

# Robust structure-based resonance assignment for functional protein studies by NMR

Dirk Stratmann · Eric Guittet · Carine van Heijenoort

Received: 26 October 2009 / Accepted: 4 November 2009 / Published online: 19 December 2009  
© The Author(s) 2009. This article is published with open access at Springerlink.com

**Abstract** High-throughput functional protein NMR studies, like protein interactions or dynamics, require an automated approach for the assignment of the protein backbone. With the availability of a growing number of protein 3D structures, a new class of automated approaches, called *structure-based* assignment, has been developed quite recently. Structure-based approaches use primarily NMR input data that are not based on J-coupling and for which connections between residues are not limited by through bonds magnetization transfer efficiency. We present here a robust structure-based assignment approach using mainly  $H^N$ – $H^N$  NOEs networks, as well as  $^1H$ – $^{15}N$  residual dipolar couplings and chemical shifts. The *NOEnet* complete search algorithm is robust against assignment errors, even for sparse input data. Instead of a unique and partly erroneous assignment solution, an optimal assignment ensemble with an accuracy equal or near to 100% is given by *NOEnet*. We show that even low precision assignment ensembles give enough information for functional studies, like modeling of protein-complexes. Finally, the combination of *NOEnet* with a low number of ambiguous J-coupling sequential connectivities yields a high precision assignment ensemble. *NOEnet* will be available under: <http://www.icsn.cnrs-gif.fr/download/nmr>.

**Keywords** NMR · Assignment · Structure-based · NOE · Network · Chemical shifts · Residual dipolar couplings · *NOEnet*

## Introduction

Since the beginning of protein NMR, the automation of the tedious assignment process of the NMR spectra has been sought. Several factors make it difficult to automate the assignment process. For instance, as NMR data yield mainly ambiguous information for the assignment, it is difficult to guarantee to find always the correct assignment for all residues of a protein. Manual verification of the obtained result is often required, for example by inspecting visually the raw NMR spectra. Additionally, despite the large number of automation solutions proposed during the last 20 years none of them became a standard in the NMR community (reviews: (Moseley and Montelione 1999; Gronwald and Kalbitzer 2004; Baran et al. 2004; Altieri and Byrd 2004; Billeter et al. 2008; Williamson and Craven, 2009; Güntert 2009)). These difficulties are reflected by the fact that the majority of the proteins studied by NMR are still assigned manually, requiring often several weeks even for an experienced spectroscopist.

In parallel to the 3D structure determination of proteins (a sometimes lengthy task, specially for larger proteins), NMR demonstrated over the years its invaluable potential for functional studies, like protein–protein and protein–ligand interactions or protein dynamics. It is highly beneficial to automate the assignment of the backbone resonances for these studies that often only require the assignment of the backbone resonances, especially if they should be done in a high-throughput manner.

**Electronic supplementary material** The online version of this article (doi:10.1007/s10858-009-9390-3) contains supplementary material, which is available to authorized users.

D. Stratmann  
NMR, Utrecht University, Padualaan 8, 3584 CH Utrecht,  
The Netherlands

E. Guittet · C. van Heijenoort (✉)  
Centre de Recherche de Gif, Laboratoire de Chimie et Biologie  
Structurales ICSN-CNRS, 1, av. de la terrasse,  
91190 Gif-sur-Yvette, France  
e-mail: carine@icsn.cnrs-gif.fr

Thanks to the growing number of available 3D structures of proteins, the 3D structure of the protein investigated in a functional NMR study is often already known in its free form. Knowing the 3D structure before any assignment allows the spectroscopist to use alternative sensitive NMR experiments instead of triple resonance experiments for the assignment of the  $^{15}\text{N}$ - $^1\text{H}$  HSQC spectrum.

We showed recently (Stratmann et al. 2009) with the development of “NOEnet” that  $^1\text{H}^{\text{N}}-^1\text{H}^{\text{N}}$  NOE networks are valuable experimental constraints in *structure-based* assignment. Other structure-based assignment approaches (Dobson et al. 1984; Bartels et al. 1996; Gronwald et al. 1998; Bailey-Kellogg et al. 2000; Pristovsek et al. 2002; Pristovsek and Franzoni 2006; Hus et al. 2002; Erdmann and Rule 2002; Pintacuda et al. 2004; Langmead et al. 2004; Langmead and Donald 2004; Apaydin et al. 2008; Xiong and Bailey-Kellogg 2007; Xiong et al. 2008) require a combination of alternative data, like residual dipolar couplings (RDCs), chemical shifts (CS),  $^1\text{H}^{\alpha}-^1\text{H}^{\text{N}}$  NOEs, solvent accessibility or TOCSY data (see (Stratmann et al. 2009) for more details). To be able to extract assignment information from these alternative data sets, incomplete optimization algorithms are mainly used giving a limited number of solutions (often only one global assignment), with the drawback that their accuracy is difficult to assess. The alternative data sets used in structure-based assignment are usually too sparse or too ambiguous to yield one unique assignment solution for all peaks. By searching for a unique assignment solution, a high amount of assignment errors are consequently introduced. For example, the *contact replacement* approach (Xiong et al. 2008) yields only an accuracy of 60–80% with highly ambiguous data. For many of the alternative data sources their ambiguity increases with the number of residues. Probably because of this, none of the existing structure-based approaches has been tested on protein sizes above 200 residues using real experimental data. With NOEnet we showed for the first time that structure-based assignment is feasible on a protein size above 200 amino acids with an accuracy near to 100%.

A guarantee of high accuracy (near 100%) is crucial for the large adoption of automated assignment approaches. NOEnet was designed to tackle specifically this problem of accuracy, through an efficient complete search algorithm that yields all assignment solutions compatible with the input data in form of an assignment ensemble.

The sparseness and quality of experimental NOE data condition the size of the assignment ensemble obtained. A fraction of the  $^{15}\text{N}$ - $^1\text{H}$  HSQC peaks are uniquely assigned, while the others have multiple assignment possibilities. Fortunately, multiple assignment possibilities can be exploited in structure-based assignment. For example, the

set of assignment possibilities can be mapped onto the known 3D structure for each peak alone or for a group of peaks. This allows a visual inspection of the possible assignment zone. In order to quantify the extension of each assignment zone, we introduced a quality factor named *spatial assignment range* (SAR).

In this article, we investigate how additional data beside the NOE network can restrict the final assignment ensemble. To achieve this goal, we introduce a general filter approach that allows the inclusion of almost any type of input data without much efforts. We establish a parameter optimization protocol that allows a first test of the data quality and an optimal restriction of the assignment ensemble. We first investigate the case of  $^{15}\text{N}$  labeled proteins and the impact of  $^{15}\text{N}$  and  $\text{H}^{\text{N}}$  chemical shifts (CS) and  $^1\text{H}$ - $^{15}\text{N}$  residual dipolar couplings (RDC). The adding of  $^1\text{H}$ - $^{15}\text{N}$  RDC appears particularly effective. For the case of doubly  $^{15}\text{N}$ ,  $^{13}\text{C}$ -labeled proteins, the sole use of additional  $^{13}\text{C}_{\alpha}$ ,  $^{13}\text{C}_{\beta}$  and  $^{13}\text{CO}$  chemical shifts that can be obtained from the two triple resonance experiments CBCA(CO)NH and HNCO markedly improves the precision of the assignment ensemble. Finally, we show that the combined use of a  $^1\text{H}^{\text{N}}-^1\text{H}^{\text{N}}$  NOE network and highly ambiguous sequential connectivities allows an accurate, uniquely defined assignment, even for large proteins like EIN (259 amino acids).

## Methods

### Conceptual bases of NOEnet

NOEnet searches to assign the backbone resonances of the  $^{15}\text{N}$ - $^1\text{H}$  HSQC spectrum to the residues of the protein, from a known 3D structure of the protein and a network of unambiguous  $^1\text{H}^{\text{N}}-^1\text{H}^{\text{N}}$  NOEs. The main idea of NOEnet is to sample all possible matches of the whole available experimental NOE network onto the connectivity network of the 3D structure. In terms of graph theory, the algorithmic problem, which belongs to the class of NP-hard problems, is to find all possible *subgraph monomorphisms* or *graph matchings*. In opposition to algorithms that search one or several assignment solutions, NOEnet searches iteratively the assignment *impossibilities*, while ensuring in general that the correct assignment is not removed. At the beginning of the search, each peak can be assigned to any residue in the  $(n_{\text{peaks}} \times n_{\text{residues}})$  *assignment table*  $A$ . During the search, impossible peak assignments are removed from  $A$ . NOEnet makes several refinement cycles, returning each time an assignment ensemble in form of the assignment table  $A$ , which will have less assignment possibilities at each cycle. This approach allows the exploitation of the current result, even if the complete search is

still not finished. In general, the first cycle will return rapidly an assignment ensemble almost as good as the final assignment ensemble. More details about the algorithmic concepts realized in *NOEnet* can be found in (Stratmann et al. 2009). Here will be explained in detail, how additional data are handled by *NOEnet*.

The minimal input for *NOEnet* is a list of unambiguous  $^1H^N$ - $^1H^N$  NOEs and the 3D structure in the Protein Data Bank (PDB) file format. Unambiguous NOEs means that each NOE cross peak can be related to exactly two unambiguous resonances of the  $^{15}N$ - $^1H$  HSQC spectrum. Peaks with degenerated [ $^{15}N$ ,  $^1H^N$ ] chemical shifts have to be identified in advance (thanks to pH, salt or temperature variations) and removed from the set of peaks to assign. The NOE cross peaks, which can be related to more than two of the remaining HSQC peaks, are also excluded.

Beside the peaks of the ( $^{15}N$ ,  $^1H$ ) atom-pairs of the protein backbone, some peaks correspond to the ( $^{15}N$ ,  $^1H$ ) atom-pairs of side-chains. Especially, the tryptophan (TRP) side-chains generate ( $^{15}N$ ,  $^1H$ ) peaks, which are not distinguishable from the peaks corresponding to the backbone of the protein. We included the TRP side-chains as additional pseudo-residues. The peak pairs corresponding to  $NH_2$  groups of side-chains are assumed to be identified by their identical  $^{15}N$  frequency or from decoupled HSQC experiments, and were not included as assignment possibility.

#### Incorporation of additional data

If available, additional data should restrict the solution space further. We added  $^{15}N$  and  $^1H^N$  chemical shifts (CS) and  $^1H$ - $^{15}N$  residual dipolar couplings (RDC). The  $^{15}N$  and  $^1H^N$  chemical shifts are readily obtained from the  $^{15}N$ - $^1H$  HSQC spectrum without additional effort, whereas the measurement of RDCs requires a protein sample dissolved in a weak alignment medium (Bax and Grishaev 2005). If a doubly  $^{15}N$ ,  $^{13}C$ -labeled sample is available,  $^{13}C$  chemical shifts ( $^{13}C_\alpha(i-1)$ ,  $^{13}C_\beta(i-1)$ ,  $^{13}CO(i-1)$ ) can be obtained from two sensitive triple resonance experiments (CBCA(CO)NH and HNCO) and associated to the HSQC peak corresponding to residue  $i$ . These carbon chemical shifts are included in *NOEnet* in the same way as chemical shifts of the HSQC spectrum.

In order to include these secondary data, we implemented a general approach based on *filters*. The assignments, which are inconsistent with the constraints imposed by additional NMR data, are rejected during the search of assignment possibilities. The *filter consistency* is tested at each elementary step of the search for assignment possibilities. The filter approach allows a straightforward extension of *NOEnet* with any type of additional peak assignment constraint.

#### The chemical shift filter

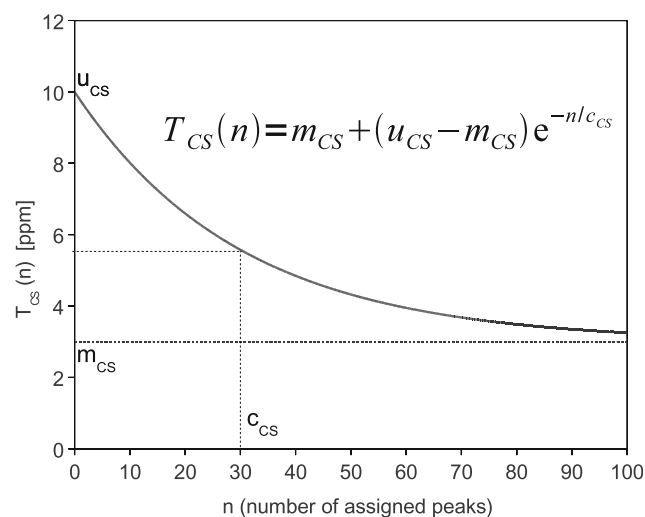
Experimental chemical shifts  $\delta^{exp}$  are converted into an assignment constraint, through comparison with theoretical CS values  $\delta^{theo}$  predicted from 3D structure by one of the available programs (Shen and Bax 2007; Neal et al. 2003; Xu and Case 2001; Meiler 2003). Each assignment of a peak  $j$  to a residue  $i$  can be evaluated by comparison of the corresponding CS values  $\delta_j^{exp}$  and  $\delta_i^{theo}$ . The general idea is to reject only assignments  $a = [a_1, a_2, \dots]$  of a set of peaks and not to reject a single peak assignment  $a_k = (peak_{j_k}, residue_{i_k})$ , since this could introduce errors in the assignment result due to the imperfections of the predicted values  $\delta^{theo}$ . The RMSD for a current assignment  $a = [a_1, a_2, \dots, a_n]$  is given by:

$$RMSD_a = \sqrt{\frac{\sum_{k=1}^n (\delta_{j_k}^{exp} - \delta_{i_k}^{theo})^2}{n}} \quad (1)$$

with  $n$  being the actual number of assigned peaks, i.e. the depth of the backtracking search (see (Stratmann et al. 2009) for more details). The current assignment  $a$  along the backtracking search is rejected if  $RMSD_a > T_{CS}(n)$ . The empirically chosen RMSD threshold function  $T(n)$  is defined as:

$$T_{CS}(n) = m_{CS} + (u_{CS} - m_{CS}) \cdot e^{-n/c_{CS}} \quad (2)$$

It decreases exponentially (see Fig. 1) from  $u_{CS}$  to  $m_{CS}$  with the size  $n$  of  $a$ , as the correct assignment of only a small number of peaks is likely to lead to a higher RMSD than the correct assignment of a high number of peaks. In



**Fig. 1** Threshold function  $T_{CS}(n)$  of the chemical shift filter. Only current assignments  $a = [a_1, a_2, \dots, a_n]$  with  $RMSD_a < T_{CS}(n)$  are accepted during the search for the assignment ensemble. The ‘upper’ threshold  $u_{CS}$  is usually fixed, while the minimum threshold  $m_{CS}$  and the decay constant  $c_{CS}$  are optimized. The threshold function  $T_{RDC}(n)$  of the RDC filter has the same shape. See text for more details

addition, the CS-filter is applied only if the current assignment  $a$  contains a minimum number of five peaks. The  $^{15}\text{N}$  and  $^1\text{H}^{\text{N}}$  chemical shifts have a range of about 30 and 3 ppm respectively, but the  $^1\text{H}^{\text{N}}$  shifts are less well predicted and require therefore proportionally higher  $u_{\text{CS}}$  and  $m_{\text{CS}}$  parameter values. The parameters  $u_{\text{CS}}$  and  $m_{\text{CS}}$  are set for the  $^1\text{H}^{\text{N}}$  chemical shifts to 2 and 0.8 ppm, respectively. For the  $^{15}\text{N}$  chemical shifts, the value of  $u_{\text{CS}}$  remains fixed to 10 ppm, while  $m_{\text{CS}}$  is optimized. The decay constant  $c_{\text{CS}}$  is the same for both nuclei and is also optimized. For  $^{13}\text{C}$  chemical shifts,  $u_{\text{CS}}$  is fixed to the same value as for  $^{15}\text{N}$  chemical shifts (10 ppm), while  $m_{\text{CS}}$  and  $c_{\text{CS}}$  are optimized independently of the values for  $^{15}\text{N}$ .

A general optimization protocol is presented in the “Parameter optimizations” section.

### The use of RDC data during the search

The prediction of theoretical RDC-values  $D^{\text{theo}}$  not only needs the 3D structure but also a good estimate of the alignment tensor  $\mathbf{A}$ . This can only be obtained from a set of at least five (in practice >15) experimental RDC-values  $D^{\text{exp}} = \{D_1^{\text{exp}}, D_2^{\text{exp}}, \dots, D_n^{\text{exp}}\}$  of NHs whose assignments are unique. The alignment tensor that yields the best fit between  $D^{\text{exp}}$  and  $D^{\text{theo}}$  of the peaks having a unique assignment is obtained by SVD (Losonczi et al. 1999). If the minimum number of uniquely assigned peaks is not reached, the RDC data are converted into a *temporary* assignment constraint: the idea is to calculate an alignment tensor  $\mathbf{A}_{a_{\text{initial}}}$  using the initial assignment  $a_{\text{initial}}$  found during the search for the first 15 peaks. Temporary theoretical RDC-values  $D_{a_{\text{initial}}}^{\text{theo}}$  can be calculated for all residues from the 3D structure with  $\mathbf{A}_{a_{\text{initial}}}$ . The comparison of  $D_{a_{\text{initial}}}^{\text{theo}}$  with the experimental RDCs  $D^{\text{exp}}$  constrains the search for the remaining peaks connected by the NOE network. If  $a_{\text{initial}}$  is incorrect, RDC-data evaluated using  $\mathbf{A}_{a_{\text{initial}}}$  will constrain the remaining peaks in an erroneous way. The goal is to find more rapidly if  $a_{\text{initial}}$  is incorrect. In average it should be more difficult to find an assignment of all peaks of the NOE network that satisfies the erroneous RDC-constraints and the network-properties simultaneously. In this sense, the addition of the RDC-data should help to prune more rapidly the search tree. If initial assignment  $a_{\text{initial}}$  is correct, the RDC-data in couple with a correct alignment tensor  $\mathbf{A}_{a_{\text{initial}}}$  should filter out assignment possibilities, which would be allowed by the NOE network constraints alone.

A *permanent* alignment tensor  $\mathbf{A}_{\text{unique}}$  is calculated once a sufficient number of uniquely assigned peaks is reached (in practice  $n \geq 15$ ).  $\mathbf{A}_{\text{unique}}$  is updated with every new unique peak assignment. With the availability of  $\mathbf{A}_{\text{unique}}$ , the *temporary* RDC assignment constraints are replaced by *permanent* constraints, independent of the temporary

assignment  $a$ . Having the predicted RDCs  $D^{\text{theo}}$  and the measured RDCs  $D^{\text{exp}}$ , the comparison  $D^{\text{theo}} \leftrightarrow D^{\text{exp}}$  is done in the same way as with the CS data, employing the RMSD threshold approach explained in the previous paragraph with:

$$\text{RMSD}_a = \sqrt{\frac{\sum_{k=1}^n (D_{j_k}^{\text{exp}} - D_{i_k}^{\text{theo}})^2}{n}} \quad (3)$$

As for the CS data, the threshold function is a decreasing exponential with the number of currently assigned nodes  $n$ :

$$T_{\text{RDC}}(n) = \Delta T_{\text{RDC}}(n_{\text{unique}}) + m_{\text{RDC}} + (u_{\text{RDC}} - m_{\text{RDC}}) \cdot e^{-n/c_{\text{RDC}}} \quad (4)$$

For a typical value range of  $D^{\text{exp}}$  of  $\pm 30$  Hz for NH RDCs,  $u_{\text{RDC}}$  is set to 10Hz.  $c_{\text{RDC}}$  and  $m_{\text{RDC}}$  are optimized (see “Parameter optimizations” section).  $T_{\text{RDC}}(n)$  is increased by an empirical additional margin:

$$\Delta T_{\text{RDC}}(n_{\text{unique}}) = 2 \times (1 - n_{\text{unique}}/N_{\text{peaks}}) \quad (5)$$

It takes into account that the quality of the *permanent* alignment tensor increases with the number of unique assignments  $n_{\text{uniques}}$ .

### Response to erroneous constraints

Erroneous constraints yield incompatibilities in the constraint framework. If the constraint framework is dense enough, an incompatibility can leave some strongly constrained peaks with no assignment possibility. This generates *holes* in the list of peak assignment possibilities. The occurrence of a hole along the matching process indicates that there must be one or more erroneous constraints in the data set. Inversely, if every peak has at least one assignment possibility at the end of the matching process, it is highly improbable to have an error in this result. Erroneous constraints can be caused by all data sources. Erroneous NOEs can be caused by the use of too small distance thresholds in the theoretical graph built from the 3D structure, artifacts from the NOESY spectra or large differences between the reference tridimensional structure and the structure of the protein in solution. For CS- and RDC-data, the choice of a too tight RMSD threshold will also cause the removal of correct assignment possibilities and in most cases the appearance of holes in the assignment list.

### Parameter optimizations

Beside the experimental data and the 3D structure, NOE $_{\text{net}}$  needs a set of threshold parameters ( $d_{\text{max}}^{\text{theo}}$ ,  $T_{\text{NOE}}$ ,  $\Delta d$ ,  $c_{\text{CS}}$ ,  $m_{\text{CS}}$ ,  $c_{\text{RDC}}$ ,  $m_{\text{RDC}}$ ) for the interpretation of the input data (see Table 1). The search for appropriate thresholds should

**Table 1** Parameters required for NOE<sub>net</sub>

Name	Description	Typical values
NOE-data		
$d_{max}^{theo}$	maximum $^1H^N-^1H^N$ distance in 3D structure, (7, 6, 5 Å) for weak, medium and strong NOEs one value for each NOE intensity class	
$T_{NOE}$	number of allowed outliers	1–10
$\Delta d$	outlier range	1–1.5 Å
CS-data		
$c_{CS}$	decay constant of RMSD filter	10–30 residues
$u_{CS}$	upper RMSD threshold	10 ppm for $^{15}N$ and $^{13}C$ ; 2 ppm for $^1H^N$
$m_{CS}$	minimum RMSD threshold	3 ppm for $^{15}N$ , 0.8 ppm for $^1H^N$ and 1.5 ppm for $^{13}C$
RDC-data		
$c_{RDC}$	decay constant of RMSD filter	10–30 residues
$u_{RDC}$	upper RMSD threshold	10 Hz
$m_{RDC}$	minimum RMSD threshold	3 Hz

begin with common values, which were experimentally found valid for most cases and which are scaled to the experimental conditions like the mixing time in the NOESY experiment. Three cases can occur: (1) Some of the thresholds are too tight, causing assignment errors. In most cases, NOE<sub>net</sub> produces an assignment list with holes. (2) The thresholds are not tight enough, causing an explosion of the search space. This can be seen by a poor convergence causing a high number of stopped possibilities (i.e. searches temporarily stopped after a given number of trials). (3) The thresholds are in an optimal range, NOE<sub>net</sub> will return relatively rapidly (some minutes to some hours, depending on the protein size and the amount of data) a well constrained result, if the given input data are sufficient.

The different input data (NOE, CS, RDC,...) and their corresponding thresholds should be tested sequentially. The general protocol for the parameter optimization is shown by the flowchart in Fig. 2. NOE<sub>net</sub> uses the notion of outliers, i.e. NOE cross peaks corresponding to distances shorter than the upper theoretical distance  $d_{max}^{theo}$  by at most a given value  $\Delta d$ . The number of allowed outliers is given by  $T_{NOE}$ . The NOE thresholds ( $d_{max}^{theo}$ ,  $T_{NOE}$ ,  $\Delta d$ ) are optimized first, then the parameters for CS and RDC data, if used.

The OPTIMIZE procedure in Fig. 2 can be done in a parallel fashion by running several trials with different parameter values simultaneously. It consists in running several trials with NOE<sub>net</sub> to optimize the threshold parameter  $x$ . If  $x$  is too tight, holes are likely to occur in the assignment list, requiring to relax  $x$ . If the assignment list contains no hole,  $x$  can be tightened. At each change of  $x$ , either the tightest successful value of  $x$  is saved in  $x_{opt}$  or the most relaxed unsuccessful value of  $x$  is saved in  $x_{hole}$ . The tests  $x = x_{hole}?$  and  $x = x_{opt}?$  prevent from retesting the same values or value ranges of  $x$  twice.

By default, a first set of theoretical distance thresholds  $d_{max}^{theo} = (5, 6, 7 \text{ \AA})$  for short, medium and long distances, respectively, is used. If this is not successful, the long distances threshold is increased by 0.5 Å, as the corresponding weak NOEs could more likely correspond to spin diffusion than the medium or strong NOEs. The NOE outlier parameters are optimized first, their initial default value are:  $T_{NOE}^{init} = 3$  and  $\Delta d^{init} = 1 \text{ \AA}$ . If available, CS and RDC data are included with loose thresholds ( $c_{CS/RDC} = 30$  and  $m_{CS} = 3.5 \text{ ppm/mRDC} = 3.5 \text{ Hz}$  (for a  $D^{exp}$  range of  $\pm 30 \text{ Hz}$ )). The OPTIMIZE function returns the optimal value  $T_{NOE}^{opt}$ . If  $T_{NOE}^{opt} = 1$ , then the  $\Delta d$  threshold can be tightened (i.e. increased).  $T_{NOE}^{opt} = -1$  indicates that even with high  $T_{NOE} > 15$  values, no hole free assignment ensemble could be obtained by NOE<sub>net</sub>.  $d_{max}^{theo}$  has to be increased in this case as described above.

Once the NOE parameters optimized, the CS and RDC parameters are also optimized sequentially using the optimal NOE parameters.

Typically about 20 trials have to be done to optimize all parameters. Several of these trials are very fast (some minutes) due to the occurrence of holes, while some trials can take several hours depending on the parameter values and the data quality.

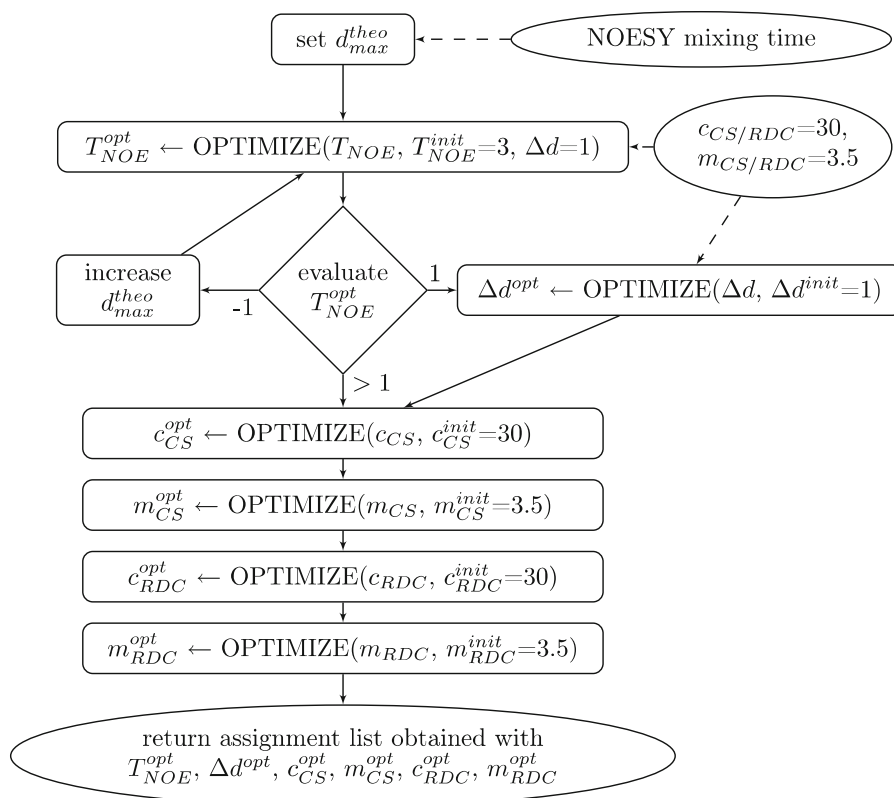
Tables S1 and S2 show all the trials performed for the thresholds optimization on lysozyme with realistic simulated and experimental NOE data, respectively. The trials for EIN are shown in Table S3.

## Results analysis

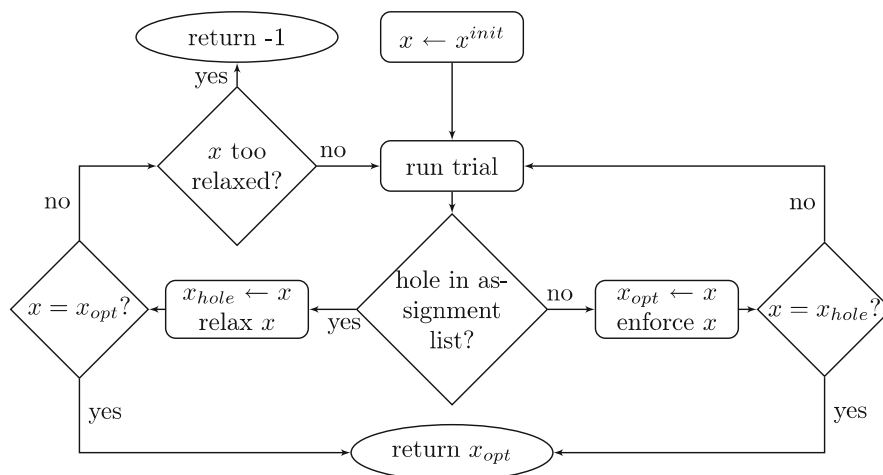
### Spatial assignment range (SAR)

As NOE<sub>net</sub> does not search for a unique assignment for all peaks, but for an *assignment ensemble* compatible with the

**Fig. 2** Flowchart of the parameter optimization protocol. First are optimized the NOE thresholds  $d_{max}^{theo}$  (maximum  ${}^1H^N-{}^1H^N$  distance in 3D structure, one value for each NOE intensity class),  $T_{NOE}$  (number of allowed outliers) and  $\Delta d$  (outlier range in Å). Then are optimized the CS thresholds  $c_{CS}$  (decay constant of RMSD filter) and  $m_{CS}$  (minimum RMSD threshold) and finally the RDC thresholds  $c_{RDC}$  and  $m_{RDC}$ . See text for further explanations



OPTIMIZE( $x, x^{init}$ ) procedure:



input data, the precision and accuracy of the ensemble can be quantified. The accuracy is defined in this context as:

$$\text{Accuracy} = 1 - \frac{N_e}{N_p} \quad (6)$$

with  $N_e$  the number of HSQC peaks that do not have the correct assignment in their list of assignment possibilities and  $N_p$  the number of peaks. An accuracy of 100% means that the assignment ensemble contains among other compatible assignments also the correct assignment. The precision of the assignment ensemble is defined in terms of

*completeness*. We define two types of completeness: First the unicity completeness describing the ratio of the number of uniquely and correctly assigned peaks to the total number of peaks:

$$C_1 = \frac{N_{unique}}{N_p}. \quad (7)$$

Second, the peaks with multiple assignment possibilities can be classified by a quality factor obtained with the available 3D structure of the protein. To obtain this quality factor, we calculate the inter-residue spatial  ${}^1H^N-{}^1H^N$

distances for all residue pairs taken from the peak assignment possibilities for a specific peak. We define the *spatial assignment range*(SAR) as the maximum of those distances and calculate it for each peak. The idea is that studies that do not require exact positioning in the 3D structure, as for example chemical shift perturbation studies for protein–protein interactions, can also exploit the peaks that are not uniquely assigned, but that have a small SAR value. We thus define a second type of completeness: the ratio between the number of peaks with a SAR-value below a given threshold (typically 10 Å) and the total number of peaks:

$$C_2(<10 \text{ \AA}) = \frac{N_{\text{SAR}<10\text{\AA}}}{N_P} \quad (8)$$

The uniquely assigned peaks are given a SAR-value of zero.

Compared to the inherent uncertainty of chemical shift perturbation data, we estimate that 10 Å is a reasonably conservative threshold. A recent study compared a large number of methods to translate chemical shift perturbation data into a predictor of the interfacial residues (Schumann et al. 2007). The results have been expressed in terms of Matthews correlation coefficients (MCC), defined as:

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{((TP + TN)(TP + FP)(TN + FP)(TN + FN))}} \quad (9)$$

TN: True Negative, TP: True Positive, FN: False Negative, FP: False Positive.

Testing 15 methods on 4 complexes, the authors obtained a large range of MCC values from 0.14 to 0.82. The average value (about 0.5) is quite far from a perfect predictor which would have a MCC value of 1.0. This shows for the first time quantitatively the high uncertainty of a chemical shift perturbation based predictor for interfacial residues. Chemical shift perturbation does not allow therefore the delineation of the interaction site with an atomic precision, but rather with a precision around 10 Å. The delineation of interaction surfaces from chemical shift perturbations (CSPs) is thus not necessarily more precise when using a unique, ‘atomic-resolution’ assignment of chemical shifts than when using an assignment ensemble with 10 Å resolution (i.e. spatial assignment range SAR <10 Å). Actually, we show in the results section that the interaction site prediction on the EIN-HPR complex is of a good quality, using the assignment ensemble with SAR values up to 30 Å.

#### Individual peak assignment refinement

The CS- and RDC-filters are routinely applied on assignments of at least five peaks, in order to prevent the

introduction of assignment errors due to outlier values of single peaks in the CS or RDC data. As done by other approaches, like the NVR algorithm (Langmead et al. 2004; Langmead and Donald 2004; Apaydin et al. 2008), assignment possibilities can also be restricted for individual peaks, yielding a higher reduction in assignment possibilities with the risk of introducing assignment errors. This risk decreases with the number of independent data available for each peak. We thus also implemented in NOEnet a refinement procedure of the assignment list operating on individual peaks. In order to minimize the risk of introducing assignment errors, solely assignment possibilities of peaks for which at least three CS/RDC data values are available can be removed. The refinement procedure uses a combined cost of the available CS ( $\delta^{exp}$ ) and RDC ( $D^{exp}$ ) data for each peak assignment possibility (peak  $i$  to residue  $j$ ):

$$C(i, j) = \frac{1}{N_{data}} \sum_{k=1}^{N_{data}} w_k |x_{k,i}^{exp} - x_{k,j}^{theo}|$$

*data*  $x_k$  = CS, RDC;  $w_k$ - normalization factors (10)

For CS data, the normalization factors  $w_k$  are set according to the expected RMSD error  $E^{RMSD}$  of the employed prediction program ShiftX (Neal et al. 2003) between experimental and predicted chemical shifts of a given atom type  $k$ :

$$w_k = 1/E_k^{RMSD} \quad (11)$$

The absolute value range of the chemical shifts of a given atom type is not the sole determinant for  $w_k$ , as differences between predicted and experimental chemical shifts are normalized and these differences depend also on the relative prediction accuracy. For example, the relative prediction accuracy is better for  $^1H^\alpha$  compared to  $^1H^N$ . Both, the absolute value range and the relative prediction accuracy determine the  $E_k^{RMSD}$  values. The  $E_k^{RMSD}$  values are taken from the ShiftX article—Table 10 (Neal et al. 2003) (validation data set):  $^{15}N$ : 2.53 ppm,  $^1H^N$ : 0.52 ppm,  $^{13}C_\alpha$ : 1.02 ppm,  $^{13}C_\beta$ : 1.10 ppm,  $^{13}CO$ : 1.17 ppm.

The RMSD error of RDC data is here estimated to be  $E_{RDC}^{RMSD} = 2 \text{ Hz}$  and the RDC normalization factor is set accordingly:  $w_{RDC} = \frac{1}{E_{RDC}^{RMSD}} = \frac{1}{2\text{Hz}}$ .

The refinement procedure removes iteratively peak assignment possibilities with high costs, until a hole occurs in the assignment list. The result of the iteration step before the occurrence of a hole is returned to minimize further the risk of introducing assignment errors.

#### Calculations

The runtimes indicated in the results section and in the Supporting Information correspond to the use of a single

core of a Intel Xenon Woodcrest CPU at 2.66 GHz with 1 gigabyte of RAM. Although *NOE<sub>net</sub>* is not programmed in a parallel manner, several cores or CPUs can be useful to test several parameters in parallel.

## Figures

The figures of protein structures in this article were prepared with the program MOLMOL (Koradi et al. 1996).

## Datasets

In this article we assume that all NOE data sets (simulated or experimental) could have been obtained by a 4D NOESY experiment. We thus removed from the NOE data sets all NOEs, which involve an ambiguous [ $^{15}\text{N}$ ,  $^1\text{H}^{\text{N}}$ ] HSQC peak, defined by the tolerance distances [tolN, tolH] equal to [0.2 ppm, 0.02 ppm].

The description and analysis of the simulated datasets are given in the supplementary material.

### Experimental data for Lysozyme

The  $H^{\text{N}}-H^{\text{N}}$  NOE-data set deposited in the PDB (1E8L) (Schwalbe et al. 2001) contains 190  $H^{\text{N}}-H^{\text{N}}$  NOE-constraints obtained from a 3D NOESY-HMQC experiment (Schwalbe et al. 2001). Ambiguous NOEs were removed as described above reducing the number of NOEs to 169. The average number of NOEs per residue is  $r = 169/132 = 1.3$ . The NOE classification given by (Schwalbe et al. 2001) was used. The  $^{15}\text{N}$  and  $^1\text{H}$  chemical shifts were considered (taken from the BMRB: bmr4831.str (Schwalbe et al. 2001) for  $^{15}\text{N}$ -CS and bmr4562.str (Wang et al. 2000) for  $^1\text{H}$ -CS) as well as the  $^1\text{H}-^{15}\text{N}$  RDC data in two alignment media (Schwalbe et al. 2001). Theoretical chemical shifts were predicted using ShiftX (Neal et al. 2003). The X-ray structure 193L (PDB code) (Vaney et al. 1996) was used as the reference 3D-structure to build the theoretical graph  $G^{\text{theo}}$  and to predict the theoretical CS and RDC values.

### Experimental data for EIN

The structure of the 28 kDa protein EIN has been determined by X-ray crystallography (PDB 1ZYM (Liao et al. 1996)) and NMR (PDB 1EZA (Garrett et al. 1997a)). The RMSD of the backbone heavy atoms between 1ZYM and 1EZA is equal to 1.55 Å. A large number of NMR experiments have been recorded on EIN (Garrett et al. 1997a), especially a 4D  $^{15}\text{N}/^{15}\text{N}$ -separated NOESY experiment on perdeuterated EIN with a mixing time of 170 ms (Garrett et al. 1997a) and a 3D  $^{15}\text{N}$ -separated NOESY with

a mixing time of 100 ms (Garrett et al. 1997a). The two experiments permitted the extraction of 555  $H^{\text{N}}-H^{\text{N}}$  NOE-constraints (PDB 1EZA). Since the X-ray structure is truncated at the C-terminal end by 10 residues, we removed by hand the NOE-constraints involving residues 250–259, which left 535 out of the 555  $H^{\text{N}}-H^{\text{N}}$  NOE-constraints. Removal of ambiguous NOEs reduced the number of NOEs to 407. The NOE data completeness is higher for EIN than for lysozyme (average number of NOEs per residue  $r = 407/250 = 1.6$ ). Experimental  $^{15}\text{N}$  and  $^1\text{H}^{\text{N}}$  chemical shifts (Garrett et al. 1997a) were included systematically as assignment constraint. Theoretical chemical shifts were predicted using ShiftX (Neal et al. 2003) on the X-ray structure 1ZYM. Carbon chemical shifts ( $^{13}\text{C}_{\alpha}(i-1)$ ,  $^{13}\text{C}_{\beta}(i-1)$ ,  $^{13}\text{CO}(i-1)$ ) were taken from the data set included in the distribution of MARS (Jung and Zweckstetter 2004). The results that make use of this carbon chemical shift data set are labeled with ‘CS<sub>carbon</sub>’, while ‘CS’ indicates the use of  $^{15}\text{N}$  and  $^1\text{H}^{\text{N}}$  chemical shifts.

NOE constraints were classified in three classes (strong, medium and weak). Crosspeaks, which appear only in the 4D NOESY with a long mixing time of 170 ms and which have an intensity below 11% of  $I_{\text{max}}(4\text{D})$  were classified as weak. All crosspeaks from the 3D NOESY with a mixing time of 100ms were classified as strong, if their intensity was greater than 14% of  $I_{\text{max}}(3\text{D})$ . All other crosspeaks were classified as medium. The 407 experimental NOEs finally consisted in 36 strong, 208 medium and 163 weak NOEs.

No RDC are available for EIN free in solution. In order to test the impact of RDCs, we simulated two sets of  $^1\text{H}-^{15}\text{N}$  RDCs that would be obtained in two independent alignment media by using two different alignment tensors. The X-ray structure of the free form of EIN (PDB 1ZYM (Liao et al. 1996)) was used to calculate the theoretical RDC values sets  $D_{\text{theo}}(1)$  and  $D_{\text{theo}}(2)$  for two different alignment tensors. A gaussian random error of  $\sigma = 1$  Hz was added to each data set for the simulated data sets:

$$\begin{aligned} D_{\text{sim}}(1) &= D_{\text{theo}}(1) + \varepsilon_{\text{gaussian}} \\ D_{\text{sim}}(2) &= D_{\text{theo}}(2) + \varepsilon_{\text{gaussian}} \end{aligned} \quad (12)$$

## Results

### Introduction

We first analyzed the results of *NOE<sub>net</sub>* for the medium size protein lysozyme using various simulated data sets, from ideal to more realistic ones (see supplementary material for details). This allows us to investigate more generally the impact of experimental data sets features, such as NOE sparseness and addition of chemical shifts and RDCs, on the capability of the 3D-structure-based assignment method.



The optimization of NOE $net$  parameters and the results are detailed in the supplementary material. The use of an ideal NOE data set (same crystal structure used for simulated data and reference structure, same thresholds for calculated NOE distances and crystal structure distance network) yields an assignment ensemble with a unique assignments completeness of 95% and an accuracy of 100%. This first shows that there exists almost only one possibility to match graph  $G^{exp}$  onto  $G^{theo}$  if the two graphs are identical. However, even in that case, it should be noted that 5% of the peaks still have several assignment possibilities. These peaks all have a SAR value below 10 Å and the multiple assignments mostly correspond to swaps in helices. As soon as more realistic data sets are used (different thresholds or/and different crystal structures for NOE data simulation and 3D-structure graph), the number of uniquely assigned peaks drops considerably. It appears then essential to use both NOEs classification (short, medium and long distances) and NOE outliers (restricted number of distances in the upper limit range) to reduce the matching possibilities down to a usable degree.

#### Sparse experimental NOE data in combination with RDC data on lysozyme

The sparseness and fragmentation of the NOE experimental network considerably degrades the quality of the assignment ensemble (case 1 of Table 2). The peaks that are uniquely assigned or that have a SAR below 10 Å are all localized in the larger experimental NOE network (black and violet/blue zone in Fig. 3a). The presence of small, disconnected NOE networks precludes here the obtaining of low SAR assignments (see (Stratmann et al. 2009)). Adding CS helps to reduce the SAR values, without increasing significantly the number of unique assignments, whereas adding CS and RDCs reduces considerably the assignment ambiguity in all NOE-networks (Fig. 3, cases 2 and 3 in Table 2). The NH bonds orientation appears to be here a key complementary information to  $H^N-H^N$  distances. RDCs can thus be efficient to circumvent the problem of fragmentation of the NOE networks. The application of the individual peak refinement procedure

(see “Methods”) applied on the assignment ensemble obtained using NOE, CS and RDC data happens to be highly efficient here (case 4, Table 2). It considerably improves the precision of the assignment (unicity completeness  $C_1$  increased from 44 to 73% and  $C_2$  (<10 Å) from 75 to 83%) without degrading its accuracy.

#### Error detection

A crucial point of any automated assignment method is its capability to assess its accuracy. In the case of NOE $net$ , the appearance of holes in the assignment list clearly brings to the fore the presence of inconsistencies between experimental data and structure-derived constraints. However, it can happen that the correct assignment for a peak is removed from the list without the appearance of a hole, as seen in the case of EIN in which two assignments are swapped, or in the case of lysozyme when RDCs are added without modifying the number of NOE outliers (see Tables S1 and S3). A way to test the capability of NOE $net$  to detect assignment errors is to introduce experimental constraints that are inconsistent with the 3D-structure. For that, we generated biased experimental data sets, which comprised the experimental data plus an increasing number of randomly simulated NOEs whose classification was incorrect (see Fig. 4). A wrong classification is only one possible type of inconsistencies using NOE data, but this example should demonstrate the general properties of the error detection in NOE $net$ . We then defined the *success rate of error detection* of the program as  $N_{hole}/N_{error}$ , with  $N_{hole}$  being the number of runs for which holes occurred in the assignment list and  $N_{error}$  the total number of runs yielding assignment errors (with holes or not) due to the badly classified NOEs. One run consisted in the random generation of badly classified NOEs and the subsequent search for assignment possibilities by NOE $net$  on this erroneous dataset. In order to obtain precise estimations of the average values of the error detection success rate, we performed 8293 runs overall. Despite the use of erroneous datasets, a small number of runs did not yield any assignment error and therefore also no hole. These runs were not counted in  $N_{error}$ . In the case of lysozyme shown in the

**Table 2** Lysozyme with experimental NOE data set from BMRB

Case	Trial # (Table S2)	$N_{peaks}$	Runtime	Additional data	$N_{NOEs}$	$d_{max}^{theo}$ (Å)	$N_{dist}$	$\frac{N_{unique}}{N_{peaks}}$ (%)	$\frac{N_{SAR < 10\text{Å}}}{N_{peaks}}$ (%)	Accuracy (%)
1	6	132	5 h		169	7	528	17	23	100
2	19	132	10 min	CS	169	7	528	20	33	100
3	31	132	30 s	CS + RDC	169	7	528	44	75	100
4	31	132	<1 s	Refined	169	7	528	73	83	100

$N_{NOEs}$  experimental unambiguous NOEs have been used. The X-ray structure 193L was used, yielding  $N_{dist}$  distances below  $d_{max}^{theo}$ . NOE classes and NOE outliers have been used. For case 4, see paragraph “Individual peak assignment refinement” in methods

**Fig. 3** Assignment results on lysozyme using experimental data obtained by (Schwalbe et al. 2001). **a** Results represented on the NMR structure 1E8L (Schwalbe et al. 2001), using NOE data only (left, case 1 of Table 2), NOE + CS data (middle, case 2 of Table 2), NOE + CS + RDC data (right, case 3 of Table 2). The black lines on the left structure correspond to the experimental NOEs network. Proline residues are shown in gray. The color code represents the spatial assignment range (SAR), as depicted in the colorbar below the structures. Unique assignments are shown in black. **b** Spatial assignment range (SAR) for each peak and for each case. The peaks are ordered by increasing SAR values. The SAR-value of 10 Å that has been chosen as maximum for the class of *exploitable* peaks is depicted (dashed red line)

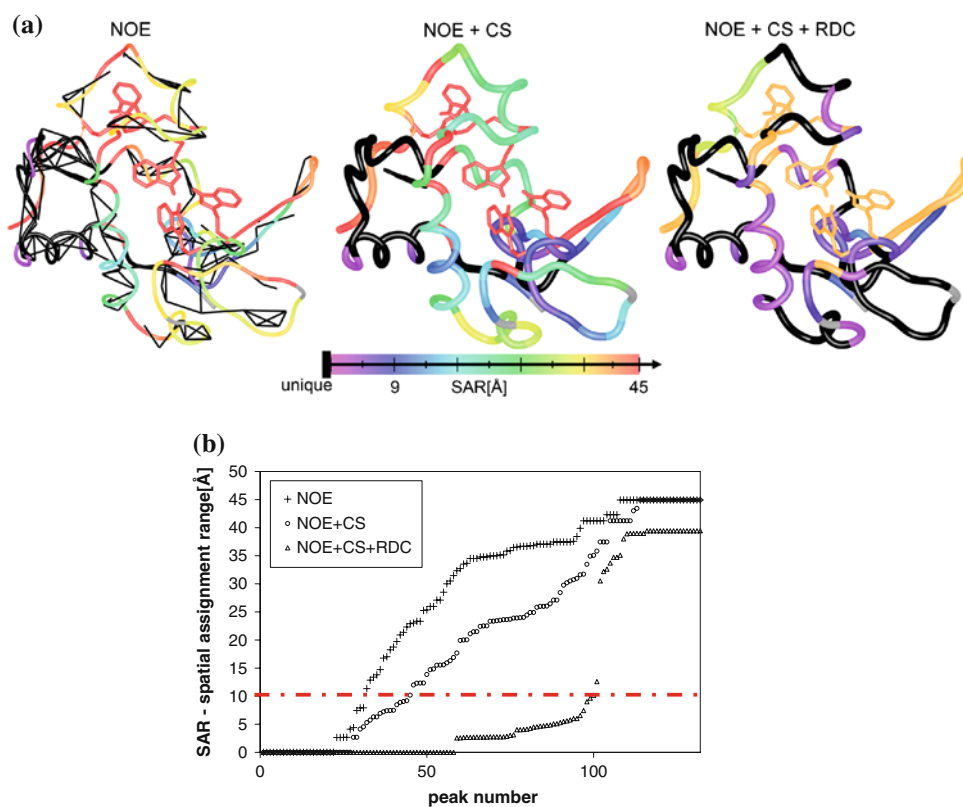


Fig. 4, the success rate of error detection increases with the number of errors and the data density, as expected.

In general the greater the number of inconsistencies, the higher will be the probability that a hole occurs in the assignment list. For a low data density (NOE only in Fig. 4), the assignment ensemble is still so large, that a limited number of inconsistencies does not necessarily lead to holes in the peak assignment lists, but simply to the removal of the correct assignment possibility in the lists of some peaks. For a high data density (NOE + CS + RDC in Fig. 4), the success rate of error detection is near 100%, even if only a low number of inconsistencies is present (Fig. 4). As already 60% of the peaks have only one assignment possibility here, the assignment ensemble cannot be adapted to the added inconsistent NOE constraints.

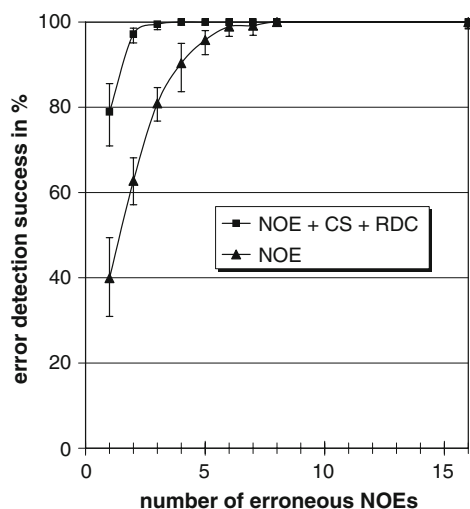
Thanks to this error detection, accuracies below 90% are very unlikely (see Fig. 5). For the rare cases where the assignment of uniquely assigned peaks is wrong, the assignment error is limited spatially: The distance between the correct residue and the residue to which a peak is wrongly assigned range from 2 to 5 Å for most cases, meaning that these peaks are mainly assigned to a neighbor residue of the correct residue. The maximum value of assignment error distances of all uniquely assigned peaks in an assignment ensemble can be used to quantify the assignment error one could make by using an erroneous

assignment ensemble. This value is named here maximum spatial assignment error ( $SAR_{max}$ ). Its distribution among the few runs for which the presence of errors was not detected by NOE<sub>net</sub> is shown in Fig. 6.

For the errors induced by too tight threshold parameters, Tables S1–S3 show that the number of assignment errors  $N_e$  (of all peaks) and  $N_{eu}$  (of uniquely assigned peaks) is quite small for the rare cases for which the inconsistencies remained undetected (Status = ‘finished’ or ‘not finished’ and  $N_e > 0$ ). The worst case (trial 20 in Table S1) yields incorrect assignments for 8 out of 132 peaks, 6 peaks were uniquely but incorrectly assigned to spatially nearby residues ( $SAR_{max} < 4.5$  Å). Moreover, the few assignment errors remain spatially restricted, like the swap of residue 207 with 208 for their corresponding peaks of EIN (see below).

The case of a larger protein: EIN

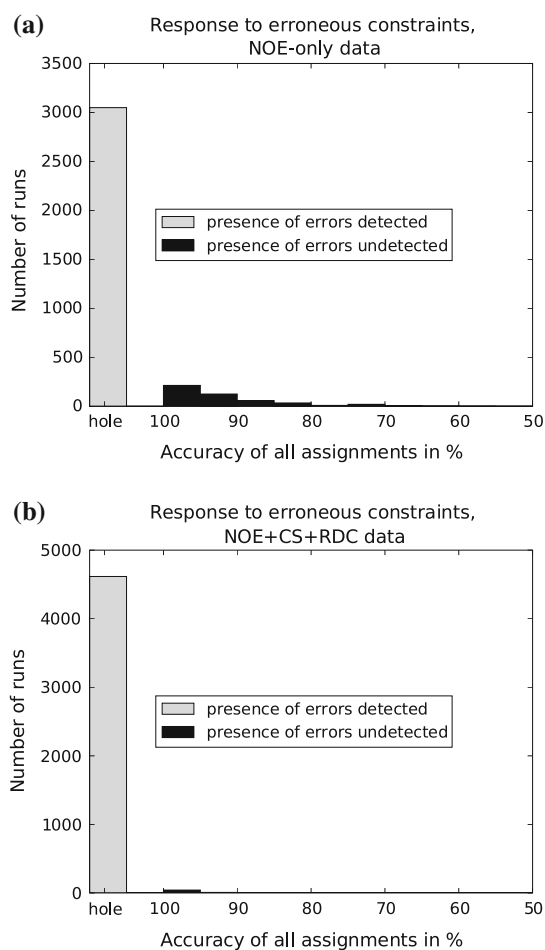
The results on EIN using only NOE data are described in (Stratmann et al. 2009). We present here the results obtained on EIN using the same NOE data set in conjunction with the  $^{15}N$ - $^1H$  chemical shift (CS) and simulated RDCs. Additionally, we show results on the inclusion of carbon chemical shifts and on the use of sequential connectivities by a combination of NOE<sub>net</sub> with MARS (Jung and Zweckstetter 2004).



**Fig. 4** Error detection success in % for lysozyme. The impact of errors in the classification of NOEs is tested here. Erroneous NOEs were generated by adding to experimental data randomly simulated NOEs corresponding to medium or long distances in the lysozyme X-ray structure 193L and classified incorrectly as short or medium NOEs, respectively. The number of erroneous NOEs introduced is shown on the *x*-axis. The error detection success, shown on the *y*-axis, is the ratio between the number of runs for which holes occurred in the assignment list and the total number of runs for which the erroneous NOEs caused the removal of correct assignment possibilities. One run consists in the random generation of erroneous NOEs, as described above, and the search for assignment possibilities by *NOE**net*

The size of EIN (28 kDa) is quite challenging for a complete search algorithm: The sampling of all possible graph matches of a NOE network with 407 edges onto the 3D structure with 1,034 edges and 243 nodes is not a trivial problem. A full convergence is only obtained after 6 days of calculation time (case 4, Table 3). On the other hand, thanks to the stop search procedure of *NOE**net* (see (Stratmann et al. 2009)), the result obtained after only 6 hours of calculation time (case 3, Table 3) is almost as good as the final result. Despite the sparseness of the input data ( $^1H^N-^1H^N$  NOEs and  $^{15}N$  and  $^1H^N$  chemical shifts only), 70% of the peaks have a SAR value below 10 Å (Table 3). This result is already sufficient to characterize the interface between the protein and its partner HPR (Fig. 7) and to obtain good docking results for the complex EIN-HPR (see next subsection).

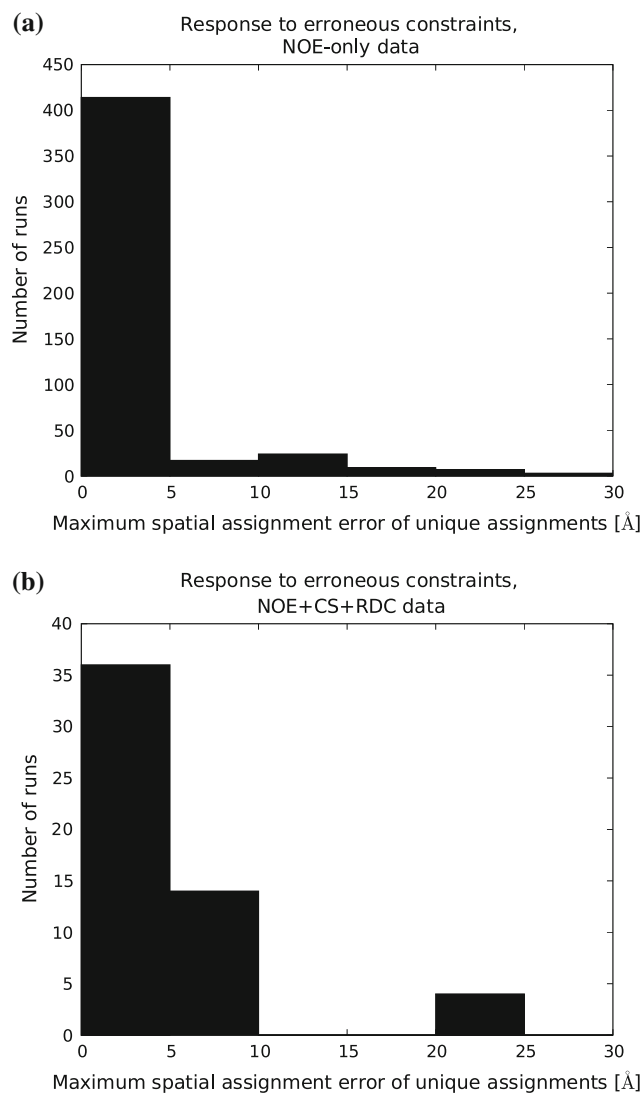
The accuracy for EIN is here below 100%, because of two assignment errors: residues 207 and 208 are interchanged for their corresponding uniquely assigned peaks. This small assignment error remained undetected as no hole occurred in the assignment list. A higher number of allowed NOE outliers of  $T_{NOE} = 3$  yields a 100% accurate but less precise assignment ensemble (see trial 4 in Table S3). The chemical shift information does not significantly improve here the assignment result, but it reduces the



**Fig. 5** The response of *NOE**net* to erroneous constraints is shown on the corrupted NOE data set introduced in Fig. 4. All runs that were done for Fig. 4 are taken together, independent of the number of erroneous NOE-constraints, as their amount is not known in advance in real situations. The *gray bar* shows the number of runs for which the presence of erroneous NOE-constraints is detected successfully through the appearance of holes in the assignment ensemble. The *black bars* show the number of runs for which the detection is not successful, as no hole occurred. The distribution of assignment accuracies of these runs is shown in form of a histogram. **a** The NOE-only data set. **b** The NOE+CS+RDC data set

runtime by a factor of three (compare case 1 with case 3 in Table 3). The chemical shift filter seems to cut branches of the search tree which do not lead to successful assignments and which have to be traversed completely if only the NOE data are available.

As for lysozyme, the RDC information greatly improves both the runtime and the quality of the assignment ensemble (case 5 in Table 3). Moreover, the accuracy of the result is now equal to 100%. This is due to the necessity to increase the number of NOE outliers from 2 to 3 to get results without holes (see Table S3). The addition of RDC data shows that the number of allowed NOE outliers  $T_{NOE}$  was too low for the given NOE data, which probably caused the assignment swap between residue 207 and 208.



**Fig. 6** The assignment errors of uniquely assigned peaks are quantified here by the maximum spatial assignment error ( $SAR_{max}$ ), i.e. the maximum distance to the correct residue among all uniquely, but erroneously assigned peaks of one assignment ensemble. All runs are taken together, independent of the number of introduced erroneous NOE-constraints, like in Fig. 5. **a** NOE-only data set. **b** NOE+CS+RDC data set

Even very loose thresholds for the RDC data already yield holes in the assignment ensemble with  $T_{NOE} = 2$  (see case 17, Table S3). Only for  $T_{NOE} = 3$  ‘hole-free’ assignment ensembles can be obtained. Without RDC data and with  $T_{NOE} = 3$ , the assignment ensemble is less precise but at least 100% accurate (see case 2 in Table 3). The compromise between precision and accuracy of the assignment ensemble can be solved by additional data. The addition of RDC data yields a precise and 100% accurate assignment ensemble at the same time. Due to the completely different nature of information brought by NOEs and RDCs (connections among peaks through proximity information vs

labeling of peaks with angular information), the conjunction of both data appears highly effective to avoid undetected assignment errors.

If a doubly  $^{15}N$ ,  $^{13}C$ -labeled protein is available, a triplet of  $^{13}C$  chemical shifts ( $^{13}C_{\alpha}$ ,  $^{13}C_{\beta}$ ,  $^{13}CO$ ) can also be used instead or together with the doublet of RDC-values. The ‘CScarbon’ data set includes only  $^{13}C(i-1)$  chemical shift values, i.e. no sequential connectivity is used at this stage. The results that were obtained using sequential connectivities (case 10 and 11 of Table 3) are discussed in the last results section below. Replacing the RDC data with ‘CScarbon’ data gives similar results (compare case 7 with case 5 in Table 3). The combination of RDC and CScarbon data allows 80% of the peaks to be assigned uniquely (case 9 in Table 3). The accuracy of the assignment ensemble is quite close to 100%, with 99.6 and 99.2%. Here, the assignment errors are not due to the number of NOE outliers  $T_{NOE} = 3$ , but to the *individual peak assignment refinement* procedure (see “Methods”). The refined assignment ensemble generally yields a much higher number of unique assignments. However, the results of this procedure should always be taken cautiously, since outlier values in the CScarbon data set can generate assignment errors, and the precision of the assignment can thus not be guaranteed.

#### Assignment ensembles and docking

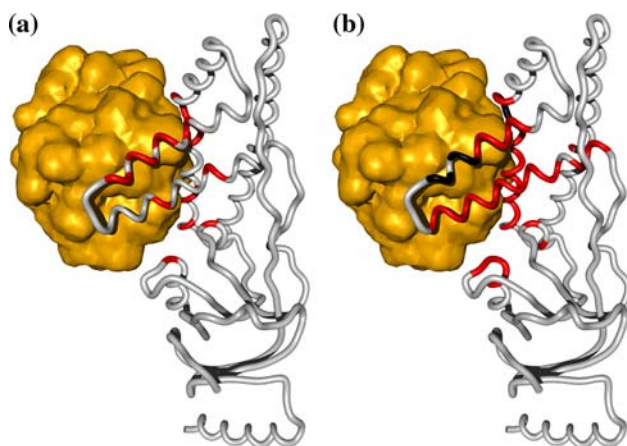
To demonstrate the usability of ambiguous assignments having a limited spatial assignment range (typically  $SAR < 10$  Å), the assignment ensemble obtained on EIN was used to model the EIN-HPR complex using the software HADDOCK (Dominguez et al. 2003; de Vries et al. 2007). The starting structures were those of the free proteins (1ZYM for EIN (Liao et al. 1996) and 1POH for HPR (Jia et al. 1993)). Interfacial residues were defined from  $^{15}N$ - $^1H$  chemical shift perturbation (CSP) data previously published for each protein (Garrett et al. 1997b; van Nuland et al. 1995), using the unique and correct assignment of chemical shifts for HPR and an assignment ensemble obtained from NOE $net$  for EIN. The less well defined assignment ensemble obtained from only NOE and  $^{15}N$ - $^1H$  chemical shifts was used (Case 4 of Table 3). All perturbed peaks of EIN that have a  $SAR$ -value up to 30 Å were used to define the interaction zone on EIN (see Fig. 7b). Only 6 out of the 21 perturbed peaks have a unique assignment (shown in black in Fig. 7b).

Two docking runs using the HADDOCK-server (default parameters) have been performed: one with the unique and correct assignment of all peaks of EIN and one with the assignment ensemble described above (see Fig. 8). The first run (Fig. 8a) is the reference case, whose best scored

**Table 3** EIN with experimental NOE data set from BMRB

Case	Trial #	Table S3	$N_{peaks}$	Runtime	$T_{NOE}$	Additional data	$\frac{N_{unique}}{N_{peaks}}$ (%)	$\frac{N_{SAR < 10\text{\AA}}}{N_{peaks}}$ (%)	Accuracy (%)
1	–	–	243	18 h	2		28	53	99.2
2	4	–	243	107 h	3	CS	19	34	100
3	5	–	243	6 h	2	CS	30	53	99.2
4	6	–	243	6 days	2	CS	31	70	99.2
5	26	–	243	50 min	3	CS + RDC	63	84	100
6	30	–	243	5 min	3	CS + CScarbon	42	80	100
7	31	–	243	1 s	3	–“–, refined	66	83	99.6
8	33	–	243	5 min	3	CS + CScarbon + RDC	67	84	100
9	34	–	243	1 s	3	–”–, refined	80	85	99.2
10	–	–	253	–	–	CA( $i \rightarrow i - 1$ ) with MARS	7	–	100
11	–	–	253	–	–	NOE + CS with NOE $_{net}$ and CA( $i \rightarrow i - 1$ ) with MARS	97.2	–	99.2

$N_{NOEs} = 407$  experimental unambiguous NOEs have been used. The X-ray structure 1ZYM was used, yielding  $N_{dist} = 1034$  distances below  $d_{max}^{theo} = 7.5\text{\AA}$ . NOE classes and NOE outliers have been used. Two sets of RDCs were simulated from 1ZYM as described in material and methods.  $T_{NOE}$ : maximum number of allowed NOE outliers for an arbitrary matching. Case 1 was reported in (Stratmann et al. 2009)



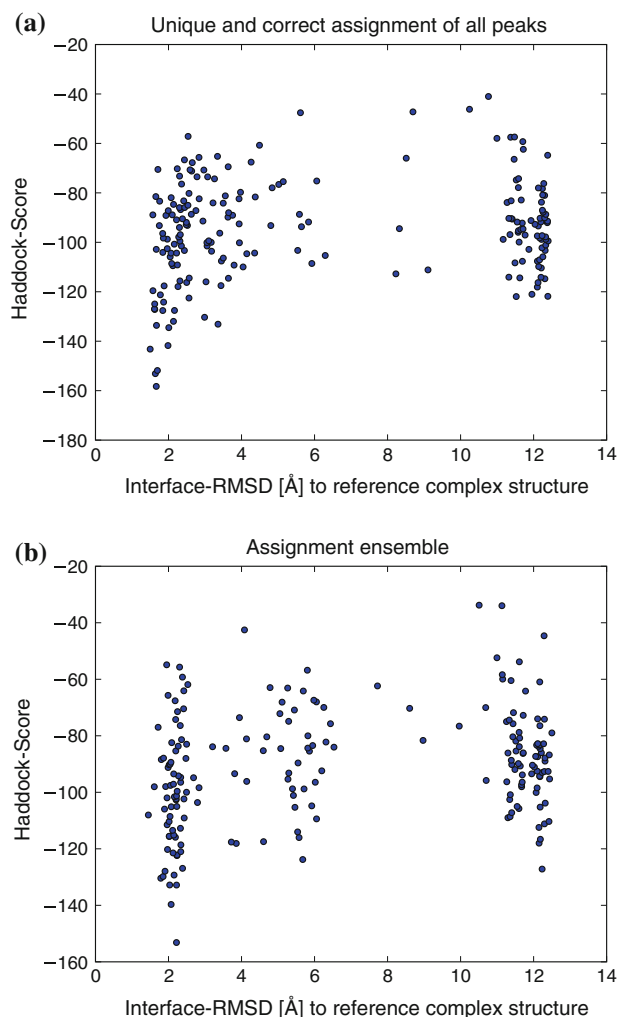
**Fig. 7** Interaction site estimation of EIN-Hpr using  $^{15}N$ - $^1H$  chemical shift perturbation (CSP) data (Garrett et al. 1997b) with **a** the correct assignment and **b** the assignment ensemble obtained by NOE $_{net}$  (Case 4 of Table 3). The NMR structure of the complex EIN-Hpr (PDB 3EZA (Garrett et al. 1999)) is shown here. EIN is shown by its backbone ribbon and Hpr is shown in yellow by its solvent accessible surface (including side chains). **a** Using the correct assignment, the corresponding residues of the perturbed peaks are colored in red. **b** The ensemble of assignment possibilities of the same perturbed peaks are colored in red and in black. The unique assignments are colored in black, while the assignment possibilities of the perturbed peaks with a SAR value below  $30\text{\AA}$  are colored in red

structure is rather close to the reference EIN-HPR complex structure (3EZA, (Garrett et al. 1999)) with an interface RMSD of  $1.8\text{\AA}$ . The second run (Fig. 8b) gives similar results, especially for the best scored structure (interface RMSD =  $2.1\text{\AA}$ ). This test demonstrates that the use of an assignment ensemble, even with low precision, does not degrade the quality of docking solutions.

### Exploiting both structure based and sequential triple resonance experiment based assignment

Some proteins are difficult to assign, because of their size or of the lack of sufficient connectivity data. The main connectivity data that is usually employed is the sequential connectivity between residue  $i$  and  $i - 1$  which is obtained from triple resonance experiments, like the CBCANH experiment. These type of experiments are often not very sensitive on large proteins, so that an important number of connectivities can be missing. It would therefore be helpful to fill the gaps by an independent data source. Structure-based assignment allows the use of other, independent data sources, like HN-HN NOEs, chemical shifts or RDCs. The integration of structure-based assignment approaches with the existing automated assignment approaches that are based on sequential connectivities can lead to a more robust assignment approach which would be less dependent on the completeness of a single data source.

As a first step towards this goal, we combined NOE $_{net}$  with the existing automated assignment approach MARS (Jung and Zweckstetter, 2004) that exploits mainly sequential  $i \rightarrow i - 1$  connectivities of CA and CB, as well as the chemical shift values of H( $i$ ), N( $i$ ), CA( $i$ ), CB( $i$ ) and C'( $i$ ). As MARS can make use of a list of reduced assignment possibilities for each peak, we gave the output of NOE $_{net}$ , i.e. the assignment ensemble, as input to MARS. We tested this approach on the protein EIN. In order to get 100% assignment of EIN in MARS, it is necessary to give an extensive set of H( $i$ ), N( $i$ ), CA( $i$ ), CB( $i$ ), CA( $i - 1$ ), CB( $i - 1$ ), C'( $i - 1$ ) chemical shifts. Fortunately, this data set is included in the distribution of MARS.



**Fig. 8** Docking results using HADDOCK on EIN-HPR. The scoring results of the final 200 water-refined EIN-HPR complex models are shown here. The interaction zone (active residues) is defined using  $^{15}\text{N}$ - $^1\text{H}$  chemical shift perturbation (CSP) data on EIN (Garrett et al. 1997b) and HPR (van Nuland et al. 1995). **a** The correct assignment of EIN and HPR is used to define the active residues from the CSP data (see Fig. 7a). **b** The assignment ensemble of EIN, obtained by NOEnet using NOE-data and  $^{15}\text{N}$ - $^1\text{H}$  chemical shifts (Case 4 of Table 3), is used to define the active residues on EIN (see Fig. 7b)

In order to simulate a case with a less complete data set of sequential connectivities, we reduced the EIN-data set to only the  $\text{H}(i)$ ,  $\text{N}(i)$ ,  $\text{CA}(i)$ ,  $\text{CA}(i - 1)$  chemical shifts. We ran MARS on it, using the default parameters and obtained assignments only for 18 out of the 253 peaks (Case 10 of Table 3). This is due to high ambiguity in sequential connectivities, if only CA without CB connectivities are available. At least all 18 assignments are the right ones, showing the robustness of MARS against assignment errors.

In a second run we added to the same reduced data set the assignment ensemble, obtained by NOEnet on EIN using NOE-data and  $^{15}\text{N}$ - $^1\text{H}$  chemical shifts (Case 4 of

Table 3). Even though this assignment ensemble contains only a low number of unique assignments (30%) and the number of unique assignments obtained by MARS in the previous run is even lower (7%), the combination of both gives very good results: 97.2% of the peaks are assigned uniquely with an accuracy of 99.2% (Case 11 of Table 3). The only assignment error is the swap between residue 207 and 208 that already occurred in the assignment ensemble.

This result shows that the combination of the structure-based and the classical sequence-based assignment methods is more robust against missing data than each approach taken alone.

## Discussion

One reason for the fact that many proteins are still assigned manually, is that the majority of NMR spectroscopists do not trust the results given by automated assignment procedures. Therefore, it is very important that an automated assignment procedure ensures a nearly 100% accurate result to be accepted widely. The input data should be translated reproducibly by the automated procedure into exactly one assignment ensemble corresponding to the information that can be extracted from the input data, not more and not less. The precision of this assignment ensemble should be the highest possible one under the maintenance of a 100% accurate result.

We demonstrate in this article that our structure-based assignment program NOEnet (Stratmann et al. 2009) translates reproducibly several types of input data (NOE, CS, RDC) into assignment ensembles having in the vast majority an accuracy of 100% and a precision near to the optimum. Erroneous assignment ensembles can be detected by NOEnet, the better the more errors they have, so that the accuracy of all ensembles that are not ruled out is guaranteed to be above at least 90%. The extensive tests of NOEnet's *error detection* feature presented in this article show first that the probability to miss errors is quite low. And secondly, they show that the undetected errors have a limited impact, as often the wrong assignment is done to spatially neighboring residues of the correct residue. Especially for structure-based assignment, such a strong guarantee of high accuracy is novel, as the input data used are often quite sparse, sparser than the input data available through triple resonance experiments. We handle the sparseness of the input data by searching for an assignment ensemble instead of a unique but possibly erroneous assignment. This ensemble was found useful for many applications, for example the localization of binding sites on the protein structures (Fig. 7) and the subsequent modeling of protein-protein complexes.

In order to obtain the most precise assignment ensemble, the various input data sources require adapted threshold-parameters which restrict the search space as well as the size of the resulting assignment ensemble. While fixed parameter values could be used, they are in general not optimal. We showed in this article that the threshold-parameters can be optimized thanks to two principles. First, NOE $net$  is capable to indicate too tight thresholds by the appearance of holes in the resulting assignment list. A hole in the list means that a peak has no assignment possibility left under the given constraints, indicating an inconsistency in these constraints. The second principle is simply that too relaxed thresholds will yield a huge search space as well as a large resulting assignment ensemble. The huge search space can be detected by too long runtimes without waiting till the end of the run. Using these two principles, we proposed an optimization protocol for the threshold-parameters used in NOE $net$ . We showed its application to different data sets (ideal and realistic simulated and experimental data) of two proteins—lysozyme and EIN.

We show here that the general filter strategy incorporated in NOE $net$  allows a straightforward incorporation of additional experimental data, such as  $^{15}N$  and  $H^N$  chemical shifts ( $CS$ ) and  $^1H$ - $^{15}N$  residual dipolar couplings (RDC). While it is possible to obtain a well constrained assignment ensemble using only NOE data (Stratmann et al. 2009), a similar approach using  $CS$  and/or RDC data only is likely to remain elusive for many reasons. First, individual  $CS$  and RDC data points are independent from each other and do not benefit from the network character of the NOEs, that constrains the assignment possibilities of several HSQC peaks simultaneously. Second, a very high correlation coefficient between experimental and predicted  $CS$  or RDC data points is not achievable, because of the structural differences between the template 3D-structure obtained from X-ray or from modeling and the dynamic solution state structure observed in NMR. The number of assignment possibilities, that satisfy the commonly found correlation coefficients for this type of data, remains then very high. RDC data have the additional problem that the alignment tensor can only be accurately estimated, once at least five (in practice >15) HSQC peaks have been assigned.  $CS$  and RDC data proved however invaluable assignment constraints in combination with the NOEs. RDCs markedly improve the assignment ensemble precision when used in complement to NOE data, especially if two or more RDC data sets are available.

The parameter optimization procedure of NOE $net$  allows the use of any kind of  $CS/RDC$  input data, independently of their quality in terms of RMSD values between predicted values from the template 3D-structure and experimental values. This unique feature of NOE $net$  is only possible through the search for an assignment

*ensemble* instead of a unique assignment for all peaks. While the precision of the assignment ensemble depends of course on the quality of the input data, the accuracy of the ensemble is generally independent from it.

Compared to RDCs,  $^{15}N$  and  $H^N$  chemical shifts do not yield a significant improvement of the assignment ensemble, because of the difficulty to predict accurately  $^{15}N$  and especially  $H^N$  chemical shifts from the 3D structure. They help at least to speed up the search process of NOE $net$  by ruling out impossible assignments at an earlier stage. On the contrary,  $^{13}C$  chemical shifts bring a clear improvement. Since  $^{13}C$  chemical shifts are more accurately predicted than  $^{15}N$  and  $H^N$  chemical shifts, their use constrains the assignment ensemble much better. The carbon chemical shifts ( $^{13}C_\alpha$ ,  $^{13}C_\beta$ ,  $^{13}CO$ ) of the preceding residue  $i - 1$  can be connected to the  $^{15}N$ - $H^N$  peaks by two of the most sensitive triple resonance experiments, the CBCA(CO)NH and the HNCO.

Other surely efficient additional constraints could come from selective labeling strategies. The inclusion of methyl-methyl or methyl-amide NOEs can also be exploited, and should be particularly helpful for the assignment of large perdeuterated, methyl protonated proteins. Finally, NOEs are an independent data source from the J-coupling used in triple resonance experiments. The combination of NOEs with sequential connectivities obtained from triple resonance experiments yields an even more robust structure-based assignment approach, especially for difficult cases. The unique ability of NOE $net$  to handle NOE networks combined with its general filter approach represents a straightforward avenue to combine both sources. This is demonstrated here in the case of the protein EIN, for which the combination of NOE $net$  with the robust automated assignment program MARS (Jung and Zweckstetter 2004) yielded a unique assignment for almost all 253 peaks (97.2%), whereas each program yielded only a low number of unique assignments when used separately (30% for NOE $net$  and 7% for MARS).

## Conclusion

The growing number of available protein 3D structures, obtained mainly by X-ray crystallography, makes the *structure-based* assignment concept particularly interesting for the protein NMR community. It is especially interesting in the case of protein functional studies involving protein-protein interactions, protein-ligand interactions or protein dynamics for which the backbone assignment is usually sufficient to bring key answers. Also the cross-validation or extension of assignments obtained by classical methods could be an application of structure-based assignment, as independent data sets compared to the classical triple

resonance experiments can be used, like in the case of NOE<sub>net</sub>.

While NOE<sub>net</sub> is primarily based on unambiguous  $^1H^N$ – $^1H^N$  NOEs (Stratmann et al. 2009), it has evolved to a quite flexible method by the introduction of the general filters concept associated with the facilitated parameter optimization protocol presented in this article. We introduce here two such filters for the inclusion of  $^{15}N$  and  $H^N$  chemical shifts (CS) and  $^1H$ – $^{15}N$  residual dipolar couplings (RDC) and demonstrate their efficiency in the assignment process. This can be particularly useful in high-throughput processes, in which proteins are usually produced  $^{15}N$  labeled. We show here that the quality of the assignment ensemble produced by NOE<sub>net</sub> is sufficient to delineate protein–protein interaction surfaces and even to obtain good quality models of protein–protein complexes. When higher precision is required, or for difficult proteins, a third carbon chemical shift filter can be used, if a ( $^{15}N$ ,  $^{13}C$ )-labeled sample is available. The assignment ensemble obtained by NOE<sub>net</sub> can then also be used as input for automated assignment procedures based on sequential connectivities. This combination yields an automated assignment approach which is highly robust against missing or erroneous data sets and should greatly improve the precision and the completeness of the assignment especially for difficult proteins.

#### Availability

The NOE<sub>net</sub> program will be available under: <http://www.icsn.cnrs-gif.fr/download/nmr>.

**Acknowledgements** We thank Ewen Lescop and François Bontems for a critical reading of the manuscript. We thank Olivier Serve and Guillaume Loire for the realization of the calculation cluster project in our lab. This work was supported by CNRS and a fellowship from the Ministère de l'Enseignement Supérieur et de la Recherche to D.S.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

#### References

- Altieri AS, Byrd RA (2004) Automation of NMR structure determination of proteins. *Curr Opin Struct Biol* 14(5):547–553
- Apaydin M, Conitzer V, Donald B (2008) Structure-based protein NMR assignments using native structural ensembles. *J Biomol NMR* 40:263–276
- Bailey-Kellogg C, Widge A, Kelley JJ, Berardi MJ, Bushweller JH, Donald BR (2000) The NOESY jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *J Comput Biol* 7:537–558

- Baran MC, Huang YJ, Moseley HNB, Montelione GT (2004) Automated analysis of protein NMR assignments and structures. *Chem Rev* 104(8):3541–3556, PMID: 15303826
- Bartels C, Billeter M, Güntert P, Wüthrich K (1996) Automated sequence-specific NMR assignment of homologous proteins using the program GARANT. *J Biomol NMR* 7:207–213
- Bax A, Grishaev A (2005) Weak alignment NMR: a hawk-eyed view of biomolecular structure. *Curr Opin in Struct Biol* 15:563–570
- Billeter M, Wagner G, Wüthrich K (2008) Solution NMR structure determination of proteins revisited. *J Biomol NMR* 42(3):155–158
- de Vries SJ, van Dijk ADJ, Krzeminski M, van Dijk M, Thureau A, Hsu V, Wassenaar T, Bonvin AMJJ (2007) HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins* 69(4):726–733
- Dobson CM, Howarth MA, Redfield C (1984) Nuclear overhauser effects and the assignment of the proton NMR spectra of proteins. *FEBS Lett* 176:307–312
- Dominguez C, Boelens R, Bonvin AMJJ (2003) HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* 125(7):1731–1737
- Erdmann MA, Rule GS (2002) Rapid protein structure detection and assignment using residual dipolar couplings. Technical Report CMU-CS-02-195, school of Computer Science, Carnegie Mellon University
- Garrett DS, Seok YJ, Liao DI, Peterkofsky A, Gronenborn AM, Clore GM (1997a) Solution structure of the 30 kDa N-terminal domain of enzyme I of the *Escherichia coli* phosphoenolpyruvate:sugar phosphotransferase system by multidimensional NMR. *Biochemistry* 36:2517–2530
- Garrett DS, Seok YJ, Peterkofsky A, Clore GM, Gronenborn AM (1997b) Identification by NMR of the binding surface for the histidine-containing phosphocarrier protein hpr on the N-terminal domain of enzyme I of the *Escherichia coli* phosphotransferase system. *Biochemistry* 36:4393–4398
- Garrett DS, Seok YJ, Peterkofsky A, Gronenborn AM, Clore GM (1999) Solution structure of the 40,000 Mr phosphoryl transfer complex between the N-terminal domain of enzyme I and HPr. *Nat Struct Mol Biol* 6:166–173
- Güntert P (2009) Automated structure determination from NMR spectra. *Eur Biophys J* 38(2):129–143
- Gronwald W, Kalbitzer HR (2004) Automated structure determination of proteins by NMR spectroscopy. *Prog Nucl Magn Reson Spectrosc* 44(1–2):33–96
- Gronwald W, Willard L, Jellard T, Boyko RF, Rajarathnam K, Wishart DS, Sönnichsen FD, Sykes BD (1998) CAMRA: chemical shift based computer aided protein NMR assignments. *J Biomol NMR* 12:395–405
- Hus JC, Prompers JJ, Brüschweiler R (2002) Assignment strategy for proteins with known structure. *J Magn Reson* 157:119–123
- Jia Z, Quail JW, Waygood EB, Delbaere LT (1993) The 2.0-Å resolution structure of *Escherichia coli* histidine-containing phosphocarrier protein HPr. A redetermination. *J Biol Chem* 268(30):22490–22501
- Jung YS, Zweckstetter M (2004) Mars—robust automatic backbone assignment of proteins. *J Biomol NMR* 30:11–23
- Koradi R, Billeter M, Wüthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph* 14:51–55, 29–32
- Langmead CJ, Donald BR (2004) An expectation/maximization nuclear vector replacement algorithm for automated NMR resonance assignments. *J Biomol NMR* 29:111–138
- Langmead CJ, Yan A, Lilien R, Wang L, Donald BR (2004) A polynomial-time nuclear vector replacement algorithm for automated NMR resonance assignments. *J Comput Biol* 11:277–298



- Liao DI, Silverton E, Seok YJ, Lee BR, Peterkofsky A, Davies DR (1996) The first step in sugar transport: crystal structure of the amino terminal domain of enzyme I of the *E. coli* PEP: sugar phosphotransferase system and a model of the phosphotransfer complex with HPr. *Structure* 4:861–872
- Losonczi JA, Andrec M, Fischer MW, Prestegard JH (1999) Order matrix analysis of residual dipolar couplings using singular value decomposition. *J Magn Reson* 138:334–342
- Meiler J (2003) PROSHIFT: protein chemical shift prediction using artificial neural networks. *J Biomol NMR* 26:25–37
- Moseley HN, Montelione GT (1999) Automated analysis of NMR assignments and structures for proteins. *Curr Opin Struct Biol* 9(5):635–642
- Neal S, Nip AM, Zhang H, Wishart DS (2003) Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *J Biomol NMR* 26:215–240
- Pintacuda G, Keniry MA, Huber T, Park AY, Dixon NE, Otting G (2004) Fast structure-based assignment of 15N HSQC spectra of selectively 15N-labeled paramagnetic proteins. *J Am Chem Soc* 126:2963–2970
- Pristovsek P, Franzoni L (2006) Stereospecific assignments of protein NMR resonances based on the tertiary structure and 2D/3D NOE data. *J Comput Chem* 27:791–797
- Pristovsek P, Rüterjans H, Jerala R (2002) Semiautomatic sequence-specific assignment of proteins based on the tertiary structure—the program st2nmr. *J Comput Chem* 23:335–340
- Schumann FH, Riepl H, Maurer T, Gronwald W, Neidig KP, Kalbitzer HR (2007) Combined chemical shift changes and amino acid specific chemical shift mapping of protein–protein interactions. *J Biomol NMR* 39(4):275–289
- Schwalbe H, Grimshaw SB, Spencer A, Buck M, Boyd J, Dobson CM, Redfield C, Smith LJ (2001) A refined solution structure of hen lysozyme determined using residual dipolar coupling data. *Protein Sci* 10:677–688
- Shen Y, Bax A (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J Biomol NMR* 38:289–302
- Stratmann D, van Heijenoort C, Guittet E (2009) NOEnet – Use of NOE networks for NMR resonance assignment of proteins with known 3D structure. *Bioinformatics* 25(4):474–481
- Vaney MC, Maignan S, Riés-Kautt M, Ducruix A (1996) High-resolution structure (1.33 Å) of a hen lysozyme tetragonal crystal grown in the APCF apparatus. data and structural comparison with a crystal grown under microgravity from SpaceHab-01 mission. *Acta Crystallogr D Biol Crystallogr* 52:505–517
- van Nuland NA, Boelens R, Scheek RM, Robillard GT (1995) High-resolution structure of the phosphorylated form of the histidine-containing phosphocarrier protein HPr from *Escherichia coli* determined by restrained molecular dynamics from NMR-NOE data. *J Mol Biol* 246(1):180–193
- Wang Y, Bjorndahl TC, Wishart DS (2000) Complete 1H and non-carbonylic 13C assignments of native hen egg-white lysozyme. *J Biomol NMR* 17:83–84
- Williamson M, Craven C (2009) Automated protein structure calculation from NMR data. *J Biomol NMR*
- Xiong F, Bailey-Kellogg C (2007) A hierarchical grow-and-match algorithm for backbone resonance assignments given 3D structure. *Bioinformatics and Bioengineering, 2007 BIBE 2007 Proceedings of the 7th IEEE International Conference*. pp 403–410
- Xiong F, Pandurangan G, Bailey-Kellogg C (2008) Contact replacement for NMR resonance assignment. *Bioinformatics* 24:i205–i213
- Xu XP, Case DA (2001) Automated prediction of 15N, 13C $\alpha$ , 13C $\beta$  and 13C $\gamma$  chemical shifts in proteins using a density functional database. *J Biomol NMR* 21:321–333