IOP Conf. Series: Journal of Physics: Conf. Series **1036** (2018) 012008 doi:10.1088/1742-6596/1036/1/012008

Fully automated assignment of methyl resonances of a 36 kDa protein dimer from sparse NOESY data

I Pritišanac*[1§], JM Würz[1§] & P Güntert*[1,2,3]

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

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

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

*E-mail: pritisanac@em.uni-frankfurt.de, guentert@em.uni-frankfurt.de

Abstract. High-resolution solution-state NMR spectroscopy studies of large proteins typically require uniform deuteration of the system and selective protonation and isotope labelling of methyl groups. Under such circumstances, the assignment of methyl resonances presents a considerable experimental challenge and automation of the process using computational algorithms has been actively sought. Through-space connectivities between the labelled methyl groups can be established through nuclear Overhauser enhancement spectroscopy (NOESY). If a high-resolution structure of the system is available, the sparse connectivity restraints derived from this information enable structure-based methyl resonance assignment. Here, we outline a protocol for full automation of the methyl resonance assignment process using the CYANA software package. We tested the protocol on three-dimensional (3D) $^{13}C/^{13}C$-separated NOESY spectra of a dimer of regulatory chains of aspartate transcarbamoylase (ATCase-r_2). We used CYPICK to detect NOE signals, followed by automatic resonance assignment with FLYA. On this dataset, FLYA generated highly similar results using either automatically or manually generated peak lists, confidently assigning ~60% of the methyl groups with high accuracy (95 ± 2% correctness). We compared this performance to two alternative automatic methyl assignment protocols, MAP-XSII and FLAMEnGO2.0, both of which, similarly to FLYA, support unassigned NOESY peak lists as input.

[§]contributed equally

1. Introduction

High-resolution nuclear magnetic resonance (NMR) spectroscopy studies of biological macromolecules occupy a central role in biophysical chemistry, structural biology, drug design, structural genomics, and interactomics [1]. Hydrogen (^{1}H, or proton) is highly abundant in biological macromolecules and, owing to its high gyromagnetic ratio, is the most sensitive NMR probe of molecular structure and dynamics. Studies of complex biomolecules are aided by isotopic labelling that introduces NMR-active atomic nuclei to a system [2]. For instance, in studies of small and medium-size protein molecules, uniform labeling with carbon-13 (^{13}C) and nitrogen-15 (^{15}N) isotopes replaces NMR-inactive ^{12}C and ^{14}N nuclei. The *J*-coupling interactions between covalently linked ^{1}H, ^{13}C and ^{15}N nuclei are routinely exploited to establish atomic connectivity along the protein backbone and side chains [3,4].

A high density of NMR-active nuclei, especially protons, in biomolecular samples is a source of nuclear relaxation, which leads to rapid signal decays. Consequently, line broadening is a hallmark of high molecular weight (>30 kDa) protein NMR spectra [5]. To simultaneously optimize resolution and sensitivity for larger molecules, partial or complete deuteration (^{2}H) of samples can be employed, followed by back-introduction of solvent exchangeable ^{1}H nuclei [6]. However, as the molecular mass of proteins or protein complexes increases, additional considerations must be given to isotopic labelling. For instance, deuteration can be combined with ^{1}H, ^{13}C-labelling of only some amino acids or side chains. A particularly successful strategy is selective protonation of ^{13}C-labeled methyl groups ($^{13}CH_3$) in otherwise ^{12}C-labeled, uniformly deuterated proteins. Cost-effective and robust biosynthetic strategies have been established for selective or simultaneous labelling of all methyl-containing amino acids in *Escherichia coli* [7,8], and of isoleucine-δ1 methyls in a eukaryote, *Pichia pastoris* [9]. One additional advantage of methyl groups is their high abundance in protein sequences, as six out of 20 amino acids have at least one methyl group in their side chain. Moreover, methyls are found both in the core and, to a lesser extent, at the surface of proteins, making them particularly useful probes of protein structure and dynamics [10,11].

Cross-correlated relaxation effects in $^{13}CH_3$ spin systems in the macromolecular limit can be exploited in heteronuclear multiple quantum correlation (HMQC) spectroscopy, offering considerable improvements in spectral resolution. Hence, the measurement of [^{1}H,^{13}C]-HMQC spectra is referred to as methyl transverse relaxation-optimized spectroscopy (methyl-TROSY) [12]. Combined with the three-fold degeneracy of methyl ^{1}H spins, spectra can be recorded with intense and sharp resonances that are well dispersed in two-dimensional ^{1}H–^{13}C correlation plots. Taken together, the combination of isotopic labelling and advantageous spectroscopic properties of methyl groups now allows atomic-level insight into the structures and dynamics of large proteins and protein oligomers up to 1 MDa in molecular mass [13–17].

The assignment of methyl resonances is an essential prerequisite for the interpretation of structural and dynamical information derived from methyl-TROSY studies. Resonance assignment refers to the attribution of ^{1}H-^{13}C correlations from [^{1}H,^{13}C]-HMQC spectra to individual methyl-bearing residues in the protein sequence. The assignment step in NMR studies can be particularly laborious and time-consuming, especially with increasing complexity of the investigated systems [17,18]. The aforementioned line-broadening and extensive overlap of backbone chemical shifts in spectra of large proteins prevent standard through-bond NMR assignment strategies. Instead, a typically employed experimental strategy for large proteins is fragmentation of the system into smaller components for which through-bond NMR spectra of sufficient quality can be obtained [14,15,19]. Assignment of methyls is then obtained using standard experiments that correlate NMR backbone and side-chain resonances [4]. This strategy can be supplemented with site-directed mutagenesis of individual methyl-bearing residues [14,17]. Although highly reliable, these approaches are laborious and come with the disadvantage that any changes introduced to the protein for assignment purposes may result in considerable shifts of correlations in [^{1}H,^{13}C]-HMQC spectra. This may compromise the reliable transfer of assignments between the spectra of protein fragments/mutants and of the intact protein.

Alternatively, if a high-resolution structure or a homology model of the protein is available, automatic, structure-based NMR resonance assignment strategies can be considered [20–25]. In a

methyl-TROSY study, isotopically labelled $^{13}CH_3$ groups are the only available source of connectivity information, and therefore, automatic assignment strategies can only make use of highly sparse and ambiguous experimental information. Currently available automatic methyl assignment approaches make use of short- and long-range distance restraints, which are obtained by nuclear Overhauser enhancement spectroscopy (NOESY) and/or measurements of paramagnetic relaxation enhancements (PREs). Both of these restraints can provide inter-methyl distance information, with PREs yielding inter-methyl distances up to 25-35 Å [20], depending on the paramagnetic tag employed, and NOEs reported on distances up to 12 Å in highly deuterated systems [26]. Importantly, NOEs do not require mutagenesis to introduce a paramagnetic tag and can reveal a network of connectivity between spatially proximal methyl groups [25].

Presently available automatic methyl resonance assignment approaches utilize these through-space restraints in different ways. An approach introduced by Venditti *et al.* [20] uses experimental ^{1}H-methyl PRE rates as primary restraints. A Metropolis Monte Carlo algorithm is employed to find assignment solutions that minimize the difference between measured PRE rates and those expected from a high-resolution protein structure. Measured chemical shifts and NOEs provide additional restraints to refine assignments. Two other programs, Methyl Assignment Prediction from X-ray Structures (MAP-XS) [21,22] and Fuzzy Logic Assignment of Methyl Groups (FLAMEnGO) [23,24] also employ Metropolis Monte Carlo optimization, but rely primarily on methyl-methyl NOEs as restraints. Both programs evaluate matches between structure-based predictions and experimentally measured NOEs and methyl chemical shifts. Additional restraint-specific terms are defined for optional incorporation of PREs, pseudocontact shifts (PCS), residual dipolar couplings (RDC), and TOCSY data in the search [21–24]. Most recently, Methyl Assignment by Graph MAtching (MAGMA) was introduced, which compares graphs derived from methyl-methyl NOEs and a protein structure using a combination of exact algorithms for graph-subgraph monomorphism and maximal common edge subgraph problems [25]. To treat the high sparsity of input restraints, MAGMA exhaustively samples all theoretically possible methyl assignment solutions that maximize the number of explained methyl-methyl NOEs. Relying solely on NOE restraints, MAGMA could achieve perfectly accurate methyl resonance assignments on a range of protein targets of various molecular sizes and shapes [25].

All currently available automatic methyl assignment strategies rely on the identification of methyl-methyl NOE cross-peaks by manual analysis of 3D or 4D ^{13}C-resolved NOESY spectra. Peak picking produces NOESY peak lists, in which each peak position is defined by proton (^{1}H) and carbon (^{13}C) frequencies of the two spatially proximal methyl groups that give rise to an NOE. In the case of a 4D ^{13}C-NOESY peak list, ^{1}H and ^{13}C frequencies from both NOE-contributing methyls are known ($^{1}H_1$,$^{13}C_1$,$^{1}H_2$,$^{13}C_2$), whereas for a 3D ^{13}C-resolved NOESY list, either the ^{13}C or ^{1}H frequency of one of the methyls is unknown (e.g. $^{1}H_1$,$^{13}C_1$,$^{13}C_2$; Fig. 1). Manually prepared peak lists are typically curated by an experienced spectroscopist before being fed to the automatic assignment programs, which introduces a user bias to the process.

A missing link in the full automation of the methyl assignment using currently available protocols [22,24,25] is the automatic detection of methyl-methyl NOEs. Recently, a contour-based automatic signal detection protocol (CYPICK) was introduced and compared to other approaches [27]. CYPICK analyses the NMR spectrum in the form of two-dimensional contour plots, mimicking the manual approach taken by a spectroscopist as far as possible. The human visual inspection of the spectrum is replaced by the analysis of certain geometric contour line properties, i.e. local extremality, approximate circularity (after appropriate scaling of spectral axes), and convexity. CYPICK has been tested on several soluble proteins up to a molecular mass of ~20 kDa, and was shown to work optimally with the FLYA automatic resonance assignment protocol, performing particularly favourably on ^{13}C-edited and ^{15}N-edited NOESY data, as compared to other existing algorithms [27]. FLYA is a generic automatic resonance assignment protocol incorporated in CYANA that supports the input of NMR peak lists from a variety of through-bond (*J*-coupling), or through-space (NOESY) experiments for the assignment of protein backbone and side-chain resonances [28,29]. To optimize the match between expected and experimentally observed peaks, FLYA employs an evolutionary algorithm that operates on a population

of assignment solutions (individuals) governed by two scoring functions that consider the fitness of the assignment of individual atoms (local) and of the complete assignment solution (global). FLYA supports structure-based resonance assignment on the basis of ^{13}C- and ^{15}N-resolved NOESY data [29]; however, it has not been tested on datasets originating from large, exclusively methyl-labelled proteins.

Here, we tested a fully automatic, user-unbiased protocol for methyl resonance assignment using CYANA. We used CYPICK to automatically detect NOE signals in a 3D ^{13}C/^{13}C-separated NOESY spectrum of a 36 kDa homodimer of regulatory subunits of aspartate transcarbamoylase (ATCase-r_2). The CYPICK-generated NOESY peak lists were directly fed to the structure-based FLYA assignment protocol. We show that, on this dataset, FLYA achieves similar levels of assignment coverage (~60%) and accuracy (~95%) as with manually prepared NOESY peak lists, and that these results compare favorably to alternative methyl resonance assignment approaches.

2. Methods

Fig. 1 outlines the CYPICK-FLYA coupled automatic methyl resonance assignment strategy. The two protocols (CYPICK and FLYA) used in the study were not modified from their original implementations [27,28]. An alteration was introduced in the consolidation step of the FLYA assignment result analysis, as detailed in section 2.2.3 below.

2.1. CYPICK calculations

The noise level (L) of the spectrum was estimated automatically by CYPICK and multiplied by a user-defined baseline factor β for defining the intensity of the lowest contour line, B:

$$B = I_0 = \beta L.$$

The intensity I_i of contour line i is calculated from

$$I_i = B\gamma^i.$$

In our study, we used β values of 2, 3, and 5 while keeping γ fixed at 1.3, since moderate changes of γ did not show any significant influence on the resulting peak lists. The scaling factors for the spectral dimensions [27] were set to σ_1 = 0.18 ppm for the $^{13}C_1$ dimension, σ_2 = 0.16 ppm for $^{13}C_2$, and σ_3 = 0.036 ppm for $^{1}H_1$. The ^{13}C/^{13}C-separated NOESY spectrum was picked using the recently implemented restricted peak picking mode of CYPICK [27]. A manually prepared 2D [^{1}H,^{13}C]-HMQC peak list was used as a frequency filter in CYPICK, restricting the peak picking of the ^{13}C/^{13}C-separated NOESY spectrum to specific spectral regions. Only local maxima in the ^{13}C/^{13}C-separated NOESY that were found within a given tolerance range of a manually identified 2D HMQC peak were further analyzed in the contour line analysis routine. This tolerance was set to 0.1 ppm for ^{13}C and 0.01 ppm for ^{1}H. Local maxima within the tolerance range that fulfilled the circularity and convexity criteria, as described in Würz *et al.* [27], were considered as peaks and stored in a peak list.

The peak picking performance was evaluated based on *find*, *artefact*, and *overall* scores with respect to a reference peak list. The definition of these scores is described in [27]. As a reference, we used a manually prepared ^{13}C/^{13}C-separated NOESY peak list [25] with a tolerance of 0.04 ppm for ^{1}H and 0.4 ppm for heavy atoms. *Overall* scores were computed using an artefact weight of 0.2.

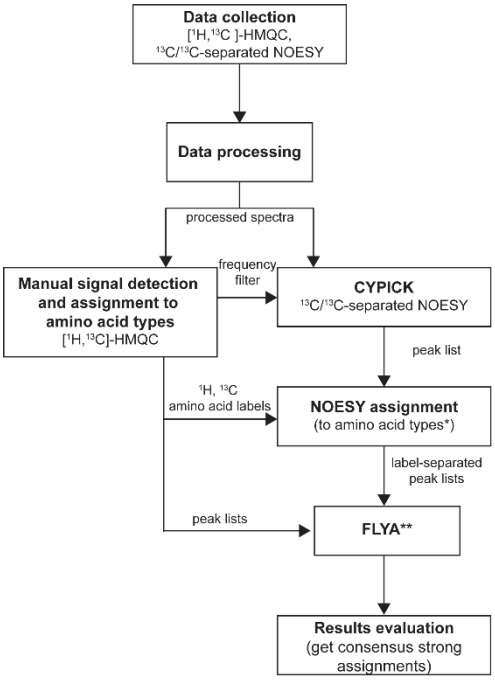

Figure 1. Outline of the CYPICK-FLYA protocol for fully automated methyl resonance assignment. *NOE peaks are assigned to types according to their amino acid assignment: Ile-Ile, Ile-Leu, Ile-Val, Leu-Ile, Leu-Leu, Leu-Val, Val-Ile, Val-Leu, Val-Val. **Automatically analysed and amino acid type-assigned NOE peak lists can be fed to FLYA, or other automatic methyl resonance assignment programs that support peak lists as input.

2.2. FLYA calculations

2.2.1. Preparation of peak lists for FLYA

FLYA supports input data from a variety of through-bond or through-space NMR experiments, which are fed to the algorithm as appropriately formatted peak lists [28,29]. The data presented in this work consisted of a 3D $^{13}C/^{13}C$-separated NOESY (CCNOESY) spectrum and a 2D [^{1}H,^{13}C]-HMQC spectrum, both of which were collected on a uniformly deuterated, Ile-[$\delta1^{13}CH_3$], Leu-[$\delta^{13}CH_3$,$^{12}CD_3$], Val-[$\gamma^{13}CH_3$,$^{12}CD_3$] labelled protein sample of the r-chains of aspartate transcarbamoylase (ATCase-r_2) [18].

To reduce the high dimensionality of the combinatorial search for methyl resonance assignments, following the strategy outlined by MAGMA [25], the 2D [^{1}H,^{13}C]-HMQC peak list was first split according to the amino acid types of the methyl groups (in this case Ile, Leu, Val); i.e. prior knowledge of this information was assumed.

The amino acid type assignment of every ^{1}H-^{13}C cross-peak from the 2D [^{1}H,^{13}C]-HMQC peak list was used to automatically assign NOE interactions in the three-dimensional NOESY peak list to amino acid types. For each entry in the NOESY peak list, the difference between $^{13}C_1$, $^{1}H_1$ shifts and the reference ^{1}H, ^{13}C frequencies in the HMQC peak list is computed:

$$\Delta_1 = \sqrt{a(C_1^{NOE} - C_{ref}^{HMQC})^2 + (H_1^{NOE} - H_{ref}^{HMQC})^2}$$

as well as the absolute difference between $^{13}C_2$ NOESY shifts and the ^{13}C shifts from the reference HMQC peak list:

$$\Delta_2 = \left| C_2^{NOE} - C_{ref}^{HMQC} \right|$$

Each NOE is then attributed the C_1-H_1 and C_2 amino acid types of the reference HMQC peak for which Δ_1 and Δ_2 are minimal, respectively. This results in the NOESY peak list type assignments as shown in Fig. 2b. The CCNOESY peak list is then split according to the types of NOE interactions into separate Ile-Ile, Ile-Leu, Ile-Val, Leu-Leu, Leu-Ile, Leu-Val, Val-Val, Val-Ile, Val-Leu NOESY peak lists. The expected peaks computed by FLYA based on the input crystal structure are also classified according to the above types. This restricts matching of measured NOESY peaks to expected NOESY peaks of the same NOE interaction type during the FLYA automatic assignment calculation (Fig. 1).

2.2.2. FLYA assignment calculations

All FLYA calculations were carried out using the same parameters. To generate expected methyl-methyl NOE peaks, ^{1}H-^{1}H distances were computed from input crystal structures. For ATCase-r_2, a representative structure of the T-state of the enzyme (PDB ID: 1TUG [30]) was used, as previously proposed by Velyvis and Kay [18]. In addition, FLYA was run with a representative structure of the R-state of the enzyme (PDB ID: 1D09 [31]).

In FLYA, the ^{1}H-^{1}H distance for a pair of methyl groups is computed as the r^{-6} sum over the nine individual ^{1}H-^{1}H distances:

$$d_{\text{eff}} = \left(\sum_{i=1}^{3} \sum_{j=1}^{3} d_{ij}^{-6} \right)^{-1/6}$$

where d_{eff} stands for the effective (FLYA-computed) distance, the sums run over the ^{1}H atoms of methyl groups 1 and 2, respectively, and d_{ij} is the Euclidean distance between the individual methyl protons i and j in the input structure. For instance, if we assume all d_{ij} distances to be approximately equal, then $d_{\text{eff}} \approx 9^{-1/6} d_{ij} = 0.693\, d_{ij}$. Considering this approximate scaling factor, the distance cut-off $d_{\text{eff}} \leq d_{\text{cut}}$ for generating expected methyl-methyl NOE peaks was set to 5 Å, corresponding to an average maximal observable ^{1}H-^{1}H distance of $5.0/0.693$ Å ≈ 7.2 Å. The optimal distance cut-off was established by running calculations with distance cut-offs of 4–10 Å in steps of 0.5 Å. The optimal value was chosen as the distance at which the highest percentage of expected peaks were assigned (Fig. A2).

In FLYA, the probability for expected methyl-methyl NOE peaks was set to 0.2 for all distances [28,29]. The optimality of this value was established in a more systematic effort on a larger benchmark of methyl-methyl NOESY data (data not shown). The chemical shift tolerances for the chemical shift assignment calculations were 0.4 ppm for ^{13}C and 0.04 ppm for ^{1}H chemical shifts. These values were reduced in the evaluation step (see next section) to 0.2 ppm and 0.02 ppm. The size of the population for the evolutionary algorithm of FLYA was set to 200, which was previously reported as optimal for NOESY-only FLYA calculations [29]. To increase the reliability of assignment solutions, 20 independent FLYA calculations were performed in parallel with the same input data but different random number generator seeds [28,29].

2.2.3. FLYA assignment consolidation. For methyl resonance assignments, a change in the FLYA shift consolidation function was made to allow for simultaneous 'consolidation' of both ^{1}H and ^{13}C chemical shifts for each assigned methyl group. Normally, FLYA computes a consensus chemical shift for every assigned atom based on the values from (typically 20) separate, independent runs of the algorithm (see above). To allow for simultaneous consolidation of ^{1}H -^{13}C chemical shift pairs, we performed k-means clustering [32,33] in two dimensions, with the number of centroids set to three. FLYA defines an atom assignment as strong if at least 80% of the chemical shift values, obtained for that atom from the

independent algorithm runs, deviate less than the allowed matching tolerance (see above) from the consensus value. Consequently, we here define a methyl assignment as strong if the largest k-means cluster has at least 16 members (80% of 20) or if multiple smaller clusters, whose centroids are separated by less than the tolerance value in ^{1}H and ^{13}C dimensions, amount to at least 16 members. Shift assignments that do not fulfil these criteria are considered weak and therefore unreliable. The k-means clustering option of assignment results for FLYA chemical shift consolidation will be available with the next release of CYANA.

2.2.4. FLYA calculation replicates. We performed an additional 20 replicates of the entire FLYA calculation (which in turn comprises 20 independent, parallel runs of the optimization algorithm) with the optimal parameter values to check for the consistency between different sets of the 20 parallel calculations.

2.3. MAP-XS and FLAMEnGO calculations

MAP-XSII and FLAMEnGO2.0 protocols were run and their results analysed according to instructions outlined in the original publications [21–24]. For both programs, a range of distance thresholds for computing theoretical inter-methyl NOE restraints based on a high-resolution protein structure were considered (Figs. A3, A4). Calculations were run with both the structure of the T-state (PDB ID: 1TUG [30]) and the R-state (PDB ID: 1D09 [31]) of the ATCase-r_2 dimer. The optimal distance thresholds for MAP-XSII and FLAMEnGO2.0 calculations were determined as 9 Å and 15 Å, respectively (Figs. A3, A4). For FLAMEnGO2.0, 10000 Monte Carlo (MC) steps were run for a range of distances to obtain the graph in Fig. A4. Subsequently, 100 parallel calculations, each with 100000 MC steps were run at the optimal distance threshold of 15 Å. With MAP-XSII, 20 parallel calculations were run with the weighting between chemical shifts and NOEs set to the default value of 0.2.

3. Results and Discussion

We tested different values of the CYPICK baseline factor parameter β on a 3D ^{13}C/^{13}C-separated NOESY spectrum of ATCase-r_2. As established previously [27], an increase in the baseline factor reduces the extent of artefact picking (Table 1), which in turn leads to a better outcome of the automatic resonance assignment with FLYA (Fig. A1). In contrast, enhancing the baseline factor reduces the number of real peaks that are identified. The effect of changes in the contour line factor γ on the peak picking scores was negligible, and this parameter was therefore kept fixed throughout this study (see Methods). It was previously shown that CYPICK performs well in combination with FLYA when applied to the assignment of backbone resonances of uniformly ^{15}N/^{13}C-labelled proteins with up to ~20 kDa molecular mass. Using exclusively ^{13}C-edited and ^{15}N-edited NOESY data analyzed with CYPICK, FLYA resonance assignments were accurate enough to calculate structure bundles in a fully automated fashion with an RMSD bias below 2 Å [27]. Similarly, we note that, in this study, FLYA achieved the best methyl assignments for ATCase-r_2 using the CYPICK peak list with the lowest artefact score (12.9%, Table 1, Fig. A1).

Table 1. CYPICK results for the ATCase-r_2 3D CCH-NOESY spectrum

Baseline factor (β)	Find score (%)	Artefact score (%)	Overall score (%)
2	86.5	40.5	74.7
3	84.7	25.3	79.0
5	76.6	12.9	74.4

The high dimensionality of the search space and the sparsity of restraints render methyl resonance assignment a difficult combinatorial problem. For optimal performance of automatic algorithms, it is essential to include the maximal amount of experimentally attainable information in the search. The FLYA search space can be considerably reduced if methyl resonances are separated into distinct groups according to their amino acid types (e.g. Ile, Leu, Val). Ile methyl resonances are readily recognized in

[^{1}H,^{13}C]-HMQC spectra due to their upfield (i.e. lower chemical shift) ^{13}C frequencies (Fig. 2). Discrimination between the overlapping regions of Leu and Val resonances can be achieved by preparing separate Leu-only or Val-only methyl-labelled protein samples [34,35]. Segregation of the search space according to amino acid type can also be applied to the interpretation of methyl-methyl NOESY data (see Methods). With this information available, FLYA then maps inter-methyl contacts (expected peaks) extracted from a protein structure only to methyl-methyl NOEs of the same type (Fig. 2b). In the protocol of Fig. 1, the classification of methyl-methyl NOEs is done automatically, based on the amino acid labelling of frequencies from the [^{1}H,^{13}C]-HMQC peak list (see Methods). As such, the protocol effectively automates both NOE peak picking and amino acid type classification (Fig 2). Although presented here on an example featuring selective Ile-δ1, Leu-δ, and Val-γ labelling, the protocol can easily be extended to any of the other available methyl labelling schemes including, for instance, labelling of Ile-γ2, Ala-β, Met-ε, or Thr-γ methyls.

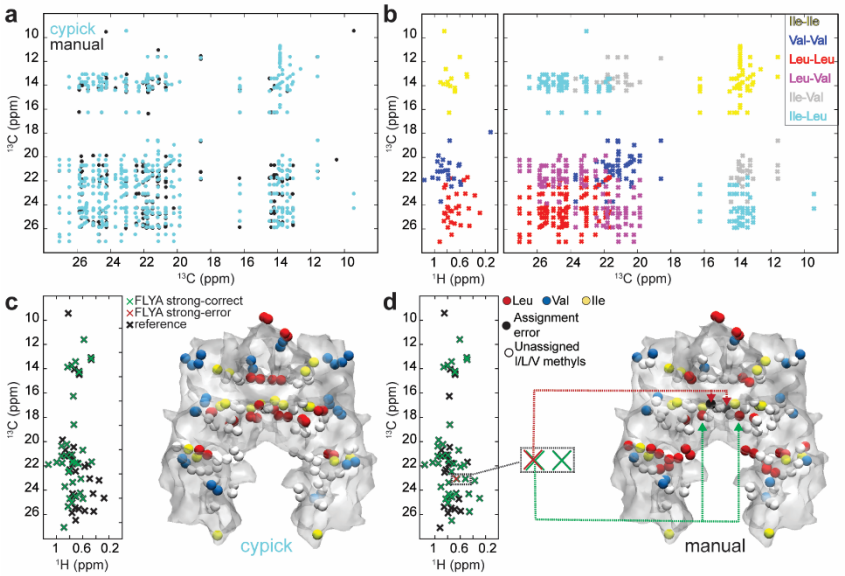

Figure 2. FLYA automatic methyl resonance assignment results using manually prepared and automatically analysed ^{13}C/^{13}C separated NOESY data. (**a)** Overlay of ^{13}C/^{13}C projections of CYPICK (cyan) and manually generated 3D NOESY peak lists (black). (**b)** Separation of methyl-methyl NOEs according to amino acid types. Positions of methyl ^{1}H-^{13}C correlations for Ile (yellow), Val (blue), and Leu (red) in the two-dimensional [^{1}H,^{13}C]-HMQC spectrum (left) and the ^{13}C/^{13}C projection of the 3D NOESY (right). (**c**, **d**) Left, overlays of the FLYA 'strong' assigned ^{1}H, ^{13}C chemical shifts with chemical shifts of ATCase-r$_2$ measured in the [^{1}H,^{13}C]-HMQC spectrum (black) using CYPICK (c) or manually analysed NOE data (d). Correct and erroneous assignments are shown in green and red, respectively. Right, representation of the FLYA assignment coverage on the T-state structure of ATCase-r$_2$. Methyl carbons are shown as balls; yellow, blue, and red balls correspond to correct strong FLYA assignments of Ile, Val, and Leu methyls, respectively. Unassigned methyls (i.e. weak FLYA assignments) are shown in white, and FLYA assignment errors in black.

The optimal distance cut-off for the computation of expected methyl-methyl NOE cross-peaks was found to be ~5 Å (Fig. A2). This threshold may appear low, considering the larger range of inter-methyl

distances shown to be measurable for selectively methyl-protonated, uniformly deuterated proteins [25,26]. However, this FLYA derived distance reflects r^{-6} summation over all ^{1}H-^{1}H distances of two methyl groups, and thus corresponds to an actual distance of ~7.2 Å between individual protons (see Methods). For comparison, for the same dataset, a 10 Å carbon-carbon distance was found to be optimal for the computation of expected NOEs by MAGMA [25]. Adding the length of the C-H bond (~1.1 Å) [36] on both ends, in the simplest manner with no geometry considerations, the optimal FLYA C-C distance cut-off is only slightly shorter than that of MAGMA (7.2 Å + 2 × 1.1 Å = 9.4 Å), showing a good agreement between the two methods.

In Fig. 2, FLYA results for CYPICK (c) and manually prepared (d) methyl-NOESY peak lists are compared. The FLYA-calculated strong chemical shift assignments (green/red) overlay well with the reference [^{1}H,^{13}C]-HMQC frequencies of ATCase-r$_2$ (black). Using a CYPICK-generated NOESY peak list, FLYA assigned 58% of methyl resonances as 'strong' (i.e. confident) with complete accuracy (Fig. 2c). The manually acquired NOESY peak list resulted in 68% of confidently assigned resonances, albeit with one assignment error (Figs. 2d, 3). Closer inspection of the result shows that the erroneously assigned methyl group (Fig. 2d, red) shares the ^{1}H-^{13}C frequency position with another methyl (Fig. 2d, green), which is the correct assignment for that position. Moreover, in this case, the two methyls are spatially proximal in the protein structure (Fig. 2d). Such 'assignment overlaps' may occur occasionally in FLYA assignment results, as FLYA allows multiple expected peaks to be mapped to a measured peak, and should be given special attention in the inspection of FLYA results.

It should be noted that FLYA uses an approximate approach (genetic algorithm) to calculate resonance assignments. Therefore, despite consolidating assignment results across a set of multiple (typically 20), independent runs of the optimization algorithm, different sets could in principle feature some assignment differences. We thus repeated a full set of 20 parallel runs multiple times to evaluate the consistency in coverage and accuracy of methyl assignments determined by FLYA (see Methods). On ATCase-r$_2$ data, on average, the fully automated CYPICK-FLYA protocol yields strong (confident) assignments for 68 ± 4% of the methyls with 95 ± 2% accuracy. Using manually prepared lists, 66 ± 6% strong methyl resonance assignments are obtained with 94 ± 1% accuracy.

For the FLYA calculations presented here (Figs. 2, 3), we used a structure of the r-chain dimer (r$_2$) extracted from a structure of the T-state of the enzyme [30]. The T-state is predominantly populated in the apo form of the enzyme, which is consistent with the work of Velyvis *et al.* [18] from which the reference methyl assignments for the ATCase-r$_2$ dimer were obtained. However, when extracted from the holoenzyme context, the isolated ATCase-r$_2$ dimer likely features some structural and dynamical differences, as suggested by a comparison of the [^{1}H,^{13}C]-HMQC spectrum of the selectively ILV-labelled r-chain in the holoenzyme with that of the isolated r-chain dimer [18]. Given that the exact solution structure of the isolated ATCase-r$_2$ is not known and, as previously concluded in the MAGMA study, it may not be sufficient to consider a single conformation in a structure-based assignment strategy [25], we also checked the performance of FLYA on ATCase-r$_2$ extracted from an R-state structure of the enzyme and analyzed the consistency of the methyl assignments across both structural forms (Fig. A5). FLYA generates more 'strong' (i.e. confident) assignments for the R-state ATCase-r$_2$, albeit with considerably lower accuracy (87% compared to 100% for the T-state; Fig. A5). If strong assignments that are consistent across both structures are considered, the coverage of strong assignments falls to ~40%; however, all of these assignments are accurate. The FLYA 'preference' for the T-state form of ATCase-r$_2$ differs from that reported for MAGMA, where completely accurate, confident assignment for 31% of methyl residues was achieved using a structure extracted from the R-state, albeit with manually prepared input NOE restraints [25]. This discrepancy is likely a combination of differences in the input restraints (automatically vs. manually picked NOEs), and algorithmic approaches (genetic vs. exact graph matching). This difference in conformational preference between MAGMA and FLYA for this dataset will therefore be subject to further investigation.

Finally, we compared the performance of FLYA on CYPICK-generated NOESY peak lists to the automatic methyl assignment protocols MAP-XSII and FLAMEnGO2.0 (Figs. 3, A5). All three protocols show comparable (FLYA), or slightly better (MAP-XSII, FLAMEnGO2.0) performance when

using CYPICK- versus manually-prepared peak lists as input (Fig. 3). Using the dimer structure extracted from the T-state of the enzyme, FLYA and MAP-XSII produced comparable levels of confident methyl assignments, 58% and 60%, respectively, with FLYA showing higher assignment accuracy (Fig. 3). With the same input data, FLAMEnGO2.0 generated considerably less confident methyl assignments (31%, 10% of which are erroneous). All three protocols show a significant drop in the fraction of confidently assigned methyls when assignment consistency between T- and R-state dimer structure is considered (Fig. A5). Under such circumstances, 39% of methyls could be confidently assigned by FLYA, followed by 34% using MAP-XSII, and 13% using FLAMEnGO2.0. Despite the drop in assignment coverage, the consideration of methyl assignment consistency over different structural forms is strongly advised, as it reflects favorably on the assignment accuracy (Fig. A5). It will be important to verify these findings in the context of methyl resonance assignment using MAGMA, which is currently the most accurate automatic methyl assignment protocol available [25]. At the present time, MAGMA does not support unassigned NOESY peak lists, but requires manually prepared, arbitrarily assigned, and appropriately filtered NOE contact lists as input [25]. The ambiguous nature of the automatically analyzed 3D NOESY data, reflected in the presence of numerous frequencies that fall within a small range in the ^{13}C (NOE) dimension, presently makes the CYPICK-derived 3D NOESY peak lists incompatible with the MAGMA input requirements. Future developments, including automatic assignment and filtering of CYPICK-derived NOESY peak lists could allow for the combined usage of CYPICK and MAGMA.

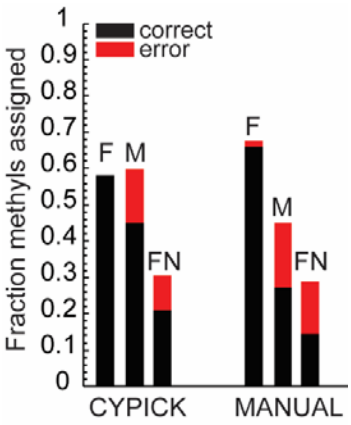

Figure 3. Comparison of automatic methyl resonance assignment protocols using CYPICK- or manually-generated NOESY peak lists. Abbreviations: F, FLYA; M, MAP-XSII; FN, FLAMEnGO2.0.
In the comparison, the input structure was the T-state form of ATCase-r$_2$ (PDB ID: 1TUG [30]).

4. Conclusions

We outlined a protocol for fully automated methyl resonance assignment based on sparse NOESY restraints collected on a uniformly deuterated, selectively methyl-labelled, high molecular weight protein. We combined the recently published contour-based CYPICK algorithm for signal identification with FLYA for automatic methyl resonance assignment. We showed that using CYPICK, high quality peak lists can be obtained for a 3D ^{13}C/^{13}C-separated NOESY spectrum of a 36 kDa homodimer. Methyl resonance assignments obtained subsequently by FLYA are of comparable quality to those obtained using manually prepared peak lists (Figs. 2, 3). Comparison to dedicated automatic methyl assignment approaches that support unassigned NOESY peak lists as input (MAP-XSII and FLAMEnGO2.0) shows that, on this dataset, FLYA could achieve at least as many assignments as the alternatives, with considerably higher accuracy (Figs. 3, A5).

In the future, the outlined protocol must be tested on a larger benchmark of experimental datasets to evaluate its robustness and general applicability. Any automatic methyl assignment protocol that relies on information from NOESY peak lists could benefit from CYPICK (Fig. 1). As such, an important future goal is to combine CYPICK with MAGMA.

5. Acknowledgements

We thank Prof. AJ Baldwin for the ATCase-r_2 data, and TR Alderson for critical reading and feedback on the manuscript.

6. Appendix

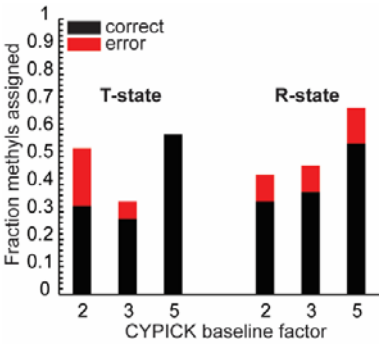

Figure A1. Performance of FLYA on CYPICK peak lists generated with different values of the baseline factor. For the highest baseline factor (5), the least artefacts were picked and the most 'strong' assignments were achieved with FLYA.

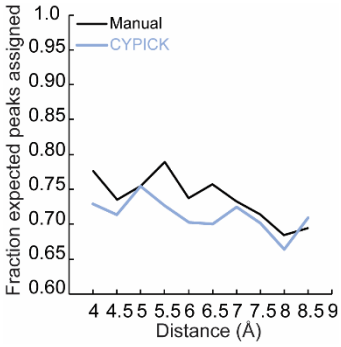

Figure A2. Determination of the distance threshold for the FLYA calculation on ATCase-r_2. The percentage of expected peaks assigned by FLYA at different distance thresholds was used to compute the expected methyl-methyl NOEs. The highest percentage is reached at a 5 Å distance with CYPICK data, or at 5.5 Å with the manually prepared NOESY peak list.

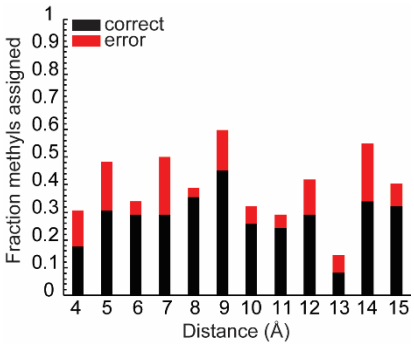

Figure A3. Determination of the optimal distance cut-off for MAP-XSII calculations on ATCase-r_2, as proposed by Xu *et al.* [21,22].

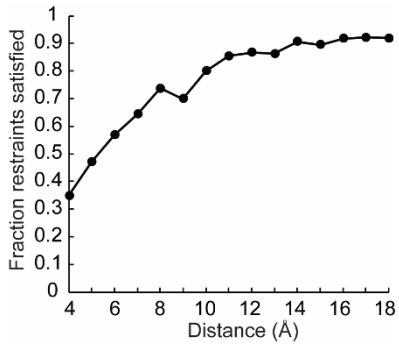

Figure A4. Determination of the optimal distance cut-off for FLAMEnGO2.0 calculations on ATCase-r_2, as proposed by Chao *et al.* [23,24]

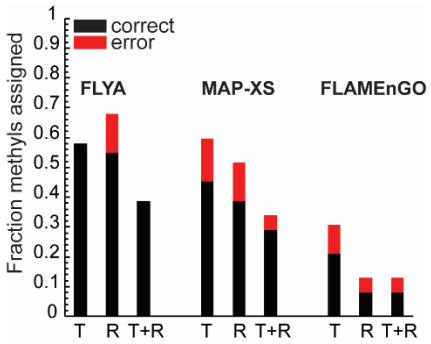

Figure A5. Comparison of automatic methyl assignment protocols on 3D $^{13}C/^{13}C$ separated NOESY data from ATCase-r_2, automatically analysed by CYPICK. The performance was compared using the T-state (PDB ID: 1TUG [30]) and R-state (PDB ID: 1D09 [31]) structures. Assignments consistent with both R- and T-states are evaluated in the T+R column. Correct assignments are indicated in black, erroneous ones in red.

7. References

[1] Wüthrich K 2003 *Angew. Chemie - Int. Ed* **42** 3340–63
[2] Kainosho M and Güntert P 2009 *Q. Rev. Biophys.* **42** 247–300
[3] Bax A 1994 *Curr. Opin. Struct. Biol.* **4** 738–44
[4] Sattler M, Schleucher J and Griesinger C 1999 *Prog. Nucl. Magn. Reson. Spectrosc.* **34** 93–158
[5] Pervushin K, Riek R, Wider G and Wüthrich K 1997 *Proc. Natl. Acad. Sci.* **94** 12366–71
[6] Gardner K H and Kay L E 1998 *Annu. Rev. Biophys. Biomol. Struct.* **27** 357–406
[7] Wiesner S and Sprangers R 2015 *Curr. Opin. Struct. Biol.* **35** 60–7
[8] Proudfoot A, Frank A O, Ruggiu F, Mamo M and Lingel A 2016 *J. Biomol. NMR* **65** 15–27
[9] Clark L, Zahm J A, Ali R, Kukula M, Bian L, Patrie S M, Gardner K H, Rosen M K and Rosenbaum D M 2015 *J. Biomol. NMR* **62** 239–45
[10] Janin J, Miller S and Chothia C 1988 *J. Mol. Biol.* **204** 155–64
[11] Ollerenshaw J E, Tugarinov V and Kay L E 2003 *Magn. Reson. Chem* **41** 843–52
[12] Tugarinov V, Hwang P M, Ollerenshaw J E and Kay L E 2003 *J. Am. Chem. Soc.* **125** 10420–8
[13] Mainz A, Religa T L, Sprangers R, Linser R, Kay L E and Reif B 2013 *Angew. Chemie - Int. Ed.* **52** 8746–51
[14] Sprangers R and Kay L E 2007 *Nature* **445** 618–22
[15] Sprangers R, Velyvis A and Kay L E 2007 *Nat. Methods* **4** 697–703
[16] Baldwin A J, Religa T L, Hansen D F, Bouvignies G and Kay L E 2010 *J. Am. Chem. Soc.* **132** 10992–5
[17] Kato H, van Ingen H, Zhou B-R, Feng H, Bustin M, Kay L E and Bai Y 2011 *Proc. Natl. Acad. Sci. U. S. A.* **108** 12283–8
[18] Velyvis A, Schachman H K and Kay L E 2009 *J. Am. Chem. Soc.* **131** 16534–43
[19] Rosenzweig R and Kay L E 2014 *Annu. Rev. Biochem.* **83** 291–315
[20] Venditti V, Fawzi N L and Clore G M 2011 *J. Biomol. NMR* **51** 319–28
[21] Xu Y, Liu M, Simpson P J, Isaacson R, Cota E, Marchant J, Yang D, Zhang X, Freemont P and Matthews S 2009 *J. Am. Chem. Soc.* **131** 9480–1
[22] Xu Y and Matthews S 2013 *J. Biomol. NMR* **55** 179–87
[23] Chao F-A, Shi L, Masterson L R and Veglia G 2012 *J. Magn. Reson.* **214** 103–10
[24] Chao F A, Kim J, Xia Y, Milligan M, Rowe N and Veglia G 2014 *J. Magn. Reson.* **245** 17–23
[25] Pritišanac I, Degiacomi M T, Alderson T R, Carneiro M G, Ab E, Siegal G and Baldwin A J 2017 *J. Am. Chem. Soc.* **139**
[26] Sounier R, Blanchard L, Wu Z and Boisbouvier J 2007 *J. Am. Chem. Soc.* **129** 472–3
[27] Würz J M and Güntert P 2017 *J. Biomol. NMR* **67** 63–76
[28] Schmidt E and Güntert P 2012 *J. Am. Chem. Soc.* **134** 12817–29
[29] Schmidt E and Güntert P 2013 *J. Biomol. NMR* **57** 193–204
[30] Stieglitz K, Stec B, Baker D P and Kantrowitz E R 2004 *J. Mol. Biol.* **341** 853–68
[31] Jin L, Stec B, Lipscomb W N and Kantrowitz E R 1999 *Proteins Struct. Funct. Genet.* **37** 729–42
[32] Arthur D and Vassilvitskii S 2007 *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* pp 1027–1025
[33] Pedregosa F and Varoquaux G 2011 *J Mach Learn Res* **12** 2825-2830
[34] Lichtenecker R J, Weinhäupl K, Reuther L, Schörghuber J, Schmid W and Konrat R 2013 *J. Biomol. NMR* **57** 205–9
[35] Lichtenecker R J, Coudevylle N, Konrat R and Schmid W 2013 *ChemBioChem* **14** 818–21
[36] Maksić Z B and Randić M 1970 *J. Am. Chem. Soc.* **92** 424–5