J Biomol NMR (2008) 40:263–276
DOI 10.1007/s10858-008-9230-x

ARTICLE

# Structure-based protein NMR assignments using native structural ensembles

**Mehmet Serkan Apaydın · Vincent Conitzer · Bruce Randall Donald**

Received: 1 October 2007 / Accepted: 15 February 2008 / Published online: 26 March 2008


**Abstract** An important step in NMR protein structure determination is the assignment of resonances and NOEs to corresponding nuclei. Structure-based assignment (SBA) uses a model structure ("template") for the target protein to expedite this process. Nuclear vector replacement (NVR) is an SBA framework that combines multiple sources of NMR data (chemical shifts, RDCs, sparse NOEs, amide exchange rates, TOCSY) and has high accuracy when the template is close to the target protein's structure (less than 2 Å backbone RMSD). However, a close template may not always be available. We extend the circle of convergence of NVR for distant templates by using an ensemble of structures. This ensemble corresponds to the low-frequency perturbations of the given template and is obtained using normal mode analysis (NMA). Our algorithm assigns resonances and sparse NOEs using each of the structures in the ensemble separately, and aggregates the results using a *voting scheme* based on maximum bipartite matching. Experimental results on human ubiquitin, using four distant template structures show an increase in the assignment accuracy. Our algorithm also improves the robustness of NVR with respect to structural noise. We provide a confidence measure for each assignment using the percentage of the structures that agree on that assignment. We use this measure to assign a subset of the peaks with even higher accuracy. We further validate our algorithm on data for two additional proteins with NVR. We then show the general applicability of our approach by applying our NMA ensemble-based voting scheme to another SBA tool, MARS. For three test proteins with corresponding templates, including the 370-residue maltose binding protein, we increase the number of reliable assignments made by MARS. Finally, we show that our voting scheme is sound and optimal, by proving that it is a maximum likelihood estimator of the correct assignments.

**Keywords** Automated NMR assignments · Normal mode analysis · NMR structural biology · Protein flexibility via structural ensembles · Structural bioinformatics

**Abbreviations**

| | |
|---|---|
| bb RMSD | Backbone root mean square distance |
| BPG | Bipartite graph |
| CS | Chemical shift |
| EIN | N-terminal domain of enzyme I |
| EM | Expectation-maximization |
| GαIP | G-α interacting protein |
| HD | Homology detection |
| MBM | Maximum bipartite matching |
| MBP | Maltose-binding protein |
| MLE | Maximum likelihood estimator |
| MR | Molecular replacement |
| NMA | Normal mode analysis |
| NMR | Nuclear magnetic resonance |
| NOE | Nuclear overhauser effect |
| NVR | Nuclear vector replacement |
| PR | Pseudoresidue |
| RDC | Residual dipolar coupling |
| SBA | Structure-based assignment |
| SPG | Streptococcal protein G |

M. S. Apaydın · V. Conitzer · B. R. Donald (✉)
Department of Computer Science, Duke University, Durham, NC 27708, USA
e-mail: brd+jbn08@cs.duke.edu

B. R. Donald
Department of Biochemistry, Duke University Medical Center, Durham, NC 27708, USA

## Introduction

One of the key steps in NMR protein structure determination is resonance and NOE assignments. The assignment problem requires mapping spectral peaks to tuples of interacting atoms in a protein. In this paper, we report a new algorithm for automated structure-based NMR assignments by exploiting an ensemble of structural templates.

Structure-based assignment (SBA) denotes automated assignment given *prior information in the form of* the putative structure ("template") of the protein. By analogy, in X-ray crystallography, the molecular replacement (MR) technique allows solution of the crystallographic phase problem when a "close" or homologous structural model is known, thereby facilitating rapid structure determination (Rossman and Blow 1962). An automated procedure for rapidly determining NMR assignments given a homologous structure will similarly accelerate structure determination. Furthermore, even when the structure has already been determined by crystallography or homology modeling, NMR assignments are valuable to probe protein–protein interactions and protein–ligand binding (via chemical shift mapping or line-broadening). Previous SBA algorithms include CAP (Al-Hashimi and Patel 2002; Hus et al. 2002), NVR (Langmead et al. 2003; Langmead and Donald 2004a), (Meiler and Baker 2003), and MARS (Jung and Zweckstetter 2004b). The idea of correlating unassigned experimentally-measured residual dipolar couplings (RDCs) with bond vector orientations from a known structure was first proposed by Al-Hashimi and Patel (2002) and subsequently demonstrated by Al-Hashimi et al. (2002) who considered permutations of assignments for RNA. In Hus et al. (2002), RDC-based maximum bipartite matching (MBM) was successfully applied to SBA. Similarly, MARS (Jung and Zweckstetter 2004b) matches RDCs to those calculated from a known structure. An SBA algorithm should be robust with respect to structural noise and handle distant structural templates: A small change in the putative structure should not change the assignments drastically and it should work even when a close structural template is not available.

NVR (Langmead et al. 2003; Langmead and Donald 2004a) is an MR-like approach for SBA of resonances and sparse NOEs. NVR computes assignments that correlate experimentally-measured $H^N$–$^{15}N$ HSQC, $H^N$–$^{15}N$ RDCs (in two media), 3D NOESY-$^{15}N$-HSQC spectra ($d_{NN}$'s) and amide exchange rates, to a given backbone structural model. The algorithm requires only uniform $^{15}N$-labeling of the protein. The NMR data used by NVR can be acquired relatively rapidly compared to the traditional suite of experiments used to perform assignments. NVR runs in minutes and assigns with high accuracy the ($H^N$,$^{15}N$) backbone resonances as well as the sparse $d_{NN}$'s from the 3D $^{15}N$-NOESY spectrum. NVR works well only when the structure of the protein is known or for close templates (less than 2 Å backbone (bb) RMSD). SBA in general and NVR in particular have had an impact on algorithms for NMR methodology (Bailey-Kellogg et al. 2004; Vitek et al. 2005), and SBA has been important in the determination of protein structures (Potluri et al. 2006, 2007).

We introduce an algorithm that extends the circle of convergence of NVR such that *distant* templates can be used to obtain high assignment accuracies. We also improve NVR's robustness with respect to structural noise. In addition, we provide a measure of confidence for individual assignments.

As in NVR, our procedure takes as input NMR data plus a single structure $P$ (Fig. 1). $P$ is called the "template" and is obtained from a putative (remote) structural homolog of the protein that originated the NMR data. We then generate an ensemble of structures from $P$ by considering its flexibility, and then make the assignments for each structure in the ensemble separately. We then aggregate the assignments of each of the models using MBM as a voting scheme, which we show is a maximum likelihood estimator (MLE). In our study, we find that this scheme generally improves the assignment accuracy and improves the robustness of assignments with respect to structural noise. The percentage of models that agree on a given assignment provides an intuitive confidence measure for the assignment. We demonstrate our algorithm on four different structural models of human ubiquitin, using HD (for homology detection) (Langmead and Donald 2004b), a variant of NVR. In contrast to the original NVR, where the

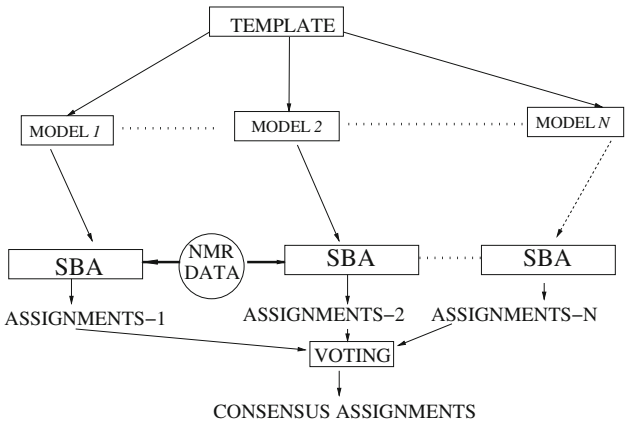


**Fig. 1** Overview of our methodology. We start from a template $P$. We use normal mode analysis (NMA) to obtain an ensemble of perturbed models around this template. Each model is in turn used as an input to a structure based assignment (SBA) algorithm (such as NVR or MARS) along with measured NMR data to compute the assignments. Each assignment is then combined using our voting scheme to obtain the "consensus" assignments

structural template must be less than 2 Å bb RMSD from the target, we use templates ranging in bb RMSD between 3.2 and 7.7 Å. The assignment accuracy of NVR for these distant structural templates ranges between 47–57% for human ubiquitin. With our new algorithm, the range of the assignment accuracy improves to 69–74%. Furthermore, *combining* the models from all ensembles raises the accuracy to 86%. Similarly, for G-α interacting protein (GαIP), the assignment accuracy increases from 65% to 77%. For streptococcal protein G (SPG), our results are mixed. However, combining the models for SPG raises the assignment accuracy.

We demonstrate the generality of our approach (of using NMA ensembles around a given template with our voting scheme) with MARS, which is a significantly different SBA tool from NVR (in terms of its algorithm and its input data). MARS can use both $^{13}$C- and $^{15}$N-labeled data and takes as input the observed intra- and inter-residual chemical shifts grouped into pseudoresidues (PR). Depending on the type of available spectra, MARS uses chemical shifts of $H^N_i$, $N_i$, $C'_{i-1}$, $C^{\alpha}_i$, $C^{\alpha}_{i-1}$, $C^{\beta}_i$, $C^{\beta}_{i-1}$, grouped into a PR with the $H^N_i$ and $N_i$ serving as an anchor, obtained from an $^{15}$N-$^1$H HSQC spectrum. In addition, when a template structure is available, MARS can use arbitrary RDCs from triple-resonance experiments to help the assignments. MARS is a hybrid assignment framework that optimizes local and global quality of fit of the amino acid sequence to the pseudoresidues. It links pseudoresidues to obtain PR segments of length five to two using sequential connectivity information in the linking stage. It then maps these segments to the amino acid sequence in the matching stage to obtain the assignments. It compares these assignments with one obtained using a global energy function and retains the consistent assignments. MARS follows an iterative procedure, where the experimental data is perturbed by adding noise to extract robust assignments. MARS computes a reliability information for each assignment, denoting each assignment as with low, medium or high confidence. It also lists all possible assignments for a given PR, along with their probabilities. We demonstrate our algorithm on three proteins that come with the MARS software distribution (Jung and Zweckstetter 2004a), and corresponding templates. The templates are close structural homologs of the corresponding target proteins, and with 100% sequence identity to the target proteins. The target proteins are: 76-residue human ubiquitin, 259-residue amino terminal domain of enzyme I from *E. Coli* (EIN), and 370-residue maltose-binding protein (MBP). With our new technique, we show that the number of correct and reliable (high confidence) assignments increases in all test cases. As in Jung and Zweckstetter (2004b), we apply our framework to MARS with varying amount of data, such as with and without sequential connectivity information, and up to three RDCs per residue. Depending on the amount of data used as input, the number of correct and reliable assignments increases by up to 23 at the expense of introducing three incorrect assignments (corresponding to a 3-fold increase in the number of correct assignments). Furthermore, the number of incorrect assignments generally does not increase.

Using an ensemble of structures in SBA is reasonable, since the structures of proteins in the PDB presumably correspond only to the ground state of these proteins (Kay 1998). The NMR data acquired from a protein in solution corresponds to a time- and ensemble-average over the many conformations assumed during data acquisition. We use NMA to perturb the template to obtain an ensemble of structures. NMA is a technique commonly used to study the low-frequency motion of proteins. It represents the energy landscape around a given energy minimum with a harmonic approximation and solves for the equations of motion within that well analytically. It has been shown that over half of the known protein movements can be modeled by displacing the protein along at most two low frequency normal modes (Krebs et al. 2002). Furthermore, NMA has been shown to reproduce the deformations in the core of homologous proteins caused by sequence differences in 35 large, diverse, and well studied superfamilies (Leo-Macias et al. 2005). Therefore, it seems reasonable to expect that the conformational differences between the template and the target protein can be modeled by NMA. In contrast to classical molecular motion simulation techniques such as molecular dynamics, NMA can very rapidly compute an ensemble of structures that correspond to the likely conformations assumed by the molecule around its energy minimum. NMA has been successful in predicting experimental quantities such as temperature factors of proteins (Bahar et al. 1997). We use coarse-grained NMA where several amino acids are grouped into a single super-residue which effectively removes the small scale fluctuations of a protein such as sidechain motions to model the slow, large-scale motions (such as backbone rearrangements) (Suhre and Sanejouand 2004a).

To the best of our knowledge, ours is the first approach that uses ensembles for structure-based resonance assignments. Note that previously ensembles have been used successfully in structure determination (given assignments) (Best and Vendruscolo 2004), and for NOE assignments (given resonance assignments) (Mumenthaler et al. 1997; Güntert 2004). Our results show that ensemble-based approaches are also useful for structure-based resonance assignments.

NMA was analogously used by Suhre and Sanejouand (2004b) for protein structure determination by MR, using X-ray diffraction data. The authors observed that although the original template did not help solve the crystallographic phase problem, there existed a structure in the NMA

ensemble that enabled the refinement of the target structure. This structure was chosen from the ensemble using a scoring function.

Our contributions in this paper are:

1. The use of NMA structural ensembles in structure-based NMR assignments,
2. "Robust" NMR assignments with respect to structural noise, by which we mean there is only a small change in assignment accuracy when the input structure changes slightly (note that this is not the case in general for maximum bipartite matching based assignment algorithms (including Langmead and Donald 2004a),
3. Increased radius of convergence of NVR with respect to the target–template structural similarity,
4. Improved assignment accuracy of NVR for distant templates (by up to 22%),
5. A confidence measure for each assignment,
6. A demonstration of the generality of our framework by improving the assignment accuracy of MARS on three test proteins (by up to 3-fold), and
7. A proof that our voting rule, which aggregates the assignments corresponding to individual models, is a maximum likelihood estimator.

## Preliminaries

### NMR data used by NVR

An assignment algorithm must determine the mapping of the resonances and NOEs to the corresponding nuclei of the protein. We can define the assignment problem as the mapping of the peaks to the corresponding residues, due to the specific set of NMR data used by our framework.

We use the following NMR data: $H^N$-$^{15}N$ HSQC, NOESY-$^{15}N$-HSQC (yielding sparse $d_{NN}$'s, observed between nearby pairs of amide protons), NH RDCs in two media (which provide global orientational restraints on NH amide bond vectors), $^{15}N$ TOCSY (for the sidechain chemical shifts), and amide exchange HSQC (to identify, probabilistically, solvent-exposed amide protons).

RDCs provide global information on the orientation of internuclear vectors. For each RDC $r$, we have the following RDC equation (Tolman et al. 1995; Tjandra and Bax 1997):

$$r = D_{max}\mathbf{v}^{\mathbf{T}}\mathbf{S}\mathbf{v}. \quad (1)$$

Here $D_{max}$ is the dipolar interaction constant, $\mathbf{v}$ is the internuclear bond vector orientation relative to an arbitrary molecular frame, and $\mathbf{S}$ is the $3 \times 3$ Saupe order matrix which describes the average substructure alignment in the weakly-aligned anisotropic phase. Equation 1 shows the quadratic dependence of $r$ on $\mathbf{v}$, thus explaining the sensitivity of RDCs (and hence, SBA algorithms that use RDCs, such as NVR) with respect to structural noise.

Only unambiguous $d_{NN}$'s are used in NVR. Typically only a few unambiguous $d_{NN}$'s (e.g., 43 for ubiquitin) can be obtained from the 3D-NOESY. These $d_{NN}$'s are automatically-assigned as a byproduct of NVR's resonance assignments (Langmead and Donald 2004a).

### NVR

NVR is an automated SBA algorithm for proteins of known structure or with a known close structural homolog. NVR uses MBM in an expectation maximization (EM) framework to compute the assignments. Each peak $p$ and residue $r$ form the nodes of a bipartite graph (BPG), where one set of vertices is the set of peaks, the other set of vertices is the set of residues, and the edges correspond to the likelihood of assigning $p$ to $r$ in the bipartite graph. The EM framework is used to iteratively select the most likely (*peak*, *residue*) assignment. More details can be found in (Langmead and Donald 2004a).

NVR integrates various NMR data as a means to increase the signal-to-noise ratio. The signal is the computed likelihood of the assignment between a peak and the (correct) residue. The noise is the uncertainty in the data, where the probability mass is distributed among multiple residues. Each line of evidence (i.e., experiment) has noise, but the noise tends to be random and thus cancels when the lines of evidence are combined. Conversely, the signals embedded in each line of evidence tend to reinforce one another, resulting in relatively unambiguous assignments.

NVR has the advantage that it only needs $^{15}N$-labeled data, which is cheaper to obtain than $^{13}C$-labeling, which is required by many automated assignment algorithms. NVR only uses unassigned data.

## Methods

An overview of our methodology is presented in Fig. 1. Our algorithm starts with a structural model. We apply NMA to this model to obtain an ensemble of structures. Then, for each member of the ensemble, we predict the backbone chemical shifts, and we also extract the NH amide bond vectors as well as proton coordinates of the amide bonds. NVR requires these, as well as the experimental NMR data. We predict the chemical shifts using the BMRB (Seavey et al. 1991), SHIFTS (Xu and Case 2001), and SHIFTX (Neal et al. 2003), following the protocol in Langmead and Donald (2004a, b). We then run NVR for each of the structural models. We combine the resulting assignments using MBM (Fig. 2). The MBM is done on a

BPG in which one set of nodes represents peaks and the other represents residues. The edge weights are simply the number of models in the ensemble that vote for the corresponding assignment. We used the Hungarian (Kuhn–Munkres) algorithm (Kuhn 1955), as implemented by N. Borlin, to solve MBM. While MBM has been used previously for NMR assignments (Hus et al. 2002; Xu et al. 2002; Langmead and Donald 2004a), edge-weights based on votes by a structural ensemble are novel.

We tested our algorithm on NVR with three proteins, and a total of seven distant templates, previously studied in Langmead and Donald (2004b). They are listed in Table 1. The three proteins are the 76-residue human ubiquitin (PDB ID 1D3Z, (Cornilescu et al. 1998)), the 56-residue streptococcal protein G (SPG) (PDB ID 3GB1, (Kuszewski et al. 1999)), and the 128-residue GαIP (PDB ID 1CMZ, (De Alba et al. 1999)). For these proteins, the NMR data (but not the actual structures) were used by our algorithm. For ubiquitin, the NH residual dipolar couplings recorded in two separate media (bicelle and phage) (Cornilescu et al. 1998), and $H^N$–$^{15}N$ HSQC and NOESY-$^{15}N$-HSQC spectra (Harris 2002) were used. For SPG and GαIP, the chemical shifts deposited into BMRB (Seavey et al. 1991) and amide-bond RDC data (Kuszewski et al. 1999; De Alba et al. 1999, resp.) were used. A set of sparse, unassigned $d_{NN}$'s were simulated for SPG and GαIP using the target structure and BMRB shifts as

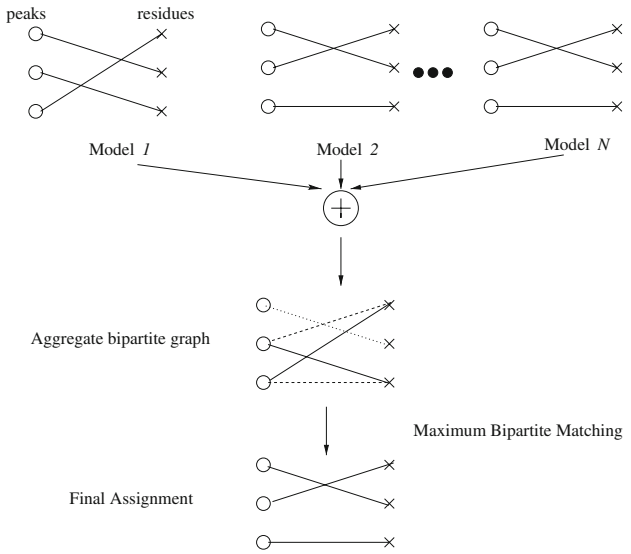


**Fig. 2** Our ensemble-based voting algorithm (maximum bipartite matching) combines the assignments for each model. The aggregated bipartite graph (BPG) combines the BPGs corresponding to each of the individual models. In the aggregate bipartite graph, the edge weight is one for the continuous lines, two for the dashed edges, and three for the dotted edge

previously described (Langmead and Donald 2004b). The template structures were obtained from the structural homologs of the target protein by homology modeling, as previously described (Langmead and Donald 2004b), using MODELLER (Sali and Blundell 1993). MODELLER was used to construct a backbone model for the target using template's backbone structure. Next, the sidechains for the model were constructed using MAXSPROUT (Holm and Sander 1991). MAXSPROUT considers rotamers for each sidechain and avoids steric clashes. Hydrogen atom coordinates were added to the template structures and these structures were energy-minimized using the PROTONATE and SANDER modules of AMBER (Pearlman et al. 1995), respectively. There is less than 30% sequence identity between each target protein and its structural homologs. We report in Table 1 the backbone RMSD as well as the CE RMSD of these distant templates. The CE RMSD refers to the RMSD of the aligned (homologous) regions between the template and the target, as computed by CE (Shindyalov and Bourne 1998). CE performs a combinatorial search to find the optimal structural alignment, and matches the vectors between Cα atoms to obtain aligned fragment pairs, which it optimizes using dynamic programming. The aligned regions measure the degree of homology between the target and the template.

We further tested our algorithm on three more proteins and a set of three structurally close templates, previously studied by Jung and Zweckstetter (2004b), and that came with the MARS software distribution (Jung and Zweckstetter 2004a). The target proteins are, human ubiquitin (PDB ID 1D3Z), the 259-residue amino terminal domain of enzyme I from *E. coli* (EIN) (PDB ID 3EZA), and the 370-residue maltose-binding protein (MBP) (PDB ID 1EZP). The set of NMR data used for these proteins, as well as the template information, is given in Table 3. Unlike our tests with NVR, in which we used a more distant ensemble of structures, the templates are structurally closer to the target structure (the bb RMSD ranges between 0.4–3.7 Å). Hence this study provides both a test of our algorithm with a significantly different SBA tool, as well as with structurally similar templates.

For both NVR and MARS, we used an NMA webserver, elNémo (Suhre and Sanejouand 2004a) to obtain an ensemble of structures around the template. We computed the five lowest-frequency normal mode displacements, with default parameters. Each of the low frequency normal modes returned 11 structures, corresponding to the motion of the template along that normal mode. We thus obtain 55 structures. We also displaced the template structure bidirectionally along its two lowest frequency normal modes, resulting in a total of 176 structures per template. The bb RMSD of the most distant structure to the starting model is less than 3 Å.

**Table 1** Assignment accuracy (% correct assignments) of NVR with distant templates for corresponding target proteins

| Target[a] protein | Homolog[b] | bb RMSD[c] (Å) | CE RMSD[d] (Å) | Sequence identity (%)[e] | Original[f] | Lowest–highest[g] | HD Score[h] | Ensemble[i] | Confident[j] |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | Accuracy (%) | | | | |
| Human ubiquitin | 1RFA | 8.0 (7.4–9.5) | 2.2 (89) | 12 | 51 | 19–67 | 57 | 73 | 97 |
| | 1EF1:A[4–84] | 6.3 (6.0–9.0) | 1.7 (38) | 10 | 57 | 17–69 | 63 | 74 | 100 |
| | 1H8C:A | 6.7 (6.4–8.4) | 1.9 (89) | 16 | 47 | 21–76 | 51 | 69 | 89 |
| | 1VCB:B | 3.5 (3.4–5.1) | 3.8 (44) | 13 | 53 | 13–73 | 64 | 74 | 95 |
| | All templates[k] | – | – | – | – | 13–76 | 64 | 86 | 100 |
| GαIP | 1DK8:A | 2.7 (2.6–2.7) | 1.9 (82) | 29 | 65 | 34–78 | 46 | 77 | 96 |
| SPG | 1HEZ:E | 5.0 (5.0–7.0) | 2.0 (94) | 13 | 60 | 25–76 | 71 | 62 | 79 |
| | 1JML:A | 8.6 (7.2–11.0) | 1.9 (75) | 13 | 65 | 33–80 | 64 | 60 | 83 |
| | All templates[k] | – | – | – | – | 25–80 | 71 | 69 | 87 |

[a] The target protein (source of the NMR data). This structure was not used by our algorithm. Instead, a template structure was used, obtained using homology modeling and energy minimization, starting from the structural homolog

[b] PDB ID for the structural homologs

[c] overall backbone (bb) RMSD between the template structure and the target protein's structure. The range of the RMSD distance of the ensemble to the target is provided in parenthesis

[d] bb RMSD of the structural alignment as computed by CE (Shindyalov and Bourne 1998). The percentage of the residues in CE alignment is shown in parenthesis

[e] The sequence identity between the sequences of the target protein and the structural homolog, as computed by CE

[f] Assignment accuracy obtained using the template

[g] The range of assignment accuracy over the NMA ensemble: the minimum and the maximum

[h] The accuracy of the structure in the NMA ensemble with the highest HD score

[i] The accuracy with our NMA ensemble-based voting scheme

[j] Assignment accuracy for the 'confident' peaks (where the confidence threshold is 0.5)

[k] Obtained by combining all corresponding NMA ensembles

Our algorithm runs in $O(mn + mn^{2.5}\log(cn))$ time, where $m$ is the number of models in the ensemble, $n$ is the number of residues in the target protein, and $c$ is the maximum edge weight in an integer-weighted bipartite graph. In comparison, NVR runs in time $O(n^{2.5}\log(cn))$, whereas HD has a time complexity of $O(pn^{2.5}\log(cn) + p\log p + pn)$, where $p$ is the number of proteins in a database of structural models. For a discussion of the complexity of NVR and HD, see Langmead and Donald (2004a, b) respectively. For reference, $c$ is a constant and is dictated by the resolution of the NMR data. NVR runs in minutes on a desktop PC to assign a protein with about 56–128 residues using one template.

## Results

We ran NVR for all three target proteins, with the corresponding templates obtained from structural homologs, for each of the ensemble of models obtained by NMA. We report the assignment accuracy for the template structure, as well as the range of accuracies in the NMA ensemble in Table 1. It can be seen that if we could choose the right template from this ensemble, we would improve the assignment accuracy of NVR. However this requires a scoring function that correlates strongly with the assignment accuracies.

### Using a scoring function to choose a model from the ensemble

Suhre and Sanejouand (2004b) used NMA to perturb the structural model, and then chose a perturbed template structure with a scoring function ("free R factor") in MR in X-ray crystallography, which allowed them to solve the target protein structure. We hypothesized that we could follow a similar methodology to choose a template from the NMA ensemble as input to NVR in NMR structure determination.

The HD score function combines the "preference list" of all the seven "voters" of NVR. These "voters" correspond to the NMR data used by NVR. They are: RDCs in two media, chemical shifts predicted using three different protocols (Langmead and Donald 2004a), amide-exchange and TOCSY data. Each "voter" has a ranked list of probabilities ("preference list") for each peak, corresponding to the likelihood of matching that peak with each residue (e.g., according to RDCs). The HD scoring function (Langmead

and Donald 2004b) simply multiplies and normalizes these probabilities to obtain an overall matrix representing the aggregated preference of all the voters for each peak. Given an assignment, the set of probabilities corresponding to individual (*peak*, *residue*) assignments are combined and returned as the HD score.

We used HD score to choose a model that has the highest HD score from the NMA ensemble. The assignment accuracy for this structure is in column entitled 'HD score accuracy' of Table 1. The correlation of the HD score with the assignment accuracies is shown in Fig. 3. Each point in the scatter plot corresponds to one of the structural models. The *x*-axis corresponds to the HD score and the *y*-axis to the assignment accuracy. The correlation between the HD score and the assignment accuracy is 0.44. It can be seen that HD score cannot be used reliably to choose a model with a higher assignment accuracy from the ensemble, with respect to the starting template.

## MBM voting over the NMA ensemble

We used MBM to aggregate the assignments corresponding to all of the models in the NMA ensemble (see Methods). The results of this scheme are in column entitled 'Ensemble accuracy' of Table 1. For all three proteins with the corresponding templates, the assignment accuracy improves in all but one of the seven protein–template pairs, with respect to the starting structural model, by up to 22%. We also combined the assignments of all the models corresponding to all four templates for human ubiquitin and both templates for SPG, obtaining an even higher accuracy (86% and 69%, respectively, shown in column entitled 'Ensemble accuracy' and row entitled 'All templates' of Table 1). For SPG using the template obtained from pdb ID 1JML (which is 8.7 Å bb RMSD from SPG), the assignment accuracy actually decreases with the consensus scheme. Note that the template obtained from 1JML is the farthest structure from its corresponding target structure in our test cases, and it may be that this starting template is outside the radius of convergence of NVR.

## Confidence measure

Given the assignments for each structural model in the NMA ensemble and the consensus assignments computed using MBM voting, we can compute the fraction of models that agree on a given resonance assignment. This ratio can be used as a 'confidence' measure for that assignment. Intuitively, the larger the number of models that agree on a (*peak*, *residue*) assignment, the less likely it is that that assignment is due to noise.

In Fig. 4, we show the ratio of the models that agree on a particular assignment (the 'confidence', shown on the *y*-axis) for each individual peak (the *x*-axis), for human ubiquitin with template obtained from 1RFA (which is 7.7 Å bb RMSD from human ubiquitin). Each blue 'circle' (resp., red 'cross') corresponds to a correct (resp., incorrect) assignment. The ratio of 'blue circles' to all signs

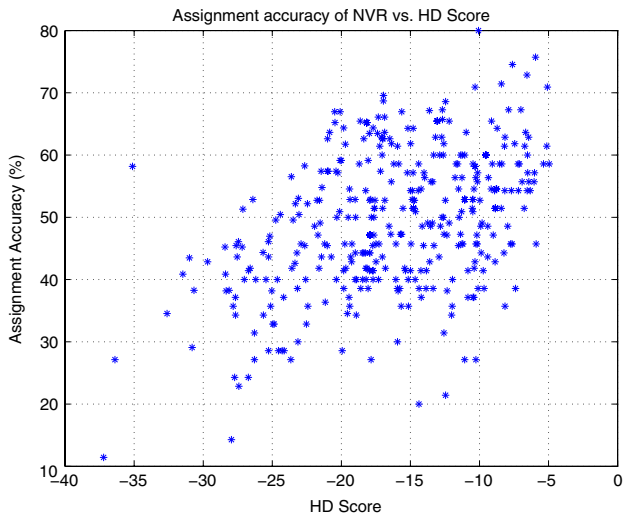


**Fig. 3** HD score vs. assignment accuracy. Each of the points correspond to a template in the normal mode analysis (NMA) ensemble. A representative set of templates for all three target proteins are shown. The *x*-axis is the HD-Score of the template structure, whereas the *y*-axis is the assignment accuracy (%). The correlation coefficient is 0.44

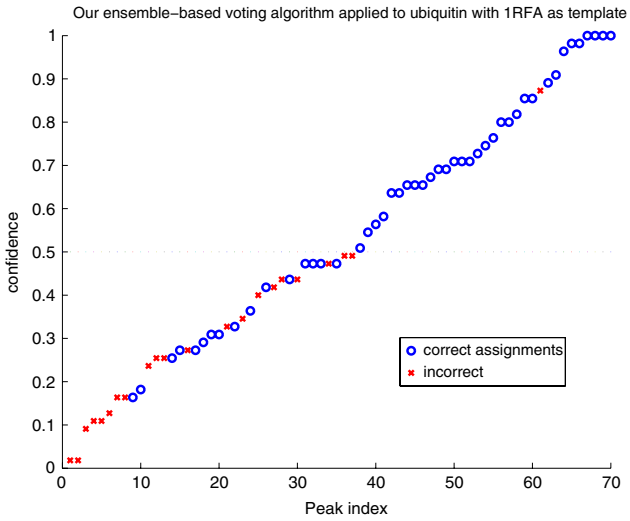


**Fig. 4** Assignments and confidences for human ubiquitin with pdb ID 1RFA as template and NVR. The diagram shows in blue ('o') the correct resonance assignments, and in red ('x') the incorrect ones. The *x*-axis corresponds to the individual peaks from the ubiquitin spectra, and the *y*-axis shows the "confidence", which is the fraction of the models that agree on the corresponding assignment for that peak over all models. The peaks are sorted in the order of ascending confidences. The assignment accuracy increases to 71% with our algorithm (compared to 51% with the single-structure based NVR)

determines the accuracy of the consensus assignments reported in column entitled 'Ensemble accuracy' of Table 1. The blue 'circles' have a higher confidence value than the red 'crosses' in general, suggesting that the 'confidence's indeed correlate with the assignment accuracy. The higher the 'confidence' for an assignment, the more likely it is to be correct.

In Fig. 4, there are very few incorrect assignments for which more than half of the models agree. Therefore, we selected a threshold of 50%, and called an individual assignment 'confident' if more than 50% of the models agree on that assignment. The assignment accuracy of the confident assignments is in the last column of Table 1. We also combined all the models and report the corresponding 'confident' assignment accuracy. The 'confident' assignment accuracy is higher than consensus assignment accuracy in all cases.

If we select a lower confidence threshold than 0.5, we can include more of the correct individual (*peak*, *residue*) assignments, at the expense of introducing some of the incorrect individual (*peak*, *residue*) assignments. This trade-off can be seen with a receiver–operator characteristic (ROC) curve. For each threshold, one can compute the sensitivity and the specificity and plot these points as in Fig. 5. For instance, for the target protein SPG with pdb ID 1HEZ as template (which is at 5.1 Å from SPG), a confidence threshold of 0.9 seems more suitable to correctly assign more than 40% of the peaks (corresponding to 13 peaks) without introducing any incorrect assignments. On the other hand, for the target protein ubiquitin with 1EF1 as template (which is at 6.2 Å from ubiquitin), a confidence threshold of 0.5 results in 25 correctly assigned peaks with no incorrect assignment among them. The trade-off between choosing a confidence threshold of 0.9 and 0.5 can also be seen in Table 2, where the absolute number of correct and incorrect peaks found using both thresholds are provided. One can select a confidence threshold to return the higher number of correct assignments, while minimizing the number of incorrect assignments.

## Robustness with respect to structural noise

We call a structure-based assignment algorithm "robust" if its result does not change significantly when the input structure changes slightly. This is a reasonable definition of robustness, since due to structural noise, there may be small perturbations in the input structure. In order to demonstrate the improved robustness in the assignment accuracies with the consensus and 'confident' assignment schemes, we chose 11 structurally-similar starting models for human ubiquitin from the NMA ensemble computed around the template obtained from pdb ID 1H8C (which is 6.7 Å bb RMSD from human ubiquitin), and computed the assignment accuracies using the original NVR (Langmead and Donald 2004a) and our ensemble-based voting algorithm. For our approach, we report both the accuracy of the

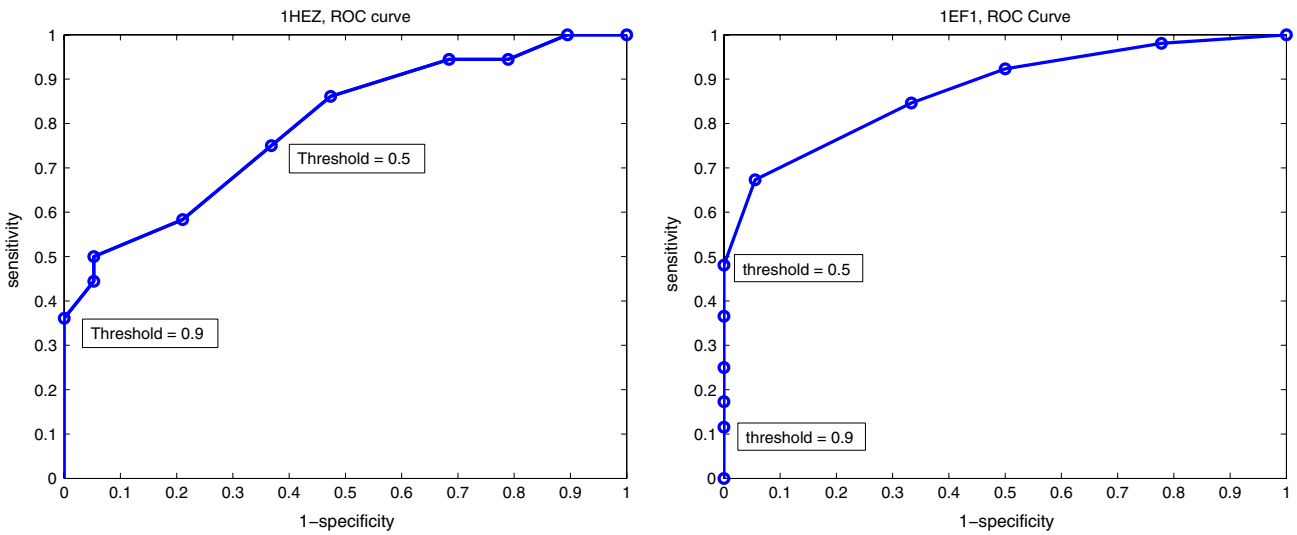


**Fig. 5** Receiver–operator characteristic (ROC) Curve for varying confidence thresholds: (Left) for SPG with the template obtained from 1HEZ; (Right) for ubiquitin with the template obtained from 1EF1. The confidence threshold is the ratio of models that must agree on a particular (*peak*, *residue*) assignment in order to include that pairing on the reported subset of assignments. For a given threshold, the *x*-axis is the ratio of reported incorrect assignments over all incorrect assignments (1-specificity). The *y*-axis is the ratio of correct assignments over all correct assignments (sensitivity). An ideal confidence threshold would be such that the returned assignments would include a maximum subset of the correct assignments and a minimum subset of the incorrect assignments. (Left) For this case, a confidence threshold of 0.9 would return about 40% of the correct assignments with no incorrect assignments. (Right) For this case, a confidence threshold of 0.5 would return about 50% of the correct assignments with no incorrect assignments

**Table 2** Effect of varying the confidence threshold: Number of correct and incorrect peaks with NVR with distant templates for corresponding target proteins, with varying confidence thresholds

| Target[a] protein | Homolog[b] | # Correct (# incorrect)[c] | # Correct (# incorrect)[d] |
|---|---|---|---|
| Human ubiquitin | 1RFA | 32 (1) | 7 (0) |
| | 1EF1 | 25 (0) | 6 (0) |
| | 1H8C | 25 (3) | 9 (0) |
| | 1VCB | 35 (2) | 8 (1) |
| | All templates[e] | 29 (0) | 5 (0) |
| GαIP | 1DK8 | 75 (3) | 38 (2) |
| SPG | 1HEZ | 27 (7) | 13 (0) |
| | 1JML | 29 (6) | 16 (0) |
| | All templates[e] | 26 (4) | 11 (0) |

[a] The target protein (source of the NMR data). This structure was not used by our algorithm. Instead, a template structure was used, obtained using homology modeling and energy minimization, starting from the structural homolog

[b] PDB ID for the structural homologs

[c] Number of correct (resp., incorrect) confident peaks with a confidence threshold of 0.5

[d] Number of correct (resp., incorrect) confident peaks with a confidence threshold of 0.9

[e] Obtained by combining all corresponding NMA ensembles

resulting assignments after voting, and the accuracy of the 'confident' assignments selected with a confidence threshold of 0.5. The 'confident' assignments comprise a significant (more than 35%) subset of all (*peak*, *residue*) assignments. Note that our approach requires constructing an ensemble around each of these 11 structural models. The results are in Fig. 6. The *x*-axis corresponds to the method used, where the first column is with the assignments using a single structural model, the second column is obtained using the consensus assignment scheme, and the third column is with the 'confident' assignment scheme. The plot shows the range of the assignment accuracies, the lower and upper quartiles, and the red line in the middle of the box is the median of the assignment accuracies. The whiskers show the extent of the data and the outliers are shown with '+'s. As can be seen, our approach not only improves the assignment accuracies, but also reduces the variance. Therefore, our ensemble-based assignment scheme improves the robustness of NVR with respect to structural noise.

## Application of our framework to MARS

We used our NMA ensemble-based voting scheme with MARS, an SBA tool that is significantly different from NVR in terms of the data it uses, as well as its algorithm. We tested our NMA ensemble-based voting scheme on three target proteins with corresponding close structural

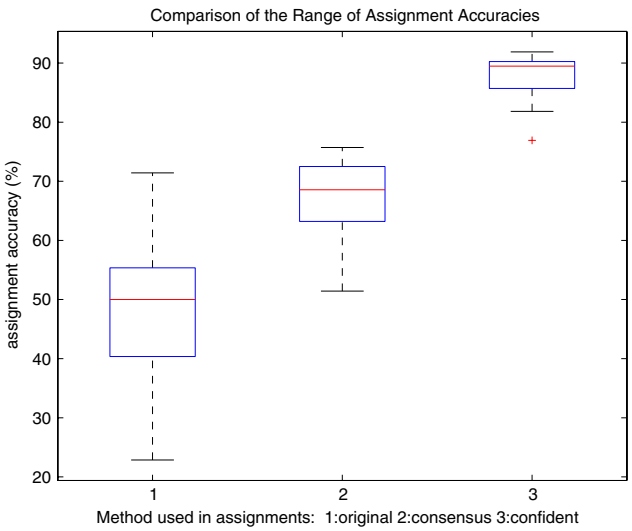


**Fig. 6** Our ensemble-based voting scheme improves the robustness of NVR against structural noise, and increases the assignment accuracies. We show the distribution of the assignment accuracies with the single-structure based NVR and our ensemble-based voting scheme with NVR, for 11 starting structures (obtained along an individual normal mode) that are structurally similar, for human ubiquitin with 1H8C as template. The first column shows the distribution of the assignment accuracies with single-structure NVR. The second column shows the accuracy with our ensemble-based voting scheme, and the third column shows the assignment accuracy of the subset of 'confident' assignments (with a confidence threshold of 0.5). The boxplot shows the lower and upper quartile, and the median in red. The whiskers show the extent of the accuracy results and the outliers are shown with '+' sign. There is only one outlier. The variance in assignment accuracies decreases with our algorithm, while the assignment accuracy increases

templates (Table 3). We only considered the subset of assignments that are labeled as 'reliable' ('H'igh and 'M'edium reliability) by MARS. MARS calls an assignment 'H'ighly reliable if it is consistent across all solutions obtained by MARS. According to Jung and Zweckstetter (2004c), 'M'edium does not fulfill all criteria for 'H' and the criterion is adjusted automatically according to the completeness of the input data. We report the number of correct and incorrect reliable assignments with each template in Table 4. We find and report the confident subset of the assignments with a confidence threshold of 0.05 in order to automatically discard incorrect assignments made by individual models in the ensemble. As in Jung and Zweckstetter (2004b), we tested our framework with MARS on different sets of data, such as with and without sequential connectivity information, and by varying the number of RDCs per residue. We report the number of correct and incorrect assignments for the original template, for the template in the NMA ensemble that has the highest number of reliable assignments, and the result of our voting scheme. Depending on the amount of data used as input, the number of correct and reliable assignments increases by

**Table 3** Proteins used with MARS

| Target[a] protein | Template (crystal) structure[b] (PDB ID) | Sequence[c] identity (%) | # of residues with data | # of residues | BMRB Code | RDCs (PDB ID) | bb[d] RMSD (Å) | CE[e] RMSD (Å) |
|---|---|---|---|---|---|---|---|---|
| Ubiquitin | 1UBQ | 100 | 76 | 72 | – | 1D3Z | 0.7 | 0.5 (100) |
| EIN | 1ZYM | 100 | 259 | 248 | 4106 | 3EZA | 3.7 | 1.2 (94) |
| MBP | 1DMB | 100 | 370 | 335 | 4354 | 1EZP | 3.4 | 3.3 (99) |

[a] The target protein (source of the NMR data). This structure was not used by MARS. Instead, a template structure was used

[b] PDB ID for the template structure. Unlike NVR, these templates were not used in homology modeling, but directly used with MARS

[c] The sequence identity between the sequences of the target and template protein, as computed by CE (Shindyalov and Bourne 1998)

[d] Backbone (bb) RMSD between the template and the target protein structure

[e] Backbone (bb) RMSD of the structural alignment as computed by CE (Shindyalov and Bourne 1998). The percentage of the residues involved in CE alignment is shown in parenthesis

**Table 4** MARS assignment accuracy improves with our NMA ensemble-based voting algorithm

| Protein name | Template | RDCs | Chemical Shifts for linking | Chemical shifts for matching | Reliable assignments # correct (# incorrect) | | |
|---|---|---|---|---|---|---|---|
| | | | | | Original model[a] | Best model[b] | NMA ensemble[c] |
| *Without sequential connectivity information* | | | | | | | |
| Human ubiquitin | 1UBQ | $^1D_{NH}$ | – | $C'_{i-1}$, $C^{\alpha}_{i-1}$, $C^{\beta}_{i-1}$ | 11 (0) | 16 (0) | 18 (0) |
| | | $^1D_{NH}$ $^1D_{NC'}$ | – | $C'_{i-1}$, $C^{\alpha}_{i-1}$, $C^{\beta}_{i-1}$ | 11 (0) | 26 (1) | 34 (3) |
| | | $^1D_{NH}$, $^1D_{NC'}$, $^1D_{CaC'}$ | – | $C'_{i-1}$, $C^{\alpha}_{i-1}$, $C^{\beta}_{i-1}$ | 51 (3) | 51 (3) | 57 (1) |
| *With sequential connectivity information* | | | | | | | |
| | | $^1D_{NH}$ | $C^{\alpha}$ | $C'_{i-1}$, $C^{\alpha}_{i-1}$, $C^{\alpha}_{i}$ | 51 (0) | 67 (2) | 66 (3) |
| | | $^1D_{NH}$, $^1D_{NC'}$ | $C^{\alpha}$ | $C'_{i-1}$, $C^{\alpha}_{i-1}$, $C^{\alpha}_{i}$ | 70 (0) | 72 (0) | 70 (0) |
| | | $^1D_{NH}$, $^1D_{NC'}$, $^1D_{CaC'}$ | $C^{\alpha}$ | $C'_{i-1}$, $C^{\alpha}_{i-1}$, $C^{\alpha}_{i}$ | 72 (0) | 72 (0) | 72 (0) |
| EIN | 1ZYM | $^1D_{NH}$ | $C^{\alpha}$, $C^{\beta}$ | $C'_{i-1}$, $C^{\alpha}_{i-1}$, $C^{\alpha}_{i}$, $C^{\beta}_{i-1}$, $C^{\beta}_{i}$ | 238 (2) | 244 (0) | 246 (2) |
| MBP | 1DMB | $^1D_{NH}$ | $C^{\alpha}$, $C^{\beta}$ | $C'_{i-1}$, $C^{\alpha}_{i-1}$, $C^{\alpha}_{i}$, $C^{\beta}_{i-1}$, $C^{\beta}_{i}$ | 323 (2) | 328 (0) | 329 (1) |
| | | $^1D_{NH}$, $^1D_{NC'}$ | $C^{\alpha}$, $C^{\beta}$ | $C'_{i-1}$, $C^{\alpha}_{i-1}$, $C^{\alpha}_{i}$, $C^{\beta}_{i-1}$, $C^{\beta}_{i}$ | 328 (0) | 331 (0) | 331 (0) |
| | | $^1D_{NH}$, $^1D_{NC'}$, $^1D_{CaC'}$ | $C^{\alpha}$, $C^{\beta}$ | $C'_{i-1}$, $C^{\alpha}_{i-1}$, $C^{\alpha}_{i}$, $C^{\beta}_{i-1}$, $C^{\beta}_{i}$ | 327 (0) | 331 (0) | 331 (0) |

MARS links fragments of pseudoresidues together (in the "linking" stage) and then maps them to the amino acid sequence (in the "matching" stage). The chemical shifts used for linking and matching are listed

[a] The number of correct (resp., incorrect) reliable (denoted as 'M'edium and 'H'igh reliability in MARS) assignments returned by MARS for the original template

[b] The results for the structure in the NMA ensemble that has the highest number of reliable assignments, as returned by MARS

[c] The number of reliable and confident assignments with our ensemble-based voting scheme. We used a confidence threshold of 0.05

8 for EIN and by up to 6 for MBP, while the number of incorrect assignments decreases or stays constant in most cases. For human ubiquitin, we compute the assignments using pdb ID 1UBQ as template, both with and without sequential connectivity information, and using up to three RDCs per residue. The number of correct reliable assignments increases by up to 23 at the expense of introducing three incorrect assignments (corresponding to a 3-fold increase in the number of correct assignments). Note that for MARS, unlike NVR, the best model in the NMA ensemble also leads to improved assignment accuracies. For instance, for the ubiquitin target without sequential connectivity information and with 2 RDCs per residue, the number of correct assignments increases by 15 while the number of incorrect assignments increases by one, for the best model in the NMA ensemble. This corresponds to 96% assignment accuracy with more than twice the original number of assignments.

## Discussion and conclusions

In this paper, we improved the assignment accuracy of NVR for distant structural models, and made it robust with respect to structural noise. On three different proteins, with distant structural homologs, we obtained an increased

assignment accuracy compared to the initial structural model for all cases but one, which used the template farthest from the target structure in our test set. However, in this case, combining the ensembles from both templates still increased the assignment accuracy. We also calculated a measure of confidence in the individual assignments. We used this measure to assign a subset of the peaks with even higher assignment accuracy. We also improved the robustness of NVR with respect to structural noise. We further demonstrated the general applicability of our approach to SBA by improving the assignment accuracy of MARS, a significantly different SBA algorithm from NVR.

Given a distant structural homolog, our methodology used NMA to obtain a set of structural models, which were then provided as input to NVR. We combined the NVR assignments for each of these structural models by maximum bipartite matching. The percentage of structural models that agreed on a given assignment provided the confidence measure. We also showed (see Appendix) that MBM is a maximum likelihood estimator of the correct assignments.

The greatest improvement with our ensemble-based assignments comes when we do not have sequential connectivity information. Nevertheless, modest improvements are seen even with sequential connectivities. Even these modest improvements are potentially useful, and our results represent a significant improvement over all previous structure-based assignment algorithms (e.g., Hus et al. 2002; Meiler and Baker 2003) for distant structural homologs (as opposed to exact crystal structure). No former SBA algorithm performs well using even slightly distant homologs. For instance, (Hus et al. 2002) was tested only on the crystal structure. In Meiler and Baker (2003), assignment accuracies in the range of only 5% to 40% were obtained using ROSETTA models with 3–6 Å RMSD from the native structure. Our approach tests whether ensembles for assignments can begin to overcome this bottleneck, and forms a basis for SBA that can be improved in the future.

Our approach demonstrates that an ensemble of structures simulating the fluctuations of a protein in its native state improves the accuracy and robustness of SBA. Furthermore, our voting scheme reinforces the signal (for the correct assignments), whereas the noise (incorrect assignments) cancels out. This is supported by the fact that we obtain high assignment accuracies despite the large fluctuations in assignment accuracies across the ensembles. Therefore, NMA is useful for both MR in X-ray crystallography (Suhre and Sanejouand 2004b) and SBA in NMR (this paper). Note that our results with MARS show that the best structure in the NMA ensemble helps improve the assignment accuracy, with respect to the starting template, analogous to (Suhre and Sanejouand 2004b). However, unlike (Suhre and Sanejouand 2004b), we also show that *an entire ensemble* is useful to improve the assignment accuracy with NVR.

It is interesting that our voting scheme obtains an assignment accuracy that is greater than or equal to the maximum assignment accuracy achieved by any individual structure in the NMA ensembles, both with MARS and NVR, for most target protein–template pairs. This suggests that our voting scheme is more likely to improve the assignment accuracy than any single-structure scoring function.

An analysis of our assignments reveals that the confident assignments (with a confidence threshold of 0.9) which have 95% or higher assignment accuracy mostly fall into regular secondary structure elements. For ubiquitin, GαIP and SPG, 1/5, 3/40 and 1/11 of the confident assignments fall into loop regions, respectively; furthermore for GαIP, the sixth helix contains most of the correct assignments, similarly, most of the confident assignments of SPG are in its alpha helix. The secondary structure elements are the similar regions between the target and the template protein, and therefore it is expected to find most of the correct assignments in those regions.

We envision three scenarios where our ensemble approach is useful. The first is for medium-sized proteins. One can perform a suite of triple-resonance experiments and use MARS with our ensemble method in order to improve MARS assignments, as was shown in this paper. Thus, we tested the hypothesis that SBA can be improved using ensembles, for medium-sized proteins. The second scenario is also for medium-sized proteins, but our NVR protocol requires only $^{15}$N-labeling and reduced spectrometer time. While RDCs must be measured, recent progress made by Tolman and co-workers (Ruan and Tolman 2005) make it more convenient to find multiple alignment media for the proposed RDC measurements. Measurement of RDCs for small- to medium-sized proteins usually only needs 2D IPAP experiments, and thus can be done in less time. The third scenario is for large proteins, where one can hopefully collect chemical shifts, $d_{NN}$'s, and RDCs (but other data might be hard to collect), and then use NVR with our ensemble-based technique. Since our algorithm requires only sparse data, this could make it less susceptible to the overlap problems that can occur with large proteins. Finally, since NVR requires only $^{15}$N-labeling, the cost of sample preparation is less for the last two scenarios.

Our approach should be valuable in pharmacology and drug design (Ferentz and Wagner 2000) by helping assign proteins for which there is no close structural homolog available. One could use our scheme to assign a subset of peaks with high confidence, and then do a few more disambiguating NMR experiments (e.g., using selective labeling) in order to assign the remaining peaks. Furthermore, it is possible to run the algorithm iteratively, setting the confident assignments found in the previous iteration to

boost the number of peaks reported with high confidence. Our method is simple and general, and can be used with other SBA algorithms, such as MARS, to improve their accuracy and robustness.

Our approach has some similarities to previous work such as Jung and Zweckstetter's (2004b) MARS and (Meiler and Baker 2003). Both of these works obtain multiple assignments for a protein, and retain the subset of peak-residue assignments that are consistent across those assignments. The difference is in how the assignments are computed. Jung and Zweckstetter (2004b) modulate the predicted chemical shifts by adding Gaussian noise and run MARS on perturbed data to obtain new assignments. Meiler and Baker (2003) start from random assignments and then use Monte Carlo search to optimize them. In contrast, we compute an ensemble of structures using NMA, and then use each structure to calculate a new assignment. Of these three approaches, ours is the only one that simulates the likely equilibrium conformations assumed by the template protein. It also has an intuitive correspondence with the NMR ensemble that generated the experimental data. As shown in section "Application of our framework to MARS", our approach can be used with MARS; it is likely that it can also be used with other (such as Meiler and Baker's 2003) SBA algorithms.

As future work, we are interested in developing a single-structure scoring function that takes into account the dependencies between various sources of NMR data. This would allow to choose a model from the ensemble that has the highest assignment accuracy. Secondly, other techniques that characterize the flexibility of protein structures such as FRODA (Wells et al. 2005) or protein ensemble method (Shehu et al. 2006) could be used and compared with NMA using the lens of SBA. Finally, NVR currently returns a single assignment for each template, even though there may be many assignments consistent with the structural model. Incorporating backtracking into the assignments as in Vitek et al. (2005) to obtain all consistent assignments could improve the accuracy and robustness.

## Availability

The NVR software as well as our scripts to run NVR on an ensemble of proteins and aggregate the results are available upon request. It is written in Matlab and Perl and is approximately 10K lines.

Our scripts to run MARS on an ensemble of templates and aggregate the results are less than 1K lines of code and are similarly available upon request.

**Acknowledgments** We thank Drs. C. Bailey-Kellogg, P. Zhou, Mr. D. Keedy, Mr. J. MacMaster, Mr. C. Tripathy, Mr. A. Yan, Mr. M. Zeng and all members of the Donald Lab for discussions and comments. This work is supported by a grant to B.R.D. from the National Institute of Health (R01 GM-65982).

## Appendix

### Analysis of MBM as a voting rule

#### *Motivation*

In this section, we justify our use of MBM as a voting rule in SBA to combine the assignments for each structure in the NMA ensemble. To that end, we show that MBM is a maximum likelihood estimator (MLE). Maximum likelihood estimation is a general technique to estimate the unknown parameters of a distribution, given a set of observed data values derived from the distribution. It returns the parameters that maximize the likelihood of observing the set of data values.

Our proof demonstrates that our voting scheme is sound and optimal, by showing that our algorithm returns the assignment that maximizes the likelihood. In our case, the set of observed data values comprises the assignments for each model in our NMA ensemble, and the unknown parameters of the distribution are the unknown correct assignments. An MLE estimator has many desirable properties: In particular, it is *consistent*, which means that it converges to the true value of the estimated parameter (Wasserman 2004). This means that, as the number of models in our NMA ensemble increases, the assignments returned by our voting scheme converge to the correct assignments. This proof depends on our assumption that the assignments computed for each model are independent and identically distributed, according to our noise model (which is described below).

First, we formulate our algorithm as a *voting scheme*. In voting, there are multiple voters and multiple candidates. Each voter may vote for one (or a subset of) the candidates, or may rank the alternatives. In our setting, a vote is the resonance assignments for a structure in our NMA ensemble. Our voting scheme aggregates these preferences to compute "consensus" assignments, which are returned by our algorithm.

The idea of using MLE in voting was first proposed by de Caritat (Marquis de Condorcet 1785), who analyzed 2- and 3-candidate elections; and was extended two centuries later to arbitrary number of candidates by Young (1995). However, none of the voting rules studied in these works corresponds to a widely-used voting rule. Conitzer and Sandholm (2005) then studied which of the well-known voting rules can be viewed as an MLE. For this purpose, they adopted the following model/assumptions: There exists an (unknown) ground truth winner (or ranking) of the election $w$, and each voter's vote is a *noisy*

measure of this ground truth. Due to noise, each voter's vote may be different from the ground truth. The *noise* models the probability of observing a vote $v_i$ for voter $i$, given the ground truth winner $w$. The votes are independent given $w$, and identically distributed. Under these assumptions, given a set of votes $v_1,\ldots,v_m$, where $m$ is the number of votes, a voting rule is an MLE of the correct winner $w$ if it returns a winner $w_o$ that maximizes the likelihood of the observed votes. That is, it returns:

$$\arg\max_{w_o} p(v_1, v_2, \ldots, v_m|w_o) = \arg\max_{w_o} p(v_1|w_o)p(v_2|w_o)\ldots p(v_m|w_o)$$

where $p(v_1,v_2,\ldots,v_m|w_i)$ is the probability of observing $v_1,v_2,\ldots,v_m$ if the (unknown) ground truth were $w_i$.

*Proof that MBM is an MLE*

We now show that our voting rule, MBM, is the MLE of the correct assignments. In our setting, there is a ground truth winner, which is the correct (and unknown) (*peak*, *residue*) assignments. The individual votes correspond to the assignments made using each of the structures in the NMA ensemble separately using an SBA algorithm (Fig. 2). The MBM is done on a BPG where one set of nodes corresponds to peaks and the other set to residues. The edge weights are the number of structures that assign ("vote for") the corresponding (*peak*, *residue*) pair.

We assume the following noise model: For each template in the ensemble ("voter"), each peak is correctly (resp., incorrectly) assigned with probability $p$ (resp., $q$) where $p > q$, independent of other peaks, and such that if the resulting assignments have more than one peak assigned to the same residue, each peak is reassigned (again with probability $p$ to the correct residue and probability $q$ to an incorrect residue). We further assume that the assignments corresponding to individual models are independent, given the correct assignments. So, the noise is independent and identically distributed.

With this noise model, the probability of a given assignment (vote) $i$ in which $k_i$ of the $n$ peaks are matched correctly is (proportional to) $p^{k_i}q^{n-k_i}$. The joint probability of all $m$ votes corresponding to all $m$ templates together is proportional to

$$\prod_i p^{k_i}q^{n-k_i} = p^a q^b \qquad a = \sum_i k_i \qquad b = nm - \sum_i k_i \tag{2}$$

where the product and the sums in (2) are from $i = 1,\ldots,m$.

An MLE of the correct assignment chooses an assignment $w_o$ such that (2) is maximized. Fix a particular protein and its NMA ensemble, so that $n$ and $m$ are constants. Then, since $p > q$ and $nm$ is a constant, (2) is maximized when $\sum_i k_i$ is maximized. $\sum_i k_i$ is the number of times each (*peak*, *residue*) assignment (for each of the structural models) coincides with the (*peak*, *residue*) assignment in $w_o$. This is maximized by MBM. Therefore MBM is an MLE of the correct assignment. □

## References


Al-Hashimi H, Gorin A, Majumdar A, Gosser Y, Patel D (2002) Towards structural genomics of RNA: rapid NMR resonance assignment and simultaneous RNA tertiary structure determination using residual dipolar couplings. J Mol Biol 318(3):637–649

Al-Hashimi H, Patel D (2002) Residual dipolar couplings: synergy between NMR and structural genomics. J Biomol NMR 22(1):1–8

Bahar I, Atılgan A, Erman B (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des 2(3):173–181

Bailey-Kellogg C, Chainraj S, Pandurangan G (2004) A random graph approach to NMR sequential assignment. In: RECOMB, San Diego, CA, pp 58–67

Best R, Vendruscolo M (2004) Determination of protein structures consistent with NMR order parameters. J Am Chem Soc 126(26):8090–8091

Conitzer V, Sandholm T (2005) Common voting rules as maximum likelihood estimators. In: Proceedings of the 21st annual conference on uncertainty in artificial intelligence (UAI-05), Edinburgh, Scotland, UK, pp 145–152

Cornilescu G, Marquardt J, Ottiger M, Bax A (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. J Am Chem Soc 120(27): 6836–6837

De Alba E, De Vries L, Farquhar MG, Tjandra N (1999) Solution structure of GaIP (galpha interacting protein): a regulator of G protein signaling. J Mol Biol 291(4):927

de Caritat (Marquis de Condorcet) MJAN (1785) Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. L'Imprimerie Royale, Paris

Ferentz AE, Wagner G (2000) NMR spectroscopy: a multifaceted approach to macromolecular structure. Q Rev Biophys 33(1): 29–65

Güntert P (2004) Automated NMR structure calculation with CYANA. Methods Mol Biol 278:353–378

Harris R (2002) The ubiquitin NMR resource page, BBSRC Bloomsbury Center for Structural Biology, http://www.biochem.ucl.ac.uk/bsm/nmr/ubq/index.html. Cited 02 Jun 2007

Holm L, Sander C (1991) Database algorithm for generating protein backbone and side-chain coordinates from a c alpha trace application to model building and detection of coordinate errors. J Mol Biol 218(1):183–194

Hus J, Prompers J, Brüschweiler R (2002) Assignment strategy for proteins of known structure. J Mag Res 157(1):119–125

Jung Y-S, Zweckstetter M (2004a) Mars—robust automatic backbone assignment of proteins. http://www.mpibpc.mpg.de/groups/griesinger/zweckstetter/_links/software_mars.htm. Cited 02 Jun 2007

Jung Y, Zweckstetter M (2004b) Backbone assignment of proteins with known structure using residual dipolar couplings. J Biomol NMR 30(1):25–35

Jung Y, Zweckstetter M (2004c) Mars—robust automatic backbone assignment of proteins. J Biomol NMR 30(1):11–23

Kay L (1998) Protein dynamics from NMR. Nat Struct Biol 5(Suppl):513–517

Krebs W, Alexandrov V, Wilson C, Echols N, Yu H, Gerstein M (2002) Normal mode analysis of macromolecular motions in a database framework: developing mode concentration as a useful classifying statistic. Proteins Struct Funct Genet 48(4):682–695

Kuhn H (1955) The Hungarian method for the assignment problem. Nav Res Logist Quart 2:83–97

Kuszewski J, Gronenborn AM, Clore GM (1999) Improving the packing and accuracy of NMR structures with a pseudopotential for the radius of gyration. J Am Chem Soc 121(10):2337–2338

Langmead C, Donald B (2004a) An expectation/maximization nuclear vector replacement algorithm for automated NMR resonance assignments. J Biomol NMR 29(2):111–138

Langmead C, Donald B (2004b) High-throughput 3D structural homology detection via NMR resonance assignment. In: Proc IEEE Comput Syst Bioinform Conf, Stanford, CA, pp 278–89. PMID: 16448021

Langmead C, Yan A, Lilien R, Wang L, Donald B (2003) A polynomial-time nuclear vector replacement algorithm for automated NMR resonance assignments. In: Proc the seventh annual international conference on research in computational molecular biology (RECOMB). ACM Press, Berlin, Germany, April 10–13, pp 176–187. Appears in: J Comput Biol 11(2–3): 277–98 (2004)

Leo-Macias A, Lopez-Romero P, Lupyan D, Zerbino D, Ortiz A (2005) An analysis of core deformations in protein superfamilies. Biophys J 88(2):1291–1299

Meiler J, Baker D (2003) Rapid protein fold determination using unassigned NMR data. Proc Nat Acad Sci USA 100(26):15404–15409

Mumenthaler C, Güntert P, Braun W, Wüthrich K (1997) Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. J Biomol NMR 10(4):351–362

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(3):215–240

Pearlman D, Case D, Caldwell J, Ross W, Cheatham T III, DeBolt S, Ferguson D, Seibel G, Kollman P (1995) Amber, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. Comp Phys Commun 91(1–3):1–41

Potluri S, Yan A, Chou J, Donald B, Bailey-Kellogg C (2006) Structure determination of symmetric homo-oligomers by a complete search of symmetry configuration space, using NMR restraints and van der Waals packing. Proteins 65(1):203–219

Potluri S, Yan A, Donald B, Bailey-Kellogg C (2007) A complete algorithm to resolve ambiguity for inter-subunit NOE assignment in structure determination of symmetric homo-oligomers. Protein Sci 16(1):69–81

Rossman M, Blow D (1962) The detection of sub-units within the crystallographic assymetric unit. Acta Crystal (D) 15:24–31

Ruan K, Tolman JR (2005) Composite alignment media for the measurement of independent sets of NMR residual dipolar couplings. J Am Chem Soc 127(43):15032–15033

Sali A, Blundell T (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234(3):779–815

Seavey B, Farr E, Westler W, Markley J (1991) A relational database for sequence-specific protein NMR data. J Biomol NMR 1(3):217–236

Shehu A, Clementi C, Kavraki LE (2006) Modeling protein conformational ensembles: From missing loops to equilibrium fluctuations. Proteins Struct Funct Bioinform 65(1):164–179

Shindyalov I, Bourne P (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11(9):739–747

Suhre K, Sanejouand Y (2004a) Elnémo: a normal mode web-server for protein movement analysis and the generation of templates for molecular replacement. Nucleic Acids Res 32(1):W610–W614. http://www.igs.cnrs-mrs.fr/elnemo/. Cited 12 Jun 2007

Suhre K, Sanejouand Y (2004b) On the potential of normal mode analysis for solving difficult molecular replacement problems. Acta Crystal (D) 60(4):796–799

Tjandra N, Bax A (1997) Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. Science 278(5340):1111–1114

Tolman JR, Flanagan JM, Kennedy MA, Prestegard JH (1995) Nuclear magnetic dipole interactions in field-oriented proteins: Information for structure determination in solution. Proc Natl Acad Sci USA 92(20):9279–9283

Vitek O, Bailey-Kellogg C, Craig B, Kuliniewicz P, Vitek J (2005) Reconsidering complete search algorithms for protein backbone NMR assignment. Bioinformatics 21(Suppl2):ii230–ii236

Wasserman L (2004) All of statistics: a concise course in statistical inference (Springer Texts in Statistics). Springer

Wells S, Menor S, Hespenheide B, Thorpe M (2005) Constrained geometric simulation of diffusive motion in proteins. Phys Biol 2(4):S127–S136

Xu XP, Case DA (2001) Automated prediction of 15N, 13C$\alpha$, 13C$\beta$ and 13C chemical shifts in proteins using a density functional database. J Biomol NMR 21(4):321–333

Xu Y, Xu D, Kai D, Olman V, Razumovskaya J, Jiang T (2002) Automated assignment of backbone NMR peaks using constrained bipartite matching. Comput Sci Eng 4(1). Life Sci Div, Oak Ridge Nat Lab, TN

Young P (1995) Optimal voting rules. J Econ Perspect 9(1):51–64