Journal of Magnetic Resonance 214 (2012) 103–110

Contents lists available at SciVerse ScienceDirect

Journal of Magnetic Resonance

journal homepage: www.elsevier.com/locate/jmr

FLAMEnGO: A fuzzy logic approach for methyl group assignment using NOESY and paramagnetic relaxation enhancement data

Fa-An Chao [a], Lei Shi [b], Larry R. Masterson [a], Gianluigi Veglia [a,b,*]

[a] Department of Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, Minneapolis, MN 55445, United States
[b] Department of Chemistry, University of Minnesota, Minneapolis, MN 55445, United States

ARTICLE INFO

Article history:
Received 10 August 2011
Revised 29 September 2011
Available online 20 October 2011

Keywords:
Methyl group assignments
Methyl-TROSY
Automated assignment
Sparse NMR data
Fuzzy logic
Monte Carlo

ABSTRACT

Building on a recent method by Matthews and co-workers [1], we developed a new and efficient algorithm to assign methyl resonances from sparse and ambiguous NMR data. The new algorithm (FLAMEnGO: Fuzzy Logic Assignment of MEthyl GrOups) uses Monte Carlo sampling in conjunction with *fuzzy logic* to obtain the assignment of methyl resonances at high fidelity. Furthermore, we demonstrate that the inclusion of paramagnetic relaxation enhancement (PRE) data in the assignment strategy increases the percentage of correct assignments with sparse NOE data. Using synthetic tests and experimental data we show that this new approach provides up to ~80% correct assignments with only 30% of methyl–methyl NOE data. In the experimental case of ubiquitin, PRE data from two spin labeled sites improve the percentage of assigned methyl groups up to ~91%. This new strategy promises to further expand methyl group NMR spectroscopy to very large macromolecular systems.

1. Introduction

Traditionally, the structural elucidation of large proteins and protein complexes at high resolution has been the realm of X-ray crystallography. However, this technique does not provide an atomic view of the molecular motions. Also, the presence of conformational transitions can only be inferred from the B-factors, rather than directly probed. On the other hand, conventional NMR approaches based on the main chain spectroscopy are not sensitive enough to study large macromolecular complexes.

In the past few years, selective methyl group labeling techniques in highly deuterated proteins, in concert with TROSY spectroscopy, enabled the analysis of large protein complexes up to 1 MDa [2]. A significant problem for this approach is the assignment of the methyl $^{1}H/^{13}C$ HMQC (methyl-TROSY) spectrum. Ideally, one can utilize methyl 'out-and-back' experiments, mapping all of the methyl groups and linking them to the protein backbone [3,4]. However, this strategy necessitates the assignment of the backbone nitrogen, C^{α}, or C' resonances, which is problematic to achieve for large systems. Moreover, the pulse sequences utilized for these experiments require high-level of deuteration and are generally very insensitive, due to fast T_2 relaxation. When the macromolecular complexes are sufficiently large or in the presence of conformational dynamics (broad resonances and overlap), this approach fails. In favorable cases, it is possible to use a *divide and conquer* strategy [5], where smaller fragments of the proteins or isolated components of the complex are expressed individually and the resonance assignment is transferred from the sub-spectra to the spectrum of the intact protein or protein complex [6,7]. Site specific mutagenesis has also been used to assign specific methyl resonances [7,8]. Though, the latter is very time-consuming and prone to generation of non-native folds.

To overcome these hurdles, Matthews and co-workers introduced an automated assignment procedure for methyl group assignment [1], which compares the experimental chemical shifts and NOE contacts with those back-calculated from an X-ray crystal structure. The procedure requires that the number of experimental NOEs is at least 50% of those back-calculated [1]. Another stringent condition is the unambiguous mapping of the NOE data onto the methyl-TROSY spectrum, which requires 4D F_2–^{13}C, F_3–^{13}C-edited NOESY experiments. Under these conditions, the approach leads to >90% correct assignments of methyl resonances for small and large systems [1]. However, large systems often display fewer NOE cross-peaks than those predicted from the X-ray structures. The lack of complete NOE networks significantly deteriorates the performance of this procedure [1]. In addition, resonance overlap makes it difficult to accurately map all of the NOE data to the donor resonances in the methyl-TROSY spectrum. While the use of a 4D NOESY spectrum alleviates this issue, this route is not robust enough for larger macromolecular systems, which have limited solubility and lower sensitivity.

Here, we present a new automated assignment algorithm, FLAMEnGO (Fuzzy Logic Assignments of MEthyl GrOups), which has high tolerance for sparse NOE information (as low as 30%),

* Corresponding author. Address: 6-155 Jackson Hall, 321 Church St SE, Minneapolis, MN 55455, United States. Fax: +1 612 625 2163.
E-mail address: vegli001@umn.edu (G. Veglia).

1090-7807/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2011.10.008

and enables the use of ambiguous methyl–methyl NOEs through the combination of Monte Carlo sampling and *fuzzy logic* [9]. FLAMEnGO can incorporate 3D NOESY data from amide-methyl contacts, 4D NOESY data, as well as paramagnetic relaxation enhancements (PREs). The latter represents a crucial aid for unambiguous assignment of the NOEs in large systems and in the presence of sparse data. Using only 30% of synthetic NOE data, we show that FLAMEnGO is able to assign ~70% of the methyl resonances of maltose binding protein (MBP) and cutinase. Finally, for experimental data acquired on ubiquitin, FLAMEnGO can achieve ~80% of the methyl assignments with only 30% of the NOE data, and up to ~91% when PRE data are included.

2. Theoretical basis of the algorithm

2.1. Global score function

The architecture of FLAMEnGO is illustrated in Fig. 1. As for the previously proposed method [1], our algorithm is based on a global score function that estimates the agreement between experimental and simulated NOE contacts. As an input, our algorithm requires a X-ray crystal structure, a peak picking of the 2D methyl-TROSY spectrum, and experimental NOESY data of the system under examination (Fig. 1A). A seed assignment is given to the 2D methyl-TROSY spectrum and no assignment is necessary for the NOESY spectrum. The global score function is defined as:

$$G(x) = \max_{a \in A}\{Match_{total}\}$$

where a is an assignment from the 2D methyl-TROSY spectrum and A is the set of all possible assignments of the methyl-TROSY spectrum , x is the NOE distance cutoff. The function $G(x)$ is used to find the maximum of a total matching function ($Match_{total}$), which is a linear combination of NOEs ($Match_{NOE}$), chemical shift ($Match_{CS}$), and PRE ($Match_{PRE}$) terms:

$$Match_{total} = Match_{NOE} + Match_{CS} + Match_{PRE}$$

Similar to the approach by Matthews and co-workers, a scaling factor is used in the $Match_{CS}$ term (here, we used 0.2), while all other terms are not scaled.

2.2. NOE matching and fuzzy logic

The NOE matching function compares the expected NOE contacts obtained from the crystal structure with the experimentally determined NOEs. The program first simulates all NOE contacts among methyl groups from a crystal structure using a given NOE distance cutoff (x), where contacts are considered to be any methyl pair less than x. Based on this NOE contact information, we use the experimental methyl group chemical shifts from the methyl-TROSY spectrum and a seed assignment to identify the expected NOESY cross peaks ($S_{NOE}(x,a)$), i.e., simulated NOESY spectrum. Note that we did not calculate the intensity of the NOEs, since we are only concerned about the number and position of expected NOE cross peaks from a 3D structure and the donor peaks in a methyl-TROSY spectrum.

In our algorithm, both the NOE distance cutoff and the peak assignment are regarded as variables, which are optimized to generate an optimal cutoff and the best-fit assignment. To compare the simulated and experimental spectrum (E), the algorithm matches the simulated NOE peak (p_2) to the closest experimental NOE peak (p_1). The NOE matching function ($Match_{NOE}$) is then defined as:

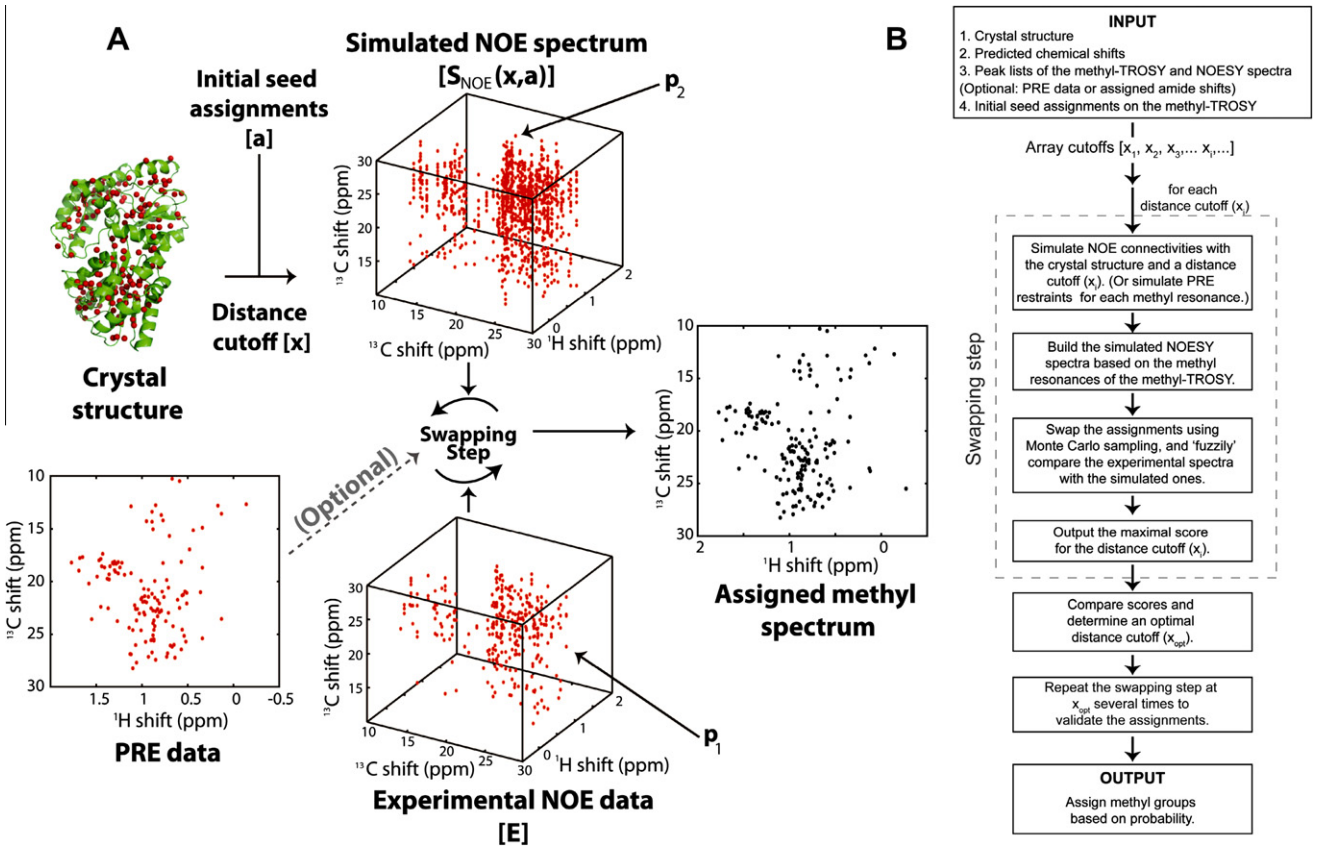

Fig. 1. Outline of the auto assignment procedure. (A) A graphical illustration of the algorithm shows that an X-ray crystal structure, experimental NOE data, and (optional) PRE data are required as input information. (B) A flowchart of the FLAMEnGO algorithm.

$$Match_{NOE}(S_{NOE}(x,a),E)=\sum_{p_1\in E}\max_{p_2\in S_{NOE}(x,a)} e^{-0.5\cdot D^2(p_1,p_2)} \tag{E1}$$

where $D(p_1,p_2)$ represents the distance between the experimental ($p_1(x_1,y_1,z_1)$ and simulated $p_2(x_2,y_2,z_2)$ peaks. $D(p_1,p_2)$ is estimated from the full width at half maximum (FWHM), assuming Gaussian shaped peak fitting with FWHM LW_x, LW_y and LW_z for the x, y, and z dimensions:

$$D(p_1,p_2)=2\sqrt{2\ln 2}\cdot\left[\left(\frac{(x_1-x_2)}{LW_x}\right)^2+\left(\frac{(y_1-y_2)}{LW_y}\right)^2+\left(\frac{(z_1-z_2)}{LW_z}\right)^2\right]^{1/2} \tag{E2}$$

The procedure starts with an initial seed assignment, which is swapped iteratively using a Monte Carlo sampling step [10]. The best-fit assignment is achieved when the match score function reaches its maximum. To compare the simulated spectra with experimental ones after each assignment swap, we used *fuzzy logic* [9]. Fuzzy logic accounts for ambiguous NOE information by providing a range of values between a perfect match ('100%' or 'yes') and a complete mismatch ('0%' or 'no'), rather than binary responses (such as 'yes' or 'no'). The fuzzy logic step allows the program to compare and score objects (*i.e.*, spectra) based on measurable criteria (*i.e.*, chemical shift differences), estimating a percentage of confidence.

Assuming a Gaussian line shape for the experimental NOE peaks, the FWHM is measured and converted in to a standard deviation of the average peak position. Then, the chemical shift difference between an experimental cross peak and a simulated peak will be divided by the standard deviation (E2). Using Eq. (E1), the program screens all simulated NOE cross peaks (p_2) to match the given experimental peak (p_1) by maximizing the Gaussian function. Thus, the smaller the difference between simulated and experimental chemical shifts, the higher the score of the Gaussian function. A score between 1 (complete match) and 0 (no match) from the Gaussian function will be returned for each peak (p_1) in the experimental NOESY spectrum (E), and the program sums all scores to provide a match score ($Match_{NOE}(S_{NOE}(x,a),E)$) for the experimental data and the simulated ones.

2.3. Predicted chemical shift information

As an option, we also included a matching function for chemical shifts predicted from the X-ray structure using software such as CH3*Shift* [11] or SHIFTX2 [12]. For the chemical shifts, the target function ($Match_{CS}$) for all of the methyl groups (i) is defined as:

$$Match_{CS}=-c\cdot\sum_i\frac{|\delta_i^r-\delta_i^p|}{\sigma_i} \tag{E3}$$

where δ_i^r represents the experimental chemical shifts, δ_i^p the predicted chemical shifts, σ_i the predicted error, and c is a scaling constant set to 0.2. The index, i, represents each methyl resonance. This option is very similar to that introduced by Matthews and co-workers [1], and minimizes the differences between experimental and predicted chemical shifts. For methyl-TROSY peaks that do not have NOESY cross-peaks with other resonances, this term can increase the percent of correct assignments.

2.4. PRE information

Finally, we included an additional function that uses PRE data ($Match_{PRE}$). The PRE data can be implemented either qualitatively or quantitatively. For *qualitative* PRE data, we used the following convention:

$$Match_{PRE}=\sum_i r_i\begin{cases} r_i=0 & \text{if the resonance } i \text{ is NOT in its PRE range}\\ r_i=1 & \text{if the resonance } i \text{ is in its PRE range}\end{cases} \tag{E4}$$

where r_i is a restraint imposed by PRE on each methyl resonance i. Methyl resonances are classified into three groups based on the quenching effects: unaffected (<20% reduction in signal intensity), slightly quenched (20–80%), and strongly quenched (above 80%). At the same time, all methyl groups in the structure are ranked based on their distance to the spin label and grouped into three categories in analogy with the quenching patterns. For all methyl resonance, $Match_{PRE}$ (E4) is used to score the agreement between experimental data and distance range. In other words, if a resonance in the strongly quenched group is assigned to a methyl group within close distance to the spin label, r_i is set to 1, indicating a good agreement of the assignment with PRE restraints. Otherwise, a penalty is added by setting r_i to 0.

Furthermore, our algorithm can also use a more quantitative interpretation of the PRE data, with explicit distance restraints between each individual methyl group and the spin label [13]. If *quantitative* PRE data are available, the distance between a methyl group and a spin label can be calculated using the following equations [13]:

$$\frac{I_{ox}}{I_{red}}=\frac{\exp(-2R_2^{ox}\tau)}{\exp(-2R_2^{red}\tau)}\cdot\left(\frac{R_2^{red}}{R_2^{ox}}\right)^2 \tag{E5}$$

$$d=\left[\frac{\beta}{R_{2,PRE}}\left(4\tau_c\frac{3\tau_c}{1+\omega_H^2\tau_c^2}\right)\right]^{1/6} \tag{E6}$$

where I_{red} is the peak intensity in the methyl-TROSY spectrum of the protein with the spin label in the reduced state and I_{ox} is the peak intensity in the oxidized state. The value τ is the time for transfer of magnetization between ^{1}H and ^{13}C spins, $\beta = 1.23\times10^{-44}\,m^6\,s^{-2}$, ω_H is Larmor frequency of protons (rad s^{-1}), and τ_c is the rotational correlation time of the protein (s) [13]. R_2^{ox} can be calculated from Eq. (E5) if R_2^{red} is known, and then $R_{2,PRE}=R_2^{ox}-R_2^{red}$ [13]. Finally, the distance d can be obtained, and an additional 2 Å uncertainty is included to form a PRE range for a particular methyl resonance [13]. Again, the same convention (Eq. (E4)) is used: if the resonance is assigned to a methyl group which is within $d-2$ and $d+2$ of the spin label, r_i is set to 1; otherwise a penalty is added by setting r_i to be 0.

2.5. Determination of the optimal NOE distance cutoff

An important step in the entire protocol is the optimization of the NOE distance cutoff (x) to calculate the NOESY data from the X-ray structure. Smaller NOE distance cutoffs are unable to take full advantage of the information in the NOE data, while larger cutoffs introduce more uncertainties, resulting in lower accuracy of the methyl assignment. In our approach, we optimize the cutoff by carrying out multiple assignment calculations with different distance cutoff values. Thus, the optimal distance cutoff is obtained when the *global score* function reaches a plateau. As a result, the simulated data with the optimal cutoff can utilize most of experimental data without overfitting. A schematic of the algorithm is reported in Fig. 1B. For a given NOE distance cutoff value, the total matching function is maximized using the Metropolis Monte Carlo method [10] by swapping the initial seed assignment on the experimental methyl-TROSY spectrum. The maximal value is called a *global score*. In the Monte Carlo sampling an annealing 'temperature' T defined as the total number of average cross-peaks plus PRE restraints is reduced to 1 in a ~1 million swapping steps. At each step, the assignments on two methyl resonances from the same residue-type (and the same

prochirality) are randomly chosen and exchanged. If the value of the target function increases or decreases within a random value, the swapping step is accepted; otherwise, it is rejected and a new pair of assignments is chosen and swapped. Depending on the system size, the number of steps can be adjusted to reach convergence. To determine the confidence for each assignment, we repeated the calculations at the optimal cutoff several times. The most probable assignment is chosen as the final assignment. The mathematical proof for the optimization of the NOE distance cutoff is provided in the Supporting Information.

3. Materials and methods

3.1. Sample preparation

The BL21(DE3) competent cells were transformed with plasmids containing the ubiquitin sequence, and inoculated into 1 l LB medium supplemented with 100 μg ampicillin. Upon reaching OD_{600} of ~1, the cells were harvested and transferred into 250 ml 100% deuterated M9 medium containing 1 g $^{15}NH_4Cl$, 4 g deuterated glucose, 70 mg/L methyl labeled α-ketoisovalerate, and 90 mg/L methyl labeled α-ketobutyrate. After 1 h of incubation at 37 °C, 1 mM IPTG was added to induce protein over-expression. The culture was harvested after 5 h of induction at 37 °C and stored at −20 °C. The frozen cell pellet was lysed by sonication in 50 mM sodium acetate buffer at pH 5.0 and centrifuged at 45,000 g at 4 °C. The supernatant was loaded into a P11 cation exchange column (WHATMAN) and eluted with a gradient of 0–1 M NaCl. The pooled fractions containing ubiquitin were further purified by size-exclusion chromatography using a Sephacryl S-200 resin (GE) with 100 mM phosphate buffer (pH 7.0). The fractions were concentrated and the sample was then dialyzed in NMR buffer (20 mM phosphate buffer, 1 mM NaN_3, and pH 6.5), and concentrated to ~1.5 mM for the NOESY experiments.

The K48C and G75C mutants of ubiquitin were generated using a QuikChange kit from Stratagene. Expression and purification were performed as described above. A fivefold excess of MTSSL was added to the mutant protein dissolved in NMR buffer at 25 °C for 4 h. The free MTSSL was dialyzed out in NMR buffer at room temperature. The final samples were concentrated to ~1.5 mM for PRE measurements.

3.2. NMR spectroscopy

A time-shared 3D HMQC-NOESY-HMQC was acquired on a Varian VNMRS instrument operating at a ^1H larmor frequency of 600 MHz. A mixing time of 800 ms was used based on the build-up in 2D planes at various mixing times (150–1000 ms). The spectrum was acquired using a spectral width of 10,000 (3500/2200) Hz for ^1H ($^{13}C/^{15}N$). The indirect dimensions were acquired with 128 increments in the carbon/nitrogen time-shared dimension, and 42 increments in the carbon dimension. For data processing, the number of points in the ^{13}C dimension was doubled by linear prediction and all dimensions were zero-filled. Before MTSSL spin labeling, the integrity of the mutant ubiquitin samples were confirmed by [^1H, ^{15}N]-HSQC and [^1H, ^{13}C]-HMQC, which showed no changes in the methyl resonances and negligible differences in the amide fingerprint (see Fig. S1). After MTSSL labeling, a [^1H, ^{13}C]-HMQC spectrum was acquired, the sample was then reduced in the presence of a 10-fold excess of DTT, and the spectrum was reacquired (Fig. S2).

4. Results

To test the performance of the algorithm in the case of sparse NOE data, we ran initial tests with two different proteins, whose structures have been determined at high resolution by X-ray and NMR: maltose binding protein (MBP) (PDB: 1DMB, BMRB: 7114) and cutinase (PDB: 1CEX, BMRB: 4101) (Fig. 2A and B). For the calculations, the stereospecificity of Leu and Val methyl groups in both proteins was assumed to be known. We generated two sets of sparse NOE data for both MBP and cutinase, with a distance cutoff of 7 Å, and randomly eliminated 70% of the back-calculated NOEs. Using the algorithm from Matthews and co-workers [1], we found that the optimal distance cutoffs were 6.1 and 5.4 Å for MBP and cutinase, respectively. As an output, we obtained 53% of correct methyl assignments for MBP (Fig. 2C) and 41% for cutinase (Fig. 2D). In contrast, FLAMEnGO found an optimal distance cutoff of 7 Å for both cases, and resulted in ~85% and ~76% correct assignments for MBP and cutinase, respectively, with an improvement of ~35% in accuracy with respect to the original approach. A higher accuracy in the determination of the NOE distance cutoff corresponds to a better tolerance for sparse NOE data. To evaluate the original algorithm against FLAMEnGO, we again calculated an assignment with the original script using the correct cutoff (7 Å). While the accuracy of the resulting assignment improved to 77% for MBP and 47% for cutinase, it was still found to be significantly lower than that obtained with FLAMEnGO. Since the same cutoff is used in this case, the improved results from our algorithm over the original are due to the more efficient sampling accessible by the Monte Carlo algorithm built in FLAMEnGO.

To test the algorithm in the presence of spectral overlap, we simulated 3D F_2–^{13}C, F_3–^{13}C-edited NOESY experiments for both MBP and cutinase using their X-ray structures with a 7 Å NOE distance cutoff. For the original method, we took the NOE data set and randomly eliminated 30% of the data. If an NOE cross-peak fell within 0.01 ppm from ^1H and 0.1 ppm from the ^{13}C frequencies in the methyl-TROSY spectrum, we assigned it to that donor resonance. Note that more than one donor resonance can be matched. In this case the donor is chosen arbitrarily. Under these conditions, the accuracy of the assignment using the original algorithm decreased by 55% and 20% for both MBP and cutinase, respectively (Fig. 2E and F). In contrast, FLAMEnGO provided ~97% and ~87% accuracy for MBP and cutinase, respectively, for cross-peaks falling within 0.025 ppm for ^1H and 0.4 ppm for ^{13}C frequency. Therefore, the performance of the original algorithm decreases with the increase of the uncertainty of the match between the donor and the cross-peak (Fig. 2E and F).

Moreover, we tested the performance of FLAMEnGO using 4D NOESY, PRE, and amide-methyl NOEs as inputs. For both MBP and cutinase, we back-calculated 3D amide-methyl and 4D methyl–methyl NOESY spectra, containing only 30% of the NOE contacts measurable in the X-ray structures. The line widths of ^1H, ^{13}C, ^{15}N in the simulated 3D NOESY data were set to 0.025 ppm, 0.4 ppm, and 0.5 ppm, respectively. For the simulated 4D NOESY data 0.8 ppm and 0.1 ppm for both ^1H and ^{13}C were used, respectively. We obtained ~10% improvement in the assignments of both MBP and cutinase when 4D data sets were used (Fig. 3A). Subsequently, we tested the effects of PRE data. For MBP, we simulated the effect of nitroxide spin labels engineered at positions 145 and 306 and back-calculated three groups of PRE effects: methyl groups within 15 Å of the spin label are strongly quenched, methyl groups between 15 and 35 Å are slightly quenched, and those beyond 35 Å are unaffected [14]. When the simulated PRE data were included in the calculations, the percentage of correct assignment reached ~86% for the 3D dataset and ~91% for the 4D dataset for MBP (Fig. 3A). Most strikingly, if the assignment of amide chemical shifts and amide-methyl NOE information were both included in the calculation, the percentage of correct assignment increased up to ~93% for MBP and ~87% for cutinase (Fig. 3B). It is worthy to note that only 15% of the simulated 3D amide-methyl NOESY data were used.

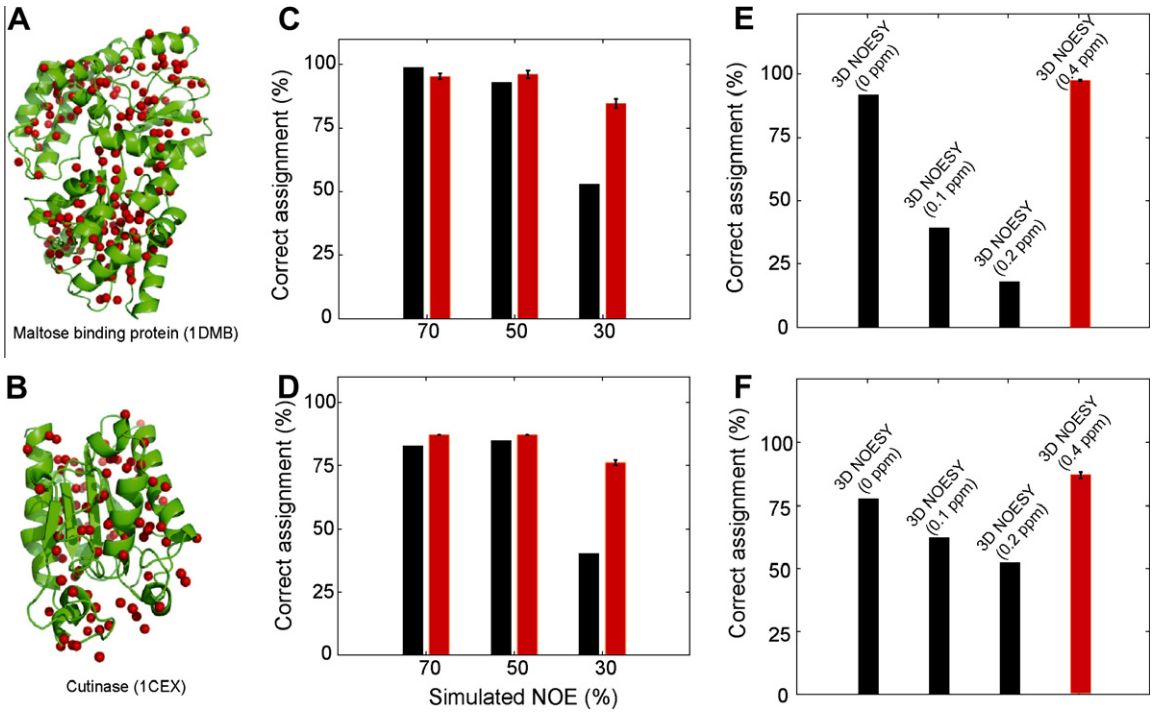

Fig. 2. Comparison of the performances of FLAMEnGO (red bars) and the original algorithm by Xu et al. [1] (black bars). Crystal structures of MBP (42 kDa, panel A) and cutinase (22 kDa, panel B). Methyl groups of Ala, Ile(δ1), Leu, and Val are marked as red spheres, for a total of 166 (92) methyl groups in MBP (cutinase). All unambiguous NOE contacts were simulated using the crystal structures and 7 Å NOE distance cutoff (panels C and D). All of the calculations were repeated three times. The error bars indicate the standard deviations. Panels E and F show the percentage of correct assignment as a function of the tolerance, i.e. chemical shift difference between NOE cross-peak and peak donor. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

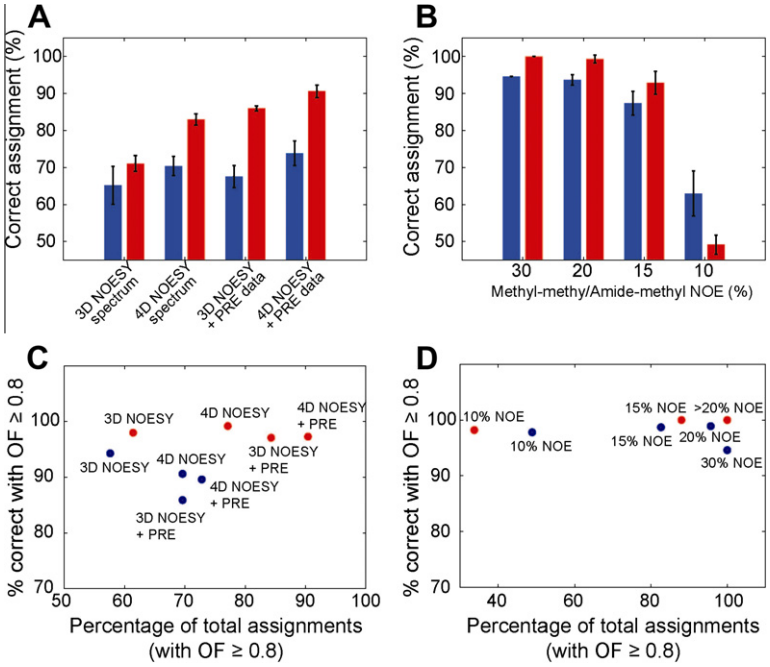

Fig. 3. Effects of different input data on FLAMEnGO's performance. Red bars indicate the calculations on MBP, while the blue bars indicate the calculations on cutinase. Only sparse and ambiguous data were used in the calculations. Data in A and B are based on five independent calculations. The error bars reflect the standard deviation among the different runs. The final assignments reported in panels C and D were based on the highest occurrence frequency (OF) for each assignment. For MBP (red), S145C and S306C were chosen as spin labeled sites, while for cutinase (blue), S54C and S135C were chosen. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Furthermore, we evaluated the probability of correct assignments by repeating the calculation five times with the same optimized NOE distance cutoff and choosing the assignment with the highest occurrence frequency (OF). In this way, we filtered out

low probability assignments using an OF of at least 0.8. The percent of correct assignments within this filtered subset (*i.e.*, % correct with OF $\geqslant$ 0.8, *y*-axis of Fig. 3C and D) was >95% and >85% for MBP and cutinase, respectively. Adding more experimental information (*i.e.*, PREs or 4D NOESY data) significantly increased the number of assignments with an OF $\geqslant$ 0.8 (Fig. 3C and D, *x*-axis). Therefore, additional experimental data provides higher confidence in the assignments, while the % correct with OF $\geqslant$ 0.8 did not change significantly (Fig. 3C and D, *y*-axis). When the 3D NOESY data with and without PRE information were used for cutinase, a ~10% drop occurs in the accuracy of the assignment with OF $\geqslant$ 0.8 (Fig. 3C). However, as shown in Fig. 3A, there is an increase in the average number of correct assignments. This is due to the fact that there are more overall assignments with an OF $\geqslant$ 0.8.

We also compared the performance of FLAMEnGO with structures of MBP derived from X-ray crystallography and solution NMR (all atom RMSD ~2 Å). Using the synthetic experimental NOE data from the crystal structure, the protocol gave a lower accuracy when different conformers were used (the difference was about 20%). However, the latter can be alleviated by providing additional stereospecific restraints. This demonstrated that relatively large structural deviations can be tolerated by this approach (Fig. 4) and highlights the importance of providing stereospecific information to the algorithm.

Finally, we tested the performance of the two algorithms with experimental data using a ubiquitin sample, which was uniformly ^{2}H, ^{15}N-labeled and selectively $^{1}H/^{13}C$-ILV methyl-labeled. We acquired a 3D (F_2–^{13}C, F_3–$^{13}C/^{15}N$-edited) time-shared NOESY at a mixing time of 800 ms. Since the original algorithm cannot handle amide-methyl NOEs, only methyl–methyl NOE data were used to provide a fair comparison. The experimental NOE data set displayed approximately 45% of the correlations expected from the X-ray crystal structure (PDB: 1UBQ, NOE distance cutoff = 5.9 Å). Since this number is likely to be smaller for large proteins, we reduced the data set to 30% of expected peaks. Ignoring the stereospecificity, FLAMEnGO provided ~70% correct assignments (Fig. S3), while the original method provided only 39% correct assignments.

To test the effect of PRE data, we expressed and purified two mutants of ubiquitin (K48C and G75C) and engineered two MTSSL spin labels. The [^{1}H, ^{15}N]-HSQC and [^{1}H, ^{13}C]-HMQC of the two mutants were essentially unaffected by the mutations (Fig. S1). The PRE effects were estimated qualitatively using the intensity of the resonances from methyl-TROSY experiments (Fig. S2). Using data

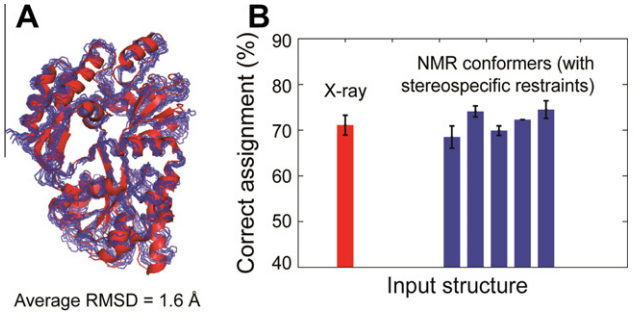

Fig. 4. Performance of FLAMEnGO using the ensemble of different conformers of MBP. (A) Overlay of a crystal structure (1DMB) and the solution structure ensemble (2H25) of MBP. The average all-atom RMSD between the crystal structure and each NMR conformer was up to 2.2 Å. (B) FLAMEnGO calculations for 5 selected NMR conformers. In addition to the seed assignments, we used residue-type information, stereospecific restraints, and 30% of the methyl–methyl NOE data as an input. Three independent calculations were carried out with an average of ~72% correct assignment.

corresponding to 45%, 30% or 15% of the expected methyl–methyl NOE cross peaks, as well as stereospecific and residue-type information, FLAMEnGO generated ~85% correct assignments (Fig. 5B) without PREs and ~91% with PREs. Remarkably, the inclusion of amide-methyl NOE cross peaks increased the correct assignments up to ~94% without PREs, and ~100% with PREs (Fig. 5D).

5. Discussion

Our approach builds on the previous work from Matthews and co-workers [1] for the assignment of methyl groups. A significant improvement with respect of the original approach is obtained with the *fuzzy logic* step and enhanced sampling by Monte Carlo method. The use of fuzzy logic allows the comparison of highly ambiguous NOEs obtained from 3D experiments and is more tolerant for sparse data. While the Matthews' approach does not perform well with sparse NOEs (less than 50% than expected), our algorithm performs well even in the presence of 30% of the NOE data. Sparse and ambiguous data sets are expected for large macromolecular systems or for situations in which conformational dynamics reduces the sensitivity of the NMR experiments. Therefore, this implementation will extend the methyl-based NMR spectroscopy approach to larger systems. Our program handles additional information such as PRE data, amide-methyl NOE data, and 4D NOESY data to improve assignment accuracy and convergence of the sampling,

As for the Matthews approach [1], our algorithm is sensitive to residue-type and stereospecific information available for the methyl groups. For instance, when applied to ubiquitin with the complete experimental NOESY data, but without residue-type information, FLAMEnGO provides ~82% of the correct assignment; in contrast to only ~55% of the correct assignment when stereospecific information is missing. Nonetheless, residue-type and stereospecific assignment can be easily obtained with selective labeling schemes [15,16], reducing the sampling space for the algorithm and increasing its convergence. For ubiquitin, the algorithm generated ~94% correct assignments when residue type information was provided.

What is the impact of the implementation of chemical shifts predicted from the X-ray structures? The protocol from Matthews and co-workers [1] relies on predicted chemical shifts as a crucial step in the assignment procedure. We found that the inclusion of predicted chemical shifts is important when NOE data are sparse. However, the inclusion of predicted chemical shifts may decrease the overall accuracy of the assignments, since discrepancies between calculated and experimental chemical shifts can be large. Furthermore, if the chemical shifts are not referenced correctly, it is possible to introduce severe errors in the assignment procedure.

As mentioned, the determination of optimal cutoffs for the NOE spectra simulated from crystal structures is very important. The NOE cutoff is optimized using several runs, arraying its value to obtain the minimal cutoff value, which accounts for most of the experimental NOE data. Although larger NOE cutoffs might appear optimal for taking into account long-range NOE information, they can also introduce more uncertainties during the assignment swapping steps of the algorithm. Long-range information can also be introduced in the algorithm using PRE effects, which we demonstrated to increase the accuracy of the assignment protocol dramatically. In fact, PRE data can provide long-range distance restraints that cannot be obtained from NOE, resulting in improvements of the assignment accuracy by up to 15% (Figs. 3A, C, 5B, and D).

Finally, the density of the NOE network influences the performance of the algorithm: the higher the number of NOE cross-peaks, the more accurate is the assignment. Moreover, pushing the limits on the range of the experimental NOE conveys more

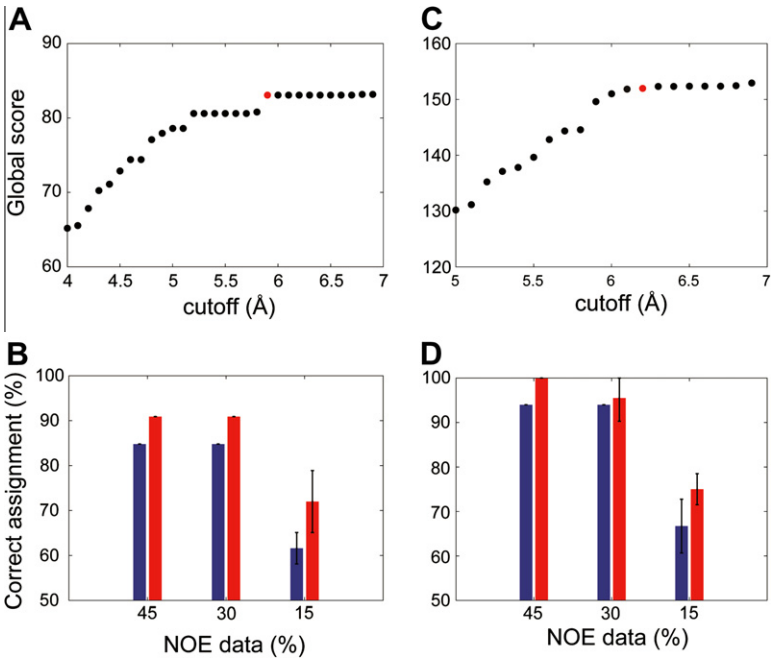

Fig. 5. Automated assignment of experimental data from ubiquitin using FLAMEnGO. Percent assignment with NOE data only is reported using blue bars, while with NOE and PRE data is reported using red bars. Three independent calculations were carried out. (A) Plot of the global score function versus the NOE distance cutoff. The optimal NOE distance cutoff (5.9 Å) is chosen at the curve plateau (red point). (B) Percentage of correct assignment obtained with only methyl–methyl NOE data (~45% of the expected NOE cross-peaks). (C) Plot of the global function versus NOE distance cutoff for both amide-methyl and methyl–methyl NOESY data. (D) Plot of the percentage of correct assignment using both amide-methyl and methyl–methyl NOESY experiments and NOE distance cutoff of 6.2 Å. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

information to the program to assign methyl groups. Thus, the mixing time in the NOESY experiment should be optimized to provide such information. In addition to Leu, Ile(δ1), and Val, new biosynthetic strategies have been introduced for producing proteins with $^{13}CH_3$ labeling at Ala, Met, Ile(γ2), and Thr [17–19]. The combination of these labeling schemes will provide more dense clusters of methyl–methyl NOE network, which should dramatically increase the performance of our approach.

6. Conclusion

In summary, we propose a new assignment strategy that is very tolerant to sparse and ambiguous NOE data. We demonstrated that the inclusion of either qualitative or quantitative PRE data, or amide-methyl NOE data, dramatically increases the convergence of the algorithm to assignment accuracy greater than 90%. These aspects make our approach more applicable to larger macromolecular systems, where sparse information due to size and intrinsic dynamics reduce the performance of NMR experiments.

Note added in Proof

A recent paper by Clore and co-workers (DOI 10.1007/s10858-011-9559-4) uses paramagnetic relaxation enhancement to speed-up the assignments on the methyl groups. While we use the same principle, our algorithm utilizes the concept of *fuzzy logic* to overcome the need of quantitative PRE data and the inclusion of 4D NOESY data.

Acknowledgments

This work was supported by the National Institute of Health (GM072701 to G.V. and T32DE007288 to L.R.M.). We also would like to thank Dr. Marco Tonelli at NMRFAM for assistance with the NMR spectroscopy. The scripts and instructions for FLAMEnGO are available at www.chem.umn.edu/groups/veglia under "downloads".

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmr.2011.10.008.

References

[1] Y. Xu, M. Liu, P.J. Simpson, R. Isaacson, E. Cota, J. Marchant, D. Yang, X. Zhang, P. Freemont, S. Matthews, Automated assignment in selectively methyl-labeled proteins, J. Am. Chem. Soc. 131 (2009) 9480–9481.
[2] R. Sprangers, L.E. Kay, Quantitative dynamics and binding studies of the 20S proteasome by NMR, Nature 445 (2007) 618–622.
[3] V. Tugarinov, L.E. Kay, Side chain assignments of Ile delta 1 methyl groups in high molecular weight proteins: an application to a 46 ns tumbling molecule, J. Am. Chem. Soc. 125 (2003) 5701–5706.
[4] V. Tugarinov, L.E. Kay, Ile, Leu, and Val methyl assignments of the 723-residue malate synthase G using a new labeling strategy and novel NMR methods, J. Am. Chem. Soc. 125 (2003) 13868–13878.
[5] A. Velyvis, H.K. Schachman, L.E. Kay, Application of methyl-TROSY NMR to test allosteric models describing effects of nucleotide binding to aspartate transcarbamoylase, J. Mol. Biol. 387 (2009) 540–547.
[6] I. Gelis, A.M. Bonvin, D. Keramisanou, M. Koukaki, G. Gouridis, S. Karamanou, A. Economou, C.G. Kalodimos, Structural basis for signal-sequence recognition by the translocase motor SecA as determined by NMR, Cell 131 (2007) 756–769.
[7] H. Kato, H. van Ingen, B.R. Zhou, H. Feng, M. Bustin, L.E. Kay, Y. Bai, From the cover: architecture of the high mobility group nucleosomal protein 2-nucleosome complex as revealed by methyl-based NMR, Proc. Natl. Acad. Sci. USA 108 (2011) 12283–12288.
[8] C. Amero, M. Asuncion Dura, M. Noirclerc-Savoye, A. Perollier, B. Gallet, M.J. Plevin, T. Vernet, B. Franzetti, J. Boisbouvier, A systematic mutagenesis-driven strategy for site-resolved NMR studies of supramolecular assemblies, J. Biomol. NMR 50 (2011) 229–236.
[9] D. Dubois, H. Prade, An introduction to fuzzy systems, Clin. Chim. Acta 270 (1998) 1–29.
[10] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller, Equation of state calculations by fast computing machines, J. Chem. Phys. 21 (1953) 1087–1092.

[11] A.B. Sahakyan, W.F. Vranken, A. Cavalli, M. Vendruscolo, Structure-based prediction of methyl chemical shifts in proteins, J. Biomol. NMR 50 (2011) 331–346.
[12] B. Han, Y. Liu, S.W. Ginzinger, D.S. Wishart, SHIFTX2: significantly improved protein chemical shift prediction, J. Biomol. NMR 50 (2011) 43–57.
[13] T.L. Religa, R. Sprangers, L.E. Kay, Dynamic regulation of archaeal proteasome gate opening as studied by TROSY NMR, Science 328 (2010) 98–102.
[14] V. Gaponenko, J.W. Howarth, G. Gasmi-Seabrook, J. Yuan, L. Columbus, W.L. Hubbell, P.R. Rosevear, Protein global fold determination using site-directed spin and isotope labeling, Protein Sci. 9 (2000) 302–309.
[15] L.Y. Lian, D.A. Middleton, Labelling approaches for protein structural studies by solution-state and solid-state NMR, Prog. Nucl. Mag. Reson. Spectrosc. 39 (2001) 171–267.
[16] M.J. Plevin, O. Hamelin, J. Boisbouvier, P. Gans, A simple biosynthetic method for stereospecific resonance assignment of prochiral methyl groups in proteins, J. Biomol. NMR 49 (2011) 61–67.
[17] M. Fischer, K. Kloiber, J. Hausler, K. Ledolter, R. Konrat, W. Schmid, Synthesis of a 13C-methyl-group-labeled methionine precursor as a useful tool for simplifying protein structural analysis by NMR spectroscopy, Chembiochem 8 (2007) 610–612.
[18] I. Ayala, R. Sounier, N. Use, P. Gans, J. Boisbouvier, An efficient protocol for the complete incorporation of methyl-protonated alanine in perdeuterated protein, J. Biomol. NMR 43 (2009) 111–119.
[19] A.M. Ruschak, L.E. Kay, Methyl groups as probes of supra-molecular structure, dynamics and function, J. Biomol. NMR 46 (2010) 75–87.