Structure-based methyl resonance assignment with MethylFLYA

Iva Pritišanac,[1] Julia Würz,[1] T. Reid Alderson,[2] and Peter Güntert*,[1,3,4]

[1]Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance, Goethe University Frankfurt am Main, 60438 Frankfurt am Main, Germany

[2] Laboratory of Chemical Physics, NIDDK, National Institutes of Health, Bethesda, Maryland 20892-0520, United States

[3]Laboratory of Physical Chemistry, ETH Zürich, 8093 Zürich, Switzerland

[4]Graduate School of Science, Tokyo Metropolitan University, Hachioji, Tokyo 192-0397, Japan

*Correspondence to: guentert@em.uni-frankfurt.de

Abstract

Methyl groups provide crucial NMR probes for investigating protein structure, dynamics and mechanisms in systems that are too large for NMR with uniform isotope labeling. This requires the assignment of methyl signals in the NMR spectra to specific methyl groups in the protein, an expensive and time-consuming endeavor that limits the use of methyl-based NMR for large proteins. To resolve this bottleneck, several methyl resonance assignment methods have been developed. These approaches remain limited with regard to complete automation and/or the extent and accuracy of the assignments. Here, we present the completely automated MethylFLYA method for the assignment of methyl groups. MethylFLYA requires as input exclusively methyl-methyl nuclear Overhauser effect spectroscopy (NOESY) peak lists. The algorithm was applied to five proteins of 28–358 kDa mass with a total of 708 isotope-labeled methyl groups. Manually made $^{1}H/^{13}C$ reference assignments were available for 674 methyls. The available experimental peak lists contained NOESY cross peaks for 614 methyls. MethylFLYA confidently assigned 488 methyls, i.e. 79% of those with NOESY data. Of these assignments, 460 agreed with the reference, 5 were different (and 23 concerned methyls without reference assignment). For three proteins of 28, 81, and 358 kDa, all confident assignments by MethylFLYA were correct. We furthermore show that, for high-quality NOESY spectra, automatic picking of NOE signals followed by resonance assignment with MethylFLYA can yield results that are comparable to those obtained for manually prepared peak lists, indicating the feasibility of unbiased, fully automatic methyl resonance assignment starting directly from the NMR spectra. This renders MethylFLYA an advantageous alternative to existing approaches for structure-based methyl assignment. MethylFLYA assigns, for most proteins, significantly more methyl groups than other algorithms, has an average error rate of 1%, modest runtimes of 0.4–1.2 h for the five proteins, and flexibility to handle arbitrary isotope labeling patterns and include data from other types of NMR spectra.

Introduction

The last decade of structural biology has seen growing interest in biologically relevant high molecular mass protein assemblies, as witnessed by an explosion of high- and low-resolution structural studies of macromolecular machines.[1-6] NMR spectroscopy is the principal experimental method for the simultaneous analysis of both the structures and dynamics of biomolecules at atomic resolution. The traditional size-limit of solution-state NMR spectroscopy, typically placed below 30 kDa, was overcome by Transverse Relaxation-Optimized SpectroscopY (TROSY).[7] The TROSY enhancement was subsequently realized also for selectively methyl-labeled proteins (methyl-TROSY).[8,9] Methyl-TROSY has since enabled studies of protein complexes in excess of 1 MDa[10,11] in unprecedented detail, revealing molecular mechanisms of RNA recognition and degradation[12], chaperone-substrate interactions,[13,14] quaternary dynamics of protein assemblies,[15] co-translational protein folding,[16] enzymatic activity,[17] allosteric communication,[18] and membrane protein dynamics associated with ligand binding.[19,20]

For optimal gains in the signal enhancement and resolution of methyl-TROSY spectra, protein samples are produced in which selectively protonated, ^{13}C-labelled methyl groups are reintroduced into an otherwise highly deuterated background.[21] To this end, cost-effective and robust biosynthetic strategies have been established for the selective or simultaneous labelling of all methyl-containing amino acids in *Escherichia coli,*[22,23] with selective labeling of methyl groups also possible in various eukaryotic expression systems.[24-26] The labeled methyl groups have favorable spectroscopic properties that render them observable also in large proteins and protein assemblies, where they serve as site-specific probes of molecular structure and dynamics (Figure 1). Distributed throughout the hydrophobic core of a protein and also across its surface, methyl groups provide faithful coverage and assessment of the fold[27] and interactions.[22]

The major bottleneck for NMR studies with selective methyl-labeled proteins is the resonance assignment, i.e. relating ^{1}H/^{13}C signals in the NMR spectra to specific methyl groups in the protein (Figure 1). In small and medium-size proteins, NMR signals from the protein backbone can be observed and used in triple-resonance, "through-bond" experiments for the sequence-specific resonance assignment of the backbone,[28] to which side-chain methyl resonances can be linked.[29] In contrast, for large proteins backbone resonances and triple-resonance spectra cannot be observed, and, unless the protein is modified, only nuclear Overhauser effects (NOEs) between methyl groups remain as NMR input data for assignment.

Assignment strategies for large proteins or proteins assemblies include divide-and-conquer approaches wherein individual protein domains or subunits are produced separately, for which backbone resonance assignments can be pursued using standard methods.[30] This approach requires that the resonance frequencies of the subsystems correspond well to those of the complete system. To complete the assignment, the approach is often supplemented with site-directed mutagenesis of individual methyl-bearing residues.[17,31] As an alternative, a high-resolution structure of the studied protein or complex can be utilized in combination with NMR experiments that reveal spatial proximity between methyl groups,[32,33] or between methyls and site-specifically attached paramagnetic probes.[34]

The laborious and time-consuming nature of these assignment strategies prompted automation efforts. Presently, two groups of structure-based, automatic assignment approaches are available: NOE spectroscopy (NOESY) or paramagnetism-based methods. Both rely on NMR-derived, sparse distance measurements that are compared to a known three-dimensional (3D) structure. Paramagnetic approaches require the site-specific introduction of paramagnetic probes, followed by calculation of the observables from the knowledge of the probes' locations and estimates of the magnetic susceptibility tensors, which are then compared to the measurements.[35-37] PRE-ASSIGN[37] uses paramagnetic relaxation enhancements (PREs) as the primary source of information, whereas PARAssign[36] relies on pseudo-contact shifts (PCSs). NOESY-based automatic approaches match a network of measured methyl-methyl distances to the network of short inter-methyl contacts predicted from the protein structure, using Monte Carlo[38-41] or graph-based[42,43] algorithms. For instance, MAGMA[42] uses exact graph matching algorithms to generate accurate assignments for a subset of well inter-connected methyls. For the remaining methyls, MAGMA reports all ambiguous assignment possibilities, which may be used for further experimental investigation.

Automated methods for structure-based methyl assignments can be characterized by the experimental requirements for measuring the input data, and by the completeness and accuracy of their assignments. An optimal algorithm functions with data that can be measured readily, tolerates experimental imperfections, is computationally efficient, and yields confident assignments for a large fraction of all methyls. To minimize the amount of error or subsequent manual checking, the algorithm (not the user) should distinguish confident assignments, which are almost certainly

correct, from other, tentative or ambiguous ones. Existing algorithms fall short of this ideal in different ways.

Therefore, we here adopt the FuLlY Automated assignment algorithm FLYA,[44] which is integrated in the CYANA structure calculation software[45] and has previously been shown to be capable to assign proteins exclusively from NOESY data,[46] for structure- and NOESY-based methyl resonance assignment. We apply the resulting MethylFLYA algorithm to a benchmark[42] of five large proteins and protein complexes and show that, on the basis of methyl-methyl NOEs alone, MethylFLYA can assign significantly more methyl resonances with high accuracy than the previously introduced methods MAGMA,[42] MAP-XSII,[39] and FLAMEnGO2.0[41] operating on the same input data. To demonstrate its robustness with respect to ambiguous and imperfect experimental information, MethylFLYA is applied also to unrefined peak lists, reduced input data sets, and peak lists obtained by automated peak picking with the CYPICK algorithm.[47]

MethylFLYA algorithm

The FLYA algorithm[44] determines resonance assignments by establishing an optimal mapping between expected peaks that are derived from the knowledge of the protein sequence, experiment types, and, if available, the 3D structure, and the observed peaks that are identified in the corresponding measured spectra. This mapping, and hence the assignments, are optimized by an evolutionary algorithm coupled to a local optimization routine.[44,48] MethylFLYA adopts the general FLYA algorithm for the assignment of methyl groups based on methyl-methyl NOEs and a known 3D structure. MethylFLYA uses the atomic positions from the input protein structure and CYANA's defined magnetization transfer routines for given input NMR experiment to compute a network of expected peaks (Figure 1). The mapping of the expected peaks to measured peaks starts from an initial population of random assignment solutions, which are optimized through successive generations. To select the best individuals for recombination a scoring function is employed, which favors the assignments that maximize the alignment of peaks assigned to the same atom as well as the assignment completeness, and minimize chemical shift degeneracy. In each generation, a local optimization routine reassigns a subset of expected peaks through a defined number of iterations. This protocol is repeated several times starting from different random initial assignments. Details of the MethylFLYA algorithm are given in the following sections.

MethylFLYA scripts. Automated methyl assignment with MethylFLYA is performed by four scripts (CYANA macros written in the INCLAN[49] programming language) that are given in the SI. The initialization macro, init.cya, is executed when CYANA starts and reads the library of residues and NMR experiment types, as well as the protein sequence. The preparation macro, PREP.cya, prepares the input data for the subsequent automated assignment calculations. This includes in particular the splitting of experimental peak lists according to amino acid type (see below) and the setup for generating the corresponding expected peaks, which is saved in the expected peak list generation macro, peaklists.cya. PREP.cya may also include other preparatory steps, such as attaching hydrogen atoms to an input 3D structure from X-ray crystallography. The calculation macro, FLYA.cya, performs the actual automated assignment calculations using the peaklists.cya macro to generate the expected peaks. FLYA.cya can be run several multiple times with different values for the NOE distance cutoff (see below). Finally, the consolidation macro, CONSOL.cya, consolidates the assignment results into a single consensus resonance assignment,[44] which is the main result of MethylFLYA.

Library of NMR experiments. The types of NMR experiments that contribute input peak lists to MethylFLYA are defined in the CYANA library[44,46] (Figure 1). For each spectrum type, the library entry defines the types of atoms that are observed in each spectral dimension and one or several magnetization transfer pathways that give rise to peaks. A magnetization transfer path is given by a probability for the observation of the corresponding peak and a linear list of atom types that defines a molecular fragment, in which atoms must be of the given type (e.g. $^{1}H_{amide}$, $^{1}H_{aliphatic}$, $^{1}H_{aromatic}$, $^{13}C_{aliphatic}$, $^{13}C_{aromatic}$, ^{15}N, etc.) and connected to the next atom in the list either by a covalent bond or by a NOE, i.e. a distance shorter than a given cutoff in the 3D structure. An expected peak is generated whenever a molecular fragment matches the covalent structure and, in case of NOEs, the 3D protein structure.

The following NMR experiments were used for MethylFLYA calculations in this paper: 2D [^{1}H,^{13}C]-HMQC (formally called C13HSQC in the CYANA library), 3D CCH-NOESY (CCNOESY3D), 3D HCH-NOESY (C13NOESY), 4D HCCH NOESY (CCNOESY), and, optionally, 4D short-mixing time HCCH NOESY (HCcCH). The latter experiment can be recorded on a doubly methyl-labelled ([$^{13}C_{\delta1}{}^{1}H_3$/$^{13}C_{\delta2}{}^{1}H_3$]-Leu, [$^{13}C_{\gamma1}{}^{1}H_3$/$^{13}C_{\gamma2}{}^{1}H_3$]-Val) protein sample to correlate the two methyl groups of Leu and Val to each other. It is formally treated as HCcCH-

TOCSY experiment in MethylFLYA. The experiment entries in the library is given in the Supporting Information (SI, cyana.lib).

Input peak lists. MethylFLYA operates on peak lists with observed peaks from the measured NMR spectra that contribute data for the resonance assignment (see above). The peak lists can be supplied in XEASY[50] format (SI, Input peak lists), or other formats supported by CYANA. If residue type-specific information is available, e.g. from appropriately isotope labeled samples, the [^{1}H,^{13}C]-HMQC peak list can be split into separate files containing only the methyl peaks of a certain residue type (called, for example, 'C13HSQC_V.peaks' for Val peaks). The NOESY peak lists can be split similarly according to the two amino acid types involved in an NOE. To this end, each NOESY peak is automatically attributed to the two [^{1}H,^{13}C]-HMQC peaks with best matching chemical shifts.[42] Separate peak lists are written for each pair of amino acid types (called, for example, 'CCNOESY_LL.peaks' and 'CCNOESY_LV.peaks' for NOEs between two Leu residues or between Leu and Val, respectively). Splitting peak lists by residue types is optional. MethylFLYA supports also joint lists for the resonances of Leu/Val type (Figure 3), as well as for any other amino acid type combinations.

Expected peak lists. Lists of expected peaks are generated by MethylFLYA for a given set of experiments based on the protein sequence, the 3D structure, the library of NMR experiments, and the isotope labeling pattern. The input 3D structure file must contain hydrogen atoms. For all calculations in this paper, hydrogens were added to the input X-ray structures using the CYANA command 'atoms attach'. If residue type-specific experimental peak lists are available, MethylFLYA generates a separate expected [^{1}H,^{13}C]-HMQC peak list for each amino acid type and separate NOESY peak lists for each pair of amino acid types. Separating the measured and expected peak lists restricts the matching of expected peaks to measured peaks of the same amino acid type(s) in the automated assignment algorithm (Figure 1).

The distance cutoff d_{cut} for NOEs is an important parameter for generating expected NOESY cross peaks because the number of expected NOEs increases approximately with the third power of the distance cutoff value. MethylFLYA computes the effective distance for a pair of methyl groups as the r^{-6} sum over the nine individual ^{1}H-^{1}H distances, such that $d_{\mathrm{eff}} = \left(\sum_{i=1}^{3} \sum_{j=1}^{3} d_{ij}^{-6}\right)^{-1/6}$, where d_{eff} stands for the effective distance, the sum includes all ^{1}H atoms of two methyl groups, and d_{ij} is the Euclidean distance between the individual methyl protons i and j that belong to two different methyl groups in the input structure. For the case that all d_{ij}

distances are assumed to be approximately equal, this yields $d_{\text{eff}} \approx 9^{-1/6} d_{ij} = 0.693\, d_{ij}$. Applying, for instance, a 5 Å cutoff to the effective distance d_{eff}, the inter-carbon distance between the two methyl groups may reach up to 5.0 / 0.693 + 2 × 1.1 ≈ 9.4 Å. To avoid giving high confidence to methyl assignments that are impacted by minor changes of the NOE distance cutoff parameter d_{cut}, MethylFLYA performs assignment calculations with the three slightly different cutoffs of d_{cut} and $d_{\text{cut}} \pm 0.5$ Å, and determines the consensus assignments from the results obtained with the three cutoffs (see below).

The observation probability was optimized (see below) and then set to 0.1 for expected NOESY peaks, and 1 for expected C13HSQC and HCcCH peaks for the calculations in this paper.

Optimization of assignments. Assignments are optimized by MethylFLYA using the same algorithm as the original FLYA method.[44] MethylFLYA uses chemical shift tolerances for the assignment calculations and results evaluation These were set to 0.4 ppm for ^{13}C and 0.04 ppm for ^{1}H chemical shifts for all calculations of this paper. The population size for the evolutionary optimization algorithm[44] was set to 200, the value was previously found to be optimal for NOESY-only FLYA calculations.[46] The number of iterations of the local optimization routine that is coupled to the evolutionary algorithm was kept at the default value of 15000. For each distance cutoff value, MethylFLYA performs 100 independent runs of the optimization algorithm with identical input data and parameters that start from different initial random assignments.

Consensus assignments. It is important for an assignment algorithm to distinguish between reliable assignments, in which the algorithm has a high confidence, from others that tentative or ambiguous. To establish the confidence of the assignment of an individual atom, MethylFLYA analyzes the chemical shift values obtained in a series of independent runs of the optimization algorithm. The global maximum of the sum of Gaussians centered at the chemical shift values of the given atom in the individual optimization runs defines the consensus chemical shift value of the atom.[44] The standard deviation of these Gaussians is set to the chemical shift tolerance value of the atom (0.4 ppm for ^{13}C and 0.04 ppm for ^{1}H). A consensus assignment is classified as "strong" (reliable) if more than 80% of the integral of the sum of Gaussians is concentrated in the region of the consensus shift ± tolerance, i.e. if more than 80% of the individual runs yielded (within the tolerance) the same chemical shift value. It has been shown for the original FLYA algorithm that strong assignments are much more accurate than the remaining "weak" ones.

In MethylFLYA, the consolidation is enhanced in three ways. (i) Three series of 100 individual runs are performed with three different distance cutoffs differing by ±0.5 Å for the generation of expected NOESY peaks (see above), and the consolidation is performed over all 3 × 100 individual runs of the optimization algorithm. This makes the algorithm less susceptible to the, necessarily somewhat arbitrary, choice of the NOE distance cutoff value, and reduces the number of inaccurate strong assignments. (ii) Special measures are necessary for the isopropyl methyls of Leu and Val, for which the stereospecific assignment is a priori unknown. In this case the chemical shift values obtained for the two methyls in the individual runs are redistributed such the consensus assignments of the first/second methyl group are determined from the smaller/larger of the two chemical shift values in each run. This simple approach implemented in the original FLYA algorithm[44] treated ^{1}H and ^{13}C assignments independently and could lead to inconsistent consensus assignments for the ^{1}H and ^{13}C resonances of a methyl group that had never occurred in the individual runs. To avoid this problem, the ^{1}H and ^{13}C chemical shifts of Leu and Val isopropyl groups are consolidated jointly in MethylFLYA. (iii) Methyl assignments are only accepted as strong if at least one methyl-methyl NOE is assigned to the methyl group. The excludes assignments for which no experimental basis exists.

MethylFLYA output. At the end of an assignment run, MethylFLYA outputs the list of consensus chemical shifts (consol.prot) and a table with assignment results (consol.tab). In the consol.tab file, strong (reliable) assignments are marked with the label "strong" but also other, tentative and ambiguous assignments are reported for possible manual inspection. Further details about the assignment process are given in the flya.txt file. It reports statistics of the expected, measured, assigned peaks in each peak list, which are useful to detect problems with individual spectra or the assignment as a whole. In addition, more detailed information about the reliability of each resonance assignment is given, and, for each assignable atom, the expected and mapped measured peaks that have been used to establish the assignment are reported.

Methods

Experimental data. MethylFLYA was applied to the five largest proteins of a benchmark data set that was originally prepared for evaluating the MAGMA algorithm for automated methyl assignment, as described in the original publication.[42] In addition, experimental data for the 20

kDa N-terminal domain (N-domain) of Heat Shock Protein 90 (HSP90), which has also been used previously with MAGMA,[51] was used for evaluating MethylFLYA in combination with automated peak picking with CYPICK.[47] The main benchmark data set comprised five proteins of varying molecular mass and shape for which NOESY data from specifically methyl-labeled samples, assignments and 3D structures are available (Table 1):[42] the N-terminal domain of E. coli Enzyme I (called EIN in this paper, molecular mass 28 kDa),[52] a dimer of regulatory chains of aspartate transcarbamoylase from *E. coli* (ATCase, 34 kDa),[34] maltose binding protein (MBP, 41 kDa),[53] malate synthase G (MSG, 81 kDa),[27,29] and the 20S "half-proteasome", a 14-mer ($\alpha_7\alpha_7$, 358 kDa).[54]

The following data were taken from the MAGMA benchmark:[42] (i) Assigned [^{1}H,^{13}C]-HMQC peak lists providing reference assignments, which were not used as input data for MethylFLYA but to evaluate the accuracy of its results. Only unassigned versions of the [^{1}H,^{13}C]-HMQC peak lists were as input for MethylFLYA. (ii) Filtered and unfiltered (see below) NOESY peak lists lists from 3D (ATCase, $\alpha_7\alpha_7$) or 4D (EIN, MBP, MSG) methyl-methyl NOESY spectra. (iii) Solution or crystal structures of the proteins, taken from the Protein Data Bank with accession codes 1EZA for EIN, 1D09 for ATCase, 1EZ9 for MBP, 1D8C for MSG, and 1YAU for $\alpha_7\alpha_7$. In addition, MethylFLYA calculations were performed for the alternative structural forms 1TUG for ATCase, 3MBP for MBP, and 1Y8B for MSG. For automated peak picking with CYPICK, processed [^{1}H,^{13}C]-HMQC and NOESY spectra in Sparky[55] format were supplied for EIN, ATCase and HSP90. Information about Leu/Val geminal methyl pairs, which was available in the MAGMA benchmark,[42] was incorporated into the MethylFLYA calculations in the form of simulated HCcCH TOCSY peak lists.

Two sets of experimental methyl-methyl NOESY peak lists were available for the five proteins from the MAGMA benchmark.[42] The first set included peak lists that were filtered for reciprocity of donor and acceptor NOE cross peaks (only the reciprocated peaks were kept), and signal-to-noise ratios (only the peaks with $S/N \geq 2$ were kept). The second set, the unfiltered peak lists, generated by manual analysis of NOESY spectra using Sparky[55] software. These peak lists are identical to those used in the earlier MAGMA study.

Optimization of MethylFLYA parameters. To establish optimal parameters for the MethylFLYA calculations, we tested a range of values for methyl ^{1}H–^{1}H distance cutoffs for the generation of expected NOESY cross peaks, d_{cut} = 3.0, 3.5, …, 8.0 Å (Figures S1, S2), observation

probabilities for expected methyl-methyl NOESY peaks, p_{NOE} = 0.1, 0.2, …, 0.9 (Figures S1, S2), and the number of independent assignment optimization runs (Figure S3).

Automated peak picking with CYPICK. The CYPICK[47] algorithm for automated peak picking was applied to the NOESY spectra of EIN, ATCase, and HSP90. CYPICK relies on analyzing 2D contour lines of the spectrum, which are calculated at intensity levels $I_i = \beta L \gamma^i$ (i = 0, 1,…), where L is the noise level of the spectrum that was estimated automatically by CYPICK. In this study, we used baseline factor values β = 2, 3, 4, 5, 10 while keeping γ fixed at 1.3. The scaling factors for the spectral dimensions[47] were set to 0.18 and 0.16 ppm for the first and second ^{13}C dimension, and 0.036 ppm for the ^{1}H dimension. The manually prepared 2D [^{1}H,^{13}C]-HMQC peak list was used as a frequency filter in CYPICK, restricting peak picking in the ^{13}C/^{13}C-separated NOESY spectrum to locations within 0.01/0.1 ppm for ^{1}H/^{13}C from a [^{1}H,^{13}C]-HMQC peak position. Local maxima within the tolerance range that fulfilled the circularity and convexity criteria[47] were considered as peaks and stored in the peak list.

The peak picking performance was evaluated by the find, artifact, and overall scores (with an artifact weight of 0.2) with respect to manually prepared reference peak lists[42] using a tolerance of 0.04 ppm for ^{1}H and 0.4 ppm for ^{13}C chemical shifts, as described in the CYPICK publication.[47]

Comparison with other assignment algorithms. The performance of the alternative structure-based methyl assignment algorithms MAGMA,[42] MAP-XSII,[39] and FLAMEnGO2.0[41] has been compared earlier.[42] Here, we used the available results and identical parameters,[42] with the exception of the MSG dataset, for which the calculations were repeated using the crystal structure (PDB ID 1D8C). The mutual agreement between the resonance assignments generated by the different methods was visualized using an online tool available at the GPCRdb web interface (http://www.gpcrdb.org/signprot/statistics).

Results

MethylFLYA parameter optimization. While most parameters of the MethylFLYA algorithm can be kept at the values that had been found optimal in earlier applications of the original FLYA algorithm,[44,46,56-59] specific optimization of a small number of parameters that are of particular relevance to structure-based methyl assignment was valuable.

MethylFLYA considers only methyl-methyl distances below a user-defined cutoff d_{cut} = for generating expected methyl-methyl NOESY cross peaks based on a protein structure (see above). In addition, each expected peak is attributed a probability value to (roughly) reflect a probability of actually observing it in the corresponding measured spectrum. For expected NOESY cross peaks, we tested a range of distance cutoffs and distance-dependent observation probabilities (Figure S1). Across these parameter values, we monitored the fraction of correct and incorrect strong (i.e. confident) methyl assignments and the percentage of explained input NMR data (methyl-methyl NOEs). Even though protein-specific profiles can be observed in Figure S1, the fractions of assigned methyl resonances generally plateaued around d_{cut} = 5 Å for EIN, ATCase, MBP, and MSG, or d_{cut} = 6 Å for $\alpha_7\alpha_7$ (Figure S1). These plateaus coincided with about 80% explained input NMR data, which was determined as optimal for these data sets. Increasing the distance-dependent probabilities generally diminished the quality of the results, as more incorrect assignments were obtained (Figure S2). Predictably, more NOEs were explained at higher distance cutoffs, but assignment errors also increased. In most cases, the assignment accuracy peaked around the plateaus of assigned methyl fractions and decreased at higher (≥7 Å) and lower (≤4 Å) distance cutoffs. Based on Figures S1 and S2, we used d_{cut} values of 5.0 ± 0.5 Å for EIN, ATCase, MBP, and MSG, and 6.0 ± 0.5 Å for $\alpha_7\alpha_7$, as well as a NOESY cross peak observation probability of 0.1 for all following MethylFLYA calculations.

Assignment completeness and accuracy. Using manually analyzed and refined NOE data,[42] MethylFLYA assigned between 63% (ATCase) and 89% ($\alpha_7\alpha_7$) of the methyl resonances for which reference assignments are available (Figure 2B), with no assignment errors for three out of five proteins in the benchmark, EIN, MSG, and α7α7. Two incorrect methyl assignments were found for MBP, and three for ATCase (Figure 2B). In the context of the structures, a more careful analysis of the errors revealed that the incorrectly assigned methyls are located in close proximity to their correct assignment positions (Figure S4, Table S1). The impact of such assignment errors is thus expected to be minor in studies that require lower resolution information, for instance, when identifying an interaction interface.

We also note that more stringent criteria can be applied to define the confident (strong) methyl assignments, which further reduce errors. For instance, increasing the requirement for self-consistency of assignments from multiple parallel runs of the algorithm from 80 to 90% (see Methods), results in a decrease in error for ATCase from 5% to 1%. This is achieved at the expense

of reducing the percentage of strong assignments on average by 6%. It is thus possible to "sacrifice" some of the strong assignments to ensure a higher accuracy.

Importantly, MethylFLYA is robust with respect to the presence of ambiguous or incorrect methyl-methyl NOEs, as judged by its comparable, or in some cases even better, performance on 'raw' NOE peak lists that were not filtered for NOE cross peak reciprocities and signal-to-noise ratios (Figure 2).

A spatial clustering of strong assignments can be discerned in the structures of EIN, ATCase, and MSG (Figure 2C). This is likely due to the low number of long-range NOEs between the clusters. In addition to the strong assignments, MethylFLYA outputs ambiguous assignment options for all resonances to which at least one inter-methyl NOE is attributed. The number of ambiguous assignment possibilities to be displayed can be specified by the user.

Reduced data sets. We tested the performance of MethylFLYA on the benchmark when experimental information provided to the algorithm was reduced (Figure 3). In the best-case scenario, both knowledge of the amino acid types of methyl resonances and linkage of two geminal methyl groups of Leu and Val is be available (Figure 3A, 3C, black). The Ile-δ_1 resonances are usually readily identified due to their upfield shifted ^{13}C frequencies. However, to discriminate between Leu and Val resonances, separate protein samples can be prepared using selective labelling schemes. For instance, by using ^{13}C-labeled α-ketoisocaproate, selective Leu labelling can be achieved[60], whereas the combined addition of unlabeled α-ketoisocaproate and labeled α-ketoisovalerate leads to exclusive labeling of Val.[61] To connect resonances from the two geminal Leu/Val methyl groups, an additional protein sample can be prepared in which both Leu/Val-methyl groups are protonated and ^{13}C-labelled. A short-mixing time NOESY experiment can then be used to record cross-peaks between the two methyl groups that belong to the same Leu or Val residue[31,42] (Fig 3A). Without discrimination between Leu and Val resonances, MethylFLYA performed very similarly as in the best-case scenario for EIN, MSG, and α7α7, confidently assigning 68, 62, and 84% of the methyl resonances, respectively, with complete accuracy (Figure 3C, dark grey). For ATCase and MBP, the percentage of accurate confident assignments slightly dropped (by 3%), whereas it increased for ATCase by the same amount (3%).

Removing the geminal Leu/Val pairing had a more significant impact in all cases, reducing the percentage of assigned methyls by ~19% for EIN, ATCase, MBP, and MSG, and up to 30% for $\alpha_7\alpha_7$ (Fig 3C, light grey). The overall accuracy, however, remained high. The critical

importance of this restraint for automatic methyl assignment was reported previously in the MAGMA study.[42] In the MAGIC study, a four-fold decrease in computational time and somewhat improved assignment accuracy were noted as benefits of the restraint.[43] As an alternative, the information about Leu/Val geminal pairs can be substituted with stereospecific labelling schemes that restrict isotopic labeling to only pro-*R* or pro-*S* methyl groups, and thus reduce the number of methyl resonances in the [^{1}H,^{13}C]-HMQC spectrum.[62] For MethylFLYA, removing both the Leu/Val-geminal pairing and discrimination between Leu/Val methyl resonances mostly resembled the outcome of the geminal pairing removal (Fig 3C, silver), and led overall only to slight further increase in erroneous assignments (1–2%). Interestingly, for ATCase, when reducing the information content, removing the Leu/Val resonance discrimination was always beneficial for accuracy (Fig 3C, dark grey, silver). Removing the Leu/Val residue discrimination, especially for smaller proteins (<80 kDa), might generally benefit methyl assignment with MethylFLYA.

In the benchmark, MethylFLYA's calculation speed roughly scaled with the number of methyl groups in the protein, and the protocol took between 22 minutes and about 1.5 h in the (Table S3). Negligible differences in speed were noted for the calculations with lower input information content (Figure 3, Table S3). This illustrates the ability of MethylFLYA to quickly deliver high-quality assignments even from considerably reduced input data, which gives it a considerable advantage over presently existing methods.

Combination with automated peak picking. All currently available automatic methyl resonance assignment strategies rely, to different extent, on a manual analysis and interpretation of the NMR data. The NOE-based methods, for instance, require manual inspection of methyl-methyl NOESY spectra to generate peak lists as input to the assignment software[38-43]. Manual analysis of NOESY data is a time-consuming and inherently user-biased task, complicated by spectral artifacts, low signal-to-noise ratios, and signal overlaps (Figure S6). We investigated whether an automatic peak picking algorithm, CYPICK,[47] which can replace human visual inspection of the NOESY spectra, could be used in combination with MethylFLYA to fully automate methyl resonance assignment. We tested the CYPICK-MethylFLYA combination on three proteins from the MAGMA study for which methyl-methyl NOESY spectra were available (Figure 4). For these spectra, CYPICK found 77–83% of the reference methyl-methyl NOEs (Figure S5, Table S2), which comparable to its performance previously reported on 3D ^{13}C-edited and ^{15}N-edited NOESY spectra.[47] The somewhat high CYPICK artifact scores for EIN (34%) and

the HSP90-N domain (46%) did not result in assignment errors, as only one methyl group misassignment was found for EIN and three for HSP90. For EIN, even sligthly more methyls were confidently and accurately assigned when the automatically generated CYPICK peak lists (78%) were used compared to the manually prepared lists (68%).

Despite the relatively large number of assignments for EIN, similar success was not found for the HSP90 and ATCase CYPICK datasets. In the case of HSP90, the considerably smaller amount of assigned methyls could be attributed to the lower percentage of explained NOE data when using the CYPICK lists (Figure 4B). When the manually generated NOE list was used, the MethylFLYA assignments explained roughly 85% of the NOE data at a 5 Å distance cutoff (Figure S5), consistent with the results presented above on the benchmark. In contrast, at the same distance cutoff, less than 60% of the NOE data were explained for the CYPICK-derived list. For ATCase, less than 40% of the methyl groups could be assigned, except for single d_{cut} value (Figure S5). The considerably lower performance of CYPICK-MethylFLYA on ATCase and HSP90-N suggests that some methyl-methyl NOEs are more critical determinants of assignment success than the others. In these cases, manual peak picking of the NOESY spectra remains the best approach for preparing the input data for MethylFLYA.

Discussion

The MAGMA study[42] included a performance comparison of the available NOE-based automatic methyl assignment software, MAP-XSII,[39] and FLAMEnGO2.0.[41] For a comparison of the available methods, we used here the results for all proteins,[42] apart from MSG, for which a different structure of the protein was used (Figure 5). The recently introduced MAGIC[43] method could not be included in the comparison because it requires the knowledge of signal intensities for all NOE cross-peaks supplied to the software, an information that was not available for three out of the five proteins of our benchmark set. Compared to the alternatives, MethylFLYA generated more confident and correct methyl assignments in all cases apart from $\alpha_7\alpha_7$ (Figure 5). For the other proteins, MethylFLYA generated on average 18% more assignments than the next best performing software. Overall, MethylFLYA generated the highest number of confident and correct methyl assignments on this benchmark (460), followed by MAGMA (343), MAP-XSII (224), and FLAMEnGO2.0 (141). Across the entire benchmark, MethylFLYA made five assignment errors and is, as such, the second most accurate method after MAGMA, which made only two assignment

errors. The latter two errors result from the use of a crystal structure for MSG (PDB 1D8C) instead of the NMR-derived structure (PDB 1Y8B)[42] that had been used in the original MAGMA benchmark. In the original study, MAGMA was reportedly sensitive to the structural difference between the two forms, likely due to the presence of the ligand in the crystal structure.[42] Here, we tested the performance of all methods exclusively on crystal structures to omit the need for NMR structures, which are anticipated to be unavailable for most proteins for which methyl resonance assignment is sought.

A comparison of the assignments found by the different methods reveals that MAGMA and MethylFLYA produce the most similar solutions, which agree on 291 of methyl assignments on this benchmark (Figures 5, S7). In contrast, MethylFLYA shares only 184 and 96 assignments with MAP-XSII and FLAMEnGO2.0, respectively. The intersection profiles are protein-specific (Figures S7). A complete overlap with MAGMA is seen in MethylFLYA solutions for $\alpha_7\alpha_7$, whereas the two methods overlap much less for MSG (Figure S7). Given that both protocols were given the same input data, a possible explanation for assignment differences could be algorithm-specific parameters. The distance cutoffs used to generate the expected NOE contacts were similar for the two methods. Nonetheless, distance cutoff for MethylFLYA is applied as an r^{-6} sum over the methyl proton distances, whereas MAGMA considers methyl carbon distances and, in addition, averages two methyl carbon positions for the isopropyl groups of Leu and Val, which are treated separately by MethylFLYA. Therefore, the exact composition of the expected NOE contacts differ between the two methods, resulting in differences in restraint matching. Furthermore, MAGMA provides assignment results for one distance cutoff, whereas, for its confident assignments, MethylFLYA requires assignment consistency over three distance cutoff values separated by 0.5 Å (see Methods). Finally, MAGMA uses exact graph comparison algorithms to exhaustively sample all assignment solutions that maximize the number of explained NOEs. In contrast, the evolutionary algorithm in MethylFLYA uses a heuristic to converge on a subset of most likely solutions, relying on differences between parallel iterations of the algorithm to assess assignment self-consistency. Despite the listed differences, the high overlap in assignment solutions between MethylFLYA and MAGMA and their high accuracy demonstrate the complementarity of these two methods. Comparing the solutions from the two methods therefore constitutes a useful cross-validation approach.

Conclusions

In conclusion, we have presented an NOE-based approach to automatic methyl resonance assignment that is a significant advance over existing methods. Even though the general FLYA algorithm underlying MethylFLYA (Figure 1) was originally designed to deal with through-bond, or a combination of through-bond and through-space information,[44] the method proved powerful also for the assignment of methyl groups exclusively from NOESY and structural data (Figure 2). This confirms earlier findings that FLYA was shown to be effective in assigning small proteins exclusively from on ^{13}C and ^{15}N-resolved NOESY data.[46] However, the assignment of methyl resonances in proteins as large as 360 kDa ($\alpha_7\alpha_7$), based on exclusively methyl-methyl NOEs, presents a considerably greater challenge because of data sparsity and minimal redundancy in data content. Nonetheless, MethylFLYA could generate as many, and in most cases significantly more, correct methyl assignments than existing algorithms (Figure 5A). Only a very small number of assignments from MethylFLYA were erroneous, and all of these were to methyls spatially proximal to the correct assignment in the 3D structure (Figure S4), thus limiting their impact on studies relying on the methyl assignments. MethylFLYA is fast and robust in coping with ambiguous and erroneous NOEs, showing nearly identical performance on raw and refined NOESY data (Figure 2, Table S3). MethylFLYA is also tolerant to ambiguity in the identity of Leu and Val resonances, whereas it significantly benefits from experimentally linking the methyl resonances from the geminal Leu/Val methyl groups (Figure 3). A high fraction of overlap in confident methyl assignments between MAGMA and MethylFLYA indicates the complementarity of the two methods and can be useful in *de novo* assignment cross-validation (Figure 5B). The utility of rapid, accurate methyl assignments is highlighted by recent studies that used NOEs between an unlabeled ligand and a methyl-labeled protein as restraints to generate models of the docked complex[42,51,63,64] and PCSs to measure reorientation of methyl groups upon ligand binding.[65] In the future, MethylFLYA could be extended to incorporate paramagnetic restraints, such as PREs or PCSs, or be combined with the existing software packages that predominantly rely on these restraints.[36,37]

Acknowledgments

We gratefully acknowledge financial support by a Eurostars grant of the Swiss Confederation and a Grant-in-Aid for Scientific Research of the Japan Society for the Promotion of Science (JSPS).

References

(1) Baker, L. A.; Sinnige, T.; Schellenberger, P.; de Keyzer, J.; Siebert, C. A.; Driessen, A. J. M.; Baldus, M.; Grünewald, K. Combined ^{1}H-detected solid-state NMR spectroscopy and electron cryotomography to study membrane proteins across resolutions in native environments. *Structure* **2018**, *26*, 161–170.

(2) Barth, K.; Hank, S.; Spindler, P. E.; Prisner, T. F.; Tampé, R.; Joseph, B. Conformational coupling and trans-inhibition in the human antigen transporter ortholog TmrAB resolved with dipolar EPR spectroscopy. *J. Am. Chem. Soc.* **2018**, *140*, 4527-4533.

(3) Deniz, A. A. Deciphering complexity in molecular biophysics with single-molecule resolution. *J. Mol. Biol.* **2016**, *428*, 301-307.

(4) Politis, A.; Schmidt, C.; Tjioe, E.; Sandercock, A. M.; Lasker, K.; Gordiyenko, Y.; Russel, D.; Sali, A.; Robinson, C. V. Topological models of heteromeric protein assemblies from mass spectrometry: application to the yeast eIF3:eIF5 complex. *Chem. Biol.* **2015**, *22*, 117-128.

(5) Clore, G. M.; Iwahara, J. Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population states of biological macromolecules and their complexes. *Chem Rev* **2009**, *109*, 4108-4139.

(6) Chiu, W.; Downing, K. H. Editorial overview: Cryo electron microscopy: Exciting advances in CryoEm herald a new era in structural biology. *Curr. Opin. Struct. Biol.* **2017**, *46*, iv-viii.

(7) Pervushin, K.; Riek, R.; Wider, G.; Wüthrich, K. Attenuated T_2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 12366–12371.

(8) Tugarinov, V.; Hwang, P. M.; Ollerenshaw, J. E.; Kay, L. E. Cross-correlated relaxation enhanced ^{1}H-^{13}C NMR spectroscopy of methyl groups in very high molecular weight proteins and protein complexes. *J. Am. Chem. Soc.* **2003**, *125*, 10420–10428.

(9) Ollerenshaw, J. E.; Tugarinov, V.; Kay, L. E. Methyl TROSY: explanation and experimental verification. *Magn Reson Chem* **2003**, *41*, 843-852.

(10) Sprangers, R.; Kay, L. E. Quantitative dynamics and binding studies of the 20S proteasome by NMR. *Nature* **2007**, *445*, 618–622.

(11) Religa, T. L.; Sprangers, R.; Kay, L. E. Dynamic regulation of archaeal proteasome gate opening as studied by TROSY NMR. *Science* **2010**, *328*, 98-102.

(12) Cvetkovic, M. A.; Wurm, J. P.; Audin, M. J.; Schütz, S.; Sprangers, R. The Rrp4-exosome complex recruits and channels substrate RNA by a unique mechanism. *Nat. Chem. Biol.* **2017**, *13*, 522–528.

(13) Sekhar, A.; Nagesh, J.; Rosenzweig, R.; Kay, L. E. Conformational heterogeneity in the Hsp70 chaperone-substrate ensemble identified from analysis of NMR-detected titration data. *Protein Sci.* **2017**, *26*, 2207-2220.

(14) Oroz, J.; Kim, J. H.; Chang, B. J.; Zweckstetter, M. Mechanistic basis for the recognition of a misfolded protein by the molecular chaperone Hsp90. *Nat. Struct. Mol. Biol.* **2017**, *24*, 407–413.

(15) Baldwin, A. J.; Hilton, G. R.; Lioe, H.; Bagneris, C.; Benesch, J. L. P.; Kay, L. E. Quaternary dynamics of αB-crystallin as a direct consequence of localised tertiary fluctuations in the C-terminus. *J. Mol. Biol.* **2011**, *413*, 310-320.

(16) Hsu, S. T. D.; Cabrita, L. D.; Fucini, P.; Christodoulou, J.; Dobson, C. M. Probing side-chain dynamics of a ribosome-bound nascent chain using methyl NMR spectroscopy. *J. Am. Chem. Soc.* **2009**, *131*, 8366–8367.

(17) Sprangers, R.; Gribun, A.; Hwang, P. M.; Houry, W. A.; Kay, L. E. Quantitative NMR spectroscopy of supramolecular complexes: Dynamic side pores in ClpP are important for product release. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 16678-16683.

(18) Csizmok, V.; Orlicky, S.; Cheng, J.; Song, J. H.; Bah, A.; Delgoshaie, N.; Lin, H.; Mittag, T.; Sicheri, F.; Chan, H. S.; Tyers, M.; Forman-Kay, J. D. An allosteric conduit facilitates dynamic multisite substrate recognition by the SCF^{Cdc4} ubiquitin ligase. *Nat. Commun.* **2017**, *8*, 13943.

(19) Toyama, Y.; Kano, H.; Mase, Y.; Yokogawa, M.; Osawa, M.; Shimada, I. Structural basis for the ethanol action on G-protein-activated inwardly rectifying potassium channel 1 revealed by NMR spectroscopy. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 3858-3863.

(20) Solt, A. S.; Bostock, M. J.; Shrestha, B.; Kumar, P.; Warne, T.; Tate, C. G.; Nietlispach, D. Insight into partial agonism by observing multiple equilibria for ligand-bound and G_s-mimetic nanobody-bound β_1-adrenergic receptor. *Nat. Commun.* **2017**, *8*, 12.

(21) Zhang, H. Y.; van Ingen, H. Isotope-labeling strategies for solution NMR studies of macromolecular assemblies. *Curr. Opin. Struct. Biol.* **2016**, *38*, 75-82.

(22) Wiesner, S.; Sprangers, R. Methyl groups as NMR probes for biomolecular interactions. *Curr. Opin. Struct. Biol.* **2015**, *35*, 60-67.

(23) Proudfoot, A.; Frank, A. O.; Ruggiu, F.; Mamo, M.; Lingel, A. Facilitating unambiguous NMR assignments and enabling higher probe density through selective labeling of all methyl containing amino acids. *J. Biomol. NMR* **2016**, *65*, 15-27.

(24) Clark, L.; Zahm, J. A.; Ali, R.; Kukula, M.; Bian, L. Q.; Patrie, S. M.; Gardner, K. H.; Rosen, M. K.; Rosenbaum, D. M. Methyl labeling and TROSY NMR spectroscopy of proteins expressed in the eukaryote Pichia pastoris. *J. Biomol. NMR* **2015**, *62*, 239-245.

(25) Suzuki, R.; Sakakura, M.; Mori, M.; Fujii, M.; Akashi, S.; Takahashi, H. Methyl-selective isotope labeling using α-ketoisovalerate for the yeast *Pichia pastoris* recombinant protein expression system. *J. Biomol. NMR* **2018**, *71*, 213-223.

(26) Kofuku, Y.; Yokomizo, T.; Imai, S.; Shiraishi, Y.; Natsume, M.; Itoh, H.; Inoue, M.; Nakata, K.; Igarashi, S.; Yamaguchi, H.; Mizukoshi, T.; Suzuki, E.; Ueda, T.; Shimada, I. Deuteration and selective labeling of alanine methyl groups of β_2-adrenergic receptor expressed in a baculovirus-insect cell expression system. *J. Biomol. NMR* **2018**, *71*, 185-192.

(27) Tugarinov, V.; Choy, W. Y.; Orekhov, V. Y.; Kay, L. E. Solution NMR-derived global fold of a monomeric 82-kDa enzyme. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 622–627.

(28) Kay, L. E.; Ikura, M.; Tschudin, R.; Bax, A. Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. *J. Magn. Reson.* **1990**, *89*, 496–514.

(29) Tugarinov, V.; Kay, L. E. 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.* **2003**, *125*, 13868-13878.

(30) Sattler, M.; Schleucher, J.; Griesinger, C. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. Nucl. Magn. Reson. Spectrosc.* **1999**, *34*, 93–158.

(31) Sprangers, R.; Velyvis, A.; Kay, L. E. Solution NMR of supramolecular complexes: providing new insights into function. *Nat. Methods* **2007**, *4*, 697–703.

(32) Gelis, I.; Bonvin, A. M. J. J.; Keramisanou, D.; Koukaki, M.; Gouridis, G.; Karamanou, S.; Economou, A.; Kalodimos, C. G. Structural basis for signal-sequence recognition by the translocase motor SecA as determined by NMR. *Cell* **2007**, *131*, 756-769.

(33) Xiao, Y.; Warner, L. R.; Latham, M. P.; Ahn, N. G.; Pardi, A. Structure-based assignment of Ile, Leu, and Val methyl groups in the active and inactive forms of the mitogen-activated protein kinase extracellular signal-regulated kinase 2. *Biochemistry* **2015**, *54*, 4307–4319.

(34) Velyvis, A.; Schachman, H. K.; Kay, L. E. Assignment of Ile, Leu, and Val methyl correlations in supra-molecular systems: An application to aspartate transcarbamoylase. *J. Am. Chem. Soc.* **2009**, *131*, 16534-16543.

(35) John, M.; Schmitz, C.; Park, A. Y.; Dixon, N. E.; Huber, T.; Otting, G. Sequence-specific and stereospecific assignment of methyl groups using paramagnetic lanthanides. *J. Am. Chem. Soc.* **2007**, *129*, 13749–13757.

(36) Lescanne, M.; Skinner, S. P.; Blok, A.; Timmer, M.; Cerofolini, L.; Fragai, M.; Luchinat, C.; Ubbink, M. Methyl group assignment using pseudocontact shifts with PARAssign. *J. Biomol. NMR* **2017**, *69*, 183-195.

(37) Venditti, V.; Fawzi, N. L.; Clore, G. M. Automated sequence- and stereo-specific assignment of methyl-labeled proteins by paramagnetic relaxation and methyl-methyl nuclear overhauser enhancement spectroscopy. *J. Biomol. NMR* **2011**, *51*, 319–328.

(38) Xu, Y. Q.; Liu, M. H.; Simpson, P. J.; Isaacson, R.; Cota, E.; Marchant, J.; Yang, D. W.; Zhang, X. D.; Freemont, P.; Matthews, S. Automated assignment in selectively methyl-labeled proteins. *J. Am. Chem. Soc.* **2009**, *131*, 9480–9481.

(39) Xu, Y. Q.; Matthews, S. MAP-XSII: an improved program for the automatic assignment of methyl resonances in large proteins. *J. Biomol. NMR* **2013**, *55*, 179–187.

(40) Chao, F.-A.; Shi, L.; Masterson, L. R.; Veglia, G. FLAMEnGO: A fuzzy logic approach for methyl group assignment using NOESY and paramagnetic relaxation enhancement data. *J. Magn. Reson.* **2012**, *214*, 103–110.

(41) Chao, F. A.; Kim, J. G.; Xia, Y. L.; Milligan, M.; Rowe, N.; Veglia, G. FLAMEnGO 2.0: An enhanced fuzzy logic algorithm for structure-based assignment of methyl group resonances. *J. Magn. Reson.* **2014**, *245*, 17–23.

(42) Pritisanac, I.; Degiacomi, M. T.; Alderson, T. R.; Carneiro, M. G.; Eiso, A. B.; Siegal, G.; Baldwin, A. J. Automatic assignment of methyl-NMR spectra of supramolecular machines using graph theory. *J. Am. Chem. Soc.* **2017**, *139*, 9523–9533.

(43) Monneau, Y. R.; Rossi, P.; Bhaumik, A.; Huang, C. D.; Jiang, Y. J.; Saleh, T.; Xie, T.; Xing, Q.; Kalodimos, C. G. Automatic methyl assignment in large proteins by the MAGIC algorithm. *J. Biomol. NMR* **2017**, *69*, 215-227.

(44) Schmidt, E.; Güntert, P. A new algorithm for reliable and general NMR resonance assignment. *J. Am. Chem. Soc.* **2012**, *134*, 12817–12829.

(45) Güntert, P.; Buchner, L. Combined automated NOE assignment and structure calculation with CYANA. *J. Biomol. NMR* **2015**, *62*, 453–471.

(46) Schmidt, E.; Güntert, P. Reliability of exclusively NOESY-based automated resonance assignment and structure determination of proteins. *J. Biomol. NMR* **2013**, *57*, 193–204.

(47) Würz, J. M.; Güntert, P. Peak picking multidimensional NMR spectra with the contour geometry based algorithm CYPICK. *J. Biomol. NMR* **2017**, *67*, 63–76.

(48) Bartels, C.; Güntert, P.; Billeter, M.; Wüthrich, K. GARANT - A general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *J. Comput. Chem.* **1997**, *18*, 139–149.

(49) Güntert, P.; Dötsch, V.; Wider, G.; Wüthrich, K. Processing of multidimensional NMR data with the new software PROSA. *J. Biomol. NMR* **1992**, *2*, 619–629.

(50) Bartels, C.; Xia, T. H.; Billeter, M.; Güntert, P.; Wüthrich, K. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *J. Biomol. NMR* **1995**, *6*, 1–10.

(51) Shah, D. M.; Ab, E.; Diercks, T.; Hass, M. A. S.; van Nuland, N. A. J.; Siegal, G. Rapid protein-ligand costructures from sparse NOE data. *J. Med. Chem.* **2012**, *55*, 10786–10790.

(52) Garrett, D. S.; Seok, Y. J.; Liao, D. I.; Peterkofsky, A.; Gronenborn, A. M.; Clore, G. M. Solution structure of the 30 kDa N-terminal domain of enzyme I of the Escherichia coli phosphoenolpyruvate:sugar phosphotransferase system by multidimensional NMR. *Biochemistry* **1997**, *36*, 2517-2530.

(53) Gardner, K. H.; Zhang, X. C.; Gehring, K.; Kay, L. E. Solution NMR studies of a 42 KDa *Escherichia coli* maltose binding protein b-cyclodextrin complex: Chemical shift assignments and analysis. *J. Am. Chem. Soc.* **1998**, *120*, 11738–11748.

(54) Tugarinov, V.; Sprangers, R.; Kay, L. E. Probing side-chain dynamics in the proteasome by relaxation violated coherence transfer NMR spectroscopy. *J. Am. Chem. Soc.* **2007**, *129*, 1743-1750.

(55) Goddard, T. D.; Kneller, D. G.; University of California: San Francisco, 2001.

(56) Schmidt, E.; Gath, J.; Habenstein, B.; Ravotti, F.; Székely, K.; Huber, M.; Buchner, L.; Böckmann, A.; Meier, B. H.; Güntert, P. Automated solid-state NMR resonance assignment of protein microcrystals and amyloids. *J. Biomol. NMR* **2013**, *56*, 243–254.

(57) Aeschbacher, T.; Schmidt, E.; Blatter, M.; Maris, C.; Duss, O.; Allain, F. H.-T.; Güntert, P.; Schubert, M. Automated and assisted RNA resonance assignment using NMR chemical shift statistics. *Nucleic Acids Res.* **2013**, *41*, e172.

(58) Krähenbühl, B.; El Bakkali, I.; Schmidt, E.; Güntert, P.; Wider, G. Automated NMR resonance assignment strategy for RNA via the phosphodiester backbone based on high-dimensional through-bond APSY experiments. *J. Biomol. NMR* **2014**, *59*, 87–93.

(59) Schmidt, E.; Ikeya, T.; Takeda, M.; Löhr, F.; Buchner, L.; Ito, Y.; Kainosho, M.; Güntert, P. Automated resonance assignment of the 21 kDa stereo-array isotope labeled thioldisulfide oxidoreductase DsbA. *J. Magn. Reson.* **2014**, *249*, 88–93.

(60) Lichtenecker, R. J.; Coudevylle, N.; Konrat, R.; Schmid, W. Selective isotope labelling of leucine residues by using α-ketoacid precursor compounds. *ChemBioChem* **2013**, *14*, 818-821.

(61) Lichtenecker, R. J.; Weinhäupl, K.; Reuther, L.; Schörghuber, J.; Schmid, W.; Konrat, R. Independent valine and leucine isotope labeling in *Escherichia coli* protein overexpression systems. *J. Biomol. NMR* **2013**, *57*, 205-209.

(62) Gans, P.; Hamelin, O.; Sounier, R.; Ayala, I.; Durá, M. A.; Amero, C. D.; Noirclerc-Savoye, M.; Franzetti, B.; Plevin, M. J.; Boisbouvier, J. Stereospecific isotopic labeling of methyl groups for NMR spectroscopic studies of high-molecular-weight proteins. *Angew. Chem. Int. Ed.* **2010**, *49*, 1958-1962.

(63) Orts, J.; Wälti, M. A.; Marsh, M.; Vera, L.; Gossert, A. D.; Güntert, P.; Riek, R. NMR-Based Determination of the 3D Structure of the Ligand-Protein Interaction Site without Protein Resonance Assignment. *J. Am. Chem. Soc.* **2016**, *138*, 4393–4400.

(64) Mohanty, B.; Williams, M. L.; Doak, B. C.; Vazirani, M.; Ilyichova, O.; Wang, G. Q.; Bermel, W.; Simpson, J. S.; Chalmers, D. K.; King, G. F.; Mobli, M.; Scanlon, M. J. Determination of ligand binding modes in weak protein-ligand complexes using sparse NMR data. *J. Biomol. NMR* **2016**, *66*, 195-208.

(65) Lescanne, M.; Ahuja, P.; Blok, A.; Timmer, M.; Akerud, T.; Ubbink, M. Methyl group reorientation under ligand binding probed by pseudocontact shifts. *J. Biomol. NMR* **2018**, *71*, 275-285.

Table 1. MethylFLYA assignment statistics

	EIN	ATCase	MBP	MSG	$\alpha_7\alpha_7$
Protein properties:					
Residues per monomer	259	153	370	731	233
Multimeric state	monomer	dimer	monomer	monomer	14-mer
Molecular mass of multimer (kDa)	28.3	34.2	40.6	81.4	358.4
Experimental NMR data:					
Labeled amino acids	AILV	ILV	ILV	ILV	ILV
NOESY dimensions	HCCH	CCH	HCCH	HCCH	CCH
Labeled methyl ^{1}H and ^{13}C resonances:					
All	292	132	246	552	194
With reference assignment	266	124	246	536	176
With NOESY peaks	232	112	236	472	176
Strongly assigned methyl resonances from filtered peak lists:					
All	202	90	172	352	160
Correct	180	72	167	346	156
Wrong	0	6	3	0	0
Strongly assigned methyl resonances from unfiltered peak lists:					
All	214	84	178	366	164
Correct	197	71	175	348	160
Wrong	0	5	3	6	0

Figures

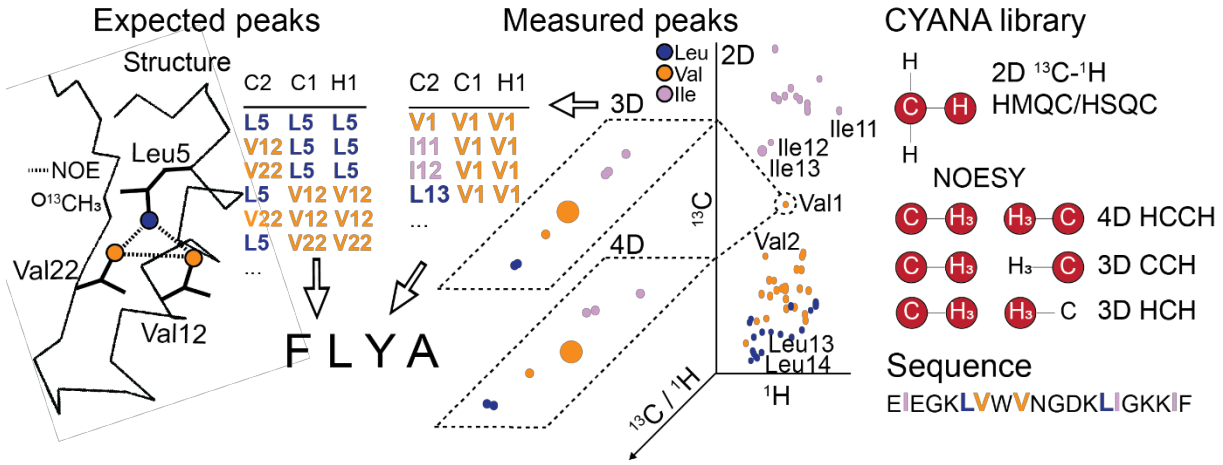

Figure 1. Input data requirements for automatic methyl resonance assignment with MethylFLYA. From *left* to *right:* the expected methyl-methyl contacts are computed from a 3D structure of the protein. The contacts are written to a list of expected NOE peaks with the amino acid type of each contact indicated (e.g. I-V, L-L). A list of measured NOESY peaks is obtained by manual (or automatic) inspection of the 3D or 4D methyl-methyl NOESY spectrum (center) guided by information from the 2D [^{1}H,^{13}C]-HMQC spectrum. If known, the amino acid types of the methyl peaks that give rise to measured NOEs are included in the peak list. The ^{1}H-^{13}C HMQC peak list and the expected and measured NOE peak lists are supplied to MethylFLYA for the automatic methyl assignment calculation. In addition to the peak lists, the MethylFLYA calculation requires a protein sequence (far right bottom) and knowledge of the magnetization transfer pathways for the employed NMR experiments, which are provided in the CYANA library (far right top).

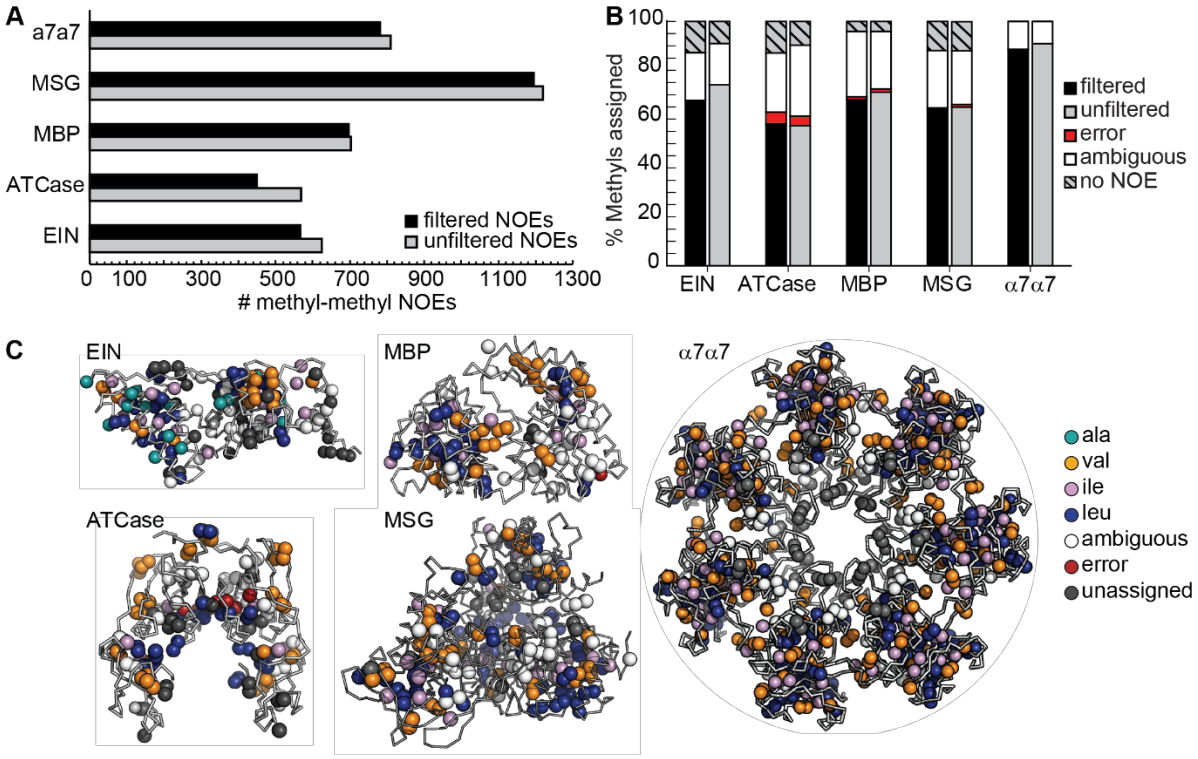

Figure 2. MethylFLYA performance on the benchmark. **A)** Number of methyl-methyl NOEs before and after filtering of manually picked NOE peak lists (see *Methods*). **B)** MethylFLYA performance on filtered (black) and "raw" (unfiltered) manually picked NOESY peak lists (grey). **C)** Methyl groups assigned as strong (confident) by MethylFLYA with filtered NOE peak lists are indicated with colored spheres in the 3D structures of the benchmark proteins. The colors indicate the amino-acid types of the assigned groups, with non-assigned groups colored white.

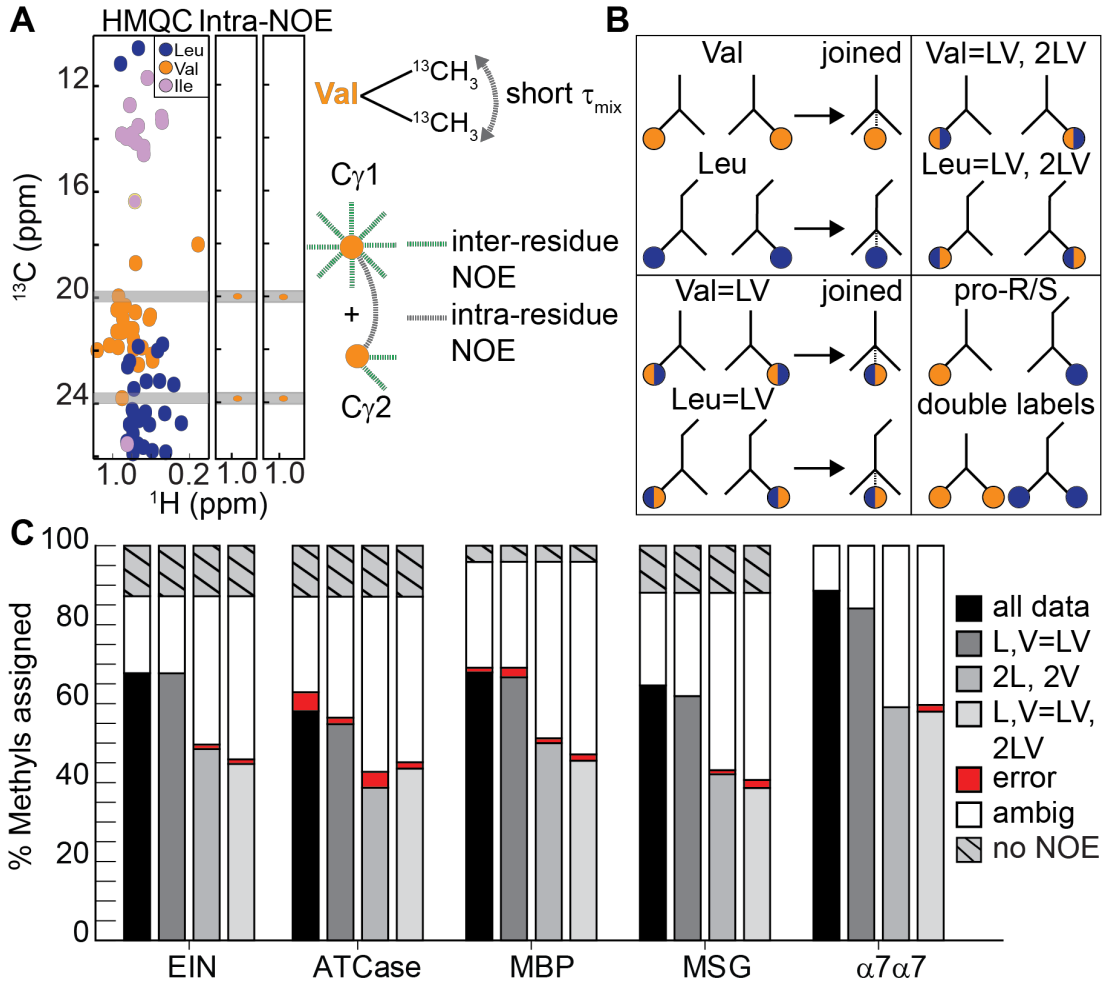

Figure 3. MethylFLYA performance on the benchmark with reduced data input. **A)** Geminal methyl groups of Leu/Val residues can be linked with a short-mixing time NOESY experiment on an exclusively doubled methyl-labeled ([$^{13}\delta_1/^{13}\delta_2$]-Leu, [$^{13}\gamma_1/^{13}\gamma_2$]-Val) protein sample. In the NOESY plane of each Leu/Val methyl resonance, a strong signal from its geminal methyl resonance is observed (right). **B)** Schematic illustration of different possibilities for treatment of methyl resonances from Leu/Val residues in MethylFLYA calculations. *Top, left* – differentiation of the Leu- and Val-type of methyl resonances and known connectivity between the geminal Leu or Val methyl groups. *Bottom, left* – no differentiation between methyl resonances of the Leu- or Val-type, but known connectivity between the geminal Leu/Val methyl groups. *Top, right* – no differentiation between methyl resonances of the Leu/Val-type, nor knowledge of the geminal methyl connectivity. *Bottom, right* – stereospecific labelling of Leu/Val-methyl groups (pro-*R* or pro-*S*), or double labelling (both pro-*R* and pro-*S*). **C)** MethylFLYA results with reduced data

input. The percentage of assigned methyl groups given knowledge of all methyl amino acid types and the geminal Leu/Val-methyl connectivity is shown in *black (all data)*. In *dark grey* (*L, V=LV*), knowledge of Ile- and Ala-methyl resonance types is assumed, but there is no discrimination between Leu or Val methyl resonances types. The geminal (Leu/Val) methyl resonances are connected. In *light grey* (*2L, 2V*), knowledge of all amino-acid types is assumed, but the geminal pairing of Leu/Val resonances is omitted. In *white* (*L, V=LV, 2LV*), neither the amino-acid type nor the geminal pairing is known for Leu and Val methyl resonances. Knowledge of the amino acid types of other resonances (Ala, Ile) is still assumed.

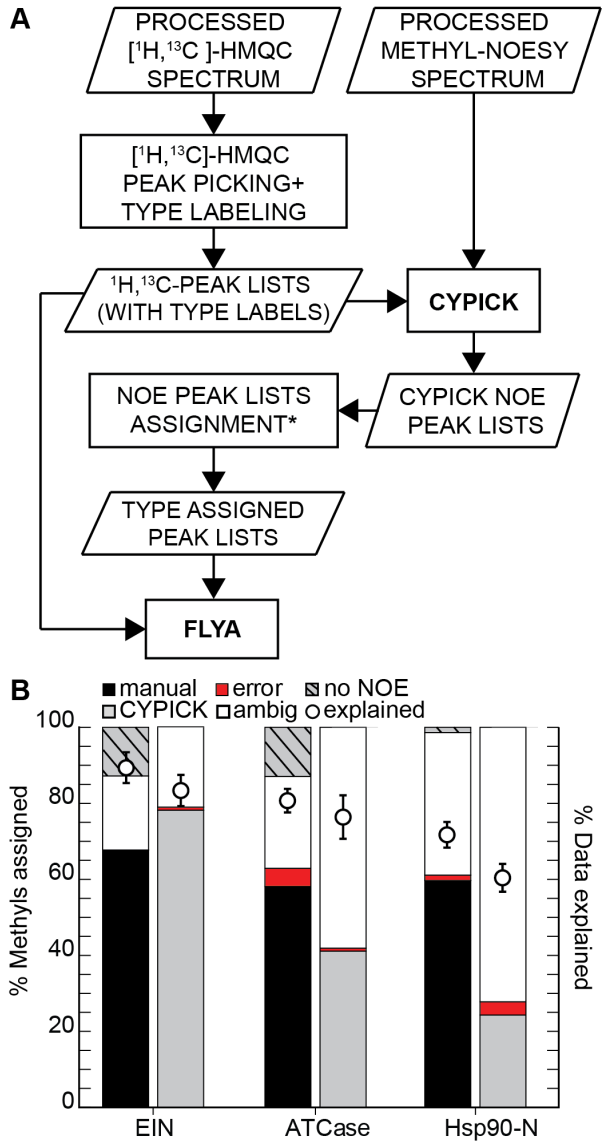

Figure 4. MethylFLYA performance on methyl-methyl NOESY peak lists generated automatically using CYPICK. **A)** An outline of the combined CYPICK-MethylFLYA assignment protocol (see *Methods*). The step marked with * refers to the attribution of methyl resonances to amino-acid types (e.g. Ala, Ile, Leu, Val). **B)** Comparison of MethylFLYA performance on manually and automatically picked NOESY data. The percentage of assigned methyl resonances and the percentage of explained input NMR data, given the MethylFLYA assignment, are shown. The error-bars are standard deviations of the percentage of data explained over the three distance cutoffs (optimal cutoff ± 0.5 Å, see *Methods*). Note that CYPICK does not assign methyl-methyl

NOEs, and hence the identity of the methyl resonances with no NOEs (“no NOE”) is not indicated for the CYPICK bars.

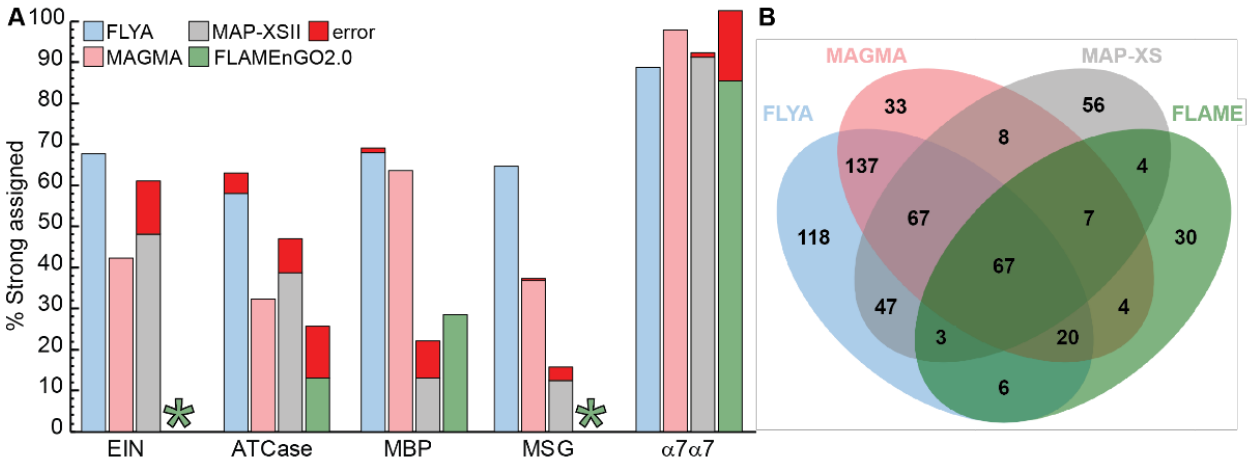

Figure 5. Comparison of MethylFLYA, MAGMA, MAP-XSII, and FLAMEnGO2.0 on the benchmark. **A)** The percentage of correctly and erroneously strongly (i.e. confidently) assigned methyl resonances for each of the cases is shown. Asterisks are given in the places where no confident (100%) assignments could be obtained with the FLAMEnGO2.0 software. **B)** Mutual agreement of the methyl resonance assignment results between the four methods. The largest intersection in strong methyl assignments is found between MethylFLYA and MAGMA.