Robust Automated Backbone Triple Resonance NMR Assignments of Proteins Using Bayesian-Based Simulated Annealing

Anthony Bishop
Texas A&M University https://orcid.org/0000-0003-4853-1073
Sravya Kotaru
Texas A&M University
Glorise Torres-Montalvo
Texas A&M University
Kyle Mimun
Texas A&M University
A. Wand (wand@tamu.edu)
Texas A&M University https://orcid.org/0000-0001-8341-0782

Article

Keywords:

Posted Date: May 3rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1554405/v1

Robust Automated Backbone Triple Resonance NMR Assignments of Proteins Using Bayesian-Based Simulated Annealing

Anthony C. Bishop[1], Sravya Kotaru[1], Glorisé Torres-Montalvo[1], Kyle Mimun[1] and A. Joshua Wand[1,2,3*]

Departments of Biochemistry & Biophysics,[1] Chemistry[2] and Molecular & Cellular Medicine[3], Texas A&M University, College Station, Texas USA 77843

Submitted to Nature Communications for consideration as an Article

Contact Information:

Professor A. Joshua Wand
Department of Biochemistry & Biophysics
MS 2128
300 Olsen Blvd
College Station, TX 77843
E-mail: josh.wand@ag.tamu.edu
Telephone: 1-979-845-2544

Abstract

The comprehensive assignment of individual resonances of the nuclear magnetic resonance spectrum of a protein to specific atoms remains a labor-intensive and often debilitating task - especially for proteins larger than 30 kDa. Recently, there have been tremendous advances in our empirical knowledge of the relationship between the structural context of a nuclear spin and its observed resonance frequency. Indeed, the expansion in the database of determined high-resolution protein structures and recent advances in structure prediction provide an enormous resource in this respect. Robust automation of the resonance assignment process nevertheless often remains a bottleneck in the exploitation of solution NMR spectroscopy for the study of protein structure-dynamics-function relationships. Here we present a new approach for the assignment of backbone triple resonance spectra of proteins. A Bayesian statistical analysis of predicted and observed chemical shifts is used to provide a pseudo-energy potential to drive the search for the most optimal set of resonance assignments. This approach has been implemented in the C++ program Bayllagio and tested against protein systems ranging in size to over 450 amino acids. Bayllagio makes almost no errors, accommodates incomplete information, is sufficiently fast to allow for real-time evaluation of data acquisition, and greatly outperforms currently employed deterministic algorithms.

Introduction

Comprehensive mapping of individual resonances comprising nuclear magnetic resonance (NMR) spectra to specific atoms is a general prerequisite for the successful analysis of the structure and dynamics of protein molecules by NMR spectroscopy. Early applications of multi-dimensional homonuclear ^{1}H NMR data to the "resonance assignment problem" presented by proteins relied heavily on human intervention. The first comprehensive approach was the sequential assignment method[1], which centered on identification of J-coupled spin systems[2] that are then assembled through connections provided by short distances revealed by the nuclear Overhauser effect (NOE) interactions between sequential residues using the identity of side chains to error-check against the primary structure[2,3]. The subsequent main chain directed (MCD) assignment strategy[4,5] formalized cyclic patterns of backbone ^{1}H-^{1}H NOE interactions and provided a more robust algorithmic framework that relieved somewhat the complexity of identifying side chain resonances[6,7]. While the MCD approach did lead to the first fully automated assignment of ^{1}H resonances to backbone hydrogens[7], automation of ^{1}H-based resonance assignments was generally frustrated by the overwhelming spectral degeneracy of multidimensional ^{1}H spectra of proteins and the interference of technical attributes such as a prominent diagonal. The introduction of heteronuclear triple resonance spectroscopy[8–13] completely changed the landscape of the resonance assignment task by providing much greater resolution, generally higher quality data, and, most importantly, definitive rules with very precise meanings for making "connectivities" between backbone resonances. In aggregate, the triple resonance connectivity rules provided a foundation for the creation of automated assignment strategies that can loosely be described as "deterministic." Triple resonance assignments of the protein backbone opens access, either directly or by tethering to side chain resonance assignments, to a wide range of dynamic phenomena[13,14] and structural information[15–17].

Automation of deterministic triple resonance algorithms have led to effectively complete backbone resonance assignments of smaller proteins with little human intervention and greatly aided the assignment of larger systems[18–20]. Yet, even with the advent of "transverse relaxation optimized spectroscopy" (TROSY)[21], the comprehensive assignment of systems larger than 30 kDa remains remarkably rare. While there are many potential contributors to this apparent ceiling, it is certainly clear that increasing ambiguity in connectivities due to degeneracy, loss of

resonances due to relaxation or artifact, and other confounding spectral attributes are simply not sufficiently accommodated by current deterministic automated assignment strategies. Here, we strive to avoid the inherent rigidity of such algorithms by appealing to the statistics of Bayes, to most effectively utilize available data via the calculation of explicit probabilities. This formalism also allows a more flexible and adaptable incorporation of chemical shift prediction and structural knowledge into the assignment process. By implementing the Bayesian analysis within a simulated annealing engine, we develop a robust and efficient search for optimal solutions. Protein assignment algorithms utilizing simulated annealing have been developed in the past[22]. However, the stochastic algorithm as described here is the first to fully utilize readily available pre-existing structural models, both experimentally determined and predicted, and in doing so exploit the rich information contained within structure-based predicted chemical shifts. We demonstrate how these invaluable restraints greatly aid the resonance assignment process, especially in cases where data may be otherwise sparse. We also compare the overall performance of this novel algorithm against two highly cited assignment algorithms on a variety of experimental datasets.

Results and Discussion

Bayllagio

We designed an algorithm, termed "Bayllagio," which utilizes a simulated annealing approach[23] to efficiently search the immense solution space for the most optimal set of assignments. The objective is to find the correct sequence specific mapping of observed spin systems in J-correlated triple resonance experiments to their originating amino acids. A spin system is defined as a backbone amide group and its associated spins (resonances) that are accessed through magnetization transfer via scalar coupling[11]. Here, a spin system is comprised of the amide N(i) and H(i), the Cα(i), Cβ(i), C'(i) and the Cα(i-1), Cβ(i-1) and C'(i-1) resonances, where i refers to any given residue position and i-1 refers to the amino acid position immediately before position i in the sequence. The simulated annealing search engine begins by randomly mapping the set of defined spin systems to specific residue positions. If there are more spin systems provided than residue positions, then the excess spin systems are placed in a cache for later use as described below. From this initial system state, a spin system is randomly chosen to either be moved to an

unoccupied residue position, to be swapped with another spin system, or added to the cache. The energy of the system state before and after this proposed swap is calculated using some effective temperature T, and the Metropolis criterion[24] is then applied (Equation 1).

$$P_{accept} = \min(1, e^{\frac{-\Delta E}{T}}) \quad (1)$$

P_{accept} is the probability of accepting the swap and ΔE is the change in energy due to the proposed swap. If the proposed swap decreases the system energy then $P_{accept} = 1$. However, if the proposed swap increases the energy then $0 < P_{accept} < 1$, with P_{accept} increasing with higher temperatures. After calculating P_{accept}, a uniformly distributed random number *r* between 0 and 1 (inclusive), is generated. If $r \leq P_{accept}$, then the swap is accepted. Otherwise, the swap is rejected and the system state is left unchanged. These random swap attempts are continued until the average energy of the sampled states does not vary significantly. At this point, T is decreased following a highly optimized schedule based on calculated system "specific heat" (see Methods). The system is further cooled and equilibrated in this manner until a set of termination criteria are achieved and the annealing protocol is ended. Finally, to ensure that the system has reached a true minimum in energy, each spin system is attempted to be swapped with every other spin system, with only decreasing energy changes being accepted. The entire procedure, starting from initialization and ending with minimization, is repeated many times and the results of each run are used to produce a consensus assignment set. The consensus assignment set is further curated to produce the final assignment set. The Bayllagio algorithm is outlined in Figure 1.

A carefully chosen energy function is necessary for the algorithm to efficiently find the correct solution. The total energy is calculated as a simple sum of the energies for the individual spin systems. Each spin system energy is composed of two terms: the "adjacency" energy and the "chemical shift" energy. The adjacency energy describes the interaction between two spin systems mapped to adjacent locations on the amino acid sequence. This energy is minimized if the Cα(i), Cβ(i), and C'(i) shifts of the spin system match the Cα(i-1), Cβ(i-1), and C'(i-1) of the spin system at the following residue in the primary sequence. Thus, spin systems are defined relative to the i^{th} residue's amide N and amide H in a completely analogous fashion to a manual assignment. In contrast, the chemical shift energy describes the interaction between a spin

system and its current residue position i.e., it is defined by the local sequence and structure. This energy is minimized when the resonances of the spin system closely match the predicted values of the current residue position, while also failing to match the predicted values at all other residue positions. The predicted chemical shift values for calculating the chemical shift energy can come from structure-based chemical shift predictions. Here we use SHIFTX+[25] if a structure is available or, if structure-based predictions are absent, predictions are generated from the amino acid dependent chemical shift statistics provided by the BMRB database[26]. A novel application of Bayes' theorem then provides a posterior probability of assigning each spin system at each location in the sequence that is based on the predicted and experimental shifts. Using this probability, the chemical shift energy is calculated (see Methods for a more detailed description).

Bayllagio is accurate, robust and fast

We tested Bayllagio against a test set of six different protein systems ranging in size: human interleukin-1 receptor antagonist C66A,C122A (IL-1Ra, 17.1 kDa), interleukin-1β (IL-1β, 17.5 kDa), *S. solfataricus* indole-3-glycerol phosphate synthase R43S (IGPS, 28.4 kDa), *E. coli* maltose binding protein (MBP, 40.8 kDa), *E. coli* yersiniabactin non-ribosomal peptide synthetase (Cy1, 51.9 kDa), and *E. coli* thymidylate synthase (ecTS, 61.0 kDa homodimer). Experimental triple resonance data was used to assemble spin systems (Table 1). Crosspeak positions used for spin system assembly were pulled from the canonical triple resonance spectra used for protein assignment (i.e., HSQC, HNCO[27], HN(CA)CO[28], HNCA[29], HN(CO)CA[29], HNCACB[30], HN(CO)CACB/CBCA(CO)NH[31]), with different implementations used depending on the system size and the labeling schemes employed (Supplementary Table 1). All but two of the assignment data sets were obtained in our laboratory. Crosspeak lists for Cy1 and ecTS were kindly provided by Professors Dominque Frueh (Johns Hopkins University) and Andrew Lee (University of North Carolina at Chapel Hill), respectively.

The results from Bayllagio were compared to reference assignments to assess program performance. Reference assignments were obtained from either the BMRB, directly from another lab, or manually determined in our lab (Table 1). Deposited assignments were manually mapped to the acquired spectra for comparison. Small movement in peak positions between the deposited

assignments and the acquired spectra were permitted to account for differences in experimental conditions. For the most part, reference assignments were considered "correct," though in a few cases Bayllagio did identify clear errors. For each residue position, Bayllagio either outputs the spin system that was assigned to that residue position or marks it as unassigned. The assignment given to each residue in the protein sequence by Bayllagio was determined to either be "matching," "missing," or "mismatching" its counterpart in the reference assignments. A residue was considered to have a matching assignment if the spin system assigned to it by the algorithm was the same as the reference. A residue was also considered to "match" the reference if it was unassigned both by Bayllagio and in the reference assignments. A residue was designated "missing" if a spin system was assigned to that location in the reference assignments, but Bayllagio designated it as unassigned. Lastly, a residue was labelled as mismatching if Bayllagio assigned a spin system, and it did not match the spin system in the reference assignments or if the residue was unassigned in the reference assignments.

For the sake of comparison, two highly-cited algorithms were executed using the same spin system datasets: MARS[32] and I-PINE[33]. Here, it is noted that I-PINE explicitly uses the BMRB database depositions in the form of a PACSY[34] query to generate its assignments and therefore systems deposited in the BMRB (i.e. IL-1β, MBP, ecTS, and CY1) cannot be used as true tests of *de novo* assignment. Bayllagio achieved >95% match against the deposited assignments in all cases and generally outperformed MARS and I-PINE (Figure 2, Supplementary Table 3). Bayllagio made more matching assignments than MARS - especially in the higher molecular weight systems. Importantly, Bayllagio made few mismatching assignments (< 2%), while I-PINE had up to 12% mismatches meaning that about 1 in 8 assignments made were incorrect. In summary, Bayllagio consistently outperformed both MARS and I-PINE particularly in situations where the non-ideality of the assignment data set hobbled that latter algorithms (see below).

Bayllagio also produced results quickly by producing a curated set of assignments from 100 annealing runs within 8 min for each system tested (see Supplementary Table 2). These run times were generally much faster than those of MARS and I-PINE. However, a direct comparison of the algorithms' speed is difficult due to MARS and I-PINE being run on public hardware, as performance could be influenced by differences in hardware, as well as extraneous machine load.

Nevertheless, with its high speed and accuracy, the advantages of Bayllagio become even more apparent when considering large proteins with suboptimal data sets.

The performance of Bayllagio with suboptimal data sets

The rather complete crosspeak lists from an extensive set of triple resonance experiments for each test protein provides valuable benchmarks for the validation of Bayllagio, but are arguably not fully illustrative of the difficult protein systems challenging modern NMR. To examine the performance of Bayllagio in cases of missing data and assess the most impactful triple resonance information, spin system data from the MBP data set were randomly discarded to generate simulated data sets emulating data collection on challenging protein systems. In this way, the relative importance of completeness in spin system definitions could be probed. In addition, a key question was to learn the extent to which structure-based chemical shifts, as opposed to general BMRB residue statistics, are able to rescue the assignment and aid the assignment process.

Figure 3 summarizes the robustness of Bayllagio when analyzing conditions of missing spin systems and spin systems of varying completeness. When structure-based predicted shifts are absent and BMRB resonance distributions are used to generate predicted shifts, Bayllagio produced matching assignments of 100%, 52% and 19% when employing simulated sets of spin systems randomly depleted 100%, 80% and 60% of the total spin systems, respectively (Figs. 3a & 3b). However, incorporation of structure-based predicted shifts allowed for nearly complete assignments of the 60% and 80% complete data set scenarios. Thus, Bayllagio can reliably assign proteins when a significant number of spin systems are absent. In addition, the data demonstrate that a Bayesian analysis of predicted chemical shifts from a protein structure serves as a powerful assignment restraint, when compared to predicted shifts from general BMRB resonance distributions. Furthermore, and most importantly, Bayllagio produced almost no mismatches. This indicates that the user can be highly confident in the assignments made, even when an extensive set of triple resonance experiments yields an incomplete set of spin systems.

To further understand the effect of missing different spin system components, the experimentally derived MBP crosspeak lists were randomly reduced to produce a multitude of distinct data sets representing a wide range of data completeness and then used as input to Bayllagio.

(Supplementary Table 4). An example of the algorithm performance is shown in Figure 3. Here data sets where only 88% and 50% of the Cα and Cβ resonances, respectively, were included in the spin system definitions with either none or 75% of the C' resonances. While reliance on BMRB database for predicted shifts (as opposed to structure-based shifts) yielded poor performance, inclusion of SHIFTX+ predictions to the algorithm input *entirely* rescues the assignment (Fig. 3e). If C' triple resonance crosspeak information is provided instead of predicted shifts, the improvement in the assignment outcome is marginal. The marginal improvement is likely because the uncertainty dominating this analysis lies in poor residue type prediction due to the missing Cβ resonances - a common issue in the assignment of spectroscopically challenging proteins. This deficiency is not mitigated by the inclusion C' data, which carries little residue type information, but is readily ameliorated by structure-based predicted shifts which are specific to not only residue type, but also local structure.

Bayllagio's performance in settings where there is a significant amount of missing data was compared against MARS and I-PINE. An experimental MBP crosspeak list with 88%, 50%, and 75% of the Cα, Cβ and C' resonances retained, respectively, was used as input for the three algorithms (Figure 3F). Bayllagio achieved nearly all matching assignments (>99%) while MARS missed almost all assignments, and PINE had 31% of assignments mismatching. Bayllagio's success clearly indicates that the availability of the structure-based chemical shift predictions can serve as a powerful constraint in protein assignment - large enough to potentially surpass the information provided by the C' experimental pair under some circumstances. Additionally, these results demonstrate that Bayllagio has excellent outcomes in circumstances where there is a large quantity of missing data – greatly outperforming existing algorithms.

A companion question to incompleteness of the definition of spin systems is: what minimum set of triple resonance experiments yield correct and maximally complete assignments? Here we compared the performance of individual pairs of triple resonance spectra (e.g., HNCO, HN(CA)CO, etc.). Without structure-predicted chemical shifts, analysis by Bayllagio of single assignment pairs (i.e., Cα, Cβ or C') fails to adequately assign the proteins (Fig. 3c). However, when provided SHIFTX+ predicted chemical shifts, Bayllagio can achieve >98% of the assignments made by the full data set, utilizing only the Cα data from the HNCA/HN(CO)CA pair alone (Fig. 3d). A high percentage of assignments can also be made by utilizing only the Cβ

data. Improvement in the percentage of matching assignments when utilizing SHIFTX+ predicted shifts is also significant in the Cβ plus C' data sets only case, but to a substantially lesser degree in the C' only case due to the high degeneracy of that resonance type. In all circumstances the mismatch rate for Bayllagio remains close to zero.

To further test Bayllagio's capabilities, we wanted to examine the minimum necessary data set for the proper assignment of CY1. The goal was to see which triple resonance pair experiments were truly necessary and what resonance information could have been skipped if an iterative assignment procedure had been adopted. Given its size, CY1 is the exemplar of the challenges facing modern protein NMR. Generation of the largely complete backbone resonance set used in this study required a battery of experimental spectra and labelling schemes[35]. Being able to assign backbone resonances with fewer data, specifically without acquiring the insensitive CB resonance pair would reduce the need for additional samples and/or spectrometer time. Furthermore, prior to the study from which this assignment data was obtained, no structure of Cy1 had been solved, with its closest homolog in the PDB having only 38% identity. Previously, generation of accurate chemical shift predictions without a high-resolution structure would be infeasible. However, recent advancements in the prediction of highly accurate structural models from a protein's sequence potentially provides an alternate path. Here we explored the utility of using a structure model predicted by the AlphaFold2 algorithm[36] to predict chemical shifts using ShiftX+ and thus examine the feasibility of assigning large proteins with unknown structure. Bayllagio was run using data from only the indicated triple resonance pairs with either predicted shifts derived from BMRB resonance distributions or from SHIFTX+ predictions based on an AlphaFold2 model (Figure 4). The use of structure-based predicted chemical shifts derived from the AlphaFold2 predicted structure clearly served as a crucial restraint – greatly improving performance. Additionally, assignments made with the Cα triple resonance experiment pair alone were 79% matching to the reference assignment. Inclusion of the C' triple resonance experiment pair raised this to 94% with only a 1% mismatch rate. These data show that CY1 could be largely assigned entirely without the least sensitive Cβ resonance pair experiments. These assignments could likely be extended by collecting the Cβ resonance pair at low signal to noise without the commitment of significant spectrometer time. To examine this point further, simulated CY1 data sets that retain only

fractions of the Cβ resonances were run through Bayllagio (Supplementary Table 5). With just 50% of the Cβ resonances present, the results had ~98% of residues matching the reference assignments.

To more fully assess the effectiveness of AlphaFold2 structures in the assignment process further, the remaining test proteins were re-assigned using predicted shifts derived from AlphaFold2 structures (Supplementary Table 6). The AlphaFold2 derived shifts performed comparably to those of the experimentally determined structures (Supplementary Table 3) - in some cases outperforming them. Taken together these data suggest that the lack of an experimental structure is unlikely to hinder the full capability of the Bayllagio algorithm.

The above results indicate that Bayllagio is clearly positioned to inform on the "real time" data acquisition side of the resonance assignment process. The speed of the algorithm and its robustness against making mistakes in assignments suggests that Bayllagio could offer the ability to guide data acquisition itself, in conjunction with reliable automated peak picking. Iterative examination by Bayllagio of sequentially acquired triple resonance spectra could, in principle, allow the user to determine if a satisfactory level of assignment can be achieved without further data acquisition and thereby save valuable spectrometer time.

Impact of selective (un)labelling data.

Selective (un)labelling is often employed during the manual assignment process to narrow down the possible residue type(s) of particular spin systems. This generally involves adding isotopically labelled amino acids to unlabeled growth media (selective labelling)[37] or unlabeled amino acids to labelled growth media (selective unlabelling)[38–40] and observe the appearance and disappearance of NMR cross peaks, respectively. This allows the simultaneous identification of all spin systems of a particular amino acid type or types. In addition to "i" residue type information, selective unlabelling, if performed in ^{13}C, ^{15}N labelled media, can be used to identify the so called "i-1" residue data - the residue type of the amino acid that precedes that of the current spin system. To utilize these increasingly common residue type data, Bayllagio allows the user to designate any arbitrary subset of the 20 standard residues as the set of possible residue types for each spin system. In this manner, both the "i" and "i-1" residue type data can be

specified for each spin system independently. Bayllagio uses these data to restrict the possible random swaps as well as in the calculation of the energy function (see Methods).

To demonstrate this capability, a MBP data set with randomly discarded data was generated such that 80% of the spin systems, 75% of the CA data, 50% of the CB data, and 60% of the CO data were retained. The residue types for K, R, N, or Q spin systems were provided as input and the "i-1" residue types were provided to spin systems that are placed immediately after a K, R, N or Q residue. This simulates a scenario in which selective unlabelling data was collected for K, R, N, and Q independently. These residue types are frequently unlabeled due to their favorable metabolic properties[39]. Figure 5 compares the Bayllagio run with the same data set run through MARS and I-PINE, both of which are capable of utilizing "i" data, but not "i-1". Therefore the "i-1" data was omitted for MARS and I-PINE runs. These results show that the selective unlabelling data had a modest effect on improving the performance of Bayllagio as well as the performance of MARS which produced few assignments overall. The unlabelling data had no effect on the performance of I-PINE, which produced unreliable assignments. The small effect of the unlabelling data on the performance of Bayllagio may be because selective unlabelling information does not provide substantially more information orthogonal to that provided by the Bayesian analysis of chemical shift information.

In summary, we have demonstrated that Bayesian-based simulated annealing combining sequential relationships derived from triple resonance spectra and chemical shift information predicted from a high-resolution structural model can greatly facilitate the triple-resonance backbone assignment of proteins. The implementation of this strategy in Bayllagio is robust to incompleteness of spin system definition (sparseness) and overall complexity of the resonance assignment challenge (protein size). Importantly, Bayllagio is relatively conservative and makes few errors. An optimized annealing strategy utilizing a "specific heat" approach to guide temperature cooling results in a very rapid analysis. The speed of analysis combined with its aforementioned robustness presents the possibility of real-time intervention during primary data collection. This becomes increasingly feasible with the utilization of automated peak picking and spin system construction routines. The Bayllagio algorithm promises the ability to easily and robustly assign previously intractable systems and simplify this otherwise challenging task. The combination of fast and robust backbone resonance assignments, made possible here, and

structure-based methyl resonance assignments[41] will reduce the resonance assignment barrier considerably and allow the full power of NMR spectroscopy to be applied in a facile manner to otherwise challenging proteins.

Methods

Spin system definition

The spin systems of proteins assigned using spectra acquired by our lab were constructed by manually peak picking in acquired spectra (i.e., not reconstructed from deposited assignments) (see Table 1). Expression and purification protocols of these proteins and NMR data acquisition parameters are described in the Supplementary Methods. Acquired spectra were processed using the NMRPipe[42] installed on NMRbox[43] and crosspeaks were picked using NMRFAM-SPARKY[44]. Crosspeak lists for ecTS and Cy1 were provided by Professor Andrew Lee (University of North Carolina, Chapel Hill) and Professor Dominique Frueh (Johns Hopkins University), respectively, and were used without further adjustment. In a standard amide ^{1}H detected triple resonance experiment, the (^{1}H,^{15}N) frequency pair is considered the resolving parent to which are attached other nuclei with bonding relationships defined by the particular experiment. Generally, a family of (^{1}H,^{15}N)-resolved triple resonance are used to attach carbons of various types (C', Cα, Cβ) with defined bonding relationships to the amino acid residue harboring the (^{1}H, ^{15}N) pair (i.e., the same or the preceding residue). In all cases described below, we begin with such a set of spin systems – unassigned as to amino acid type, position relative to any other particular spin system, or absolute position in the amino acid sequence (except where otherwise indicated).

Simulated annealing algorithm and refinement

The assignment run is initialized by randomly assigning the spin systems to the protein sequence. If there are more spin systems than sequence positions (e.g., spin systems correlated to a side-chain amide group and not the backbone are also present in the data set) the extra spin systems are placed in the "cache" and may be assigned to the sequence over the course of the run. In the case of multiple conformers of a given spin system, the user can indicate to the algorithm that two apparent spin systems are conformers of the same spin system. The algorithm will then consider both conformers when calculating the energy function at each swap, choosing the conformer that yields a lower energy. Simulation temperature is initialized at 1000 units and a spin system is chosen at random to swap with another random spin system or empty site within in the sequence. The change in energy of the system due to the swap is calculated using the

energy function described below. The swap is then accepted or rejected at a frequency corresponding to a probability generated by applying the Metropolis criterion (Equation 1). Spin systems in the cache can swap with empty sites and spins systems in the sequence. In addition, spin systems mapped to the amino acid sequence can be added to the cache leaving behind an empty site, permitting the size of the cache to change freely over the course of the annealing run. After each successful swap, the energy of the state is recorded and stored as a part of a sample of energy values. Once the sample reaches a particular size, the sample mean and standard error are calculated and an additional sample of equal size is generated by continued swapping. A Student's two tailed t-test is performed to compare the sample means of the two samples. The system is considered to have "equilibrated" at the current temperature if the p-value of the t-test is greater than a user supplied value (usually $p > 0.5$). If equilibration has not been reached, more swaps are performed to generate an additional sample and the t-test is repeated with the two most recent samples. If equilibration has been reached, then the energy values of the last sample are used to estimate the specific heat at the current temperature:

$$\frac{d<E>}{dT} = \frac{<E^2> - <E>^2}{T^2} \tag{2}$$

Where T is the current temperature, E is the energy of the system and the angled brackets indicate the sample mean. The specific heat is then used to calculate the next temperature:

$$T_{next} = T + \frac{\Delta<E>_{target}}{d<E>/dT} \tag{3}$$

Where $\Delta < E >_{target}$ is a user controlled parameter and is kept at -2000 for this study. If $T - T_{next}$ is greater than 10, then the temperature decrease is limited to 10 units to prevent overcooling the system. After decreasing the temperature, the annealing run will terminate if any of the following criteria are met: the temperature is less than 1, the product of the temperature and the last specific heat calculated is less than 200, or the ratio of unsuccessful swaps to successful swaps while collecting the last sample is greater than 10,000. If termination is not achieved a new sample size is defined using the following equation:

$$S = N \frac{\left(\frac{d<E>}{dT}\right)^{1.5}}{300} \tag{4}$$

Samples are then drawn at the new temperature to determine equilibration and the cycle is continued. Upon termination of the annealing protocol, a steepest-descent type search is performed to locally minimize the system energy and refine the assignments, discarding potentially bad assignments that were left over from the run. This is done by attempting to place (or swap) every spin system at every possible location in the sequence (including the cache). Only spin system swaps/placements that decrease the system energy are accepted.

To assign a protein, simulated annealing is independently repeated with 100 different random starting conditions. A consensus set of assignments is generated by calculating the frequency with which each spin system is placed at each amino acid location. A curated set of assignments is generated from this consensus analysis. The curation procedure is as follows: the spin system that was assigned to a given residue in a majority of runs was kept as the tentative assignment for that particular residue. Residues that did not have the same spin system assigned to it at least 50% of the time were marked as unassigned. Tentatively assigned spin systems are then evaluated by the probabilities of the assignment (posterior and prior) as well as the number of connectivities defined as a matching resonance between adjacent spin systems. Assignments were accepted if they met any of the following criteria: 1) the "assigned" spin system has at least two connectivities with adjacent spin systems, 2) spin system has at least 1 connectivity with adjacent spin systems and a posterior probability at least three times higher than the prior probability (1/N), or 3) spin system has no connectivities, but has a posterior probability > 50%. Residues with tentative assignments that did not satisfy any of these criteria were then marked as "unassigned."

Energy Function

The total energy (E_{tot}) equals the sum of all the energies of the constituent spin systems (Eqn 5). At any given step during the annealing protocol, spin systems are either tentatively assigned to a position in the sequence or unassigned and placed in the "cache." Cached spin systems are defined as having zero energy (i.e., $E_m = 0$).

$$E_{tot} = \sum_{m=0}^{M} E_m \quad (5)$$

The energy of each spin system tentatively assigned to a specific place in the amino acid sequence is comprised of the adjacency energy (E_m^{adj}) and the chemical shift energy (E_m^{cs}):

$$E_m = E_m^{adj} + E_m^{cs} \quad (6)$$

The adjacency energy is related to the degree of correspondence between the connecting Cα, Cβ and C' resonances of the current spin system and Cα(i-1), Cβ(i-1) and C' (i-1) resonances the spin system assigned to the subsequent position in the sequence. E_m^{adj} therefore, captures the process of evaluating spin system adjacency and is based on the number of potential connectivities between adjacent spin systems tentatively assigned to the sequence. For example, if spin system m is assigned to residue position (i) and spin system l at residue position i+1, the adjacency energy is given by:

$$E^{adj} = \sum_k c_0 e^{-0.5\left(\frac{\delta^m_{k(i)} - \delta^l_{k(i-1)}}{\sigma_k}\right)} + c_1 \quad (7)$$

Where $\delta^m_{k(i)}$ is the chemical shift of resonance k(i) (either Cα(i), Cβ(i) or C'(i)) of spin system m and $\delta^l_{k(i-1)}$ is the chemical shift of resonance k(i-1) (either Cα(i-1), Cβ(i-1) or C'(i-1)) of spin system l. σ_k refers to the estimated precision of the measured chemical shifts. The E^{adj} is set to the form of an inverted Gaussian when $c_0 < 0$. Previous assignment algorithms have used functions of this form to good effect for estimating adjacency[22]. In the limit of well-matched connectivities, the sum of inverted Gaussian functions will have a minimum value of Kc_0 where K is the number of connectivities whereas, for poorly matched putative connectivities, the adjacent energy will tend to a limit of Kc_1. Importantly, when an expected element of spin system m or l is missing, that contribution to the adjacency energy is set to zero. Similarly, if the subsequent position in the primary sequence is not currently assigned a spin system, then $E^{adj} = 0$. In this report, c_0 and c_1 were set to -100 and +50, respectively. The value σ is influenced by the properties of the NMR spectra from which the spin systems are defined. In this study, σ was typically chosen so that the function has an abscissa-intercept at a chemical shift difference of 0.2 ppm.

The second term of the spin system energy, E_m^{cs}, evaluates the degree of correspondence of the observed chemical shifts to those predicted. It is this term that makes use of the ability of

Bayesian statistics to incorporate diverse degrees of knowledge of the local structure of the protein. These include relatively structureless information encoded in the simple empirical distributions of chemical shifts of the amino acids observed in proteins or specific chemical shift predictions based on the high-resolution structure of the protein being examined. For the former, we utilize the BMRB[26] database. For the latter, we use crystallographic structures available in the PDB[45] or structures predicted by AlphaFold2[36]. E_m^{cs} is ultimately calculated from the Bayesian posterior probability of a proposed assignment given the observed chemical shifts:

$$P(A_{n,m}|B_m) = \frac{P(B_m|A_{n,m})P(A_{n,m})}{P(B_m)} \quad (8)$$

The subscript *n* and *m* index over the non-proline locations in the sequence that can be assigned an (^{1}H,^{15}N)-based spin system and the provided spin systems, respectively. The condition $A_{n,m}$ refers to where spin system *m* is correctly assigned to sequence position *n*. The condition B_m refers to the observed chemical shifts of spin system *m*. The prior probability $P(A_{n,m})$ refers to the initial probability of the assignment of spin system *m* to residue *n* being correct. $P(A_{n,m})$ is then taken as:

$$P(A_{n,m}) = \frac{1}{N} \quad (9)$$

Where *N* is the number of non-proline residues in the sequence. If residue type data is provided for the spin system, this equation changes to:

$$P(A_{n,m}) = \begin{cases} \frac{1}{N_{pos}}\ ;\ if\ n\ is\ included\ in\ the\ constraints \\ 0;\ if\ n\ is\ not\ included\ in\ the\ constraints \end{cases} \quad (10)$$

where N_{pos} is the number of sequence locations where the spin system can be placed, given the provided residue type constraints.

The likelihood of assignment $A_{n,m}$ (i.e., the probability of observing the chemical shifts of spin system *m* given the assignment $A_{n,m}$) is given by Eqs. 11 & 12:

$$P(B_m|A_{n,m}) = 1 - CDF(X^2_{n,m}) \quad (11)$$

$$X^2_{n,m} = \sum_{r=1}^{R} \left(\frac{\delta^m_{obs,r} - \delta^n_{pred,r}}{\sigma^n_r} \right)^2 \quad (12)$$

Where $\delta^n_{pred,r}$ is the predicted chemical shift of nucleus *r* at sequence position *n*; $\delta^m_{obs,r}$ is the observed chemical shift of nucleus *r* of spin system *m* and σ^n_r is the standard error for the chemical shift prediction of nucleus *r* at sequence position *n*. The nuclei, represented by variable r, are the following {H, N, Cα, Cβ, C', Cα(i-1), Cβ(i-1), C'(i-1)}. Assuming that the random variable $\delta^n_{pred,r}$ is normally distributed about $\delta^m_{obs,r}$ with standard deviation σ^n_r and that the error in the chemical shift measurement is much less than the error in the prediction, then the random variable $X^2_{n,m}$ would be chi-square distributed with *R* degrees of freedom, where *R* is equal to the number of nuclei for which both predictions and data are provided. The likelihood is then calculated as the value of the complementary cumulative distribution function of the chi-square distribution at $X^2_{n,m}$.

The likelihood of all other assignments of spin systems *m* are considered via the calculation of the marginalization, P(B_m):

$$\mathrm{P}(B_m) = \sum_{n=1}^{N} \left[P\left(B_m \middle| A_{n,m}\right) P\left(A_{n,m}\right) \right] \quad (13)$$

Where the summation term acts over all possible residue positions *n.* Using Bayes' theorem as expressed above, the posterior probability $\mathrm{P}\left(A_{n,m} \middle| B_m\right)$ (i.e., the probability of a particular assignment being correct given the observed data) can be calculated. E^{cs}_m is then calculated via:

$$E^{cs}_m = \frac{-E^{cs}_{min}}{\log 1/N} \log \frac{P(A_{n,m}|B_m)}{1/N} \quad (14)$$

To avoid numerical instability in the evaluation of logarithms of numbers near zero and to prevent a dominating influence of inaccurate chemical shift predictions on the energy function, instances where $E^{cs}_m > E^{cs}_{max}$ are fixed at E^{cs}_{max}. E^{cs}_{max} and E^{cs}_{min} are set to 100 and -50 respectively, for this study. The user however has full control over these parameters to change the influence of the chemical shift predictions on the course of the annealing run.

The source of predicted shifts for each resonance can be from any source, so long as the precision of the prediction algorithm is accurately estimated. Here we have used SHIFTX+

(version 1.11).[25] In the absence of an acceptable high-resolution structural model, the average and standard deviation of the BMRB distribution of chemical shifts for a given atom of a given residue type are used as the predicted shift and prediction error respectively. This is also used in regions where the sequence of interest contains a tag that is absent in the structural model used to predict chemical shifts as well as regions that are not resolved.

Test proteins and data sets

Bayllagio was run on datasets for the proteins IL-1β, IL-1Ra, IGPS, MBP, CY1 and ecTS (Table 1). The source of these data sets varied. Four were collected by the authors and two were provided by other researchers. The details of the acquisition parameters for the data sets collected by the authors can be found in Supplementary Table 1. Predicted H, N, Cα, Cβ, and C' chemical shifts were generated via SHIFTX+ using either PDB entries or AlphaFold2[24] predicted structures (Table 1) Chemical shift prediction errors for H, N, Cα, Cβ, and C' were taken from the reported root mean squared deviations of SHIFTX+ predictions: 0.45, 2.4, 0.8, 0.95, 0.9 ppm, respectively. Sequence regions present in the NMR sample but not resolved or present in the provided structure (e.g., loops or expression tags) were given predicted values from their corresponding average values in the BMRB. Prediction errors associated with BMRB-derived values were taken as the standard deviation of the corresponding resonance distribution for the particular amino acid type in the BMRB.

Bayllagio was compared to two triple resonance assignment algorithm that are highly utilized by the NMR community. The assignment algorithm MARS was run on NMRbox utilizing the same input triple resonance and chemical shift data as described above. Runs were also provided secondary structural prediction information generated via PSIPRED[46]. Cutoffs for Cα, Cβ and C' connectivity were chosen to maximize the quantity of "high" and "medium" confidence assignments and ranged between 0.2 and 0.5 ppm. Assignments made that were labelled as high and medium confidence were considered assigned while residues assigned with low confidence were considered unassigned. The assignment algorithm I-PINE was also compared to Bayllagio and was run using the I-PINE server and the same spin system data used for Bayllagio and MARS. For each algorithm, proposed assignments were compared to reference assignments. At each residue position, the proposed assignment was determined to have either matched,

mismatched, or been missing when compared to the reference assignments (see Results and Discussion). Note that I-PINE explicitly searches the BMRB database for deposited assignments and thus could only be benchmarked for IGPS and IL-1Ra which are not present in the BMRB.

To assess the performance of Bayllagio on datasets of lower quality, the acquired MBP crosspeak lists were curated to randomly discard a fraction of spin systems and/or resonances. For each data quality condition where error bars are reported, 10 different curated data sets were randomly generated and Bayllagio was run on each of them. The results from each of these executions of Bayllagio were generated from the curation of 11 independent annealing runs. The bar heights report on the average among the 10 independent data sets and the error bars report the standard deviation of the indicated quantities.

Implementation

Bayllagio was implemented in C++ and can be built on all major computing platforms (MacOS, Linux and Windows). Bayllagio possesses a command line interface, as well as a GUI implemented using the wxWidgets library and utilizes the Boost and dsrpdb libraries. For this study, the simulations were run on 2019 6-core MacBook Pro (Intel) with up to 12 annealing runs running in parallel. Each annealing run usually took less than 1 min to complete (see Supplementary Table 2). Bayllagio will be made generally available for non-commercial use through the NMRbox resource.

Data Availability

Resonance assignments for IGPS and IL-1Ra have been deposited to the BMRB under accession numbers 51347 and 51352, respectively.

Acknowledgements

We are grateful to Dominque Frueh and his colleagues for graciously providing us with crosspeak for Cy1 and for fruitful discussions, and to Andrew Lee for providing crosspeak lists for ecTS. We would also like to thank the Texas A&M High Performance Research Computing center for access to computational resources for the prediction of the Cy1 structure used for the assignment, and NMRbox for access to NMRPipe and other data processing packages. This work

was supported by grants from the Mathers Foundation (MF-1809-00155), the National Institutes of Health (GM129076) and Texas A&M University.

Contributions

A.C.B and A.J.W. conceived the algorithm. A.C.B. wrote the computer code to implement the algorithm. A.C.B., S.K., G.T.-M. and A.J.W. tested Bayllagio. A.C.B., S.K., and G.T.-M. prepared isotopically enriched protein, collected, processed and analyzed NMR data for IL-1β, IGPS, and IL-1Ra, respectively. S.K. and G.T.-M. manually assigned IGPS and IL-1Ra, respectively. K.M prepared isotopically enriched MBP and analyzed MBP NMR data. A.C.B collected and processed MBP NMR data. A.C.B. and A.J.W. wrote the manuscript with input from all authors.

Competing interest

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material

available at @

Correspondence and requests for materials should be addressed to A.J.W.

References

1. Wüthrich, K. Sequential individual resonance assignments in the ^{1}H-NMR spectra of polypeptides and proteins. *Biopolymers* **22**, 131–138 (1983).
2. Wüthrich, K., Wider, G., Wagner, G. & Braun, W. Sequential resonance assignments as a basis for determination of spatial protein structures by high resolution proton nuclear magnetic resonance. *J. Mol. Biol.* **155**, 311–319 (1982).
3. Billeter, M., Braun, W. & Wüthrich, K. Sequential resonance assignments in protein 1H nuclear magnetic resonance spectra. Computation of sterically allowed proton-proton distances and statistical analysis of proton-proton distances in single crystal protein conformations. *J. Mol. Biol.* **155**, 321–346 (1982).
4. Englander, S. W. & Wand, A. J. Main chain directed strategy for the assignment of 1H NMR spectra of proteins. *Biochemistry* **26**, 5953–5958 (1985).

5. Di Stefano, D. L. & Wand, A. J. Two-dimensional 1H NMR study of human ubiquitin: A main chain directed assignment and structure analysis. *Biochemistry* **26**, 7272–7281 (1987).
6. Wand, A. J. & Nelson, S. J. Refinement of the main chain directed assignment strategy for the analysis of 1H NMR spectra of proteins. *Biophys. J.* **59**, 1101–1112 (1991).
7. Nelson, S. J., Schneider, D. M. & Wand, A. J. Implementation of the main chain directed assignment strategy. Computer assisted approach. *Biophys. J.* **59**, 1113–1122 (1991).
8. Ikura, M., Kay, L. E. & Bax, A. A novel approach for sequential assignment of 1H, 13C, and 15N spectra of larger proteins: Heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin. *Biochemistry* **29**, 4659–4667 (1990).
9. Montelione, G. T. & Wagner, G. Conformation-independent sequential NMR connections in polypeptides by H1-C13-N15 triple-resonance experiments. *J. Magn. Reson.* **87**, 183–188 (1990).
10. Driscoll, P. C., Marius Clore, G., Marion, D., Wingfield, P. T. & Gronenborn, A. M. Complete resonance assignment for the polypeptide backbone of interleukin 1ß Using three-dimensional heteronuclear NMR spectroscopy. *Biochemistry* **29**, 3542–3556 (1990).
11. Sattler, M., Schleucher, J. & Griesinger, C. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Progr. Nucl. Magn. Reson.* vol. 34 93–158 (1999).
12. Frueh, D. P. Practical aspects of NMR signal assignment in larger and challenging proteins. *Progr. Nucl. Magn. Reson.* vol. 78 47–75 (2014).
13. Gardner, K. H. & Kay, L. E. The use of 2H, 13C, 15N multidimensional NMR to study the structure and dynamics of proteins. *Annu. Rev. Biophys. Biomol. Struct.* **27**, 357–406 (1998).
14. Palmer, A. G. Chemical exchange in biomacromolecules: Past, present, and future. *J. Magn. Reson.* **241**, 3–17 (2014).
15. Tjandra, N. & Bax, A. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science (80-.).* **278**, 1111–1114 (1997).
16. Salmon, L. & Blackledge, M. Investigating protein conformational energy landscapes and atomic resolution dynamics from NMR dipolar couplings: A review. *Reports Prog. Phys.* **78**, (2015).
17. Clore, G. M. & Gronenborn, A. M. Applications of three- and four-dimensional heteronuclear NMR spectroscopy to protein structure determination. *Prog. Nucl. Magn. Reson. Spectrosc.* **23**, 43–92 (1991).
18. Zimmerman, D. E. *et al.* Automated analysis of protein NMR assignments using methods from artificial intelligence. *J. Mol. Biol.* **269**, 592–610 (1997).
19. Moseley, H. N. B., Monleon, D. & Montelione, G. T. Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. *Methods Enzym.* **339**, 91–108 (2001).
20. Baran, M. C., Huang, Y. J., Moseley, H. N. B. & Montelione, G. T. Automated analysis of protein NMR assignments and structures. *Chem. Rev.* **104**, 3541–3555 (2004).
21. Pervushin, K., Riek, R., Wider, G. & Wüthrich, K. Attenuated T2 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. U. S. A.* **94**, 12366–12371 (1997).
22. Hitchens, T. K., Lukin, J. A., Zhan, Y., McCallum, S. A. & Rule, G. S. MONTE: An

automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins. *J. Biomol. NMR* **25**, 1–9 (2003).
23. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. Optimization by simulated annealing. *Science* **220**, 671–680 (1983).
24. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953).
25. Han, B., Liu, Y., Ginzinger, S. W. & Wishart, D. S. SHIFTX2: Significantly improved protein chemical shift prediction. *J. Biomol. NMR* **50**, 43–57 (2011).
26. Ulrich, E. L. *et al.* BioMagResBank. *Nucleic Acids Res.* **36**, D402–D408 (2008).
27. Clubb, R. T., Thanabal, V. & Wagner, G. A constant-time three-dimensional triple-resonance pulse scheme to correlate intraresidue 1HN, 15N, and 13C′ chemical shifts in 15N13C-labelled proteins. *J. Magn. Reson.* **97**, 213–217 (1992).
28. Grzesiek, S. & Bax, A. Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein. *J. Magn. Reson.* **96**, 432–440 (1992).
29. Bax, A. & Ikura, M. An efficient 3D NMR technique for correlating the proton and15N backbone amide resonances with the α-carbon of the preceding residue in uniformly15N/13C enriched proteins. *J. Biomol. NMR* **1**, 99–104 (1991).
30. Wittekind, M. & Mueller, L. HNCACB, a high-sensitivity 3D NMR experiment to correlate amide-proton and nitrogen resonances with the alpha- and beta-carbon resonances in proteins. *J. Magn.Reson. SerB* **101**, 201–205 (1993).
31. Grzesiek, S. & Bax, A. Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. *J. Am. Chem. Soc.* **114**, 6291–6293 (1992).
32. Jung, Y. S. & Zweckstetter, M. Mars - Robust automatic backbone assignment of proteins. *J. Biomol. NMR* **30**, 11–23 (2004).
33. Lee, W. *et al.* I-PINE web server: an integrative probabilistic NMR assignment system for proteins. *J. Biomol. NMR* **73**, 213–222 (2019).
34. Lee, W. *et al.* PACSY, a relational database management system for protein structure and chemical shift analysis. *J. Biomol. NMR* **54**, 169–179 (2012).
35. Mishra, S. H. *et al.* Global dynamics as communication sensors in peptide Synthetase Cyclization Domains. *bioRxiv* (2021) doi:10.1101/2021.10.06.461881.
36. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
37. McIntosh, L. P. & Dahlquist, F. W. Biosynthetic incorporation of N-15 and C-13 for assignment and interpretation of nuclear magnetic resonance spectra of proteins. *Q. Rev. Biophys.* **23**, 1–38 (1990).
38. Krishnarjuna, B., Jaipuria, G., Thakur, A., D'Silva, P. & Atreya, H. S. Amino acid selective unlabeling for sequence specific resonance assignments in proteins. *J. Biomol. NMR* **49**, 39–51 (2011).
39. Prasanna, C., Dubey, A. & Atreya, H. S. Amino acid selective unlabeling in protein NMR spectroscopy. *Methods Enzym.* **565**, 167–189 (2015).
40. Sugiki, T., Furuita, K., Fujiwara, T. & Kojima, C. Amino acid selective 13C labeling and 13C scrambling profile analysis of protein α and side-chain carbons in Escherichia coli utilized for protein nuclear magnetic resonance. *Biochemistry* **57**, 3576–3589 (2018).
41. Nerli, S., De Paula, V. S., McShan, A. C. & Sgourakis, N. G. Backbone-independent NMR resonance assignments of methyl probes in large proteins. *Nat. Commun.* **12**,

(2021).
42. Delaglio, F. *et al.* NMRPipe: A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).
43. Maciejewski, M. W. *et al.* NMRbox: A resource for biomolecular NMR computation. *Biophys. J.* **112**, 1529–1534 (2017).
44. Lee, W., Tonelli, M. & Markley, J. L. NMRFAM-SPARKY: Enhanced software for biomolecular NMR spectroscopy. *Bioinformatics* **31**, 1325–1327 (2015).
45. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
46. Buchan, D. W. A. & Jones, D. T. The PSIPRED protein analysis workbench: 20 years on. *Nucleic Acids Res.* **47**, W402–W407 (2019).
47. Gardner, K. H., Zhang, X. C., Gehring, K. & Kay, L. E. Solution NMR studies of a 42 KDa Escherichia coli maltose binding protein/β-cyclodextrin complex: Chemical shift assignments and analysis. *J. Am. Chem. Soc.* **120**, 11738–11748 (1998).
48. Sapienza, P. J. & Lee, A. L. Backbone and ILV methyl resonance assignments of E. coli thymidylate synthase bound to cofactor and a nucleotide analogue. *Biomol. NMR Assign.* **8**, 195–199 (2014).

Table 1 | Proteins employed to evaluate Bayllagio

Protein	MW (kDa)	Sequence Length	Data Source	Reference Assignments Source	Reference Assignments (%)*	Reference Structure
IL-1β	17.5	154	This work	BMRB:434[10]	95.9	PDB: 9ILB
IL1Ra C66A,C122A	17.1	152	This work	This work	92.4	PDB: 2IRT†
IGPS R43S	28.4	248	This work	This work	88.0	PDB: 1IGS
MBP	40.8	371	This work	BMRB:4354[47]	95.1	PDB: 1DMP
CY1-6xHis	51.9	453	Freuh lab correspondence	Freuh lab correspondence	90.0	Freuh lab correspondence
ecTS‡	61.0	264	Lee lab correspondence[48]	BMRB:19082[48]	91.2	PDB: 1AOB

*percentage of non-proline residues assigned a spin system in the reported or obtained manual assignment; † PDB file was modified using the PYMOL mutation wizard to incorporate the C66A and C121A amino acid substitutions prior to SHIFTX+ prediction; ‡ symmetric homodimer under experimental conditions for assignment

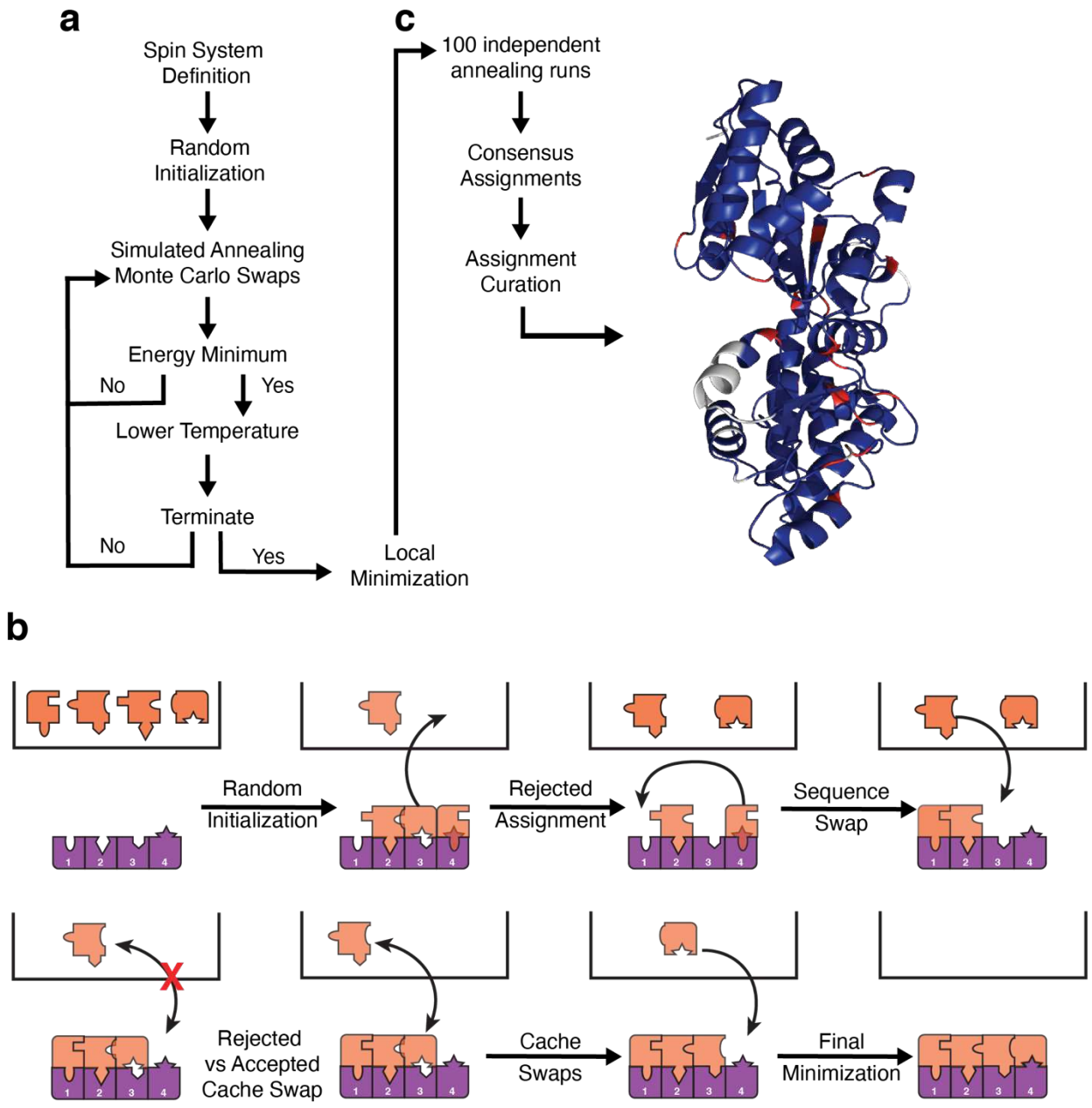

Fig. 1 | Outline of the Bayllagio triple resonance assignment algorithm. a, The search engine rests on a Bayesian-based simulated annealing protocol using a specific-heat mechanism to guide cooling of the system. The search begins with a random assignment of available spin systems to specific positions within the primary sequence of the protein. Sequential adjacency in the primary sequence of the protein is provided by triple resonance NMR spectra. Predicted chemical shift information, based on a high-resolution structural model or simply gleaned from empirical amino acid specific distributions, completes the system energy. **b**, The annealing step procedure involves Monte Carlo swapping of putative spin system assignments to specific locations in the primary sequence. Spin systems (orange puzzle pieces) begin in the cache (black box) and are initialized by random assignment to the sequence (purple pieces). Spin systems can then be swapped with others or moved to other locations of the sequence or the cache. Swaps are

accepted or rejected with a probability based on the change in energy of the proposed swap. The energy of each individual spin system is a function of how well it “fits” with the adjacent spin system (adjacency energy) as well as how well it fits the predicted shifts for that residue location (chemical shift energy). See main text and supplementary material for details of the calculation of pseudo-energy and other parameters. **c**, Curation of the final resonance assignments from multiple independent simulated annealing runs. Shown is a ribbon representation of maltose binding protein (PDB code: 1DMB) color-coded according to assignment status following analysis by Bayllagio: correctly assigned residues (blue); unassigned residues (white), prolines (red). There are no incorrectly assigned residues.

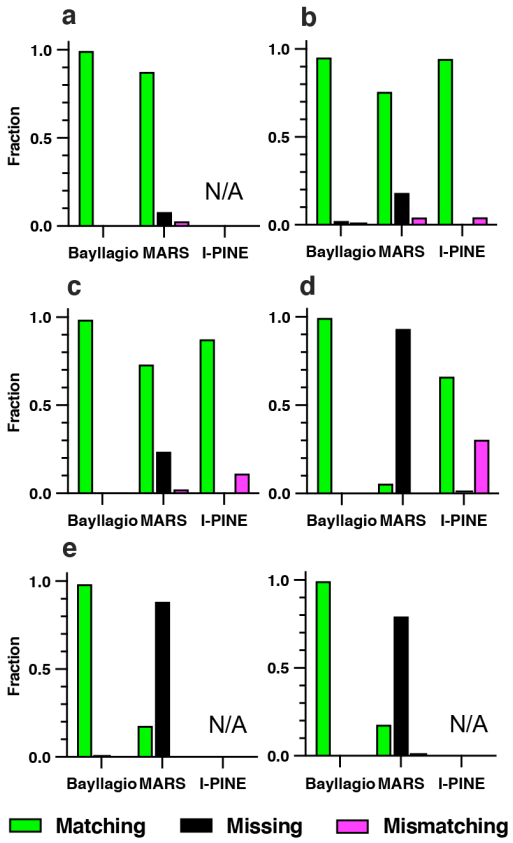

Fig. 2 | Comparison of automated assignment algorithms. Performance of Bayllagio, MARS and I-PINE against manual/deposited triple resonance assignments of six protein systems: **a**, IL-1β; **b**, IL-1Ra; **c**, IGPS; **d**, MBP; **e**, CY1; **f**, ecTS. Shown are the fractions of residues that are accurately matched to the reference assignments, incorrectly matched or unassigned. I-PINE searches the BMRB directly for a matching protein and those proteins with deposited assignments are not valid tests. All but IL-1Ra and IGPS have deposited resonance assignments. I-PINE was therefore excluded for comparison with MARS and Bayllagio for those four proteins.

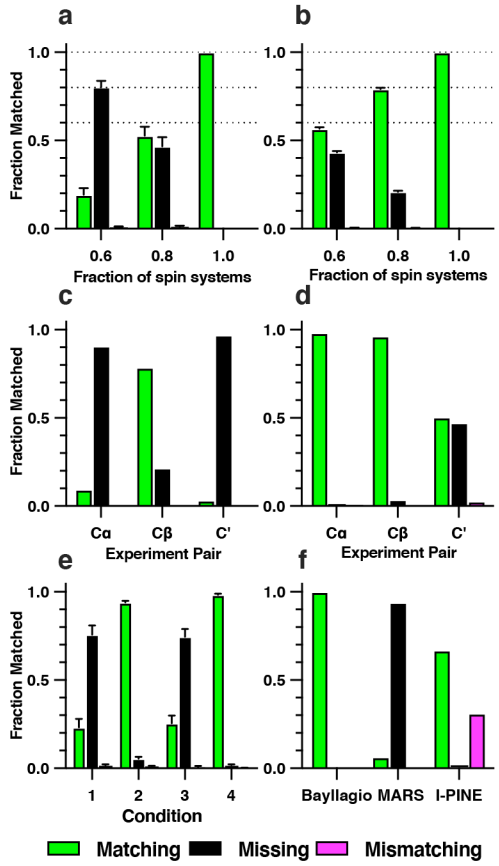

Fig. 3 | Performance of Bayllagio with incomplete triple resonance crosspeak lists. MBP crosspeak lists where randomly reduced fractions of the total spin systems are retained (**a** and **b**). MBP data sets where only a single resonance pair was used in the assignment (**c** and **d**). **a** and **c** were run without structure-based predicted chemical shifts while **b** and **d** were run using structure-based chemical shift predictions. **e**, Sparse MBP data sets with the following characteristics: 88% of CA resonances and 50% of CB resonances randomly retained, conditions 1 and 2 were run without any C' information while conditions 3 and 4 randomly retained 75% of the C' information. Conditions 1 and 3 were run without chemical shift prediction, whereas conditions 2 and 4 were run with chemical shift prediction. **f**, A data set from condition 4 of **e** was used to challenge algorithms Bayllagio, MARS and I-PINE. Bar heights show the fraction of residues that belong to each category. Error bars are from using 10 distinct, independently generated data sets by randomly discarding data from the MBP spin system data to achieve the characteristics indicated.

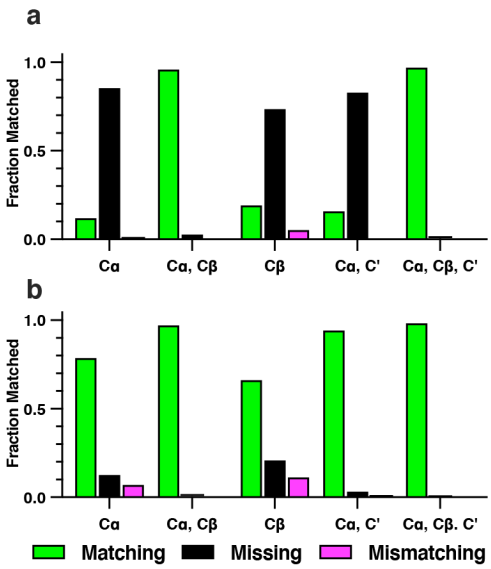

Fig 4 | Dependence of the performance of Bayllagio on availability of chemical shift predictions. Using the Cy1 data set, the assignment was carried out using the indicated triple resonance data set combinations and **a**, BMRB chemical shift statistics for residue type or **b**, using chemical shifts based on a structural model provided by AlphaFold2. Bar heights show the fraction of residues that belong to each category. Triple resonance data sets include HNCA/HN(CO)CA (Cα), HNCB/HN(CO)CB (Cβ) and HNCO/HN(CA)CO (C')

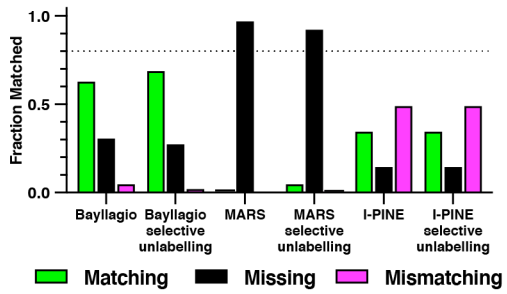

Fig. 5 | Comparison of the effect of selective unlabelling data on algorithm performance. An MBP data set with missing data was generated (see text) and KQRN residue type data was provided to the algorithms ("selective unlabelling" conditions) or the algorithm was run without the selective unlabelling the data. The dashed line at 0.8 indicates the maximum height of the "matching" bars due to only 80% of the spin systems being present.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- WandNatCommunSISubmitted.pdf
- Wandnrsoftwarepolicy.pdf