J Biomol NMR (2014) 59:147–159
DOI 10.1007/s10858-014-9833-3

ARTICLE

Automatic NOESY assignment in CS-RASREC-Rosetta

Oliver F. Lange

Received: 8 February 2014 / Accepted: 19 April 2014 / Published online: 16 May 2014

Abstract We have developed an approach for simultaneous structure calculation and automatic Nuclear Overhauser Effect (NOE) assignment to solve nuclear magnetic resonance (NMR) structures from unassigned NOESY data. The approach, autoNOE-Rosetta, integrates Resolution Adapted Structural RECombination (RASREC) Rosetta NMR calculations with algorithms for automatic NOE assignment. The method was applied to two proteins in the 15–20 kDa size range for which both, NMR and X-ray data, is available. The autoNOE-Rosetta calculations converge for both proteins and yield accurate structures with an RMSD of 1.9 Å to the X-ray reference structures. The method greatly expands the radius of convergence for automatic NOE assignment, and should be broadly useful for NMR structure determination.

Keywords Nuclear magnetic resonance · Automatic NOE assignment · Structure determination

Electronic supplementary material The online version of this article (doi:10.1007/s10858-014-9833-3) contains supplementary material, which is available to authorized users.

O. F. Lange (✉)
Biomolecular NMR and Munich Center for Integrated Protein Science, Department Chemie, Technische Universität München, Lichtenbergstrasse 4, 85747 Garching, Germany
e-mail: oliver.lange@tum.de

O. F. Lange
Institute of Structural Biology, Helmholtz Zentrum München, Neuherberg, Germany

Introduction

Structure determination by nuclear magnetic resonance (NMR) spectroscopy is largely driven by distance information gathered through Nuclear Overhauser Effect (NOE) spectroscopy. To use NOE data as distance restraints, the NOE crosspeaks in multidimensional spectra have to be assigned to individual atoms of the biomolecular system. The NOE crosspeak assignment and structure generation steps are usually performed in an integrated manner over several iterations to maximize the number of conformational restraints, while guaranteeing self-consistency of all distance restraints (Wüthrich 1986).

Chemical shift assignments of individual spins and the positions of cross peaks in NOE spectra (peak-picking) can often be obtained accurately without explicit 3D structural modeling, whereas resolving the high ambiguity in NOE cross peak assignments requires structural models. The main challenge is thus, to obtain initial 3D structures despite the high ambiguity and low-fidelity of initial automatic NOE cross peak assignments. If accurate enough, these initial 3D models can be used to start further iterations of refinement of assignments and 3D models.

Resolution Adapted Structural RECombination (RASREC) is an iterative sampling strategy for restraint guided structure determination in ROSETTA (Lange and Baker 2012). As shown previously, RASREC requires less data than standard algorithms to converge (Lange and Baker 2012) and has been shown to allow structure determination for proteins up to 20 kDa from RDC and expert-assigned backbone NOE data (Raman et al. 2010a). Using additional ILV methyl–methyl NOE data, RASREC can determine structures of proteins up to 40 kDa (Lange et al. 2012). Most importantly, RASREC requires less NOE data and is more robust against inaccurate restraints (Warner et al.

2011; Lange et al. 2012). These properties make RASREC an ideal partner for automatic NOE assignment methods, as a considerable number of initial assignments are wrong and ambiguity is high.

We sought to combine RASREC-Rosetta structure determination with automatic NOE assignment methods within the ROSETTA3 software suite. As starting point for the NOE assignment algorithm, we took established algorithms such as ARIA (Nilges 1993; Nilges et al. 1997; Linge et al. 2003), AutoStruct (Huang et al. 2003) and CANDID (Herrmann et al. 2002). Whereas CANDID is implemented in the popular programs CYANA and UNIO, ARIA and AutoStruct are part of program packages of the same name. The general approach of these algorithms is very similar. First, atoms are initially assigned to 2D, 3D or 4D NOESY crosspeaks based on the known chemical shift resonances. This yields on the order of 10–20 initial assignments for a typical 3D NOESY cross peak. Subsequently, all initial assignments of a given peak are ranked according to different descriptors, including the chemical shift compatibility, network anchoring (Herrmann et al. 2002), symmetry considerations, and compatibility with preliminary structural models. The highest-ranking assignment of a given cross peak yields a distance restraint. If multiple high-ranking assignments get selected, they are combined into an ambiguous restraint (Nilges 1993). In the first structure calculation stage, all peaks for which assignments can be found yield distance restraints. In subsequent rounds of iterative refinement, peaks are excluded if none of their assignments are compatible with the preliminary 3D models.

To determine the upper-distance bound of a restraint, various peak-calibration strategies have been proposed. The simplest strategy fixes the upper distance bound to the minimum distance required for NOE cross peaks to appear in the spectrum. Most programs, however, set the upper distance bound proportional to the inverse sixth root of the peak intensity. The necessary proportionality constants are iteratively fitted using either the preliminary structures or fixed target values (Güntert et al. 1991; Herrmann et al. 2002). ARIA further uses a spin-diffusion correction (Linge et al. 2004) to determine the upper distance bound from the intensity.

Typically in structure calculations, a quadratic penalty term is used for conformers violating the upper-distance bound of a restraint (Brunger et al. 1998; Linge et al. 2003). CS-Rosetta calculations with NOE restraint data were shown to perform better, however, if the penalty term switches to a linear slope at larger deviations (Raman et al. 2010b). Recently, a log-harmonic penalty term was shown to improve results in ARIA calculations (Bernard et al. 2011).

For iterative refinement of NOE assignments usually all assigned distances that are significantly violated by the preliminary models are removed (Nilges 1993; Herrmann et al. 2002; Huang et al. 2003; Linge et al. 2003). The obvious danger with this approach is to remove correct information prematurely. To fully exploit the ability of RASREC-Rosetta to obtain correct structures even with limited and erroneous restraint data, we took care to give RASREC sufficient time before non-fitting restraints are explicitly removed. Erroneous restraints are only problematic in the final stages of RASREC (Stages V–VI), where a converged fold undergoes rounds of refinement in ROSETTA's all-atom energy function. For all-atom refinement nearly any number of wrong restraints will cause severe frustration to the optimization process. During the earlier, low-resolution RASREC stages (Stages I–IV), however, erroneous restraints cause less frustration. Thus, we assigned the following NOE assignment Phases to the RASREC sampling stages: Phase I, consisting of Stages I–IV, restraints are never removed; Phase II, a repetition of Stages III–IV, restraints are removed if for a given atom-pair the low-energy conformers converged on a precise distance, which violates the restraint; Phase III, Stages V–VI, all restraints that violate more than 50 % of the conformers are removed.

The NOE assignment algorithm is described in detail in section "Automatic NOE Assignment" and its integration with RASREC is described in section "Integration with RASREC sampling". An extensive benchmark of the method on 50 data sets is published elsewhere (Zhang et al. 2014). Whereas the benchmark study (Zhang et al. 2014) focuses on the final accuracy of the structure calculations, we focus here on analyzing how the algorithm proceeds through the various stages of the calculation. To this purpose we picked two illustrative structure calculations performed for the two proteins, CtR107 and PsR293, whose chemical shift assignments and NOE peak-lists had been published by the North East Structural Genomics Consortium (Mao et al. 2011), and were downloaded from their website (http://psvs-1_4-dev.nesg.org/MR/dataset.html). Note, that the two cases discussed here have been picked to showcase the possible failure-cases of CYANA calculations and the improvement on such challenging cases obtained with autoNOE-Rosetta. Clearly, these results are not representative for the overall performance of the two programs and the reader should thus refer to the full benchmark study (Zhang et al. 2014).

Results and discussion

Automatic NOE assignment

The assignment process is illustrated as flow-chart in Fig. 1. NOE crosspeaks are automatically matched against

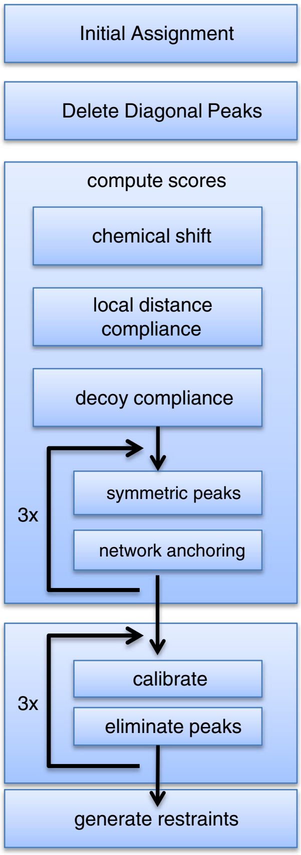


Fig. 1 Illustration of the procedure used in the new NOE-assignment module implemented in Rosetta 3.6

Table 1 Parameters used for automatic assignment

Parameter name	Phase I	Phase II	Phase III
chemshift	0.5	0.5	0.5
symmetry	10	1	1
covalent	10	1	1
calibration_target	3.8 Å	0.1	0.1
calibration:max_nudging	–	1.1	1.1
calibration:start_nudging	–	0.1	0.1
calibration:stop_nudging	–	0	0
calibration:cycles	1	3	1
calibration:ignore_eliminated_peaks	False	True	False
calibration:max_noe_dist	–	–	5.5
network:vmin	0.1	0.1	0.1
network:vmax	1.0	1.0	1.0
network:reswise_high	4.0	4.0	4.0
network:reswise_min	1.0	0.5	0.5
network:atomwise_min	0.25	0.4	0.4
local_distviol:range	–	90 %	–
local_distviol:cutoff	–	8.0 Å	–
local_distviol:cutoff_buffer	–	2.0 Å	–
elim:distviol	–	50 %	50 %
elim:dcut	–	–	0.1
elim:vmin	0.01	0.501	0.501
elim:max_assign	20	20	20

given chemical shift assignments to obtain initial assignments. Scoring with various descriptors ranks assignments of each cross peak to identify the most likely assignments. From these, distance restraints are generated with upper bounds proportional to calibrated peak intensities.

The automatic NOE assignment method is similar to and inspired by the CANDID algorithm, originally published in Ref. (Herrmann et al. 2002). To nevertheless provide full documentation of the implementation in Rosetta, we describe the algorithm here in detail. We choose here a mathematical notation, which allows conveying the algorithm exactly, without involving unnecessary details of its implementation. Parameters that can be specified via the command-line interface of ROSETTA are denoted in the form *P*(cmdline-flag) where *cmdline-flag* will be replaced by the strings of characters required to specify this parameter on the command-line. The standard settings of these parameters are given in Table 1 and a glossary of terms is provided in Table 2.

Input data

Assignment starts from 2 to 4 dimensional NOESY crosspeak lists, $\eta^d_{l,p}$where d denotes the dimension, l the peak list and p the cross peak within list l. Moreover, a list of assignable resonances δ_i is given as input data. Methyl protons with identical chemical shifts are combined into a single entry in the resonance list. Thus, we denote the set of protons assigned to resonance i by atoms $\mathcal{A}(i)$. To work with 3D and 4D spectra, we denote the resonance index of the label resonance (i.e., the heavy atom bound to proton(s) of resonance i) by $L(i)$. For convenient notation, we further define $D_l(d, i, j)$ as map of proton and label indices to the respective dimension of the cross peak list l

$$D_l(d,i,j) = \begin{cases} i & d = d_{\mathrm{P1}} \\ L(i) & d = d_{\mathrm{L1}} \\ j & d = d_{\mathrm{P2}} \\ L(j) & d = d_{\mathrm{L2}} \end{cases} \qquad (1)$$

where d_{P1}, d_{P2}, d_{L1} and d_{L2} denote the first and second proton dimensions as well as the first and second label dimension. The values of d_{P1}, d_{P2}, d_{L1} and d_{L2} are determined by the header of each individual peak-file.

Initial assignments

The goal of the automatic assignment module is to find assignments $A_{l,p}(i, j)$ to protons in the resonance list such that the frequencies $\eta^d_{l,p}$ match the respective proton and label resonances. The match of resonance $\eta^d_{l,p}$ of the dth spectral dimension of crosspeak p to chemical shift δ_i of resonance i is given by the following equation

$$M^d_{l,p}(i) = \frac{\left|\mathrm{fold}^d_l(\delta_i) - \eta^d_{l,p}\right|}{\max(\Delta\eta^d_l, \Delta\delta_i)} \tag{2}$$

where the operator $\mathrm{fold}^d_l(\delta) = (\delta - o_{l,d})\mathrm{modulo}\ \mathrm{sw}_{l,d} + o_{l,d}$ maps the chemical shift into the recorded spectral window that starts at offset frequency $o_{l,d}$ and has the sweep width $\mathrm{sw}_{l,d}$. The tolerance $\Delta\eta^d_l$ is specified for each spectral dimension in the NOESY peak list and the tolerance $\Delta\delta_i$ is defined in the chemical shift list for each resonance individually. Defining the $\widehat{M}$ as $\widehat{M} = \{1: M < 1;\ 0\colon \mathrm{else}\}$ an assignment is given as logical AND over *matches* $\widehat{M}$ in each dimension d of the peak-list l

$$A_{l,p}(i,j) = \prod_d \widehat{M}^d_{l,p}(D_l(d,i,j)). \tag{3}$$

Crosspeaks with diagonal assignment, i.e. l,p for which $A_{l,p}(i,i) = 1$ for any i, are excluded from further consideration.

Volume contribution

The volume contribution $W_{l,p}(i,j)$ of assignment (i,j) to peak (l,p) is given as

$$\begin{aligned} W_{l,p}(i,j) = {} & A_{l,p}(i,j)C_{l,p}(i,j)D_{l,p}(i,j) \\ & \min(S_{\max}, S_{l,p}(i,j)V(i,j)N_{l,p}(i,j)), \end{aligned} \tag{4}$$

where C denotes the *chemical shift score, D* the *decoy compatibility* score, S the *symmetry* score, V the covalent compliance score, and N the network anchoring score which are defined below. Further, $S_{\max}$ denotes a global weight and an ensemble average over the given input structures if available. The volume contribution is normalized over all assignments of a crosspeak l,p.

$$\overline{W_{l,p}(i,j)} = \begin{cases} \dfrac{W_{l,p}(i,j)}{\sum_{i,j} W_{l,p}(i,j)} & A_{l,p}(i,j) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Since both, the network anchoring and the symmetry score, depend on $\overline{W_{l,p}(i,j)}$, we iterate the computation of the scores four times.

Chemical shift score The chemical shift score of an assignment is given by

$$C_{l,p}(i,j) = \exp\left(-\frac{1}{2w_C^2}\sum_d M^d_{l,p}(D_l(d,i,j))^2\right), \tag{6}$$

where w_C is a globally specified weight P(chemshift).

Symmetric peaks The same proton–proton contact may appear in several spectra or opposite the diagonal in the same spectrum. If this is the case, the confidence into matching assignments of corresponding crosspeaks should be increased. To reflect this we compute

$$s_{L,P}(i,j) = \sum_{(l,p)\neq(L,P)} \overline{W_{l,p}(j,i)} + \sum_{\substack{l,p \\ l \neq L}} \overline{W_{l,p}(j,i)}. \tag{7}$$

As the symmetry score depends on the volume contribution itself, we compute the symmetry score iteratively. The final score is given as

$$S_{L,P}(i,j) = \max(1, P(\mathrm{symmetry})s_{L,P}(i,j)). \tag{8}$$

Covalent compliance This score evaluates to P(covalent) for a pair of protons if their distance is below 5.0 Å according to some prior structural knowledge or 0 otherwise. If no explicit structural knowledge is known (default), the bonus P(covalent) is given for all intra-residue protons and for the pair HA(k) − H (k + 1), where k denotes a residue position.

$$V(i,j) = \begin{cases} P & \mathrm{dist}(i,j) < 5\text{Å}(\text{removed}) \\ P & \text{i and j in same residue} \\ P & \text{HA}(k) \text{ and } \text{H}(k+1) \\ 0 & \text{else} \end{cases} \tag{9}$$

Network anchoring The network anchoring score reflects how many alternative pathways (i,k,j) exist to connect resonances i and j via a third resonance k, whereas k has to be in the same or a neighboring residue of either resonance i or j (Herrmann et al. 2002).

The resonances in the same or neighboring residues of i are denoted by set

$$\mathcal{N}(i) = \{k\colon \Delta\mathrm{seq}(i,k) \leq 1 \wedge k \neq i\}, \tag{10}$$

where $\Delta\mathrm{seq}(i,j) \equiv |\mathrm{resnum}(i) - \mathrm{resnum}(j)|$, and resnum(i) denotes the residue number of the atom assigned to resonance i. The direct connectivity

$$v(i,k) = \sum_{l,p} \overline{W_{l,p}(i,k)} + \overline{W_{l,p}(k,i)} \tag{11}$$

accumulates all normalized volume contributions to peaks that have assignments connecting resonances i and k. To reduce noise we count only connectivities with a minimum contribution $N_{\min} \equiv P(\text{network:vmin})$

$$\tilde{v}(i,k) = v(i,k)\Theta(v(i,k) - N_{\min}) \tag{12}$$

where Θ denotes the Heavyside function

$$\Theta(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{13}$$

Additionally, a network anchoring contribution $N_{\max} \equiv P(\text{network:vmax})$ is given for NOEs that should be present according to prior knowledge, i.e., $V(i,k) > 0$, and $N_{\min}$ for all other resonances that are sequential or intra-residue. The network anchoring score $N_{l,p}$ is thus given by

$$N_{l,p}(i,j) = \sum_{k \in \mathcal{N}(i) \bigcup \mathcal{N}(j)} \sqrt{n(i,k)n(k,j)}, \tag{14}$$

where we defined

$$\begin{aligned} n(i,k) = \max(&\tilde{v}(i,k), N_{\max}\Theta(V(i,k)) \\ &+ N_{\min}(1 - \Theta(V(i,k)))\Theta(2 - \Delta\text{seq}(i,k))) \end{aligned} \tag{15}$$

Decoy compatibility If no structures are available as input this score is set to 1. Otherwise, the compatibility score is computed for each input structure and averaged. To obtain the score for a single input structure the proton–proton distances are computed as

$$\text{dist}(i,j) = \left(\sum_{\substack{t \in A(i) \\ s \in A(j)}} \text{dist}(t,s)^{-6} \right)^{-\frac{1}{6}}, \tag{16}$$

where t and s denote atoms assigned to resonance i and j, respectively. To normalize the score, we also compute the cumulative distances of all assignments to the cross peak

$$\text{cumdist}_{l,p} = \left(\sum_{i,j} A_{l,p}(i,j)\text{dist}(i,j)^{-6} \right)^{-\frac{1}{6}}. \tag{17}$$

This yields the decoy compatibility score as

$$D_{l,p}(i,j) = \left\langle \left(\frac{\text{dist}(i,j)}{\text{cumdist}_{l,p}} \right)^{-6} \right\rangle \tag{18}$$

Restraint generation

For each cross peak with at least 1 assignment a restraint can be generated. Excluded are restraints that have been eliminated (see below), or have at least 1 assignment $A_{l,p}(i,j)$ with a volume contribution $\overline{W_{l,p}(i,j)} > 0.1$ with $\Delta\text{seq}(i,j) < P(\text{out:min\_seq\_sep})$. The restraint distance is computed from a conformer as follows

$$d_{\text{l,p}} = \left[\sum_{i,j} \Theta\left(\overline{W_{l,p}(i,j)} - P(\text{Vmin}) \right) \sum_{\substack{t \in A(i) \\ s \in A(j)}} \text{dist}(t,s)^{-6} \right]^{-\frac{1}{6}} \tag{19}$$

The corresponding restraint energy is computed using the ROSETTA flat-bottom potential used previously for NOE based structure determination (Lange et al. 2012). No energy penalty is applied for $1.5\text{Å} < d_{\text{l,p}} < u(l,p)$, where $u(l, p)$ denotes the upper-distance bound defined in section “Peak calibration”.

For larger distances, the penalty grows first quadratically and then linearly

$$V_{\text{restraint}} = \begin{cases} \left(\dfrac{d-u}{\sigma} \right)^2 & \text{if} \quad d < u + \dfrac{\sigma}{2} \\ \dfrac{d-u}{\sigma} - \dfrac{1}{4} & \text{otherwise} \end{cases}, \tag{20}$$

where indices l,p have been omitted for clarity. The strength of the potential is given by the number of generated restraints $N_{\text{restraints}}$ and an overall parameter

$$\sigma = \sqrt{N_{\text{restraints}}}/P(\text{cst\_strength})\cdot \tag{21}$$

If restraints are applied to a conformation in the ROSETTA centroid representation instead of all-atom representation all side chain protons are mapped to the CEN interaction center of the centroid model and the upper bound is padded by 0, 1, or 2 Å, if 0, 1, or 2 side chain atoms were mapped to CEN. For cross peaks with multiple assignments, the maximum number of CEN-mappings (max_maps) for any of its assignments is used to determine the padding. Additionally to the padding we reduce the weight of a mapped restraint relative to the other restraints by setting σ_{mapped} to σ, 2σ or 4σ for 0, 1 or 2 max_maps, respectively.

Peak calibration

To compute the upper-distance bound of a peak from the peak’s intensity, suitable proportionality constants have to be defined. We set these in an iterative calibration procedure. To allow for systematic differences in intensity between different types of protons, we categorize as *backbone, beta, methyl,* and *sidechain.* HA and HN are *backbone,* CB-bound protons are *beta* but for alanine residues, in which case they are considered as *methyl.* All other non-methyl protons are considered as *sidechain.* Calibration is carried out separately for each peak-list. The upper distance bound of a cross peak *l,p* is computed as

$$u_{l,p} = \left[I_{l,p} \sum_{i,j} \overline{W_{l,p}(i,j)} \left(\sqrt{C(T_i)C(T_j)} \right)^{-1} \right]^{-\frac{1}{6}}, \tag{22}$$

where $I_{l,p}$ denotes its intensity and $C(T)$ the calibration constant associated with proton-class T. The calibration constants for classes *backbone* and *beta* are determined by minimizing the deviation from a structure dependent or structure independent calibration target.

For structure independent calibration, constants are chosen such that the average distance bound reaches 3.8 Å (specified by $P(\text{calibration\_target})$). The calibration target for structure dependent calibration is that 10 % (specified by P(calibration_target)) of distance bounds are violated by $d_{\text{restraint}}(l,p)$ computed from the conformers. As suggested previously, the calibration constants of classes *methyl* and *sidechain* are set to 3.0 and 1.5 times the calibration constant of *backbone* and *beta,* respectively (Herrmann et al. 2002). After structure dependent calibration is finished, some distance bounds are too tight by a small margin when compared to the existing conformers. If more than $P(\text{calibration:start\_nudging})$ of the conformers are violated structural strain is avoided by nudging the distance bounds up in steps of 0.1 Å until less than $P(\text{calibration:stop\_nudging}\ )$ of conformers are violated. A maximum correction of $P(\text{calibration:max\_nudging})$ of the original distance bound can be applied in this way and no correction at all is applied if more nudging than the maximum limit would be required.

All generated distance bounds are capped by $P(\text{calibration:max\_noe\_dist})$ to avoid unreasonably high values obtained for low-intensity peaks. This parameter can be overwritten in individual peak-files to account for increased mixing times (#MAX_NOE_DIST).

Elimination of spurious crosspeaks

Several filters are applied to eliminate crosspeaks that might be spurious and should not be considered for restraint generation. Crosspeak l, p is eliminated if none of its assignments i, j reaches the minimum peak volume $\overline{W_{l,p}(i,j)} > P(\text{elim : vmin})$. It is also eliminated if the number of assignments $n(l,p)$ exceeds $P(\text{elim : max\_assign})$. As suggested previously, peaks are also eliminated if the network anchoring score of their assignments remains low (Herrmann et al. 2002). Accordingly, we compute

$$B(l,p) = \sum_{i,j} \overline{W_{l,p}(i,j)} N_{l,p}(i,j) \tag{23}$$

and

$$R(l,p) = \sum_{i,j} \overline{W_{l,p}(i,j)} R^o(\text{resnum}(i), \text{resnum}(j)), \tag{24}$$

where $R^o(t,\ s)$ denotes all network contributions between residues t, s. The latter is computed as

$$R^o(t,s) = \sum_{\substack{i \in R(t) \\ j \in R(s)}} \sum_{l,p} A_{l,p}(i,j) N_{l,p}(i,j), \tag{25}$$

where $\mathcal{R}(t) = \{i : \text{resnum(i) = t}\}$denotes the resonances of residue t. For crosspeak l, p to pass the network filter, we require either $R(l,p) > P(\text{network:reswise\_high})$ or $R(l,p) > P(\text{network:reswise\_min}) \wedge B(l,p) > P(\text{network:atomwise\_min})$.

Finally, we also eliminate cross peaks whose upper distance bounds are violated by too many decoys. Since this filter is dependent on the peak calibration we run 3 rounds of calibration followed by cross peak elimination. Two different algorithms are used for cross peak violation, *local distance violation* and *global distance violation*, and their main difference consists of the choice of allowable violation per distance bound. In the former algorithm the allowable violation is determined by the variance in distances within the considered conformations, whereas in the latter the allowable violation is set globally.

For *global distance violation*, a cross peak is eliminated if more than $P(\text{elim:dist\_viol})$ % of the conformers violate the distance bound by more than $P(\text{dcut})$. For *local distance violation* we compute the lower quartile of distances in the conformers $Q1(d_{l,p})$ and the crosspeak is eliminated if $Q1(d_{l,p}) > P(\text{local\_distviol:cutoff})$ or if $Q1(d_{l,p}) > P(\text{local\_distviol:cutoff\_buffer}) + u_{l,p}$. If the crosspeak is not yet eliminated, the largest difference d_{spread} is computed which allows to bracket the fraction $P(\text{local\_distviol:range})$ of the conformer's distances. Using $d_{\text{spread}} + P(\text{local\_distviol:global\_buffer})$ as cutoff value for distance violations we eliminate crosspeaks with more than $P(\text{elim:dist\_viol})$ of conformers violated.

Integration with RASREC sampling

To generate structural models, we want to use the CS-Rosetta methodology (Shen et al. 2008). That is, fragments are selected based on backbone chemical shifts $(C, C_\alpha, C_\beta, N, H_N, H_\alpha)$ and sequence information (Vernon et al. 2013) and subsequently assembled using a fragment assembly approach (Bowers et al. 2000). To this end, a Monte Carlo optimization is applied, where the moves are given by replacing all backbone torsion angles within a window of 3 or 9 residues, with the torsion angles of a specific fragment that has been selected for the specific window by the fragment picker method (Rohl et al. 2004). During fragment assembly a low-resolution energy function is applied. Subsequently, structures are relaxed in ROSETTA's all-atom energy function (Kuhlman et al. 2003; Raman et al. 2010a).

Identifying the lowest energy structures using CS-Rosetta sampling in the rugged landscape created by the ROSETTA energy function is challenging, and tens of thousands of conformers have to be generated for adequate sampling even for relatively small proteins (Shen et al.

2008). RASREC is an iterative sampling strategy that generates batches of 200–2,000 conformers and stores the 100–500 all-time best conformers in a structural pool (Lange and Baker 2012). Simulations of subsequent batches are seeded with structural information from the common structural pool, which allows intensifying the search in the most promising regions of conformational space. This intensification is necessary to optimize the ROSETTA all-atom energy sufficiently well for obtaining high-resolution structures when restraint data is sparse (Lange et al. 2012).

A flow-chart (Fig. 2) illustrates how automatic NOE assignment is coupled with RASREC. The NOE assignment algorithm (right side) is run once before RASREC sampling commences (left side) to yield initial *sampling restraints*, as well as a set of *filter restraints*. The *sampling restraints* are used as restraint energy during structure generation, the *filter restraints* are used to decide which newly generated conformers are accepted into the structural pool and which are discarded.

To avoid distortions and frustrations due to miss-assigned NOEs, it is beneficial to update the *sampling restraints* regularly with the hope that improved structural information in the pool allows improving the quality of assignments. For the *filter restraints*, however, we found that the opposite is true, and discrimination of native from non-native conformations actually deteriorates if they are updated alongside the structural pool. This fact is maybe not too surprising, as wrong conformations in the structural pool would disfavor assignment of those restraints that would penalize the same wrong conformations.

To generate the updated *sampling restraints* we use three distinct sets of parameters, which we call Phase I-, Phase II- and Phase III-parameters (Table 1). Phase I-parameters are rather permissive, yield highly ambiguous restraints and do not explicitly eliminate restraints that are violated by preliminary conformations. Parameters of Phase II and Phase III are less permissive, do not produce ambiguous restraints and eliminate restraints that do not fit to preliminary conformations. In contrast to Phase III, however, Phase II does not eliminate non-fitting restraints in structural regions where convergence is still low. Restraints with <4 residues of sequence separation are not present in the set of *sampling restraints*. This choice reflects our strategy to rely as much as possible on the short-range information that is encoded in the chemical shift selected fragments. This renders the approach more robust against spurious assignments.

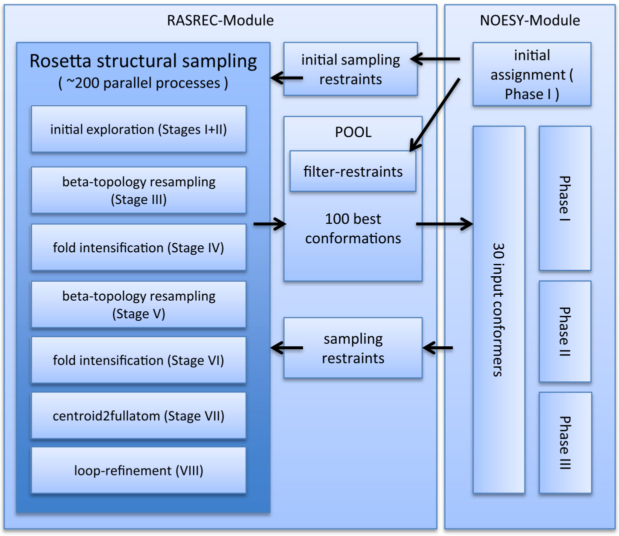

RASREC-Module
NOESY-Module
Rosetta structural sampling
(~200 parallel processes)
initial exploration (Stages I+II)
beta-topology resampling (Stage III)
fold intensification (Stage IV)
beta-topology resampling (Stage V)
fold intensification (Stage VI)
centroid2fullatom (Stage VII)
loop-refinement (VIII)
initial sampling restraints
POOL
filter-restraints
100 best conformations
sampling restraints
initial assignment (Phase I)
30 input conformers
Phase I
Phase II
Phase III

Fig. 2 Illustration of the integration of the new NOESY-module with the RASREC module of Rosetta3. The left-most block with *dark blue* background reflects ∼200 worker processes that run individual Rosetta fragment assembly and relax calculations. Sampling in these is biased by the *sampling restraints* generated by the NOESY Module (*right*). A pool of conformations is filled and updated with conformations generated by the Rosetta sampling processes. The conformations in the pool will be ranked by a combination of Rosetta score, experimental restraints (RDC and chemical shifts), and the *filter restraints* generated by the NOE module. The 30 best-ranked structures in the pool are used as structural input for subsequent rounds of NOE assignment, and thus influence the *sampling restraints.* Depending on the strange of sampling in the RASREC module (Stage I–VIII) different parameter sets for NOE assignment are used (Phase I–Phase III)

AutoNOE-Rosetta calculations of CtR107 and PsR293

In the following, we demonstrate autoNOE-Rosetta on two protein targets from the North East Structural Genomics Consortium. Both proteins have an X-ray structure as reference and their NMR data was taken from a previously published benchmark study (Mao et al. 2011) made available through an NESG website (http://psvs-1_4-dev.nesg.org/MR/dataset.html). After trimming flexible termini (Methods) the lengths are 118 and 147 residues, for PsR293 and CtR107, respectively. The targets were chosen from our full benchmark of 50 proteins published in Ref. (Zhang et al. 2014) to highlight how the new method behaves on large proteins with challenging input data. We derive that the input data is challenging from the fact that CYANA does not yield good structures with this data. Moreover, we selected these two data sets, because an Xray structure is available. The reader should be aware that the results presented here are cherry-picked from the larger benchmark (Zhang et al. 2014) and are not representative for overall behavior of CYANA or autoNOE-Rosetta. The purpose of the following analysis is merely to illustrate the progress of the algorithm throughout its different stages and to give some insight into algorithmic choices that have been made.

For PsR293 4 peak lists are available, a 4D methyl–methyl, a 3D aliphatic ^{13}C, a 3D aromatic ^{13}C, and a 3D

Table 2 Glossary of symbols used throughout the manuscript

Symbol	English definition
$\eta_{l,p}^{d}$	Frequency in dimension d of peak p in peak-list l
δ_i	ith resonance assignment
$\mathcal{A}(i)$	Set of atoms assigned to resonance i
$D_l(d, i, j)$	Return proton or label atom depending on peak-list dimension d
$M_{l,p}^{d}(i)$	Match of resonance i to peak $\eta_{l,p}$ at dimension d
$A_{l,p}(i, j)$	1 if peak $\eta_{l,p}$ can be assigned to proton resonances i and j (respecting any label resonances $L(i)$ if 3D or 4D peak)
$W_{l,p}(i, j)$	Volume contribution of resonance assignment i,j to peak $\eta_{l,p}$
$C_{l,p}(i, j)$	Chemical shift score of resonance assignment i,j to peak $\eta_{l,p}$
$D_{l,p}(i, j)$	Decoy compatibility score of resonance assignment i,j to peak $\eta_{l,p}$
$S_{l,p}(i, j)$	Symmetry score of resonance assignment i,j to peak $\eta_{l,p}$
$V_{l,p}(i, j)$	Covalent compliance score of resonance assignment i,j to peak $\eta_{l,p}$
$N_{l,p}(i, j)$	Network anchoring score of resonance assignment i,j to peak $\eta_{l,p}$
dist(i, j)	Average distance (r^{-6}-weighting) between atoms $\mathcal{A}(i)$and $\mathcal{A}(j)$
$d_{l,p}$	Measured distance in conformation given the (ambiguous) assignments to peak $\eta_{l,p}$
$u_{l,p}$	Calibrated upper bound of peak $\eta_{l,p}$
$B(l, p)$	Cumulated network anchoring score of peak $\eta_{l,p}$
$R(l, p)$	Cumulated residue-wise network anchoring of peak $\eta_{l,p}$
$Q1(x)$	Lower quartile of distribution in observable x

See text for precise definitions

^{1}H-^{15}N HSQC-NOESY. In total 2114 4D and 4247 3D peaks. For target CtR107, a single 3D peak-list is available with both ^{13}C and ^{15}N NOE cross peaks comprising 7,180 entries, of which 3,705 have non-zero intensity. Note, that the program ignores zero-intensity peaks. Chemical shift assignments are downloaded from the BMRB and CYANA 3.0 specifies their completeness to 95.9 and 93.1 % for PsR293 and CtR107, respectively.

The autoNOE-Rosetta calculations converge for both targets (Fig. 3) to accurate structures that both superimpose with 1.9 ± 0.2 Å C_α-RMSD on to the respective X-ray structure. Assignment and structural statistics support that high-quality NMR solution structures have been obtained (Table 3). As shown in Fig. 3, multiple RASREC stages are required for both targets to converge to the correct fold. Accordingly, the C_α- RMSD to the reference structure improves stage to stage (Fig. 4), with the largest improvement concentrated on Phase I (stages I–IV). Phase II improves structural models for these targets, but Phase III (stage VII–VIII) yields a further refinement, which is necessary to reach atomic accuracy.

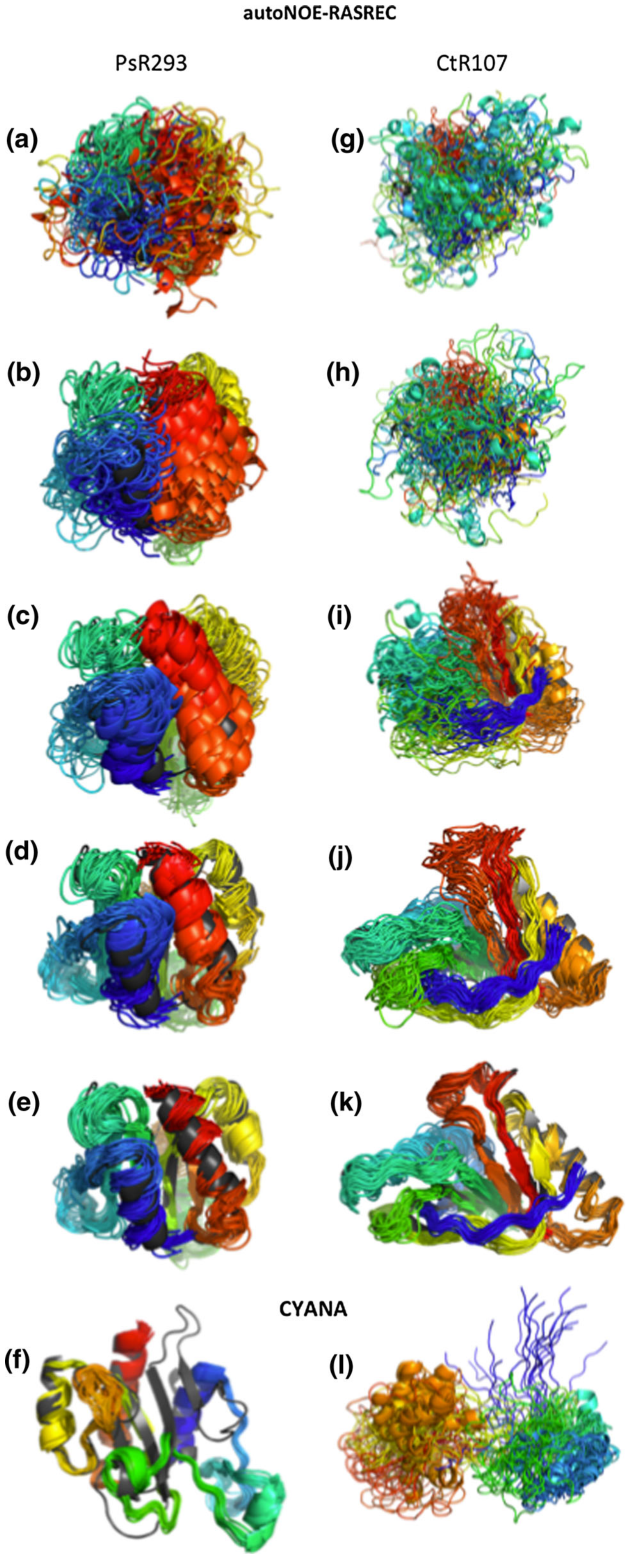
As pointed out above, the CANDID algorithm (Herrmann et al. 2002) is quite similar to the automatic NOE assignment employed here. CANDID has been the basis of popular programs for automatic NOE assignment, such as CYANA and UNIO. The main difference between autoNOE-Rosetta and CANDID is thus the different methods employed for structural sampling.

To illustrate the benefit from RASREC-ROSETTA sampling over torsional molecular dynamics, we have also run CYANA (Version 3.0) to calculate structures for PsR293 and CtR107 from the same input data. As shown in Fig. 4f, CYANA converges for PsR293 to a tight structural bundle (average backbone RMSD to mean: 0.6 Å) but the resulting structure is only accurate in parts, and the overall C_α-RMSD with respect to the Xray structure is with 10 Å rather high. Accordingly, typical quality criteria for successful CYANA calculations (P. Güntert, private communication) are borderline bad: after cycle 1, the target-function has a value of 211.83 and the average backbone RMSD to the mean of 3.4 Å (both rather high). For CtR107 (Fig. 4l), CYANA does not converge and yields an average backbone RMSD to the mean of 7.1 and 8.4 Å after cycle 7, when run with and without RDC data. As the CtR107 data set contained many zero-intensity peaks, which are ignored by auotNOE-Rosetta, we also tested whether the bad performance of CYANA resulted from these zero-intensity peaks and removed them manually. The calculation now converges in cycle 7 (cycle 1) to a bundle with an average RMSD of 2.2 Å (4.3 Å) to the mean structure and target-function values of 1.2 (25.9). The final structures, however, do not superimpose well with the reference structure (C_α-RMSD 7.1 ± 0.4 Å).

The results obtained with CYANA illustrate the difficulty of obtaining reasonable NOE assignments from the presented data sets with available methods. Apparently, the computational investment for the enhanced structural modeling of RASREC-ROSETTA has paid off and a high-quality structure can be obtained despite challenging input data. A more comprehensive benchmark of 50 data sets has been carried out subsequently, and its results confirm these observations on a broad basis (Zhang et al. 2014).

Analysis of NOE restraint quality at intermediate stages of AutoNOE-Rosetta calculations

To illustrate the progress in automatic NOE assignment during autoNOE-Rosetta, we have rescored all the decoys generated throughout the RASREC calculation with NOE restraints derived from assignments at its various stages. This procedure reveals the energy landscape generated by the restraints. Note, however, that the near-native conformations used here for rescoring are not yet sampled at early stages of RASREC but will be generated only in later stages (Fig. 4). Figure 5 shows the correlation between C_α-



◂**Fig. 3** Structural ensembles obtained with automatic NOE assignment. The reference structure (*dark gray*) is superimposed with NOE-based models depicted with a color gradient reflecting the sequence position from N-terminus (*blue*) to C-terminus (*red*). **a–e**, **g–k** Ensembles during autoNOE-Rosetta calculation of targets PsR293 (**a–e**) and CtR107 (**g–k**), respectively. Shown are (*top* to *bottom*) the 30 lowest energy conformations after RASREC stage I, stage II, stage III, stage IV and stage VIII, respectively. **f, l** Final CYANA models of PsR293 (**f**) and CtR107 (**l**), respectively. Note, that a different orientation than in (**a–e**) or (**g–k**), respectively, has been chosen for these panels

RMSD and NOE restraint energy for CtR107 with (right panels) and without (left panels) random restraint combination (section "Restraint Combination"). The corresponding plots for target PsR293 look very similar (Supp. Figure S1).

Remarkably, the initial restraints (panel a), show a good correlation with C_α-RMSD, and just selecting the lowest energy conformations would yield accurate structures (~2.0 Å). However, one also notes that the lowest energy is rather high with ~1,400 Rosetta energy units (REU). This high energy results from spurious NOE restraints that violate the correct conformation. The high offset in the energy illustrates the main problem with these initial restraints: If used for structure generation, they result in strong frustration and distortion. As is evident from Fig. 4, RASREC cannot generate conformations close to the lowest scoring decoys at ~2.0 Å RMSD in initial stages.

To avoid the strong frustration in the initial restraints we use restraint combination (Herrmann et al. 2002) for sampling in Phase I (Methods). Indeed, this drastically reduces the NOE restraint energy of the near-native conformations to ~100 REU (Fig. 5e–f). This reduction of distortive effects by factor 10–20 is sufficient for RASREC to converge towards near-native conformations by the end of stage IV (Fig. 4). To refine the structures further, however, random restraint combination is not sufficient and erroneous restraints have to be removed properly. Thus, in Phases II and III restraints are explicitly removed, if they are incompatible with the preliminary structures. This reduces the minimum energy to 5.1 and 1.5 REU, for Phase II and III, respectively (Fig. 5c–d).

Continuing to combine restraints even in Phases II and III would result in significant loss of discriminative power of the NOE restraint energy (Fig. 5g–h). Accordingly, we switch off restraint combination during Phases II–III and use the NOE restraints directly for sampling. An important observation from Fig. 5, is that the discrimination of conformations is best achieved with the initial assignments that were obtained without any feedback from structures. Accordingly, this set of restraints is used to select those conformations that are retained in the RASREC-pool for resampling. The gradual convergence towards the correct fold during RASREC stages I–IV (Figs. 3, 4), is thus mainly driven by the discriminative power of the initial NOE assignments.

Table 3 NMR and structural statistics for final autoNOE-Rosetta models of targets PsR293 and CtR107

	PsR293	CtR107
Peaks and assignments		
Picked	6,870	7,180
Zero intensity	0	3,400
Diagonal assignment	418	75
Assigned	3,658	2,762
With ≥5 initial assignments	2,497	2,271
Unassigned…	2,784	1,012
Without assignment possibility	1,925	208
Eliminated due to Network	6	6
Eliminated due to MinPeakVol	90	72
Eliminated due to MaxAssign	277	383
Eliminated due to DistViol….	486	343
Between 0.1 and 0.5 A	37	22
Between 0.5 and 2.0 A	130	116
Between 2.0 and 5.0 A	164	89
Above 5.0 A	155	116
Distance restraints		
Intraresidue	1,729	1,535
Sequential ($\|i - 1\| = 1$)	753	670
Medium-range ($1 < \|i - j\| \leq 4$)	462	145
Long-range ($\|i - j\| \geq 5$)	714	412
Other restraints		
Total HN RDCs	0	183
Dihedrals restraints (used as fragments)		
Residues with *good* talos prediction	98	117
Violations (RMSD, SD)		
Distance restraints (CING)	0.04 ± 0.02	0.08 ± 0.014
Ramachandran statistics (CING)		
Residues in most favored regions	89.9 %	92.8 %
Residues in allowed regions	10.0 %	6.9 %
Residues in generously allowed regions	0.0 %	0.3 %
Residue in disallowed regions	0.1 %	0 %
Average RMSD to mean structure (Å)		
Backbone atoms (CING)	0.71 ± 0.23	1.13 ± 0.22
Heavy atoms (CING)	1.08 ± 0.26	1.46 ± 0.19

Statistics calculated with the CING-Server (Doreleijers et al. 2012) are marked as such

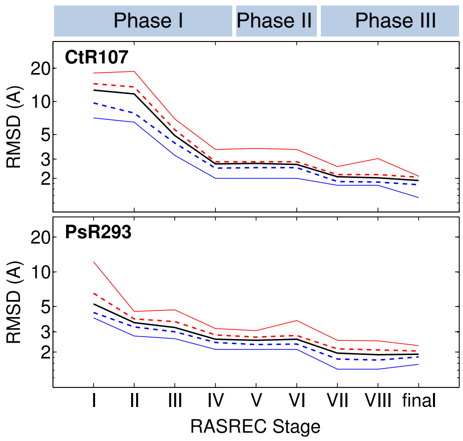


Fig. 4 Gradual improvement of preliminary models in autoNOE-Rosetta. Shown are the median (*black*), lower- and upper quartile (*dashed*, *blue* and *red*, respectively) and the lowest and highest (*solid*, *blue* and *red*, respectively) RMSD of the 30 lowest energy models in RASREC after stages I–VIII with respect to the reference structure. At *final* the RMSD statistics of the 10 lowest models selected purely by ROSETTA energy from the stage VIII RASREC pool (comprising 100 models) are shown

Conclusions

We have implemented an algorithm for automatic NOE assignment in ROSETTA and coupled it with the iterative conformational sampling method RASREC. RASREC has been shown previously to yield highly accurate structures even from sparse NMR data (Raman et al. 2010b; Lange et al. 2012), but required assigned NOEs (Raman et al. 2010a) or that an initial fold can be determined with CYANA (Lange et al. 2012). By tightly coupling automatic NOE assignment (as in CYANA) with RASREC we were able to unlock synergies between both approaches and make a significant step in automatic NOE assignment. Here, the method has been shown to yield accurate structures (1.9 Å) for two data sets of larger proteins, for which CYANA does not yield an accurate initial fold. Thus the method considerably extends the range of data sets for which automatic NOE structure determination leads to accurate results. In line with these results, a more comprehensive benchmark on 50 NOE data sets shows significantly improved performance for the new method, an improved accuracy compared to PDB-deposited NMR structures and good performance with automatically picked peak-lists (Zhang et al. 2014). The method is available within the ROSETTA3 software suite (www.rosettacommons.org) and will be released in versions 3.6 and higher. We recommend running autoNOE-Rosetta using the supporting tool-chain (CS-Rosetta toolbox version 2.0 or higher). The toolbox, additional documentation and user support can be found on the CS-Rosetta portal (www.csrosetta.org).

Methods

Automatic NOE assignment of PsR293 and CtR107

Fragments were picked by the Rosetta3 fragment picker (Vernon et al. 2013) using the chemical shift data from the BMRB. Homologous proteins using an e-value cutoff of 0.05 (sequence identity >20 %) were excluded from

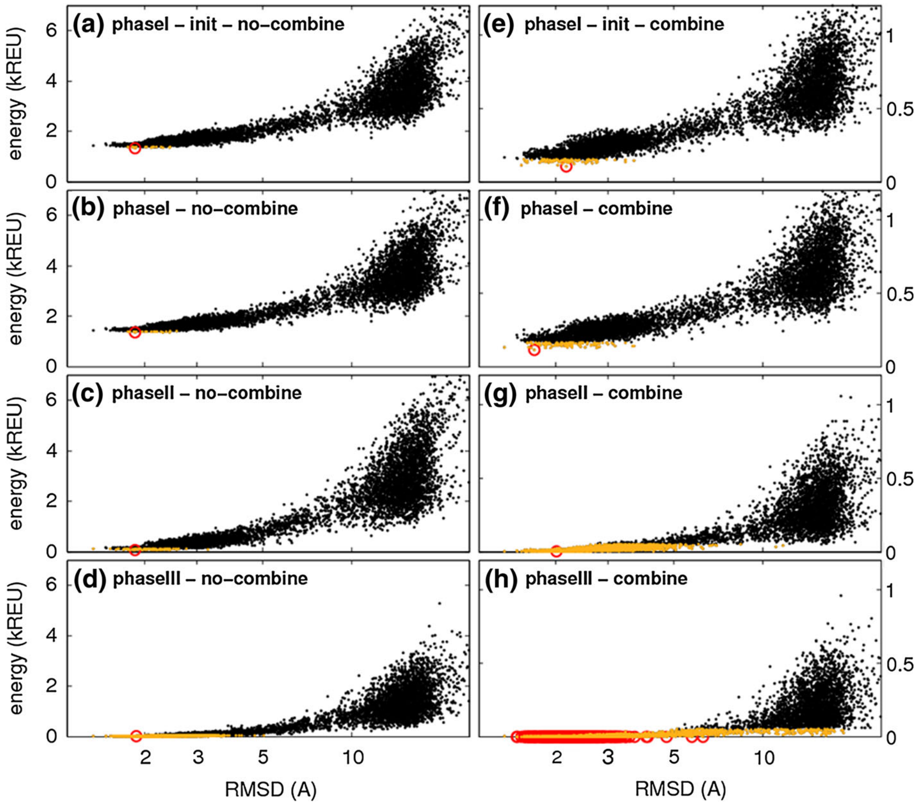


Fig. 5 Automatic NOE restraints obtained at different phases of the structure calculation with protein target CtR107 (Suppl. Figure S1 for target PsR293). All energies within 50 Rosetta Energy Units (REU) of the lowest energy are plotted in *yellow*. The lowest energy conformation is marked with a *red-circle* (multiple circles if minimum is degenerate). **a–d** restraints are used individually as in Phase II and Phase III. **e–h** In Phase I of the structure calculation, restraints are combined in random pairs for each individual decoy. To simulate this effect in this rescoring exercise, ten different sets with NOE restraints randomly paired are obtained, and the mean energy across the ten pairings is computed for each conformer

fragment picking. After removal of a C-terminal His-Tag, the target PsR293 (PDB accession 3h9x) was 118 residues long. For target CtR107 the first 8 residues, as well as the C-terminal His-Tag are removed due to flexibility (according to TALOS + computed RCI-S2 < 0.7). The remaining sequence from residue 9-155 of the NMR construct is 147 residues long. AutoNOE-Rosetta has been run on both targets with $P(\text{cst\_strength}) = 25$ using 4 HPC compute nodes equipped with four 2.6 GHz AMD Opteron processors (12 core) each. For CtR107, the structure calculation was completed in 5 h, and PsR293 in 2.5 h.

AutoNOE-Rosetta

RASREC structure calculations (Lange and Baker 2012) were run with a reduced pool-size of 100 conformers (command-line flag *-iterative:pool_size 100*) compared to the standard protocol (Lange and Baker 2012). This speeds up convergence considerably and reflects the reduced need for structural exploration when NOE data is present. *Recombination-Stages* were terminated when the acceptance ratio into the pool dropped below 10 % (-*iterative:accept_ratio* 0.1) and the cycle factor was set to 2.0(-*increase_cycles* 2). Chemical shift pseudo-energies contribute to RASREC pool evaluation with a weight of 5.0 (van der Schot et al. 2013). The original RASREC algorithm comprised of stages I–VI. The first four stages are run with Phase I parameters for NOE assignment. Subsequently, we re-run stages III and IV of the original RASREC algorithm with Phase II parameters, followed by stages V and VI with Phase III parameters. Renumbering the stages we thus get stages I–VIII as shown in Fig. 2.

Restraints in RASREC

Restraint combination

Restraints are combined into random pairs to avoid distortion due to spurious assignments as suggested previously (Herrmann et al. 2002). For each individual trajectory of conformational sampling (i.e., many thousand times), the restraint combination is re-randomized. To this end, restraints between residues i and j are first classified according to their sequence separation $|i - j|$ into restraints with $|i - j| < 5$, 20, and 50 and $|i - j| \geq 50$. Subsequently, pairs are drawn randomly from the restraints of the same sequence separation class and combined into a new ambiguous restraint. If the number of restraints of a sequence separation class is odd, the last remaining unpaired restraint is combined with one of the restraints already used in that class. For ambiguous restraints with multiple possible values for their sequence separation, one of the possible values is chosen at random.

Sequence separation

Restraints with a high sequence separation can frustrate sampling in Rosetta fragment assembly considerably (Rohl and Baker 2002). To mitigate this effect, all restraints with a sequence separation $|i-j| > \Delta_{SS}$ are switched off at first. Starting with $\Delta_{SS} = 3$ the threshold is ramped up during fragment assembly such that all restraints are activated only in the last 25 % of the sampling cycles of a fragment assembly trajectory. For ambiguous restraints with multiple possible values for their sequence separation, a sequence separation is chosen randomly for their classification each time a fragment assembly trajectory is started. Broken-chain fold-trees (Bradley and Baker 2006) are considered and an effective sequence separation is computed which reflects the shortest path between residue *i* and *j* of the restraint in the chosen broken-chain fold-tree (Lange and Baker 2012).

Redundancy removal

Automatic NOE assignment generates a large number of restraints, and we have observed a considerable slow-down of fragment assembly calculations in cases with large peak-lists. In autoNOE-Rosetta the number of restraints is thus drastically reduced during fragment assembly by removing redundancy with the following scheme:

A residue–residue contact matrix is initialized by setting elements to the following values: 2, for residue pairs with at least one unambiguous restraint between them, and 1, for residue pairs that are part of an ambiguous restraint and 0, for residue pairs that are not part of any restraint. Now restraints are drawn in random order without replacement. If an ambiguous restraint is drawn, it is accepted if any of the corresponding matrix elements are still 1. If accepted, the elements are set to 3. If an unambiguous restraint is drawn, it is accepted if the corresponding matrix element is either 2 or 3, and the corresponding matrix element is set to 4. This procedure makes certain that each residue pair is restrained by at most one unambiguous restraint and that each ambiguous restraint contains at least one sub-restraint that shares its residue pair with at most one other unambiguous restraint.

Acknowledgments I am indebted to the work of the Northeast Structural Genomics Consortium, which has made available NMR data sets for a large number of proteins. Furthermore, I have to thank Peter Güntert and Paolo Rossi for many helpful discussions. This work was supported by DFG grant LA 1817/3-1.

References

Bernard A, Vranken WF, Bardiaux B, Nilges M, Malliavin TE (2011) Bayesian estimation of NMR restraint potential and weight: a validation on a representative set of protein structures. Proteins 79:1525–1537

Bowers P, Strauss C, Baker D (2000) De novo protein structure determination using sparse NMR data. J Biomol NMR 18:311–318

Bradley P, Baker D (2006) Improved beta-protein structure prediction by multilevel optimization of nonlocal strand pairings and local backbone conformation. Proteins 65:922–929

Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS et al (1998) Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Cryst D 54:905–921

Doreleijers JF, Sousa da Silva AW, Krieger E, Nabuurs SB, Spronk CAEM, Stevens TJ, Vranken WF, Vriend G, Vuister GW (2012) CING: an integrated residue-based structure validation program suite. J Biomol NMR 54:267–283

Güntert P, Braun W, Wüthrich K (1991) Efficient computation of three-dimensional protein structures in solution from nuclear magnetic resonance data using the program DIANA and the supporting programs CALIBA, HABAS and GLOMSA. J Mol Biol 217:517–530

Herrmann T, Guntert P, Wüthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J Mol Biol 319:209–227

Huang YJ, Swapna GVT, Rajan PK, Ke H, Xia B, Shukla K, Inouye M, Montelione GT (2003) Solution NMR structure of ribosome-binding factor A (RbfA), a cold-shock adaptation protein from Escherichia coli. J Mol Biol 327:521–536

Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302:1364–1368

Lange OF, Baker D (2012) Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. Proteins 80:884–895

Lange OF, Rossi P, Sgourakis NG, Song Y, Lee H-W, Aramini JM, Ertekin A, Xiao R, Acton TB, Montelione GT et al (2012) Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. Proc Natl Acad Sci USA 109:10873–10878

Linge JP, Habeck M, Rieping W, Nilges M (2003) ARIA: automated NOE assignment and NMR structure calculation. Bioinformatics 19:315–316

Linge JP, Habeck M, Rieping W, Nilges M (2004) Correction of spin diffusion during iterative automated NOE assignment. J Magn Reson 167:334–342

Mao B, Guan R, Montelione GT (2011) Improved technologies now routinely provide protein NMR structures useful for molecular replacement. Structure 19:757–766

Nilges M (1993) A calculation strategy for the structure determination of symmetric dimers by 1H NMR. Proteins 17:297–309

Nilges M, Macias MJ, Odonoghue SI, Oschkinat H (1997) Automated NOESY interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from beta-spectrin. J Mol Biol 269:408–422

Raman S, Huang YJ, Mao B, Rossi P, Aramini JM, Liu G, Montelione GT, Baker D (2010a) Accurate automated protein NMR structure determination using unassigned NOESY data. J Am Chem Soc 132:202–207

Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini JM, Liu G, Ramelot TA, Eletsky A, Szyperski T et al (2010b) NMR structure determination for larger proteins using backbone-only data. Science 327:1014–1018

Rohl CA, Baker D (2002) De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. J Am Chem Soc 124:2723–2729

Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein structure prediction using Rosetta. Meth Enzymol 383:66–93

Shen Y, Lange OF, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A et al (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA 105:4685–4690

van der Schot G, Zhang Z, Vernon R, Shen Y, Vranken WF, Baker D, Bonvin AMJJ, Lange OF (2013) Improving 3D structure prediction from chemical shift data. J Biomol NMR 57:27–35

Vernon R, Shen Y, Baker D, Lange OF (2013) Improved chemical shift based fragment selection for CS-Rosetta using Rosetta3 fragment picker. J Biomol NMR 57:117–127

Warner LR, Varga K, Lange OF, Baker SL, Baker D, Sousa MC, Pardi A (2011) Structure of the BamC two-domain protein obtained by Rosetta with a limited NMR data set. J Mol Biol 411:83–95

Wüthrich K (1986) NMR of proteins and nucleic acids

Zhang Z, Porter J, Lange OF (2014) Robust and highly accurate automatic NOE assignment and structure determination with Rosetta. J Biomol NMR accepted