Automatic structure-based NMR methyl resonance assignment in large proteins

Iva Pritišanac,[1] Julia M. 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

Isotope-labeled methyl groups provide NMR probes that can be observed in very large, otherwise deuterated systems and enable investigations of protein structure, dynamics and mechanisms. However, the assignment of resonances to specific methyls in the protein is expensive and time-consuming, which limits the use of methyl-based NMR for large proteins. To resolve this bottleneck, methyl assignment methods have been developed. However, these remain limited regarding complete automation, computational feasibility, and/or the extent and accuracy of the assignments. Here, we present the automated MethylFLYA method for the assignment of methyl groups that is based on methyl-methyl nuclear Overhauser effect spectroscopy (NOESY) peak lists. MethylFLYA was applied to five proteins (28–358 kDa) comprising a total of 708 isotope-labeled methyl groups, of which 674 had manually determined $^{1}H/^{13}C$ reference assignments and 614 showed cross peaks in the available NOESY peak lists. MethylFLYA confidently assigned 488 methyl groups, i.e. 79% of those with NOESY data. Of these, 460 agreed with the reference, 5 were different, and 23 concerned methyls without reference assignment. For high-quality NOESY spectra, automatic NOESY peak picking followed by resonance assignment with MethylFLYA can yield results comparable to those obtained from manually prepared peak lists, indicating the feasibility of unbiased, fully automatic methyl resonance assignment starting directly from the NMR spectra. Overall, MethylFLYA assigns significantly more methyl groups than other algorithms, has an average error rate of 1%, modest runtimes of 0.4–1.2 h, and flexibility to handle arbitrary isotope labeling patterns and include data from other types of NMR spectra.

The last decade of structural biology has seen growing interest in biologically relevant large protein assemblies, as witnessed by an explosion of high- and low-resolution structural studies of macromolecular machines.[1] 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).[2] The TROSY enhancement, initially established for amide groups, was subsequently also realized for selectively methyl-labeled proteins (methyl-TROSY).[3, 4] Methyl-TROSY has since enabled studies of protein complexes in excess of 1 MDa[5] in unprecedented detail, revealing the mechanisms of dynamic molecular machines.[6-8]

For optimal gains in the signal enhancement and resolution of methyl-TROSY spectra, selectively protonated, ^{13}C-labelled methyl groups are introduced into an otherwise highly deuterated background.[9] 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.*[10, 11] Selective labeling of methyl groups is also possible in eukaryotic expression systems.[12-14] The labeled methyl groups have favorable spectroscopic properties that render them observable also in large proteins and protein assemblies. Methyl groups are effective site-specific probes of molecular dynamics, structure, and interactions, as they are found both throughout the hydrophobic core of a protein and on its surface.[10,15]

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 (Fig. 5).[16] 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,[17] to which side-chain methyl resonances can be linked.[18] 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 accessible as NMR input data for assignment.

Assignment strategies for large proteins or proteins assemblies include divide-and-conquer approaches wherein sufficiently small individual protein domains or subunits are produced separately, such that their backbone resonance assignment can be determined using standard methods.[19] This approach requires that the resonance frequencies of the subsystems coincide

closely with those of the complete system. To complete the assignment, the approach is often supplemented with site-directed mutagenesis of individual methyl-bearing residues.[20, 21] 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,[22, 23] or between methyls and site-specifically attached paramagnetic probes.[24]

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) and 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 and estimates of the magnetic susceptibility tensors. These approaches define the optimal methyl assignments as those that minimize the difference between the measured and the calculated paramagnetic observables.[25-27] For instance, PRE-ASSIGN[27] uses paramagnetic relaxation enhancements (PREs), whereas PARAssign[26] 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[28-31] or graph-based[32, 33] algorithms. For example, MAGMA[32] uses exact graph matching algorithms to generate confident 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 resonance 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,[34] which is integrated in the CYANA structure calculation software,[35] and has been shown capable to assign proteins exclusively from NOESY data,[36] for structure- and NOESY-based methyl resonance

assignment. We apply the resulting MethylFLYA algorithm to a benchmark[32] 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,[32] MAP-XSII,[29] and FLAMEnGO2.0[31] 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.[37]

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,[34, 36, 38-41] specific optimization of a small number of parameters that are of particular relevance to structure-based methyl assignment was advantageous.

MethylFLYA considers only methyl-methyl distances shorter than a user-defined cutoff d_{cut} for generating expected methyl-methyl NOESY cross peaks based on a protein structure (see Methods). In addition, each expected peak is attributed a probability value to (roughly) reflect the 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 (Fig. 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 Fig. 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 $\alpha_7\alpha_7$ (Fig. S1). These plateaus coincided with about 80% explained input NMR data, which was determined as optimal for these data sets. Increasing the observation probabilities generally diminished the quality of the results, as more incorrect assignments were obtained (Fig. S2). Predictably, more of the observed NOEs were assigned using higher distance cutoffs for generating the expected NOEs, 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 Figs. 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 filtered NOE data,[32] MethylFLYA assigned between 63% (ATCase) and 89% ($\alpha_7\alpha_7$) of the methyl resonances for which reference assignments are available (Fig. 1B, Tables 1, S1), with no assignment errors for EIN, MSG, and α7α7. Two incorrect methyl assignments were found for MBP, and three for ATCase (Fig. 1B). In the 3D structures, all incorrectly assigned methyls are located in proximity to their correct assignment positions (Fig. S4, Table S2). Such spatially localized assignment errors are expected to have minor impact on 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 ensure a higher accuracy by "sacrificing" some of the strong assignments.

On the other hand, using a single distance cutoff (5 Å for EIN, ATCase, MBP, MSG; 6 Å for $\alpha_7\alpha_7$) instead of three distance cutoffs spaced by 0.5 Å for generating the expected NOESY cross peaks in MethylFLYA increases the overall number of assignment errors about five-fold (Table 1). It is thus not advisable even though the total number of strong assignments rises by about 15%.

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 (Fig. 1).

A spatial clustering of strong assignments can be discerned in the structures of EIN, ATCase, and MSG (Fig. 1C). 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.

Table 1. Methyl resonance 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	108	236	472	176
Methyl resonances confidently assigned by MethylFLYA:					
All	202	90	172	352	160
Correct	180	72	167	346	156
Erroneous	0	6	3	0	0
Methyl resonances confidently assigned by MethylFLYA using unfiltered peak lists:					
All	214	84	178	366	164
Correct	197	71	175	348	160
Erroneous	0	5	3	6	0
Methyl resonances confidently assigned by MethylFLYA using a single distance cutoff:					
All	248	96	204	407	171
Correct	208	78	178	370	163
Erroneous	8	11	18	12	1
Methyl resonances assigned by MAGMA:[32]					
Correct	112	40	156	198	180
Erroneous	0	0	0	4	0
Methyl resonances assigned by MAP-XSII:[29]					
Correct	128	48	32	66	158
Erroneous	34	10	22	18	2
Methyl resonances assigned by FLAMEnGO2.0:[31]					
Correct	0	16	70	0	142
Erroneous	0	16	0	0	38

Filtered input NOESY peak lists were used, unless noted otherwise. See text for details.

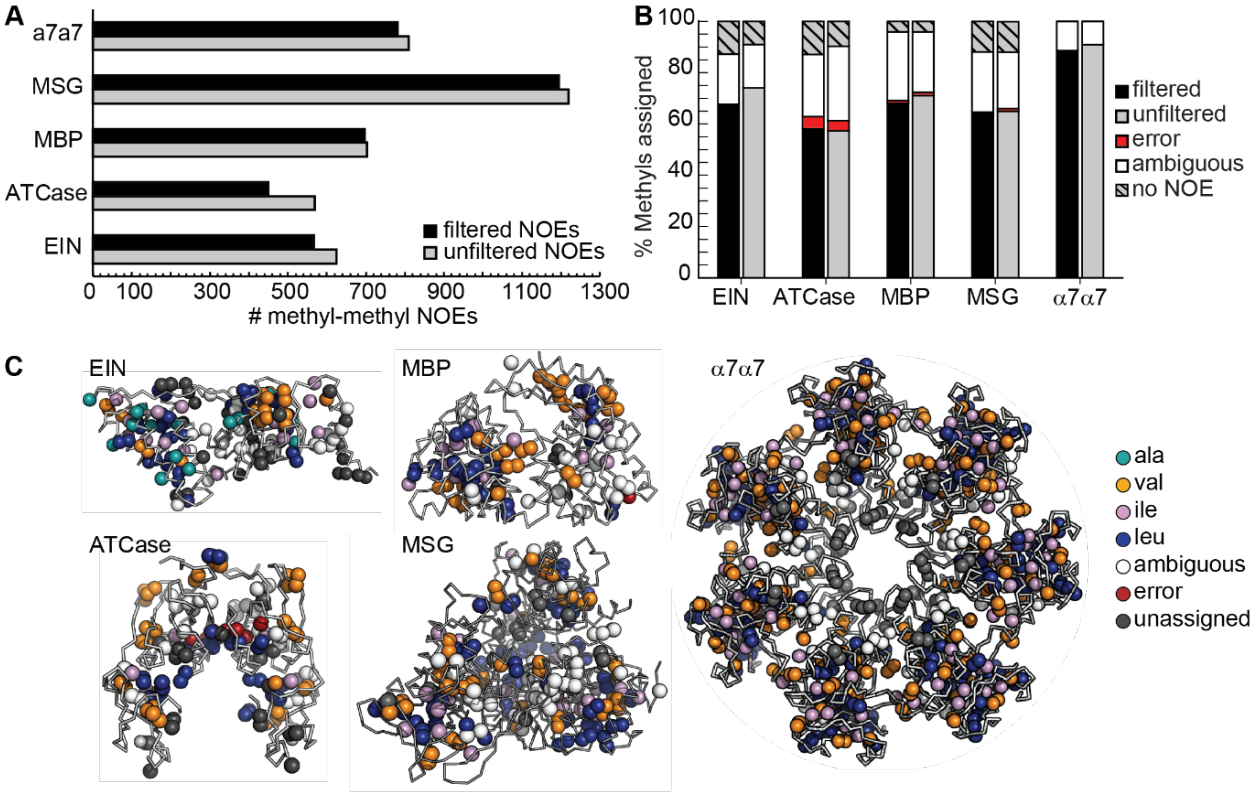

Fig. 1. MethylFLYA performance on the benchmark. **A)** Number of methyl-methyl NOEs before and after filtering of manually picked NOE peak lists (see *Methods*). **B)** MethylFLYA performance on filtered (black) and "raw" (unfiltered) manually picked NOESY peak lists (grey). **C)** Methyl groups assigned 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.

Reduced data sets. We tested the performance of MethylFLYA on the benchmark when experimental information provided to the algorithm was reduced (Fig. 2). In the best-case scenario, both knowledge of the amino acid types of methyl resonances and linkage of the two geminal methyl groups of Leu and Val is available (Fig. 2A, 3C, black). The Ile-δ_1 resonances are usually readily identified due to their upfield shifted ^{13}C frequencies. To discriminate between Leu and Val resonances, separate protein samples can be prepared using selective labelling schemes. For instance, selective Leu labelling can be achieved by using ^{13}C-labeled α-ketoisocaproate,[42] whereas the combined addition of unlabeled α-ketoisocaproate and labeled α-ketoisovalerate leads to exclusive labeling of Val.[43] 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 geminal methyl groups[21, 32] (Fig. 2A). 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 (Fig. 2C, dark grey). For ATCase and MBP, the percentage of accurate confident assignments decreased by 3%. However, for ATCase the percentage of errors was reduced simultaneously by 3%.

Removing the geminal Leu/Val pairing had a more significant impact, reducing the percentage of assigned methyls by ~19% for EIN, ATCase, MBP, and MSG, and up to 30% for $\alpha_7\alpha_7$ (Fig. 2C, 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.[32] In the MAGIC study, a four-fold decrease in computational time and a 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.[44] For MethylFLYA, removing both the Leu/Val-geminal pairing and discrimination between Leu/Val methyl resonances led to a similar outcome as geminal pairing removal alone (Fig. 2C, silver), and led overall only to a slight further increase in erroneous assignments (1–2%). Interestingly, for ATCase, removing the Leu/Val resonance discrimination always improved the accuracy (Fig. 2C, dark grey, silver). We conclude that, especially for smaller proteins (<80 kDa), Leu/Val residue discrimination is not crucial for MethylFLYA.

The computation time of MethylFLYA scaled approximately linearly with the number of methyl groups in the protein. The complete protocol took between 0.36 and 1.53 h (Table S3). Negligible differences in speed were noted for the calculations with lower input information content (Fig. 2, Table S3). This illustrates the ability of MethylFLYA to efficiently deliver high-quality assignments even from considerably reduced input data.

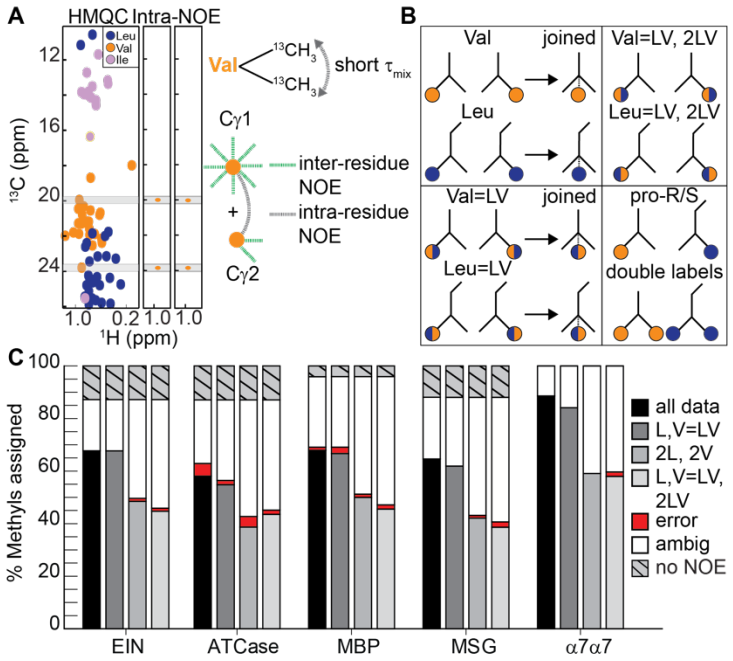

Fig. 2. MethylFLYA performance with reduced data input. **A)** Geminal methyl groups of Leu/Val residues can be linked with a short-mixing time NOESY experiment on an exclusively double methyl-labeled ([$^{13}\delta_1/^{13}\delta_2$]-Leu, [$^{13}\gamma_1/^{13}\gamma_2$]-Val) protein sample. In the NOESY plane of each Leu/Val methyl resonance, a strong signal from its geminal methyl resonance is observed (right). **B)** Different possibilities for treating Leu/Val methyl resonances in MethylFLYA. *Top, left*: differentiation of the Leu- and Val-type of methyl resonances and known connectivity between geminal Leu or Val methyl groups. *Bottom, left:* no differentiation between methyl resonances of the Leu- or Val-type, but known connectivity between geminal 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/Val methyl resonances. Knowledge of the amino acid types of other resonances (Ala, Ile) is still assumed.

Combination with automated peak picking. All currently available automatic methyl resonance assignment strategies rely, to different extents, 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.[28-33] 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 (Fig. S6). We investigated whether an automatic peak picking algorithm, CYPICK,[37] 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 (Fig. 3). For these spectra, CYPICK found 77–83% of the manually identified methyl-methyl NOEs (Fig. S5, Table S4), which is comparable to its performance previously reported on 3D ^{13}C-edited and ^{15}N-edited NOESY spectra.[37] The somewhat high CYPICK artifact scores for EIN (34%) and HSP90 (46%) did not result in assignment errors, as only one methyl group misassignment was found for EIN and three for HSP90. Moreover, for EIN, even slightly 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 (Fig. 3B). When the manually generated NOE list was used, the MethylFLYA assignments explained roughly 85% of the NOE data at a 5 Å distance cutoff (Fig. S5), consistent with the results presented above for the five proteins of 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 a single d_{cut} value (Fig. 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.

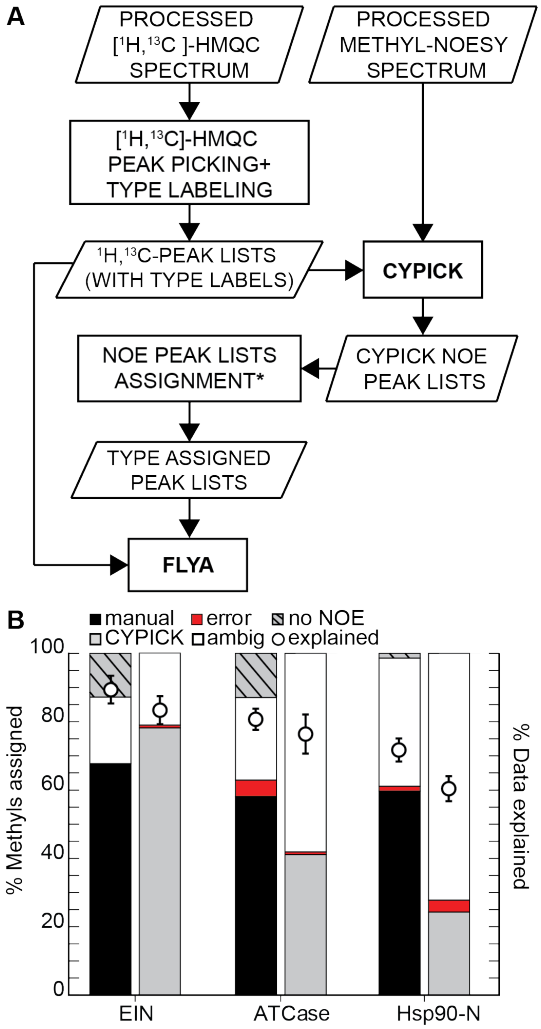

Fig. 3. MethylFLYA performance on methyl-methyl NOESY peak lists obtained by automatic peak picking using CYPICK. **A)** An outline of the combined CYPICK-MethylFLYA assignment protocol (see *Methods*). The step marked with * refers to the attribution of methyl resonances to amino-acid types (e.g. Ala, 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 NOESY 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.

Discussion

The MAGMA study[32] included a performance comparison with the available NOE-based automatic methyl assignment software packages, MAP-XSII,[29] and FLAMEnGO2.0.[31] For a comparison of the available methods, we used here the results for all proteins,[32] apart from MSG, for which a different structure of the protein was used (Fig. 4, Table S1). The recently introduced MAGIC[33] method could not be included in the comparison because it requires the knowledge of signal intensities for all NOESY cross peaks, 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 except for $\alpha_7\alpha_7$ (Table 1, Fig. 4), where all methods assigned more than 90% of the methyls. 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 ^{1}H and ^{13}C resonance assignments on this benchmark (921), followed by MAGMA (686), MAP-XSII (432), and FLAMEnGO2.0 (228). Across the entire benchmark, MethylFLYA made assignment errors for 5 methyl groups and is, as such, the second most accurate method after MAGMA, which made assignment errors only for 2 methyl groups. The latter two errors result from the use of a crystal structure for MSG (PDB 1D8C) instead of the NMR-derived structure (PDB 1Y8B) that had been used in the original MAGMA benchmark.[32] In the original study, MAGMA was reportedly sensitive to the structural difference between the two forms, which is likely due to the presence of the ligand in the crystal structure.[32] 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 the methyl assignments on this benchmark (Figs. 5, S7, Table S1). In contrast, MethylFLYA shares only 184 and 96 assignments with MAP-XSII and FLAMEnGO2.0, respectively. The intersection profiles are protein-specific (Figs. 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 (Fig. 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

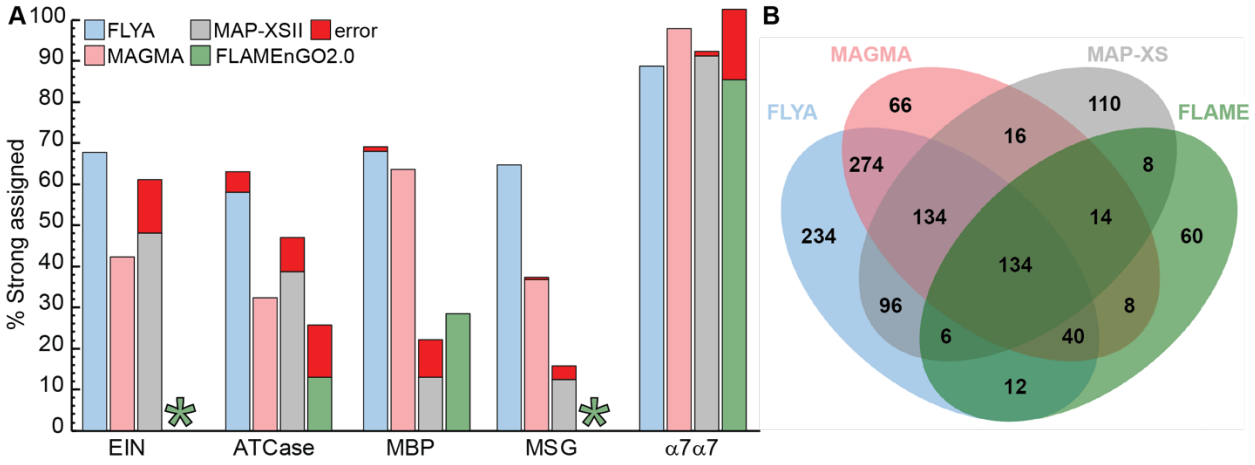

Fig. 4. Comparison of MethylFLYA, MAGMA, MAP-XSII, and FLAMEnGO2.0 on the benchmark. **A)** The percentage of correctly and erroneously strongly (i.e. confidently) assigned methyl resonances for each of the cases is shown. Asterisks are given in the places where no confident (100%) assignments could be obtained with the FLAMEnGO2.0 software. **B)** Mutual agreement of the methyl group resonance assignment results among the four methods.

---

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 differs 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.

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 (Fig. 5) was originally designed to deal with through-bond, or a combination of through-bond and through-space information,[34] the method proved powerful also for the assignment of methyl groups exclusively from NOESY and structural data (Fig. 1). This confirms earlier findings showing that FLYA is effective in assigning small proteins exclusively from ^{13}C and ^{15}N-resolved NOESY data.[36] 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 (Fig. 4A). 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 (Fig. 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 (Fig. 1, Table 1). 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 (Fig. 2). 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 (Fig. 4B). 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[32, 45-47] and PCSs to measure reorientation of methyl groups upon ligand binding.[48] In the future, MethylFLYA could be extended to incorporate paramagnetic restraints, such as PREs or PCSs, or be combined with existing software packages that predominantly rely on these restraints.[26, 27] Furthermore, MethylFLYA can straightforwardly be used to assign methyl resonances in solid-state NMR spectra.[49]

Methods

Overview of the MethylFLYA algorithm. The FLYA algorithm[34] determines resonance assignments by establishing an optimal mapping between expected peaks that are derived from knowledge of the protein sequence, experimental types, and, if available, 3D structure, and the observed peaks that are identified in the corresponding measured spectra. This mapping, and hence

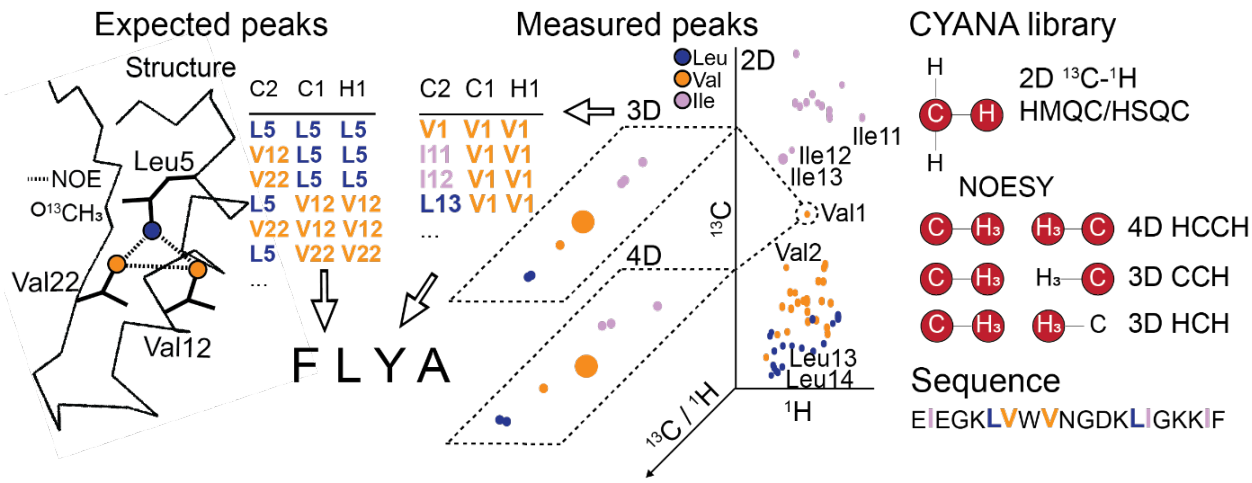

Fig. 5. Input data requirements for automatic methyl resonance assignment with MethylFLYA. The expected methyl-methyl contacts are computed from a 3D structure of the protein (left). The contacts are written to a list of expected NOE peaks with the amino acid type of each contact indicated (e.g. I-V, 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 the protein sequence and knowledge of the magnetization transfer pathways for the employed NMR experiments, which are provided in the CYANA library (right).

the assignments, are optimized by an evolutionary algorithm coupled to a local optimization routine.[34, 50] 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 atom positions from the input protein structure and magnetization transfer pathways defined for each NMR experiment type to compute a network of expected peaks (Fig. 5). The mapping of expected peaks to measured ones starts from an initial population of random assignment solutions, which are optimized through successive generations by an evolutionary algorithm. To select the best individuals for recombination, a scoring function is employed, which takes into account the alignment of peaks assigned to the same atom, the completeness of the assignment, and the minimization of chemical shift degeneracy.[34] In each generation, a local optimization routine reassigns a subset of expected peaks through a defined number of iterations. This protocol is

repeated multiple 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[51] programming language) as described in the Supporting Information (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 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 with different values for the NOE distance cutoff (see below). After completion of the automated assignment calculations, the consolidation macro, CONSOL.cya, consolidates the assignment results from all individual optimization runs into a single consensus resonance assignment,[34] 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[34, 36] (Fig. 5). For each spectrum type, a 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 observing the corresponding experimental peak and a linear list of atom types that defines a molecular fragment, in which atoms must be of a 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 an 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; $^{13}C_1$, $^{13}C_2$, $^{1}H_2$ dimensions), 3D HCH-NOESY (C13NOESY; $^{1}H_1$, $^{13}C_2$, $^{1}H_2$ dimensions), 4D HCCH NOESY (CCNOESY; $^{1}H_1$, $^{13}C_1$, $^{13}C_2$, $^{1}H_2$ dimensions), 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 geminal methyl groups of Leu and Val to each other. It is formally treated as an HCcCH-TOCSY experiment in MethylFLYA. The experiment entries in the library are given in the 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. The peak lists can be supplied in XEASY[52] 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. In the MAGMA study, this information was available from manually assigned NOESY peak lists.[32] Here, unassigned NOESY peak lists are used as input, and each NOESY peak is automatically attributed to the amino acid types of the two [$^{1}H,^{13}C$]-HMQC peaks with the closest matching chemical shifts. 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 also supports joint lists for the resonances of Leu/Val type, 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. Splitting the measured and expected peak lists by residue type(s) restricts the matching of expected peaks to measured peaks of the same amino acid type(s) in the automated assignment algorithm (Fig. 5).

The distance cutoff d_{cut} for NOEs is an important parameter for generating expected NOESY cross peaks because the number of expected NOEs is approximately proportional to d_{cut}^3.

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, i.e. $d_{\text{eff}} = \left(\sum_{i=1}^{3} \sum_{j=1}^{3} d_{ij}^{-6}\right)^{-1/6}$, 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 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 $d_{\text{eff}} \approx 9^{-1/6} d_{ij} = 0.693\ d_{ij}$. It should be noted that applying, for instance, a 5 Å cutoff to the effective distance d_{eff}, allows inter-carbon distances between the two methyl groups of up to 5.0 / 0.693 + 2 × 1.1 ≈ 9.4 Å, which includes twice the C–H bond length of 1.1 Å. To avoid giving high confidence to methyl assignments that are affected by minor changes of the NOE distance cutoff parameter d_{cut}, MethylFLYA performs assignment calculations with the three slightly different cutoffs of d_{cut} – 0.5 Å, d_{cut}, and d_{cut} + 0.5 Å, and determines the consensus assignments from the results obtained with the three cutoffs (see below).

For the calculations in this paper, the NOESY cross peak observation probability was optimized (see below) and then set to 0.1 for expected NOESY peaks, and to 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.[34] 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[34] was set to 200, the value was previously found to be optimal for exclusively NOESY-based FLYA calculations.[36] The number of iterations of the local optimization routine that is coupled to the evolutionary algorithm was kept at the default value of 15,000. 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 reliable assignments, in which the algorithm has a high confidence, from others that are 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.[34] 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 chemical 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.[34]

In MethylFLYA, consolidation into consensus assignments is enhanced in three ways over the original FLYA algorithm. (i) Three series of 100 individual runs are performed with three slightly different distance cutoffs 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, thereby reducing the number of erroneous strong assignments. (ii) Special measures are necessary for the isopropyl methyls of Leu and Val, for which the stereospecific assignment is unknown *a priori*. In this case, the chemical shift values obtained for the two methyls in the individual runs are redistributed such that the consensus assignments of the first/second methyl group are determined from the smaller/larger of the two chemical shift values in each run, and FLYA does not attempt to determine stereospecific assignments. In the original FLYA algorithm[34] this approach was applied independently to the ^{1}H pair and the ^{13}C pair of a Leu or Val isopropyl group. This could result in inconsistent consensus assignments for the ^{1}H and ^{13}C resonances of Leu and Val isopropyl groups, even though the underlying ^{1}H and ^{13}C assignments from the individual runs were always consistent with each other. 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. This 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'. Other, tentative and ambiguous assignments are also reported for possible manual inspection. Further assignment statistics are given in the flya.txt file. It reports the number of expected, measured, and assigned

peaks for 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 its assignment are reported.

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.[32] In addition, experimental data for the 20 kDa N-terminal domain of Heat Shock Protein 90 (HSP90), which has also been used previously with MAGMA,[47] was used for evaluating MethylFLYA in combination with automated peak picking with CYPICK.[37] 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):[32] the N-terminal domain of *E. coli* Enzyme I (called EIN in this paper; molecular mass 28 kDa),[53] a dimer of regulatory chains of aspartate transcarbamoylase from *E. coli* (ATCase; 34 kDa),[24] maltose binding protein (MBP; 41 kDa),[54] malate synthase G (MSG; 81 kDa),[15, 18] and the "half-proteasome" 20S core particle, a 14-mer ($\alpha_7\alpha_7$; 358 kDa).[55]

The following data were taken from the MAGMA benchmark:[32] (i) Assigned [^{1}H,^{13}C]-HMQC peak lists providing reference assignments, which were not used as input data for MethylFLYA but only to evaluate the accuracy of its results. Unassigned versions of these [^{1}H,^{13}C]-HMQC peak lists were supplied to MethylFLYA. (ii) Filtered and unfiltered (see below) NOESY peak 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. Automated peak picking with CYPICK was performed for NOESY spectra in Sparky[56] format for EIN, ATCase, and HSP90. Information about Leu/Val geminal methyl pairs, which was available in the MAGMA benchmark,[32] 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. The first set comprised peak lists from the MAGMA study 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).[32] The second set comprised unfiltered peak lists, generated by manual analysis of NOESY spectra using Sparky[56] software.

Optimization of MethylFLYA parameters. To establish optimal parameters for the MethylFLYA calculations, we tested a range of values for the methyl ^{1}H–^{1}H distance cutoffs for the generation of expected NOESY cross peaks, d_{cut} = 3.0, 3.5, …, 8.0 Å (Figs. S1, S2), observation probabilities for expected methyl-methyl NOESY peaks, p_{NOE} = 0.1, 0.2, …, 0.9 (Figs. S1, S2), and the number of independent assignment optimization runs (Fig. S3).

Automated peak picking with CYPICK. The CYPICK[37] 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 placed at intensity levels $I_i = \beta L \gamma^i$, where i = 0, 1,… and L is the noise level of the spectrum that is estimated automatically by CYPICK. In this study, we used baseline factors β = 2, 3, 4, 5, 10 while keeping γ fixed at 1.3. The scaling factors for the spectral dimensions[37] 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 ^{1}H/^{13}C chemical shift from a [^{1}H,^{13}C]-HMQC peak position. Local maxima within the tolerance range that fulfilled the circularity and convexity criteria[37] were considered as peaks and stored in the peak list.

The peak picking performance was assessed by computing the find, artifact, and overall scores (with an artifact weight of 0.2) with respect to manually prepared reference peak lists[32] 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.[37]

Comparison with other assignment algorithms. The performance of the alternative structure-based methyl assignment algorithms MAGMA,[32] MAP-XSII,[29] and FLAMEnGO2.0[31] has been compared earlier.[32] Here, we used the available results and identical parameters,[32] 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).

Data availability

Data for MethylFLYA automated methyl assignment calculations is available at http://www.cyana.org/methylflya.tgz.

References

1. Steven, A. C., Baumeister, W., Johnson, L. N. & Perham, R. N. *Molecular Biology of Assemblies and Machines*. Garland Science (2016).
2. 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* **94**, 12366–12371 (1997).
3. 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.* **125**, 10420–10428 (2003).
4. Ollerenshaw, J. E., Tugarinov, V. & Kay, L. E. Methyl TROSY: explanation and experimental verification. *Magn Reson Chem* **41**, 843-852 (2003).
5. Religa, T. L., Sprangers, R. & Kay, L. E. Dynamic regulation of archaeal proteasome gate opening as studied by TROSY NMR. *Science* **328**, 98-102 (2010).
6. Rosenzweig, R. & Kay, L. E. Bringing dynamic molecular machines into focus by methyl-TROSY NMR. *Annu. Rev. Biochem.* **83**, 291-315 (2014).
7. Boswell, Z. K. & Latham, M. P. Methyl-based NMR spectroscopy methods for uncovering structural dynamics in large proteins and protein complexes. *Biochemistry* **58**, 144-155 (2019).
8. Xing, Q., Shi, K., Portaliou, A., Rossi, P., Economou, A. & Kalodimos, C. G. Structures of chaperone-substrate complexes docked onto the export gate in a type III secretion system. *Nat. Commun.* **9**, 1773 (2018).
9. Zhang, H. Y. & van Ingen, H. Isotope-labeling strategies for solution NMR studies of macromolecular assemblies. *Curr. Opin. Struct. Biol.* **38**, 75-82 (2016).
10. Wiesner, S. & Sprangers, R. Methyl groups as NMR probes for biomolecular interactions. *Curr. Opin. Struct. Biol.* **35**, 60-67 (2015).
11. 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* **65**, 15-27 (2016).

12. Clark, L. et al. Methyl labeling and TROSY NMR spectroscopy of proteins expressed in the eukaryote Pichia pastoris. *J. Biomol. NMR* **62**, 239-245 (2015).
13. 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* **71**, 213-223 (2018).
14. Kofuku, Y. et al. Deuteration and selective labeling of alanine methyl groups of β_2-adrenergic receptor expressed in a baculovirus-insect cell expression system. *J. Biomol. NMR* **71**, 185-192 (2018).
15. 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* **102**, 622–627 (2005).
16. Gorman, S. D., Sahu, D., O'Rourke, K. F. & Boehr, D. D. Assigning methyl resonances for protein solution-state NMR studies. *Methods* **148**, 88-99 (2018).
17. Kay, L. E., Ikura, M., Tschudin, R. & Bax, A. Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. *J. Magn. Reson.* **89**, 496–514 (1990).
18. 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.* **125**, 13868-13878 (2003).
19. 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.* **34**, 93–158 (1999).
20. 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* **102**, 16678-16683 (2005).
21. Sprangers, R., Velyvis, A. & Kay, L. E. Solution NMR of supramolecular complexes: providing new insights into function. *Nat. Methods* **4**, 697–703 (2007).
22. Gelis, I. et al. Structural basis for signal-sequence recognition by the translocase motor SecA as determined by NMR. *Cell* **131**, 756-769 (2007).
23. 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* **54**, 4307–4319 (2015).
24. 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.* **131**, 16534-16543 (2009).

25. 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.* **129**, 13749–13757 (2007).
26. Lescanne, M. et al. Methyl group assignment using pseudocontact shifts with PARAssign. *J. Biomol. NMR* **69**, 183-195 (2017).
27. 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* **51**, 319–328 (2011).
28. Xu, Y. Q. et al. Automated assignment in selectively methyl-labeled proteins. *J. Am. Chem. Soc.* **131**, 9480–9481 (2009).
29. Xu, Y. Q. & Matthews, S. MAP-XSII: an improved program for the automatic assignment of methyl resonances in large proteins. *J. Biomol. NMR* **55**, 179–187 (2013).
30. 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.* **214**, 103–110 (2012).
31. 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.* **245**, 17–23 (2014).
32. Pritisanac, I. et al. Automatic assignment of methyl-NMR spectra of supramolecular machines using graph theory. *J. Am. Chem. Soc.* **139**, 9523–9533 (2017).
33. Monneau, Y. R. et al. Automatic methyl assignment in large proteins by the MAGIC algorithm. *J. Biomol. NMR* **69**, 215-227 (2017).
34. Schmidt, E. & Güntert, P. A new algorithm for reliable and general NMR resonance assignment. *J. Am. Chem. Soc.* **134**, 12817–12829 (2012).
35. Güntert, P. & Buchner, L. Combined automated NOE assignment and structure calculation with CYANA. *J. Biomol. NMR* **62**, 453–471 (2015).
36. Schmidt, E. & Güntert, P. Reliability of exclusively NOESY-based automated resonance assignment and structure determination of proteins. *J. Biomol. NMR* **57**, 193–204 (2013).
37. Würz, J. M. & Güntert, P. Peak picking multidimensional NMR spectra with the contour geometry based algorithm CYPICK. *J. Biomol. NMR* **67**, 63–76 (2017).
38. Schmidt, E. et al. Automated solid-state NMR resonance assignment of protein microcrystals and amyloids. *J. Biomol. NMR* **56**, 243–254 (2013).
39. Aeschbacher, T. et al. Automated and assisted RNA resonance assignment using NMR chemical shift statistics. *Nucleic Acids Res.* **41**, e172 (2013).

40. 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* **59**, 87–93 (2014).
41. Schmidt, E. et al. Automated resonance assignment of the 21 kDa stereo-array isotope labeled thioldisulfide oxidoreductase DsbA. *J. Magn. Reson.* **249**, 88–93 (2014).
42. Lichtenecker, R. J., Coudevylle, N., Konrat, R. & Schmid, W. Selective isotope labelling of leucine residues by using α-ketoacid precursor compounds. *ChemBioChem* **14**, 818-821 (2013).
43. 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* **57**, 205-209 (2013).
44. Gans, P. et al. Stereospecific isotopic labeling of methyl groups for NMR spectroscopic studies of high-molecular-weight proteins. *Angew. Chem. Int. Ed.* **49**, 1958-1962 (2010).
45. Orts, J. et al. NMR-based determination of the 3D structure of the ligand-protein interaction site without protein resonance assignment. *J. Am. Chem. Soc.* **138**, 4393–4400 (2016).
46. Mohanty, B. et al. Determination of ligand binding modes in weak protein-ligand complexes using sparse NMR data. *J. Biomol. NMR* **66**, 195-208 (2016).
47. 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.* **55**, 10786–10790 (2012).
48. 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* **71**, 275-285 (2018).
49. Huber, M. et al. A proton-detected 4D solid-state NMR experiment for protein structure determination. *Chemphyschem* **12**, 915-918 (2011).
50. Bartels, C., Güntert, P., Billeter, M. & Wüthrich, K. GARANT - A general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *J. Comput. Chem.* **18**, 139–149 (1997).
51. Güntert, P., Dötsch, V., Wider, G. & Wüthrich, K. Processing of multidimensional NMR data with the new software PROSA. *J. Biomol. NMR* **2**, 619–629 (1992).
52. 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* **6**, 1–10 (1995).
53. 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* **36**, 2517-2530 (1997).

54. 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.* **120**, 11738–11748 (1998).
55. 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.* **129**, 1743-1750 (2007).
56. Goddard, T. D. & Kneller, D. G. Sparky 3. (ed^(eds). University of California (2001).

Acknowledgments

We thank Prof. Andrew Baldwin for help with the MAGMA benchmark data set. 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) is gratefully acknowledged.

Author contributions

I.P. and P.G. designed and performed research. J.M.W. implemented and performed automated peak picking. I.P., T.R.A., and P.G. wrote the paper. All authors contributed to data interpretation and commented on the manuscript.