NOEnet—Use of NOE networks for NMR resonance assignment of proteins with known 3D structure

Dirk Stratmann, Carine van Heijenoort* and Eric Guittet*

Laboratoire de Chimie et Biologie Structurales, ICSN-CNRS, Gif-sur-Yvette, France

Received on September 11, 2008; revised on November 18, 2008; accepted on December 9, 2008

ABSTRACT

Motivation: A prerequisite for any protein study by NMR is the assignment of the resonances from the $^{15}N - ^1H$ HSQC spectrum to their corresponding atoms of the protein backbone. Usually, this assignment is obtained by analyzing triple resonance NMR experiments. An alternative assignment strategy exploits the information given by an already available 3D structure of the same or a homologous protein. Up to now, the algorithms that have been developed around the structure-based assignment strategy have the important drawbacks that they cannot guarantee a high assignment accuracy near to 100%.

Results: We propose here a new program, called NOEnet, implementing an efficient complete search algorithm that ensures the correctness of the assignment results. NOEnet exploits the network character of unambiguous NOE constraints to realize an exhaustive search of all matching possibilities of the NOE network onto the structural one. NOEnet has been successfully tested on EIN, a large protein of 28 kDa, using only NOE data. The complete search of NOEnet finds all possible assignments compatible with experimental data that can be defined as an assignment ensemble. We show that multiple assignment possibilities of large NOE networks are restricted to a small spatial assignment range (SAR), so that assignment ensembles, obtained from accessible experimental data, are precise enough to be used for functional proteins studies, like protein–ligand interaction or protein dynamics studies. We believe that NOEnet can become a major tool for the structure-based backbone resonance assignment strategy in NMR.

Availability: The NOEnet program will be available under: http://www.icsn.cnrs-gif.fr/download/nmr

Contact: carine@icsn.cnrs-gif.fr; eric.guittet@icsn.cnrs-gif.fr

Supplementary Information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

In order to resolve the molecular mechanism of proteins function, the structural biology community has made huge efforts to determine 3D structures of proteins at atomic resolution. Proteins function is not solely inferred from the tridimensional structure of the protein, but also from its interactions with other partners and from its dynamical and energetic properties. Consequently, even if the tridimensional structure of the free protein is known, it is necessary to study protein–ligand or protein–protein interactions and protein dynamics. Solution nuclear magnetic resonance (NMR) is a method of choice for those studies, through its unique ability to probe local dynamics and local environment changes. However, to obtain information at atomic resolution, NMR resonances first have to be assigned to their corresponding atoms in the protein.

With the availability of 3D structures for a large number of proteins, structure-based assignment strategies have been proposed. Several types of experimental NMR data can be correlated with the 3D structure. NOESY cross peaks (NOEs) between resonances correspond to spatially neighbouring atoms. Residual dipolar couplings (RDCs) measured in weakly aligning media are correlated with the orientation of the inter-atomic vector with respect to the magnetic field. Chemical shifts (CS) are nowadays more and more accurately predicted from the knowledge of the 3D structure (Neal et al., 2003; Shen and Bax, 2007). The comparison of these experimental data with the 3D structure yields the assignment constraints. This exploitation of the data cannot be done manually due to the important combinatorics to accomplish.

The existing structure-based assignment strategies require several different NMR data sources. For example, the nuclear vector replacement (NVR) algorithm (Apaydın et al., 2008; Langmead and Donald, 2004; Langmead et al., 2004) uses unambiguous $^{15}N - ^1H$ NOE-connectivity information, $^{15}N - ^1H$ RDCs obtained in two different alignment media, $^{15}N$ and $^1H$ chemical shifts and results from H-D exchange experiment. It has been tested on two small proteins and one moderate size protein (lysozyme, 129 amino acids). Another approach (Xiong et al., 2008), called contact replacement, is not based on RDCs but on $^1H - ^1H$ and $^{15}N - ^1H$ NOEs for the identification of the secondary structures. It also requires a TOCSY experiment for the estimation of the amino acid class of each $^{15}N - ^1H$ HSQC peak. The TOCSY data are analyzed by the program RESCUE (Pons and Delso, 1999) that classifies each $^{15}N - ^1H$ HSQC peak in one of 10 amino acid classes with an accuracy higher than 88%. The contact replacement approach has been tested on three moderate size proteins (up to 166 amino acids) with experimental data. The overall assignment accuracy varied from 60% to 80%, which is not sufficient in our opinion, as a typical NMR user will only trust automated assignment strategies, if they yield accuracies near to 100%. The low accuracy of the contact replacement approach is probably due to the use of highly ambiguous NOE data obtained from a 3D NOESY experiment, as well as to the translation of the RESCUE output into hard constraints.

*To whom correspondence should be addressed.
Both approaches, the NVR and the contact replacement, will be difficult to apply to large proteins, the first because of increased ambiguity of RDC assignment constraints and the second because of the increased overlap in TOCSY spectra. The PEPMORPH approach (Erdmann and Rule, 2002) has been tested on synthetic datasets of proteins with a size up to 414 amino acids. It combines NOE and RDC data, relies on unambiguous \(^1H^1\)-\(^13C\) NOEs and two RDCs per residue of the spin pairs \(^15N\rightarrow\(^13C^\omega\), \(^15N\rightarrow\(^13C^\delta\). The authors represent the NOE data and the 3D structure as mathematical graphs and the algorithm matches the NOE graph onto the 3D structure graph to obtain the assignment. Unfortunately, the PEPMORPH approach has only been tested on synthetic datasets that were highly idealized in comparison to realistic experimental datasets.

The complete list of structure-based assignment strategies is to our knowledge given by the following references: Dobson et al. (1984), Bartels et al. (1996), Gronwald et al. (1998), Bailey-Kelllogg 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-Kelllogg (2007) and Xiong et al. (2008). Despite the long list, we failed to find an approach that is applicable to large proteins and that always yields high accuracies near to 100%. The structure-based assignment problem can be translated into either a constrained bipartite matching problem (like for NVR) or a subgraph matching problem (like for contact replacement or PEPMORPH). Both problems are non-polynomial (NP)-hard, so that exact complete search algorithms solving the structure-based assignment problem can require a runtime increasing exponentially with the protein size. Because of this fact, the majority of structure-based assignment strategies use incomplete optimization algorithms giving a limited number of solutions (often only one global assignment), with the drawback that their accuracy is difficult to assess. As it is crucial to guarantee high accuracies near to 100%, we decided to explore the feasibility of an exact complete search algorithm for the structure-based assignment of large proteins.

We present here a new structure-based NMR assignment program, called NOEnet that exploits the network feature of the NOE-connectivities of unambiguous \(^1H^1\)-\(^13C\) NOEs and the 3D structure of the protein. Instead of searching for a unique global assignment, whose accuracy is difficult to assess, our goal was to determine an assignment ensemble, containing all possible assignments compatible with the available NMR data. The complete search algorithm of NOEnet ensures that the correct assignment is always part of the obtained assignment ensemble, if no erroneous constraint is introduced along the search process. Multiple assignment possibilities are spatially restricted by the assignment constraints. We thus introduce a quality factor for multiple assignments through the concept of SAR. We demonstrate here that restrained multiple assignments can be exploited for numerous NMR applications, like protein–ligand or protein–protein interaction.

2 METHODS
The assignment problem: the \(^15N\)-HSQC assignment problem consists in finding the assignment of the set of \(N_p\) \((^15N,^1H)\) resonances (also called peaks) \(P = \{p_1, p_2, \ldots, p_{N_p}\}\) to the corresponding amino acid \(R = \{r_1, r_2, \ldots, r_{N_R}\}\) in the protein sequence. A peak assignment defines the assignment of a single peak \(p_k \rightarrow r_j\) or, in a simplified notation, \(k \rightarrow j\). The set of peak assignment possibilities for a peak \(p_k\) is \(k \rightarrow i_1, i_2, \ldots, i_{N_k}\), with \(N_k\) being the number of assignment possibilities for the peak \(p_k\). The list of peak assignment possibilities contains the peak assignment possibilities for all peaks: \(k \rightarrow i_1, i_2, \ldots, i_{N_k}\), for \(k = 1, \ldots, N_p\). Two peaks \(k\) and \(l\) can share the same assignment possibility \(i_j = i_l\) in the list of peak assignment possibilities. A global assignment defines an assignment of all peaks \(p_k \rightarrow r_j\) for \(k = 1, \ldots, N_p\) with \(i_k \in \{1, \ldots, N_r\}\) and \(i_k \neq i_l\) for all \(k, l\). As the number of global assignments increases exponentially with the number of multiple peak assignments, in it is in general impossible to give the full list of global assignments, especially for sparse data. Only the list of peak assignment possibilities can always be obtained. It characterizes the assignment ensemble that should comprise all global assignments compatible with the data. Only one of the compatible global assignments is the correct global assignment or correct assignment. Using sparse data, it is in general impossible to reliably find only the correct assignment. It is nevertheless possible to obtain an accuracy of 100%, meaning that the assignment ensemble contains among other compatible assignments also the correct assignment. The accuracy is defined in this context as \(1 - N_r/N_p\) with \(N_r\) the number of peaks that do not have the correct assignment in their list of assignment possibilities and \(N_p\) the number of peaks. The quality of an assignment ensemble is not only given by its accuracy, but also by its 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: \(C_1 = N_{\text{unique}}/N_p\). 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 \(^1H^1\)-\(^1H^1\) distances for all residue pairs taken from the peak assignment possibilities for a specific peak \(p_k\). We define the SAR as the maximum of those distances and calculate it for each peak. This allows us to define a second type of completeness: the ratio between the number of peaks with a SAR-value below a given threshold (typically 10 Å) to the total number of peaks: \(C_2(\leq 10\,\text{Å}) = N_{\text{SAR}\leq 10}\,\text{Å}/N_p\). The uniquely assigned peaks are given a SAR-value of zero.

Input NMR data: the minimal input for NOEnet is a list of unambiguous \(^1H^1\)-\(^13C\) 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 \(^15N\rightarrow\(^1H\) HSQC spectrum. Beside the peaks of the \((^15N,^1H)\) atom-pairs of the protein backbone, some peaks correspond to the \((^15N,^1H)\) atom-pairs of side-chains. Especially, the tryptophan (TRP) side-chains generate \((^15N,^1H)\) peaks that 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 doublets corresponding to NH2 groups of side-chains are assumed to be identified by their identical \(^15N\) frequency, and were not included as assignment possibility.

The conceptual bases of NOEnet: the main idea of NOEnet is to sample all possible matches of the whole NOE network onto the connectivity network of the 3D structure. In terms of graph theory, the algorithmic problem is to find all possible subgraph monomorphisms or graph matchings; it belongs to the class of NP-hard problems. No polynomial-time algorithm has been found for NP-hard problems. Previously used algorithms for structure-based assignment are mostly incomplete search algorithms that yield a solution in polynomial time, but do not guarantee the correctness of this solution. NOEnet employs a complete search algorithm, giving no guarantee for a limited runtime, but ensuring the correctness of the obtained assignment ensemble. In opposition to algorithms that search one or several assignment solutions, NOEnet searches iteratively the assignment impossibilities, while ensuring that the correct assignment is not removed. At the beginning of the search, all peak assignments are in the \((N_{\text{peaks}} \times \text{Residues})\) 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 \(A\) that will have less assignment
Fig. 1. (a) Simplices and (simple) polytopes; (b) multiple neighbours; (c) generalized polytope.

Fig. 2. (a) Simple example of polytope matching possibilities: the experimental NOE-network (in blue) is matched onto the theoretical distance-network (in green) in different ways. (b) Example of an assignment table.

As neighbours, it is sufficient that they share at least one node (Fig. 1c). Contrary to the PEPMORPH algorithm that considers only matches between perfect polytopes, the two extensions—multiple neighbours and general neighbourhood—enable the use of the whole available NOE network as constraint, without any decomposition into perfect polytopes, for the full exploitation of the experimental data. Additionally, we also extended the matching possibilities: a n-simplex can be matched on all simplices with a dimension d ≥ n. The restriction of the matching possibilities to polytopes with the same dimension can indeed induce errors, due to the lower density of the experimental NOE network compared with the 3D structure network.

Graph matching: the assignment problem is solved by identifying all peak assignment possibilities that are compatible with the assignment constraints. Therefore, NOEnet tries to find all possible graph matches of \( G^{\text{exp}} \) onto \( G^{\text{theo}} \) (Fig. 2a). The graph \( G^{\text{exp}} \) is in general composed of several disconnected fragments. A fragment \( F_i^{\text{exp}} = \{V_i^{\text{exp}}, E_i^{\text{exp}}\} \) is a connected subgraph from \( G^{\text{exp}} \). The matching process is done sequentially for each fragment \( F_i^{\text{exp}} \), beginning with the largest fragment \( F_1^{\text{exp}} \). The intermediate assignment table A (Fig. 2b) is then used as a constraint for the matching processes of the next smaller fragments \( F_i^{\text{exp}}, i \leq 1 \).

Fragment matching: the matching fragment is done by a backtrack algorithm (Dechter, 2003). A start simplex \( S^{\text{start}} \) is chosen randomly among the simplices of the current fragment \( F_i^{\text{exp}} \). Every simplex of \( F_i^{\text{exp}} \) becomes one time the start simplex, as explained at the end of this subsection. Before each backtracking search, a variable ordering is determined, the variables being the simplices \( S_y^{\text{exp}} \) of \( F_i^{\text{exp}} \). Beginning at the start simplex \( S^{\text{start}} \), the next simplices \( S_2^{\text{exp}}, S_3^{\text{exp}}, \ldots, S_n^{\text{exp}} \) are determined by a breadth first search (BFS) (Sedgewick, 2002) on the graph of simplices \( G^{\text{simpl}} = \{V^{\text{simpl}}, E^{\text{simpl}}\} \). \( G^{\text{exp}} \) is the graph of the peaks connected by NOEs, whereas \( G^{\text{theo}} \) is the graph of the simplices connected by the general neighbourhood definition. The obtained variable ordering remains fixed during the whole fragment matching for the specific start simplex \( S^{\text{start}} \). We also tested a variable ordering obtained by a depth first search (DFS) (Sedgewick, 2002), but found that the ordering obtained by a breadth first search resulted in a much better performance on average.

The goal of the fragment matching is to find all possible matches of the start simplex \( S_1^{\text{exp}} \) on the graph of theoretical simplices \( G^{\text{simpl}} = \{V^{\text{simpl}}, E^{\text{simpl}}\} \). One possible matching is found in the following way: first, the n-dimensional simplex \( S_1^{\text{theo}} \) is matched to any \( S_y^{\text{theo}} \) with the same or higher dimension. For each simplex to simplex matching all possible permutations on the nodes level are tested \( (n+1)! \) node permutations for the n-Dimensional simplices. A particular match \( m_1 \) of the start simplex \( S_1^{\text{exp}} \) is the assignment of the nodes of \( S_1^{\text{exp}} \) to the nodes of \( S_1^{\text{theo}} \) with a specific permutation on the nodes level. \( m_1 \) is validated, if all other simplices of fragment \( S_i^{\text{exp}} \) can be matched onto \( G^{\text{theo}} \) without overlapping and under the constraint of the initial matching \( m_1 \) of \( S_1^{\text{exp}} \). Therefore, \( m_1 \) is extended with the predefined variable ordering of the simplices \( S_2^{\text{exp}}, S_3^{\text{exp}}, \ldots S_n^{\text{exp}} \) by a backtrack algorithm. If \( m_1 \) can be validated, the found fragment
match \( M = \{m_1, m_2, \ldots, m_3\} \) is registered as possible and is reused in further iterations to reduce the runtime. The corresponding peak to residues assignments are also registered as possible in the assignment table \( A \). If \( m_i \) cannot be validated, \( m_i \) is registered as impossible but not the corresponding peak to residues assignments. Only the simultaneous assignment of all nodes of \( S_{ij}^{exp} \) defined by \( m_i \) can be excluded and not the independent assignment at the peak-level. A change of the assignment of only one peak of \( S_{ij}^{exp} \) can therefore generate a possible matching \( m_i \neq m_i \).

Once a fragment match \( M = \{m_1, m_2, \ldots, m_3\} \) is found, NOEnet backtracks directly to the first variable, the startsimplex \( S_{ij}^{exp} \), to find the number of nodes of fragment \( F_{ij}^{exp} \) containing only n\( \text{res} \) of the whole fragment \( F_{ij}^{exp} \). The size of \( f_{ij}^{exp} \) is grown iteratively from \( n_{\text{init}} \) to \( N_i \), the number of nodes of fragment \( F_{ij}^{exp} \). A complete run of the fragment matching algorithm described above is performed for each value of \( n \) on the subset fragment \( f_{ij}^{exp} \). The iterative growth permits a progressive search for matching impossibilities of \( f_{ij}^{exp} \): first the easily detectable matching impossibilities are found with the smaller subset fragments. As the subset fragments are growing, more matching impossibilities are detected. Therefore, the assignment possibilities of each simplex of \( f_{ij}^{exp} \) are retested again for each value of \( n \). On the other hand, the already excluded assignment possibilities are not retested.

\textbf{Stop search:} NOEnet limits further the runtime by stopping temporarily the current search after a certain number of trials. For each initial matching \( M_1 \) of the startsimplex \( G_{ij}^{theo} \) to an arbitrary \( G_{ij}^{theo} \) the number of trials \( \text{singleTrials} \) to validate \( M_1 \) is limited by the threshold \( \text{maxSingleTrials} \) (set to one million by default). If the backtracking search for the extension of \( M_1 \) has been stopped, the corresponding peak assignments of \( M_1 \) are marked as \textit{stopped possibilities} in the assignment table \( A \). Stopped possibilities are neither impossible assignments nor confirmed possible assignments. They are retested again each time the subset fragment \( f_{ij}^{exp} \) has grown to a number of simplices \( n \) equal to a multiple of 10. If stopped possibilities remain in the assignment table at the end of the matching of all fragments \( F_{ij}^{exp} \), all matchings are redone with a higher \( \text{maxSingleTrials} \) threshold (by default a factor of 10 larger). This procedure allows a first result to be obtained rapidly. It is refined automatically during the following iterations.

\textbf{Pseudocode:} the pseudocode of NOEnet with its most important subfunctions is shown in the Supplementary Material (Fig. S5). It gives a simplified overview of the more complicated NOEnet source code of about 20,000 lines written in C++.

\textbf{NOE classes:} the relation between the intensity of a NOESY cross-peak and the distance between the corresponding residues depends on numerous parameters that can be difficult to evaluate precisely (local dynamics, mixing time of the experiment and proton density). NOESY cross-peaks intensities are thus usually grouped into three distance classes for protein structure determination (Wüthrich, 1986): strong, medium and weak that are associated with short, medium and long range upper distances, respectively. NOEnet can exploit this classification by using a different theoretical upper distance bound \( d_{\text{theo}} = (d_{\text{max,short}}, d_{\text{max,medium}}, d_{\text{max,long}}) \) for each class of NOEs. The NOE classes are actually taken into account through additional filters on the current assignments, obtained by the matching of \( G_{ij}^{theo} \) to \( G_{ij}^{theo} \). \( G_{ij}^{theo} \) and \( G_{ij}^{theo} \) comprise all NOEs and contacts, respectively. The edges of \( G_{ij}^{theo} \) are labelled with NOESY cross-peak intensity class [label \( L_{\text{class}} = 0 \) (weak), 1 (medium), 2 (strong)] and the edges of \( G_{ij}^{theo} \) are labelled according to the three distance classes [label \( L_{\text{dist}} = 0 \) (long), 1 (medium), 2 (strong)]. For each matching the relationship \( L_{\text{dist}} \geq L_{\text{class}} \) is checked.

\textbf{NOE outliers:} NOEs of a certain intensity class correspond to different distance values in the 3D structure. These distances are not equally distributed over the interval of allowed distance values \( [0, d_{\text{max}}] \), as the upper distance bound \( d_{\text{max}} \) should include all NOEs. Only a few NOEs correspond to the higher distances \( [d_{\text{max}}−\Delta d, d_{\text{max}}] \) and can thus be considered as outliers of the distance distribution. The outlier range is defined by \( \Delta d \). In order to reflect this feature of the distance distribution associated to NOESY, NOEnet can apply a NOE outlier filter during the search. For example, instead of allowing a 7 Å threshold for all NOE to distance matchings, all NOEs matched to the upper range between 6 Å and 7 Å are considered as outliers. During the fragment matching, the current assignment is rejected if the number of NOE outliers is above a chosen threshold \( T_{\text{NOE}} \).

\textbf{Detection of 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 \textit{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 dataset. 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. Under the correct peak \( v_{\text{exp}} \) to residue \( v_{\text{theo}} \) assignment, every NOE constraint \( v_{\text{theo}} = (v_{\text{theo}}^{\text{exp}}) \) in \( G_{ij}^{theo} \) must have a corresponding contact \( v_{\text{theo}} = (v_{\text{theo}}^{\text{exp}}) \) in \( G_{ij}^{theo} \). Otherwise assignment errors will occur during the matchings of \( G_{ij}^{theo} \) onto \( G_{ij}^{theo} \). Erroneous NOEs can be caused by too small distance thresholds for building \( G_{ij}^{theo} \), artefacts from the NOESY spectra or large differences between the reference tridimensional structure and the structure of the protein in solution.

3 MATERIALS

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, 1997b)]. The RMSD of the heavy backbone atoms between IYZM and IESA is equal to 1.55 Å. A large number of NMR experiments have been recorded on EIN (Garrett et al., 1997b), especially a 4D \( 15N \)\( 13C-N\) separated NOE experiment on perdeuterated EIN with a mixing time of 170 ms (Garrett et al., 1997b) and a 3D \( 15N\) - separated NOE experiment with a mixing time of 100 ms (Garrett et al., 1997b). The two experiments permitted the extraction of 555 \( H^N-H^N\) 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 \( H^N-H^N\) NOE-constraints. We assumed that the NOE dataset could have been obtained by a single 4D NOESY experiment. We thus removed from the NOE dataset all NOEs that involve an ambiguous \( 15N, H^N \) HSQC peak, defined by the tolerance distances [\( \text{tolN, tolH} \) equal to \( 0.2 \text{ ppm, } 0.02 \text{ ppm} \). Removal of ambiguous NOEs reduced the number of NOEs from 535 to 407. The average number of NOEs per residue is then \( r = 407/250 = 1.6 \) for EIN.

NOE constraints were classified in three classes (strong, medium and weak). Cross-peaks that appear only in the 4D NOESY with a long mixing time of 170 ms and that have an intensity below 11% of \( I_{\text{max}}(4D) \) were classified as weak. All cross-peaks from the 3D NOESY with a mixing time of 100 ms were classified as strong, if their intensity was >14% of \( I_{\text{max}}(3D) \). All other cross-peaks were classified as medium. The 407 experimental NOEs are thus divided into 36 strong, 208 medium and 163 weak NOEs.

\textbf{Calculations:} the results indicated in the results section correspond to the use of a single core of an Intel Xeon Woodcrest CPU at 2.66 GHz with 1 GB of RAM. While several cores or CPUs allows the user to test several parameters in parallel, the runtime of a single trial with NOEnet is not reduced, as NOEnet is not programmed in a parallel manner.
4 RESULTS

Introduction: the results of using NOEnet on the BMRB dataset of EIN, a large protein of 28 kDa, are shown in Figures 3 and 4, as well as in Table 1. The assignment precision is given by SAR values. A SAR value is defined for each HSQC peak as the maximum distance in the ensemble of possible assignments for the peak (see Section 2). The idea behind our SAR concept is that studies that do not require an 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 (SAR = 0 Å), but that have a small SAR value. Typically, the assignment ensemble of a peak, whose SAR value is lower than 30 Å, remains quite well spatially restrained. This concept is illustrated for three peaks of EIN HSQC in Figure S1.

Optimization of the TNOE parameter: in order to restrict the assignment possibilities in the best possible way, the parameter TNOE has to be optimized (see NOE outliers-part in Section 2). We performed five runs in parallel with different TNOE values (Table 1). Large TNOE values allow more assignment possibilities than low TNOE values. The optimal TNOE value is the smallest possible TNOE value that does not introduce a high number of assignment errors. Assignment errors can be detected by the appearance of holes in the assignment table (Section 2). For the dataset of EIN, holes appear for TNOE = 1, so that TNOE = 2 has been chosen as the optimal value. In the absence of holes, a limited number of assignment errors can still be present. For TNOE = 2, the correct assignment was not established for two peaks (a swap of assignment possibilities: residue 207 ↔ 208) without being detected. For TNOE values higher than two, no correct assignment has been removed, but the assignment ensemble is less well restricted.

Results for the protein EIN: Figure 3 shows the SAR curves of the two runs (case 4 and 6 in Table 1) with the optimal NOE outlier threshold, here TNOE = 2. The first run (case 4 in Table 1, crosses in Fig. 3) do not use the overlap check, described in Section 2, while the second run does (case 6 in Table 1, circles in Fig 3). The overlap check does not increase here the number of unique assignments (SAR-value = 0), but it increases significantly the number of peaks having a SAR value below 10 Å. The comparison of Figure 4a with Figure 4b shows that mainly the peaks corresponding to the longest helix of EIN see their assignment possibilities reduced thanks to the overlap check. Due to the calculation overhead caused by the overlap check, the runtime yielding full convergence increased from 18 h without overlap check to 44 h with overlap check.

The size of EIN (28 kDa) is still quite challenging for a complete search algorithm: the sampling of all possible graph matches of a NOE network with 407 edges (Fig. 4a) onto the 3D structure with 1034 edges (Fig. 4d) and 243 nodes is not a trivial problem. Even without the overlap check, full convergence is only obtained after 18 h of calculation time (case 4, Table 1). Despite the sparseness of the input data, 53% of the peaks have a SAR value below 10 Å (Fig. 4a and Table 1). This is a sufficiently good result for the use in protein–protein interaction studies, for example, as shown in Figure 5. In Figure 4c, the two largest NOE network fragments of EIN are shown. The NOE network fragment of the β-sheets of EIN constrains the assignment possibilities of the β-sheet backbone resonances very well (Fig. 4a). The NOE fragments of the α-helices constrains less than the fragments of the β-sheets the assignment possibilities of their peaks: the number of unique assignments is lower, but the SAR values are still small (Fig. 4a). The accuracy for EIN is below 100% (99.2%), 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.

Fig. 3. Quality of the assignment represented by the SAR for all HSQC peaks of EIN. The peaks are ordered by increasing SARs. The crosses correspond to case 4 in Table 1. The overlap check (see Section 2) has been deactivated for this case, so that the runtime remained reasonable (18 h). The results of case 4 are also represented in Figures 4 and 5. The circles correspond to case 6 in Table 1. The activated overlap check resulted in a higher number of peaks with a low SAR value at the expense of a longer runtime (44 h).

Fig. 4. Assignment results on EIN. (a, b) The SAR values are mapped on the NMR structure using the correct assignment and the indicated color code. Unique assignments are shown in black [NMR structure 1EZA (Garrett et al., 1997b)]. (a) Without overlap check (case 4 in Table 1 and crosses in Fig. 3) (b) With overlap check (case 6 in Table 1 and circles in Fig. 3). Thanks to the overlap check, the SAR values of the peaks corresponding to the longest helix of EIN are reduced significantly. The NOE fragment of the longest helix is disconnected from the two largest fragments shown in (e). (c) The two largest disconnected NOE fragments (NMR structure 1EZA is shown). All assignment possibilities of one fragment are colored by the color of the fragment. (d) Theoretical contact graph. The 1034 theoretical contacts are represented on the X-ray structure 1ZYM (Liao et al., 1996) by blue, green and red lines corresponding to the three distance classes, short (d < 4.5 Å), medium (d < 6 Å) and long (d > 7.5 Å), respectively. The NOE connectivities are represented by black lines in (a–c).
Table 1. Tested $T_{NOE}$ values on EIN using only NOE data

| No. | $T_{NOE}$ | $\Delta d$ [Å] | Check overlap | Runtime | Status     | Errors | $N_{unique}$ | $N_{peaks}$ (%) | $N_{SAR<10Å}$ | Accuracy (%) |
|-----|-----------|-----------------|---------------|---------|------------|--------|--------------|----------------|--------------|--------------|
| 1   | 5         | 1               | No            | 4 days  | Not finished |        | 5            | 24             | 100          |              |
| 2   | 4         | "              | No            | 2.5 days| Not finished | 8      | 30           | 100            |              |              |
| 3   | 3         | "              | No            | 6 days  | Not finished | 11     | 32           | 100            |              |              |
| 4*  | 2         | "              | No            | 18 h    | Finished    |        | 28           | 53             | 99.2         |              |
| 5   | 1         | "              | No            | 1 h     | Hole        |        |              | 69             | 99.2         |              |
| 6   | 2         | Yes            |               | 44 h    | Finished    |        |              | 28             |              |              |

The optimization of the $T_{NOE}$ parameter (see NOE outliers-part in Section 2) using only NOE data for EIN is shown in this table. The theoretical distance thresholds for the three NOE classes are here $d_{thres} = \{4.5, 6, 7.5\}$ Å, yielding $N_{dist} = \{391, 320, 323\}$ distances in each class (short, medium, long) using the X-ray structure 1ZYM. The number of experimental NOEs is here $N_{peaks} = \{36, 208, 163\}$ for strong, medium and weak NOEs, respectively. The number of HSQC peaks is here $N_{peaks} = 243$. Columns: $T_{NOE}$: maximum number of permitted NOE outliers for an arbitrary matching. $\Delta d$: the theoretical distance range $[d_{thres} - \Delta d, d_{thres}]$ in Ångstrom for which a NOE is considered as outlier. Check overlap: indicates whether the overlap between the disconnected fragments is tested during the search. Runtime: the calculation time required for the presented result. Status: 'hole' indicates the presence of peaks that have no assignment possibility left in the assignment table; 'finished' and 'not finished' indicates whether the run converged or not for the given runtime. $N_{unique}$: number of uniquely assigned peaks. $N_{SAR<10Å}$: number of peaks having a SAR-value below 10 Å including the uniquely assigned peaks. The optimized $T_{NOE}$ parameter that has been retained is marked by an asterisk.

Fig. 5. EIN-Hpr interaction site. Only the X-ray structure 1ZYM (Liao et al., 1996) of EIN is shown here. (a) In red are shown the residues corresponding to the peaks with a significant chemical shift perturbation (CSP) due to the interaction EIN-Hpr (Garrett et al., 1997b). This plot require the knowledge of the correct assignment possibility for each peak. In comparison, the plot (b) is using the assignment ensemble shown in Figure 4a (case 4 in Table 1, without overlap check). The assignment possibilities of the same perturbed peaks as in (a) are plotted in (b). The unique assignments are plotted in black, while the assignment possibilities of all perturbed peaks with a SAR value below 30 Å are plotted in red.

5 DISCUSSION

Philosophy of NOEnet: the design goal of NOEnet was to find all assignments compatible with the input NMR data and a 3D structure. We implemented therefore a complete search algorithm specifically developed and highly optimized for this purpose. NOEnet sets at the beginning all peak assignments as possible and removes step by step assignment impossibilities. This approach ensures to obtain always a fully accurate assignment ensemble at each step of the complete search. We found that the use of sparse or poorly constraining NMR data for the structure-based assignment cannot lead to a unique assignment of all peaks. Fortunately, the ambiguous assignments can also be exploited. We introduced a quality factor for multiple assignment possibilities by a SAR value. The peaks that belong to large NOE networks have in general a low SAR value below 10 Å using NOEnet. Even if the number of uniquely assigned peaks is quite low in the shown case, the total number of uniquely assigned and low SAR peaks is the important number for applications that do not need an exact localization for all peaks. Beside the reference protein 3D structure, NOEnet requires only one data source: unambiguous $^{1}H^{N} - ^{1}H^{N}$ NOEs. It uses the local and global structural properties of the entire NOE network in comparison to the available tridimensional protein structure, to obtain assignment constraints. The graph matching procedure of NOEnet can handle efficiently large NOE graphs, as shown on the 28 kDa protein EIN, requiring only 1 day of calculation time.

Input NMR data: as NOEnet requires unambiguous NOEs, peaks with degenerated $^{15}N, ^{1}H^{N}$ 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 that can be related to more than two of the remaining HSQC peaks also have to be excluded. To reduce the number of those ambiguous NOEs, modern 4D $^{15}N/^{15}N$-separated NOEs experiments should be employed, in general, for NOEnet, although the 3D version could be sufficient for small proteins. The spectrometer time for the NOE experiment can be reduced by the implementation of recent advances in projection methods, like GFT (Shen et al., 2005), or simply the choice of asymmetric digital resolutions for the classical 4D NOESY (Morshauer and Zuiderweg, 1999). As only the amide protons are needed, partially or completely perdeuterated proteins can be used for NOEnet. The perdeuteration will lead to an increase in resolution for large proteins and reduce the effects of spin diffusion, allowing longer mixing times for the generation of larger fragments inside the NOE network.

$^{1}H^{N} - ^{1}H^{N}$ NOESY are much less crowded than $^{1}H - ^{1}H$ NOESY. Although superpositions certainly becomes a drawback for large proteins, amide chemical shifts of well-structured proteins exhibit usually better dispersion than carbonyl or α carbons. This dispersion is actually used for triple resonance experiments, since these experiments all use the $^{1}H, ^{15}N$ plane as a basis for the assignment procedure. The density of peaks in 3D cubes extracted from 4D $^{15}N/^{15}N$ separated NOESY experiments is actually expected to be comparable with that of a 3D-HNCA-CB experiment used for sequential backbone assignment, and prior sample optimizations performed to minimize $(^{1}H^{N}, ^{15}N)$ overlaps are thus expected to be similar in both strategies.

Comparison of structure-based and triple resonance sequential assignment procedures: backbone assignment of proteins is usually performed using triple resonance experiments. In ideal cases, this procedure can be automated and yields a unique assignment for each backbone atom of the protein (see, Buran et al., 2004; Moseley and Montelione, 1999 for review). However, for many cases, it...
still needs manual intervention and may require several weeks of analysis, possibly with experimental condition or even protein optimization. The key differences between the structure-based and the triple resonance assignment procedures rely on the one hand on the parameters that determine sensitivity of the experiments performed and on the other hand on the type of constraints used. The triple resonance assignment is based on polarization transfer through bonds (relying on J-couplings), whose efficiency decreases with increasing transverse relaxation rates $R_2$. Especially, HNCA and HNCACB, which are key experiments for sequential assignment, are based on polarization transfer between $^{15}N$ of residue $i$ and $C_{\alpha}$ of residues $i$ and $(i-1)$. The coupling constants between these two nuclei are relatively weak (7–12 Hz). The balance between the delay needed to transfer the polarization, governed by the inverse of the coupling constants, and the NMR signal lifetime, governed by the transverse relaxation becomes here a critical issue. Increased $R_2$ can come from larger size, but also from local dynamic processes or local unfavourable geometry. This can hinder some magnetization transfer processes, and thus break up the assignment course which is in this strategy sequential.

Compared with this, our structure-based assignment approach relies on a network of amide protons spatial proximities, revealed by NOE data. NOE cross-peak intensities depend on longitudinal relaxation rate constants $R_1$ which are not sensitive to the same dynamic processes as $R_2$. Furthermore, in the special case of NOE $net_1$, only one NOE experiment is required. This eliminates the problems of correspondence and adjustments between experiments that are often critical in the triple resonance strategy. Finally, if the continuous gain in spectrometer sensitivity will hold, our method could even be applied using $^{15}N$ natural abundance and would allow the analysis of proteins with no heterologous expression.

As a conclusion, the two approaches use completely orthogonal data, and the combination of both sources of information should help to fill assignment gaps, occurring if only one of both sources is available. Especially the assignment of large or difficult proteins will clearly benefit from the combination of both approaches.

**Assignment ensembles for functional NMR studies:** peaks with low SAR values can be exploited for the localization of dynamical zones or the localization of interaction interfaces, as shown with the assignment ensemble obtained on EIN (Fig. 5). Despite the low percentage of unique assignments (30%) for EIN, the SAR concept leads to useful results for the localization of the interaction site with Hpr (Fig. 5). The comparison of the Figure 5a with Figure 5b shows that it is possible to correctly define the interaction site, even if the number of unique assignments is quite low among the peaks with chemical shift perturbation. In order to test the generality of this result, we first looked if the quality of the assignment ensemble obtained for EIN allowed the characterization of other potential interactions sites of the protein. The results obtained show that even the assignment ensemble obtained without overlap check indeed allows the localization of almost any interaction site on EIN (see Supplementary Figs S2–S4).

More generally, a critical parameter for the obtaining of good results is the fragmentation of the NOE network: larger fragments have a lower number of possible matches in the 3D structure. EIN is a good test case as it is composed of two sub-domains connected by loops, one containing only $\alpha$-helices and the other one mainly $\beta$-sheets. Instead of a large connected fragment spanning the whole protein, the resulting NOE network is composed of two disconnected fragments—one for each sub-domain (Fig. 4c). Despite this fragmentation of the NOE network, the assignment ensemble obtained by NOE $net_1$ is sufficiently well restrained to allow the localization of interaction sites. A lower fragmentation of the NOE network can be expected for more globular and thus more compact proteins, yielding even better restrained assignment ensembles than for the case of EIN. We thus think that the result we obtain with the protein EIN is quite general and that at least equally good results should be obtained on globular proteins of smaller or similar size than EIN and with NOE data of comparable quality.

### 6 CONCLUSION AND PERSPECTIVES

The availability of the tridimensional structures of a large number of proteins, obtained mainly by X-ray crystallography, can be of help for the assignment of the $^{15}N-^1H$ NMR spectra. We show here that a network of $^1H^{15}N-^1H^{15}N$ NOEs is a highly valuable NMR data source for the structure-based assignment, up to the point that additional data sources are not mandatory. The complete search algorithm implemented in the program NOE $net_1$ ensures to always obtain a high assignment accuracy, even with very sparse input data. The unambiguous NOEs alone yield already an important constraint for structure-based assignment. While our approach does not allow to reduce the measurement time in comparison to the classical approach using triple resonance experiments, it uses completely orthogonal NMR data, based on the NOE and not the J-coupling. Our approach is not meant to replace the classical triple resonance approach, but it demonstrates the power of NOE-networks for the structure-based assignment problem. As J-couplings revealed by triple resonance experiments are highly complementary to NOEs, the combination of both data sources will yield an even more robust assignment approach. NOE $net_1$ is a first step towards a highly robust automated assignment approach, integrating triple resonance data with NOE data. Further developments of NOE $net_1$ will focus on the possibility to include more diverse data sources. The inclusion of methyl-methyl NOEs can help for the assignment of large perdeuterated, methyl protonated proteins. The capability of NOE $net_1$ to use homology models as the reference 3D structure will also be evaluated.

A growing number of NMR studies focus on other aspects than structure determination, like protein interactions or dynamics. We show that they will clearly benefit from the introduced facilitation of the resonance assignment stage, even in cases where it is not unique.

### ACKNOWLEDGEMENTS

We thank Ewen Lescop and François Bontems for a critical reading of the article. We thank Olivier Serve and Guillaume Loire for the realization of the calculation cluster project in our lab.

**Funding:** CNRS; fellowship from the Ministère de l’Enseignement Supérieur et de la Recherche (to D.S.).

**Conflict of Interest:** none declared.

### REFERENCES

Apaydın, M. et al. (2008) Structure-based protein NMR assignments using native structural ensembles. *J. Biomol. NMR*, 40, 263–276.
Bailey-Kellogg, C. et al. (2000) The NOESY jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *J. Comput. Biol.*, 7, 537–558.

Baran, M. et al. (2004) Automated analysis of protein NMR assignments and structures. *Chem. Rev.*, 104, 3541–3556.

Bartels, C. et al. (1996) Automated sequence-specific NMR assignment of homologous proteins using the program garnet. *J. Biomol. NMR*, 7, 207–213.

Dechter, R. (2003) *Constraint Processing*. 1st edn. Morgan Kaufmann, San Francisco, CA.

Dobson, C. M. et al. (1984) Nuclear overhauser effects and the assignment of the proton NMR spectra of proteins. *FEBS Lett.*, 176, 307–312.

Erdmann, M. A. and Rule, G. S. (2002) Rapid protein structure detection and assignment using residual dipolar couplings. *Technical Report CMU-CS-02-195*, Carnegie Mellon University, Pittsburgh, PA.

Garrett, D. S. et al. (1997a) 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, D. S. et al. (1997b) 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.

Gronwald, W. et al. (1998) Camra: chemical shift based computer aided protein NMR assignments. *J. Biomol. NMR*, 12, 395–405.

Hus, J. C. et al. (2002) Assignment strategy for proteins with known structure. *J. Magn. Reson.*, 157, 119–123.

Koradi, R. et al. (1996) Molmol: a program for display and analysis of macromolecular structures. *J. Mol. Graph.*, 14, 51–5, 29–32.

Liang, H. et al. (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.

Morshauer, R. C. and Zuiderweg, E. R. (1999) High-resolution four-dimensional HMQC-NOESY-HSQC spectroscopy. *J. Magn. Reson.*, 139, 232–239.

Moseley, H. N. and Montelione, G. T. (1999) Automated analysis of NMR assignments and structures for proteins. *Curr. Opin. in Struct. Biol.*, 9, 635–642.

Neal, S. et al. (2003) Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *J. Biomol. NMR*, 26, 215–240.

Pisticci, G. et al. (2004) Fast structure-based assignment of 15N HSQC spectra of selectively 15n-labeled paramagnetic proteins. *J. Am. Chem. Soc.*, 126, 2963–2970.

Pons, J. L. and Deluc, M. A. (1999) Rescue: an artificial neural network tool for the NMR spectral assignment of proteins. *J. Biomol. NMR*, 15, 15–26.

Pristovsek, P. and Franzon, 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. et al. (2002) Semiautomatic sequence-specific assignment of proteins based on the tertiary structure—the program st2nmr. *J. Comput. Chem.*, 23, 335–340.

Sedgewick, R. (2002) *Algorithms in C++ Part 5: Graph Algorithms*. 3rd edn. Addison-Wesley Professional, San Francisco, CA.

Shen, Y. and Bax, A. (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J. Biomol. NMR*, 38, 289–302.

Shen, Y. et al. (2005) G-matrix fourier transform noesy-based protocol for high-quality protein structure determination. *J. Am. Chem. Soc.*, 127, 9085–9099.

Wuthrich, K. (1986) *NMR of Proteins and Nucleic Acids*. 1st edn. Wiley-Interscience, New York, NY.

Xiong, F. and Bailey-Kellogg, C. (2007) A hierarchical grow-and-match algorithm for backbone resonance assignments given 3D structure. In *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering*, 2007, BIBE 2007, Boston, MA, pp. 403–410.

Xiong, F. et al. (2008) Contact replacement for NMR resonance assignment. *Bioinformatics*, 24, 1205–1213.