free regions and the strong water artifacts. For reasons of comparison, a similar region was taken for the analysis in the 2D spectra. Signals were automatically identified with the routines integrated within AUREMOL. True protein signals were separated from noise and artifacts using Bayesian analysis and employing a Bayesian cutoff value of 0.8 for all spectra leading for HPr to 507 and 842 experimental signals in the 3D and 2D spectra, respectively. For RalGDS-RBD 450 signals were identified in the 3D spectrum while 1008 signals were obtained for the 2D data set. The difference in the number of identified signals between the 2D and 3D spectra reflects mainly the increased sensitivity of the 2D spectra measured at 800 MHz and the fact that the region corresponding to the aromatic signals is not present in the 3D spectra. As described in the algorithm section an automated structure based assignment was performed with these signals.

For the simulations necessary for the calculation of the R-factors the same parameters, e.g. relaxation delay between scans, mixing time etc., that were described for the corresponding experimental spectra were used. The global correlation times τ_c = 5.62 ns for HPr (Schubel et al., to be published) and τ_c = 6.66 ns for RalGDS-RBD (Döker et al., to be published) that were used in the simulations were obtained from relaxation measurements performed on uniformly ^{15}N enriched samples at 298 K. From the possible spectral densities as defined in Görler and Kalbitzer (1997) LIPARI_1, which is a simplification of the original spectral density defined by Lipari and Szabo (Lipari and Szabo, 1982a, b) was selected for all atom pairs not including a methyl group or an aromatic ring. For all atom pairs containing protons from a methyl group a fast-jump approximation was used for the spectral density. For atom pairs containing members from aromatic rings a slow jump approximation was made for the spectral density. For all atom pairs containing only backbone atoms an average order parameter S^2 of 0.95 has been experimentally determined for HPr (Schubel et al., to be published) while for the backbone atoms of RalGDS-RBD in regular secondary structure elements and loop regions S^2 values of 0.96 and 0.74 were determined, respectively. For all atom pairs containing side-chain and main-chain atoms an S^2 of 0.80 was used while for side-chain side-chain interaction an S^2 value of 0.65 was used. The latter two values were not experimentally determined but taken from the literature (Brünger, 1993). Also it is possible to automatically correct for deviations of the molecule from spherical shape. However, in case this option is activated, in the current version of RELAX the molecule is treated as a rigid body and since the three-dimensional structures of HPr and RalGDS-RBD can be approximated fairly

Table 1. R-factors calculated from 2D- and 3D-NOESY-spectra and their combination[a]

Protein	R-factor from						
		N^b	2D-NOESY	N^b	3D-NOESY-HSQC	N^b	Combined
RalGDS-RBD	R_{PWAUR}	291	0.26	117	0.30	408	0.27
	R_{PWFAR}	972	0.47	432	0.56	1404	0.50
HPr	R_{PWAUR}	195	0.31	74	0.32	269	0.31
	R_{PWFAR}	828	0.50	458	0.45	1286	0.48

[a] R-factors R_{PWAUR} and R_{PWFAR} ($\alpha = -1/6$) were calculated for the final solution structures of RalGDS-RBD and HPr. Only peaks with a signal probability $p_i > 0.8$ and with chemical shifts between 11.82 and 5.66 ppm in δ_2 dimension (2D-NOESY) and δ_3 dimension (3D-NOESY) were used. The detection limit defining the U_1-list was set to 0.55 nm. For R_{PWAUR} the A-list contains only signals from amino acids separated by more than four amino acids in the sequence, while for R_{PWFAR} all assigned signals of the A- and U_1-list were used. [b] Number of signals used for the calculation.

well with a sphere this option was not used for the tests shown.

There are many factors besides the quality of the obtained structure that can influence the absolute value of the R-factor with a given set of experimental restraints $\mathbf{R}_e$. Most important is the definition of the R-factor itself. In addition, the simulation of the spectra itself enters the calculation. Even with the use of the full relaxation matrix formalism the choice of the motional model may influence the result considerably. To test this influence on the calculated R-factors a detailed motional models was compared to an all rigid model with no internal motions. For this test the HPr protein was selected. Of the set of water refined HPr structures the lowest energy structure in terms of total energy was selected to automatically assign the 2D NOESY spectrum of HPr. Using the all rigid model and R_{PWAR} ($\alpha = -1/6$) (Equation 1) for example for the medium-range signals an R-factor of 0.28 was obtained, while when using detailed motional models the corresponding R-factor decreased to a value of 0.25 (Table 2). As to be expected, in general the more detailed motional model gives slightly smaller R-factors. However, quantitatively the difference is not very large.

Creation of a set of trial structures

A set of trial structures S^i with increasing distance $F(\mathbf{S}_1, S^i)$ to the initial structural set $\mathbf{S}_1$ was created for both proteins used in this study, the HPr protein and RalGDS-RBD by the simulated annealing procedure described above. The final lowest energy NMR structures in terms of total energy of these two proteins were subjected to 10000 5 fs steps of unrestrained molecular dynamics simulations *in vacuo* at room temperature to obtain trajectories of increasingly disordered structures where every 100 steps a 3D structure was saved. For this purpose again CNS 1.1 with the corresponding CNS force field was employed. The same standard simulated annealing protocol that was used for generating the original NMR trial structures was also used. The only exception was the removal of all experimental restraints and that the calculations were performed in Cartesian space instead of torsion angle space. Electrostatic terms were not used in decoy creation. The sets of resulting structures were ordered with respect of the corresponding rmsd values to the final solution structures. Of these structures for each test-protein 16 structures were selected for NMR R-factor calculation to cover the whole range between the original solution structures and almost totally disordered structures (Table 3). These structures define the decoy set 1.

Also, to investigate the influence of decoy generation on the resulting R-factors an additional method for obtaining increasingly disordered structures was used for RalGDS-RBD. In this case, restraints from the original NOE distance restraint list were randomly deleted and new structures were calculated using the reduced restraint lists. Reduced restraint lists contained approximately 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20% and 10% of the original data. Using the reduced distance restraint lists and no additional experimental data 50 structures were calculated from each restraint list of which in each case the best in terms of total energy was selected for further analysis (Table 4) defining decoy set 2.

Relation of the R-factors to the quality of the structure

Figures 1 and 2 show for HPr and RalGDS-RBD, respectively, the correlation between obtained NMR R-factors and rmsd differences to the

Table 2. The influence of the motional model[a]

Type of motional model	$R^{\mathrm{b}}_{\mathrm{PWAUR}}$	Intra-residual[c]	Sequential[c]	Medium-range[c]	Long-range[c]	Inter-residual[c]
Rigid	0.35	0.13	0.14	0.28	0.20	0.19
Detailed	0.35	0.13	0.14	0.25	0.20	0.18

[a] The set of 10 final HPr solution structures was selected for R-factor calculation together with the corresponding 2D NOESY spectrum. [b] R-factors were calculated using R_{PWAUR} ($\alpha = -1/6$) (Equation 2). For R_{PWAUR} the A-list contains only signals from amino acids separated by more than four amino acids in the sequence. [c] R-factors were calculated using R_{PWAR} ($\alpha = -1/6$) (Equation 1). Separate R-factors were calculated for all intra-residual signals, all sequential signals, all medium-range signals (medium range signals are defined as inter-residual signals which are arising from amino acids i and j which are not further apart in the sequence than four residues ($i < j$, $j - i \leq 4$)), all long-range signals and all inter-residual signals.

Table 3. Test set for the correlation between RMSD-values and R-factors employing unrestrained molecular dynamics calculations (decoy set 1)

Structure #	MD simulation steps of 5 fs	rmsd (nm) HPr[a]	rmsd (nm) RalGDS-RBD[1]
1	0	0.000	0.000
2	100	0.025	0.022
3	200	0.038	0.040
4	400	0.066	0.077
5	600	0.100	0.111
6	800	0.137	0.148
7	1000	0.179	0.189
8	1500	0.270	0.289
9	2000	0.370	0.385
10	2500	0.457	0.487
11	3000	0.544	0.601
12	3500	0.622	0.703
13	4000	0.698	0.805
14	5000	0.840	1.022
15	6000	0.985	1.236
16	7000	1.117	1.454

[a] As reference for the rmsd calculations the final solution structures of HPr and RalGDS-RBD were used. Rmsd values were calculated for the C^{α} atoms.

Table 4. Test set for the correlation between RMSD-values and R-factors employing reduced NOE distance restraint lists (decoy set 2)

Structure #	Number of NOE distance restraints	rmsd (nm) RalGDS-RBD[1]
1	1511	0.000
2	1364	0.120
3	1205	0.136
4	1080	0.142
5	894	0.147
6	746	0.232
7	593	0.221
8	435	0.345
9	300	0.492
10	201	1.154

[a] As reference for the rmsd calculations the final solution structure of RalGDS-RBD were used. Rmsd values were calculated for the C^{α} atoms.

original structures. Note that for these tests decoy set 1 was used. Displayed are the *PWAUR* R-factors ($\alpha = -1/6$) according to Equation 2, please note that the *A*-list contains only assigned long-range signals. Long-range signals are defined as inter-residual signals which arise from amino acids *i* and *j* that are separated by more than four residues in the primary sequence. They were considered for R-factor calculation since the presence of assigned long-range signals is strongly correlated to the correct fold of a molecule (Gronwald et al., 2000). R-factors obtained using the three-dimensional spectrum (upper part) compared to the results obtained from 2D data (middle part) and the combined values obtained from both spectra (lower part). As it can be easily seen, these R-factors are strongly discriminating with values ranging from one for structures possessing rmsd values above 1 nm to values around 0.30 for the original structure. This is true for the R-factors obtained from the 2D and 3D data and for the averaged R-factors. For the highly disordered structures the R-factor R_{PWAUR} (Equation 2) is mostly influenced by the signals of class U_1 and is approaching a value of one, while for a correct structure only few signals remain unassigned and the R-factor is mostly dominated by the difference between experimental and simulated long-range signals of class A.

From Figures 1 and 2 it is obvious that for small and medium rmsd values up to 0.8 nm an almost linear relationship can be established to our NMR R-factors. Differences between the individual R-factors of the 2D and 3D cases can be mainly seen for the R-factors where the corresponding rmsd values exceed 0.2 nm where marked deviations from the trend-lines (solid black lines) shown in Figures 1 and 2 exist. Trend-lines were obtained by a fit of the data to a fourth order polynomial. Here, more reliable results can be obtained by increasing the available database for the R-factor calculation by using an averaged R-factor (lower part of Figures 1 and 2).

Next we investigated also using decoy set 1 if the new R-factor definition given by Equation 3 (R_{PWFAR}) where additionally the signals of the U_1-list have been assigned as defined in the materials and methods section is also sufficient to establish a clear relationship between R-factors and RMSD values. As a test case RalGDS-RBD together with a ^{1}H 2D NOESY spectrum was used. We tested two versions of Equation 3 first where the *A*-list and the U_1-list contain both all assigned signals (case 1) and second where both lists contain only the subset of assigned long-range signals (case 2). Figure 3 shows the results for RalGDS-RBD employing the ^{1}H 2D-NOESY spectrum. As it can be clearly seen for both cases a

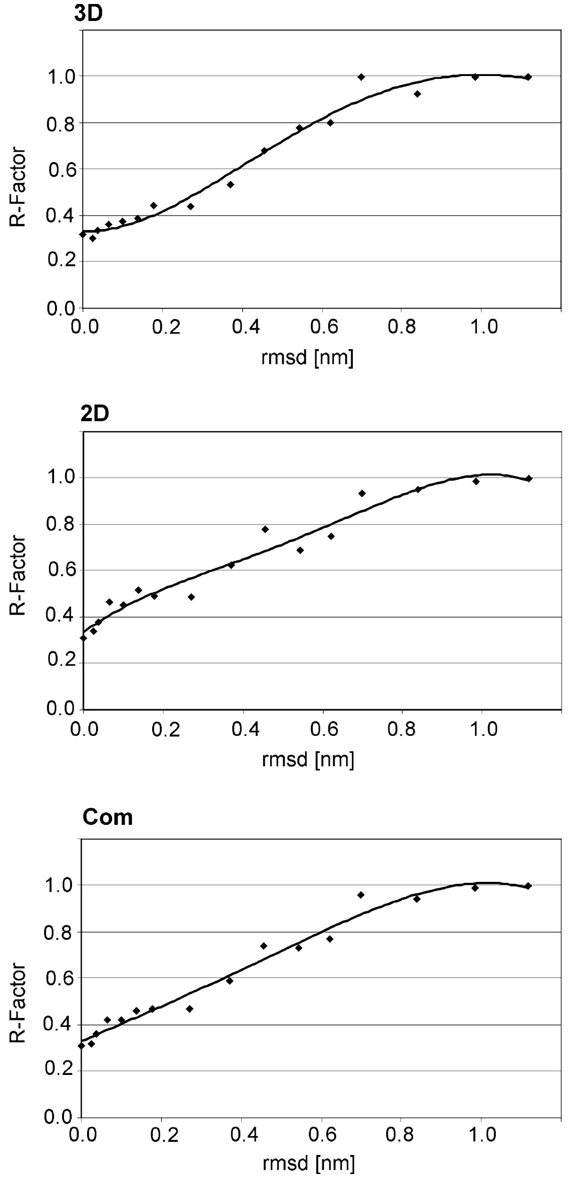

Figure 1. R-factors for HPr. In the upper part, the results obtained using a 3D ^{15}N edited NOESY-HSQC spectrum (measured at 500 MHz) are shown. In the middle, the results from a 2D ^{1}H NOESY spectrum (measured at 800 MHz) are displayed. Only signals in the range from 11.82 to 5.66 ppm in F2 were taken for the R-factor calculation. In the lower part of the figure, combined R-factors were calculated from the 2D and 3D data. Displayed are the PWAUR R-factors (α = $-1/6$) according to Equation 2, please note that the *A*-list contains only assigned long-range signals. Only peaks with a signal probability $p_i > 0.8$ were used. The results are obtained using the assigned long-range signals together with the remaining non-assigned signals. Data were fitted by a 4th order polynomial.

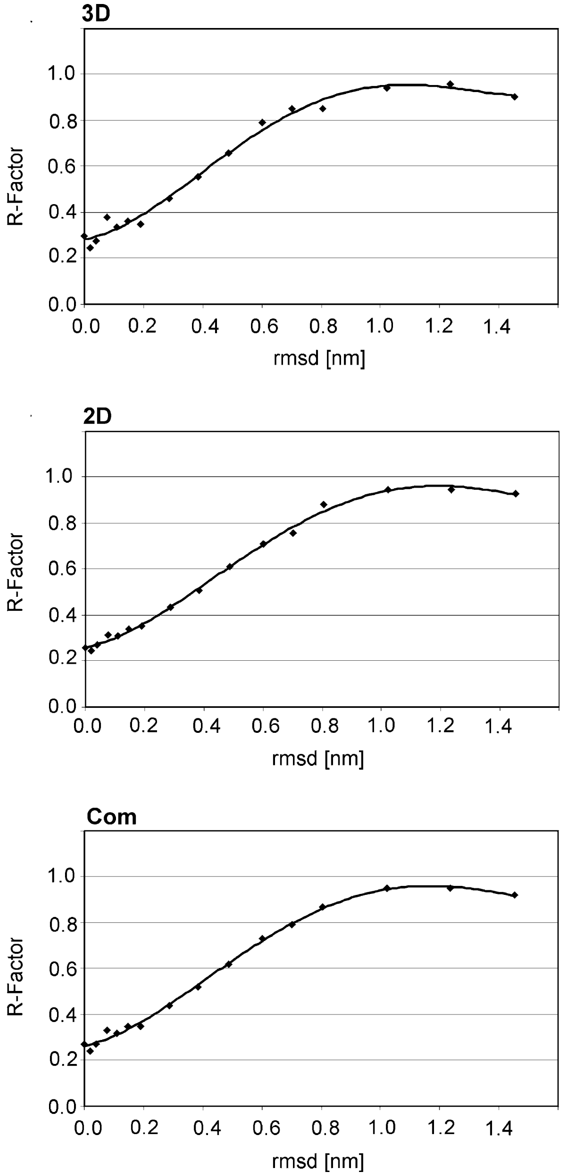

Figure 2. R-factors for RalGDS-RBD. Increasingly disordered structures for RalGDS-RBD were taken from decoy set 1. In the upper part, the results obtained using a 3D ^{15}N edited NOESY-HSQC spectrum (measured at 600 MHz) are shown. In the middle, the results from a 2D ^{1}H NOESY spectrum (measured at 800 MHz) are displayed. Only signals in the range from 10.00 to 6.2 ppm in F2 were taken for the R-factor calculation. In the lower part of the figure, weighted average R-factors were calculated from the 2D and 3D data. Displayed are the PWAUR R-factors (α = $-1/6$) according to Equation 2, please note that the *A*-list contains only assigned long-range signals. Only peaks with a signal probability $p_i > 0.8$ were used. The results are obtained using the assigned long-range signals together with the remaining non-assigned signals. Data were fitted by a 4th order polynomial.

clear relationship is visible between R-factors and rmsd values. As noted in the materials and methods section for substantially disordered structures in both cases 1 and 2 the R-factors exceed a maximum value of 1. This is especially true for case 2 where a maximum R-factor of 3.9 is obtained.

Employing a 2D NOESY spectrum of RalGDS and decoy set 1 we also analyzed the R-factor definition given by Equation 4 (R_{PWFASR}). Here the A-list and the U_1-list contain both all assigned signals and the U_c-list contains all simulated signals without corresponding experimental signals and whose volumes are also above the detection limit. The results (Figure 4) show that this R-factor also shows a clear and almost linear correlation between rmsd values and R-factors. Due to the used large statistical base, all obtained R-factors are very close to the fitted curve in Figure 4. For R_{PWFASR} a maximum value of 1.49 was obtained. In this context, we determined for $V^{-1/6}_{\mathrm{noise\_c}}$ the optimal value of $V^{-1/6}_{\mathrm{noise\_c}}$ that specifies the distance limit above which the fraction of the number of experimental signals to the number of simulated signals of the same distance class substantially decreases. For this purpose, a histogram was employed using distance classes of 0.05 nm width (Figure 5). The results show that for the last 2 distance classes from 0.55 to 0.65 nm only very few experimental signals were found. As a consequence the maximum detection limit used throughout this contribution was set to 0.55 nm since similar values were obtained for the 3D spectra and the HPr test case (data not shown). For R_{PWFASR} the value of $V^{-1/6}_{\mathrm{noise\_c}}$ was set to a slightly smaller value of 0.50 nm to ensure that for a reasonable amount of the simulated signals experimental counterparts are available. For these tests, also decoy set 1 was used.

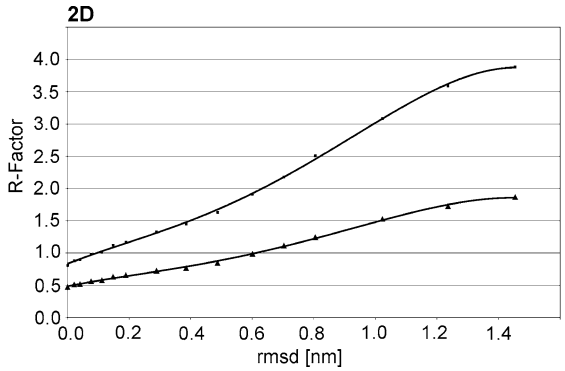

Figure 3. R-factors for RalGDS-RBD employing Equation 3 (R_{PWFAR}) and a ^{1}H 2D-NOESY spectrum. Increasingly disordered structures for RalGDS-RBD were taken from decoy set 1. For this R-factor definition additionally the signals of the U_1-list have been assigned as defined in the materials and methods section. The black triangles denote the case where the A-list and the U_1-list contain both all assigned signals and the black squares refer to the case where both lists contain only the subset of assigned long-range signals.

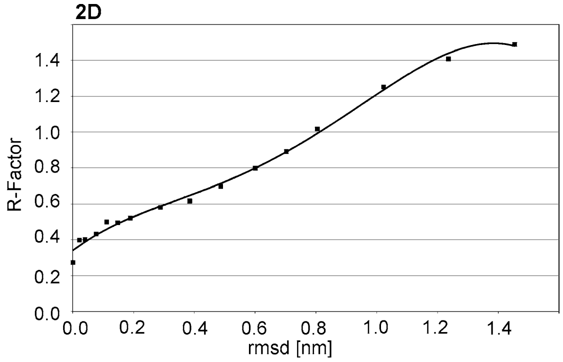

Figure 4. R-factors for RalGDS-RBD employing Equation 4 (R_{PWFASR}) and a ^{1}H 2D-NOESY spectrum. Increasingly disordered structures for RalGDS-RBD were taken from decoy set 1. For this R-factor definition, additionally the signals of the U_c-list have been used as defined in the materials and methods section. The A-list and the U_1-list contain both all assigned signals. The U_c-list contains all simulated signals without corresponding experimental signal that are above the experimental detection limit.

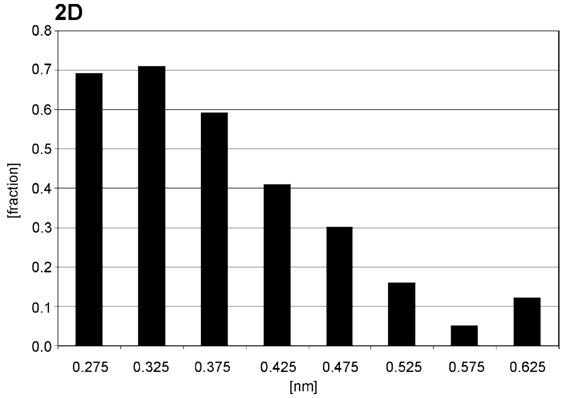

Figure 5. Histogram representation of the fraction of the number of experimental signals to the number of simulated signals of the same distance class for RalGDS-RBD employing a ^{1}H 2D-NOESY spectrum. Increasingly disordered structures for RalGDS-RBD were taken from decoy set 1. The first 2 classes ranging from 0.15 to 0.20 nm and from 0.20 to 0.25 nm were omitted in the histogram representation since the statistical basis for these two classes was rather small. A distance class width of 0.05 nm was used. For both simulated and experimental signals, only signals above the diagonal and in the range from 10.00 to 6.2 ppm in F2 were taken.

Next, the influence of decoy generation on the rmsd R-factor relationship was investigated. For this purpose, a second decoy set (decoy set 2) for RalGDS-RBD was generated were increasingly disordered structures were obtained as described above by randomly deleting increasing amounts of restraints from the original NOE distance restraint list. Figure 6 displays the *PWAUR* R-factors ($\alpha = -1/6$) according to Equation 2 obtained using a 2D NOESY spectrum. As it can be seen in Figure 6 again a linear relationship between R-factors and rmsd values is obtained. However, in comparison to the results obtained using decoy set 1 (Figure 2) the slope of the line of best fit is reduced. For the last structure possessing for the Cα atoms a rmsd value of 1.1 nm to the original structure an R-factor of 0.438 is obtained.

Discussion

Combination of R-factors

Our results show that the automated determination of 3D-R-factors implemented in AUREMOL works as reliable as that implemented earlier for 2D-NOESY spectra. The obtained values do not depend significantly on the type of spectra. This is also true for the combination of 2D with 3D data and (data not shown) for the combination of different 2D spectra. The successful combination of different NOESY data sets is an important step to integrate all experimental data in a measure for the agreement of experimental data with the obtained structures. The use of all available, structurally relevant data for a quality assessment of structures is mandatory from first principals. The calculation of a single R-factor from a set of different spectra is not trivial. We present here a solution which at least fulfils the plausible condition that different amplitude scaling in a set of spectra should not influence the calculation of the averaged R-factor. However, more complicated definitions of combined R-factors are thinkable which for example would additionally include the information content of the type of NMR spectrum or the qualities of the spectra used. We are actually exploring these possibilities; however, the solutions envisaged can hardly be called R-factors in the traditional sense.

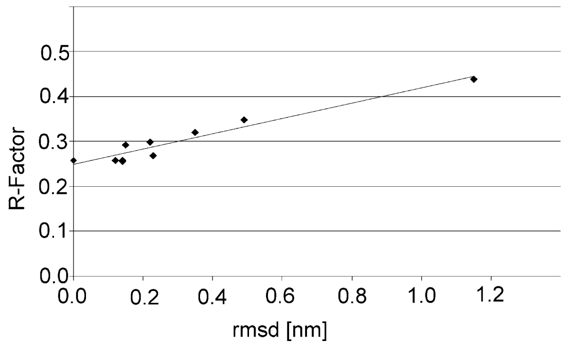

Figure 6. R-factors for RalGDS-RBD using decoy set 2. Increasingly disordered decoys for RalGDS-RBD were obtained using partial distance restraint lists. Results from a 2D ^{1}H NOESY spectrum (measured at 800 MHz) are displayed. Only signals in the range from 10.00 to 6.2 ppm in F2 were taken for the R-factor calculation. In the lower part of the figure, weighted average R-factors were calculated from the 2D and 3D data. Displayed are the PWAUR R-factors ($\alpha = -1/6$) according to Equation 2, please note that the *A*-list contains only assigned long-range signals. Only peaks with a signal probability $p_i > 0.8$ were used. The results are obtained using the assigned long-range signals together with the remaining non-assigned signals. Data were linearly fitted.

The use of a combined R-factor promises to better judge the quality of less well defined structures and or to give more reliable values if the quality of the experimental spectra is suboptimal. This can be important for the analysis of partially flexible peptides and folding intermediates and for the progress analysis during an iterative automated structure determination process using for example the program AUREMOL-KNOWNOE (Gronwald et al., 2002). Such an effect can be seen in Figures 1 and 2 where the averaged R-factor better discriminates between different intermediate structures. Since the whole process of the R-factor calculation is fully automated within AUREMOL different structures can be compared in a fast manner making it ideally suited for high-throughput approaches as used in structural genomics projects.

The use of the proper motional model

For the calculation of R-factors, the quality of the back calculation is important. As to be expected the calculated R-factors are sensitive to the employed motional models and that best results are obtained when as detailed as possible motional models are used. Therefore, for the calculation of the R-factors one should also include all available information describing the dynamics of the protein of interest.

Choice of the R-factor

The R-factors obtained by Equation 2 (Figures 1 and 2) are normalized to a maximum value of 1 for a structure that bears no similarity with the true solution structure. However, the R-factors obtained by R_{PWFAR} (Figure 3) and R_{PWFASR} (Figure 4) can adopt substantially larger values depending on the size and quality of a given trial structure making the interpretation of an obtained R-factor value in terms of structural quality more difficult. Using for Equation 3 all signals of the A-list and the U_1-list provides a larger statistical basis for the calculation of the R-factors while the R-factors obtained using only the long-range signals are more sensitive to structural changes. One important prerequisite when using R_{PWFAR} is a nearly complete sequential resonance line assignment. Otherwise signals of the U_1-list might be wrongly assigned in case that an experimental signal is explained by more than one proton pair due to chemical shift ambiguity and the resonance line assignments of the proton pair corresponding to the shortest structural distance are missing. In this case, the given experimental signal will be assigned to the proton pair with the next shortest distance and matching chemical shifts. As a consequence drastically incorrect distances might be used for R-factor calculations which in turn will lead to increased R-factor values. Therefore, R_{PWFAR} should only be used with care and R_{PWAUR} is more general applicable. The same is true for R_{PWFASR} (Equation 4) that is basically an extension of Equation 3 to take also simulated signals without corresponding experimental signals into account. A comparison of Figures 3 and 4 shows that in general the obtained relationship between R-factors and rmsd values using R_{PWFASR} (Equation 4) and R_{PWFAR} (Equation 3) (case 2) is similar. However, R_{PWFAR} (Equation 3) (case 2) is more sensitive to structural changes with R-factors ranging from 0.81 to 3.88 than R_{PWFASR} with values ranging from 0.27 to 1.49 (Equation 4). This is mainly due to the fact that for the more disordered structures also fewer signals are simulated that are above the experimental detection level. Therefore, the influence of the simulated signals without corresponding experimental signals on the resulting R-factors is relatively small. Also it is obvious from Figures 1–4 that the R-factors obtained for the final solution structures adopt values substantially different from 0 as expected in an ideal case. This discrepancy can be attributed to factors such as a limited precision of the NOE back-calculation due to unknown order parameters for the side-chain atoms, incorrect experimental volumes due to factors such as baseline rolls, the presence of artifacts, and the fact that the final solution structures that we have determined might be still away from the true structure present in solution. It can although mean that the optimization procedures commonly used to calculate the structures do not find the optimal structure or that a single solution is not sufficient to explain the data.

As a consequence the R-factors in their present definitions allow comparing different structures with each other and allows determining which of these structures explains the experimental data best. In general, one can say that the different R-factors defined by Equations 2, 3, and 4 allow a sensible discrimination between different structures. As explained above R_{PWAFR} (Equation 3) and R_{PWAFSR} (Equation 4) might be sensitive to the completeness of the resonance line assignment where in R_{PWAUR} (Equation 2) the influence of missing resonance line assignments is limited by the application of the V_{noise} term.

Use of R-factors to distinguish between different structural models

As soon as the general resonance line assignment is available our fully automated NOE cross peak assignment in combination with the also automated R-factor calculation allows the comparison of a large number of structures in a short amount of time. One obvious application is the direct comparison of structures obtained from a variety of sources, e.g. NMR, X-ray, and or homology modeling. In technical terms, the R-factor calculation described here is an extension to our previous work (Gronwald et al., 2000) and has several new aspects compared to already published work by us, the calculation of R-factors from three-dimensional NOESY-HSQC-spectra, a proper handling of signal superposition, and the use of multiple spectra to calculate weighted average R-factors. Using three-dimensional spectra it is now possible to calculate automatically reliable R-factors for larger proteins. Although, fundamentally the similar information is also contained in the 2D-spectra of large proteins, problems may arise by peak superposition. Especially, the probability is higher that non-recognized artifact peaks are misinterpreted or that overlapping of resonances includes incompletely assigned peaks.

R-factors, rmsd values and the expansion of the structure

The use of R-factors to assess the quality of a structure after or during the structure calculation has been recognized by several authors in the past see for example the paper by Gonzalez et al. (1991). Huang et al. (2005) have related their RPF and DP scores with rmsd differences between reference and intentionally distorted structures. Also Hubner et al. (2004) have shown a linear relationship between classical rmsd values between the atomic coordinate vectors and the rmsd values between corresponding distances in the set structures (dRMS values), that is the dRMS measures the agreement of inter-atomic distances in a reference structure and a trial structure. Therefore, the dRMS is related to NMR R-factors. However, to our knowledge an almost linear relationship between real NMR R-factors and rmsd values for small rmsd-values has not been noticed before. This is conceptionally a different measure since it compares essentially experimental distances with distances in calculated structures. At the moment, it is not really clear how this dependence can be further used and the apparent linear behavior of the R factor is not understood yet, considerable additional work is required which is, however, beyond the scope of this contribution.

A comparison of the R-factors obtained using different decoy sets as shown by a comparison of Figures 2 and 6 for RalGDS-RBD demonstrates that for both cases a nearly linear relationship between R-factors employing the R-factor R_{PWAUR} (Equation 2) and rmsd values is obtained. However, in comparison to the results obtained using decoy set 1 (Figure 2) the slope of the line of best fit is reduced when decoy set 2 is used. For the last structure in decoy set 2 possessing for the Cα atoms a rmsd value of 1.1 nm to the original structure an R-factor of 0.438 is obtained. A closer analysis of this last structure of decoy set 2 shows that the structure almost retains the compactness of the original structure with a radius of gyration of 1.40 nm (original structure 1.22 nm) but all secondary structure elements are completely disordered. In contrast, a corresponding structure from decoy set 1 that shows a similar rmsd value of 1.24 nm to the original structure possess an R-factor of 0.95. However, in comparison this structure is much more expanded as reflected by a radius of gyration of 2.32 nm. Therefore, one can conclude that the R-factor is related in a linear fashion to both the measured rmsd values and the expansion of the structure. In summary, one can say that the NMR R-factor is a sensitive measure to the quality of a structure that reflects the true closeness to the experimental data, more than rmsd values or radii of gyration, which are not necessarily related to each other. It should be noted that similar linear relationships with a reduced slope of the line of best fit were found for decoy set 2 (data not shown) when R_{PWFAR} (Equation 3) and R_{PWFASR} (Equation 4) were applied.

Conclusion

Besides R-factors other measures for the agreement of the experimental data were proposed, e.g. the RPF-value defined by Huang et al. (2005). None of them is ideal in all respects. However, R-factors are suitable measures for the quality of structure calculation. Compared to a simple NOE violation analysis from user prepared and filtered NOE lists R-factors as defined in the following are probably more objective. There are several points which appear to be important to the use of NMR R-factors in the quality assessment of NMR structures and are implemented in AUREMOL: (1) the calculation of the NMR factor should have been automated as much as possible and thus not depend on the user, (2) all available, structurally relevant NMR data have to be used for the calculation of the R-factor that is usually a set of NMR spectra has to be taken, (3) artifacts have to be recognized and dismissed as it is done by a Bayesian analysis in AUREMOL, (4) unassigned true signals have to be taken in to account, (5) the simulation of the NOESY spectra has to be as realistic as possible including internal mobility effects and saturation effects, (6) superposition effects have to be considered properly, and (7) the choice of the R-factor should be adapted to the problem that is for the assessment of the fold accuracy it is important to consider the non-assigned signals. In addition, we could show an almost linear relationship between R-factors and rmsd values and the compactness of the structure that clearly demonstrates the relevance of NMR R-factors for the quality assessment of protein solution structures.

Acknowledgements

We thank W. Hengstenberg, C. Hermann and A. Wittinghofer for providing HPr from *S. carnosus* and RalGDS-RBD, respectively. This work is supported by the Deutsche Forschungsgemeinschaft, the Fonds der Chemischen Industrie and the European Commission (SPINE-consortium).

References

Antz, C., Neidig, K.-P. and Kalbitzer, H.R. (1995) *J. Biomol. NMR*, **5**, 287–296.

Baleja, J.D., Moult, J. and Sykes, B.D. (1990) *J. Magn. Reson.*, **87**, 375–384.

Bonvin, A.M.J.J., Boelens, R. and Kaptein, R. (1991) *J. Biomol. NMR*, **1**, 305–309.

Borgias, B.A., Gochin, M., Kerwood, D.J. and James, T.L. (1990) *Prog. NMR Spectrosc.*, **22**, 83–100.

Borgias, B.A. and James, T.L. (1990) *J. Magn. Reson.*, **87**, 475–487.

Brünger, A.T. (1993) XPLOR Manual Version 3.1, Yale University Press, New Haven. Ref Type: Computer Program.

Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grossekunstleve, R.W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T. and Warren, G.L. (1998) *Acta Crystallogr.*, **D54**, 905–921.

Brünger, A.T., Campbell, R.L., Clore, G.M., Gronenborn, A.M., Karplus, M., Petsko, G.A. and Teeter, M.M. (1987) *Science*, **235**, 1049–1053.

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

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

Cullinan, D., Korobka, A., Grollman, A.P., Patel, D.J., Eisenberg, M. and de los Santos, C. (1996) *Biochemistry*, **35**, 13319–13327.

Geyer, M., Hermann, C., Wohlgemuth, S., Wittinghofer, A. and Kalbitzer, H.R. (1997) *Nat. Struc. Biol.*, **4**, 694–699.

Geyer, M., Neidig, K.-P. and Kalbitzer, H.R. (1995) *J. Magn Reson. B*, **109**, 31–38.

Gonzalez, C., Rullmann, J.A.C., Bonvin, A.M.J.J., Boelens, R. and Kaptein, R. (1991) *J. Magn. Reson.*, **91**, 659–664.

Görler, A., Gronwald, W., Neidig, K.P. and Kalbitzer, H.R. (1999) *J. Magn Reson.*, **137**, 39–45.

Görler, A., Hengstenberg, W., Kravanja, M., Beneicke, W., Maurer, T. and Kalbitzer, H.R. (1999) *Appl. Magn. Reson.*, **17**, 465–480.

Görler, A. and Kalbitzer, H.R. (1997) *J. Magn Reson.*, **124**, 177–188.

Gronwald, W. and Kalbitzer, H.R. (2004) *Prog. NMR Spectrosc.*, **44**, 33–96.

Gronwald, W., Kirchhofer, R., Görler, A., Kremer, W., Ganslmeier, B., Neidig, K.P. and Kalbitzer, H.R. (2000) *J. Biomol. NMR*, **17**, 137–151.

Gronwald, W., Moussa, S., Elsner, R., Jung, A., Ganslmeier, B., Trenner, J., Kremer, W., Neidig, K.P. and Kalbitzer, H.R. (2002) *J. Biomol. NMR*, **23**, 271–287.

Gupta, G., Sarma, M.H. and Sarma, R.H. (1988) *Biochemistry*, **27**, 7909–7919.

Huang, Y.P., Powers, R. and Montelione, G.T. (2005) *J. Am. Chem. Soc.*, **127**, 1665–1674.

Hubner, I.A., Shimada, J. and Shakhnovich, E.I. (2004) *J. Mol. Biol.*, **336**, 745–761.

Lane, A.N. (1990) *Biochim. Biophys. Acta*, **1049**, 189–204.

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

Lefevre, J.-F., Lane, A.N. and Jardetzky, O. (1987) *Biochemistry*, **26**, 5076–5090.

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

Lipari, G. and Szabo, A. (1982a) *J. Am. Chem. Soc.*, **104**, 4546–4559.

Lipari, G. and Szabo, A. (1982b) *J. Am. Chem. Soc.*, **104**, 4559–4570.

Mertz, J.E., Güntert, P., Wüthrich, K. and Braun, W. (1991) *J. Biomol. NMR*, **1**, 257–269.

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

Nikonowicz, E.P., Meadows, R.P. and Gorenstein, D.G. (1990) *Biochemistry*, **29**, 4193–4204.

Nilges, M., Habazettl, J., Brünger, A.T. and Holak, T.A. (1991) *J. Mol. Biol.*, **219**, 499–510.

Ried, A., Gronwald, W., Trenner, J.M., Brunner, K., Neidig, K.-P. and Kalbitzer, H.R. (2004) *J. Biomol. NMR*, **30**, 121–131.

Schulte, A.C., Görler, A., Antz, C., Neidig, K.P. and Kalbitzer, H.R. (1997) *J. Magn Reson.*, **129**, 165–172.

Sippl, M.J. (1993) *Proteins*, **17**, 355–362.

Thomas, P.D., Basus, V.J. and James, T.L. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 1237–1241.

Xu, Y., Sugar, I.P. and Krishna, N.R. (1995) *J. Biomol. NMR*, **5**, 37–48.