Model-based assignment and inference of protein backbone nuclear magnetic resonances.

Olga Vitek[*] Jan Vitek[†] Bruce Craig[*] Chris Bailey-Kellogg[†‡]

Abstract

Nuclear Magnetic Resonance (NMR) spectroscopy is a key experimental technique used to study protein structure, dynamics, and interactions. NMR methods face the bottleneck of spectral analysis, in particular determining the resonance assignment, which helps define the mapping between atoms in the protein and peaks in the spectra. A substantial amount of noise in spectral data, along with ambiguities in interpretation, make this analysis a daunting task, and there exists no generally accepted measure of uncertainty associated with the resulting solutions. This paper develops a model-based inference approach that addresses the problem of characterizing uncertainty in backbone resonance assignment. We argue that NMR spectra are subject to random variation, and ignoring this stochasticity can lead to false optimism and erroneous conclusions. We propose a Bayesian statistical model that accounts for various sources of uncertainty and provides an automatable framework for inference. While assignment has previously been viewed as a deterministic optimization problem, we demonstrate the importance of considering all solutions consistent with the data, and develop an algorithm to search this space within our statistical framework. Our approach is able to characterize the uncertainty associated with backbone resonance assignment in several ways: 1) it quantifies of uncertainty in the individually assigned resonances in terms of their posterior standard deviations; 2) it assesses the information content in the data with a posterior distribution of plausible assignments; and 3) it provides a measure of the overall plausibility of assignments. We demonstrate the value of our approach in a study of experimental data from two proteins, Human Ubiquitin and Cold-shock protein A from *E. coli*. In addition, we provide simulations showing the impact of experimental conditions on uncertainty in the assignments.

Keywords: Nuclear Magnetic Resonance (NMR) spectroscopy; uncertainty in NMR spectra; Bayesian modeling; statistical inference; protein structure; structural genomics

---

[*]Dept. of Statistics, Purdue University.
[†]Dept. of Computer Sciences, Purdue University.
[‡]Corresponding author: `cbk@cs.purdue.edu`; 250 N. Univ. St., West Lafayette, IN 47907.

1 Introduction

Knowledge of the three-dimensional structure of proteins is essential for understanding and controlling their function. The emerging field of *structural genomics* [6] requires new methods that provide structural information at a genomic scale. Nuclear magnetic resonance (NMR) spectroscopy is an experimental technique that is particularly suitable for this task. It is capable of determining atomic detail about protein structures in physiological conditions, and allows rapid and cost-efficient screening of foldedness, internal dynamics and ligand binding. Some 15%-20% of new protein structures are currently determined by NMR, and the rate is likely to grow [25].

A key step in NMR-based analysis is *sequential backbone resonance assignment*, which determines the resonance values associated with specific atoms of the protein backbone. Assignment is essential for determination of protein structure, since, for example, the distance restraints between atoms in the structure are derived from the corresponding peaks in particular (NOESY) spectra. It is also a necessary step in studies of protein dynamics and intermolecular interactions. As a major bottleneck in these approaches, efficient, confident assignment protocols are a necessary prerequisite to genomic-scale NMR studies. However, because of the significant noise present in the spectra, it often takes weeks and sometimes months to complete the assignment of a protein by hand.

Development of automated methods for backbone resonance assignment has recently become a very active area of research [26]. Numerous search algorithms and scoring functions, as well as methods that include different spectral information, have been proposed (see Sec. 6 for further discussion, as well as a comparison with our method). None of these existing methods, however, provides a statistically sound method to measure uncertainty in the assigned resonance values. Although it is generally agreed that NMR spectra are stochastic in nature, there exists no formal inferential procedure regarding assigned backbone resonances. As a result, many researchers still choose to complete assignments by hand in order to feel confident in the quality of their results.

This paper presents the first approach that provides formal statistical inference for backbone NMR assignment. By carefully modeling the sources of variability associated with NMR spectra and employing an appropriate algorithm to search for feasible assignment solutions, our method enables rigorous quantification of the uncertainty in the resonance assignment. The Bayesian framework [21] is particularly attractive here as it allows incorporation of all available *a priori* knowledge. In the Bayesian context, we quantify the uncertainty in an entire assignment by the posterior distribution over all feasible solutions, and we quantify the uncertainty in the individual assigned resonance values by their posterior standard deviations. Since this involves investigating more than the single "best" assignment, we develop a new assignment algorithm that exhaustively explores a large portion of the assignment space and finds *all* assignments that are consistent with the data within the explored space. Therefore, uncertainty in the results is a property of the information content in the data, and not of the algorithm for finding the assignments.

Our model-based approach is fully automated and provides useful information that other approaches do not. First, it measures the overall plausibility of an assignment and shows
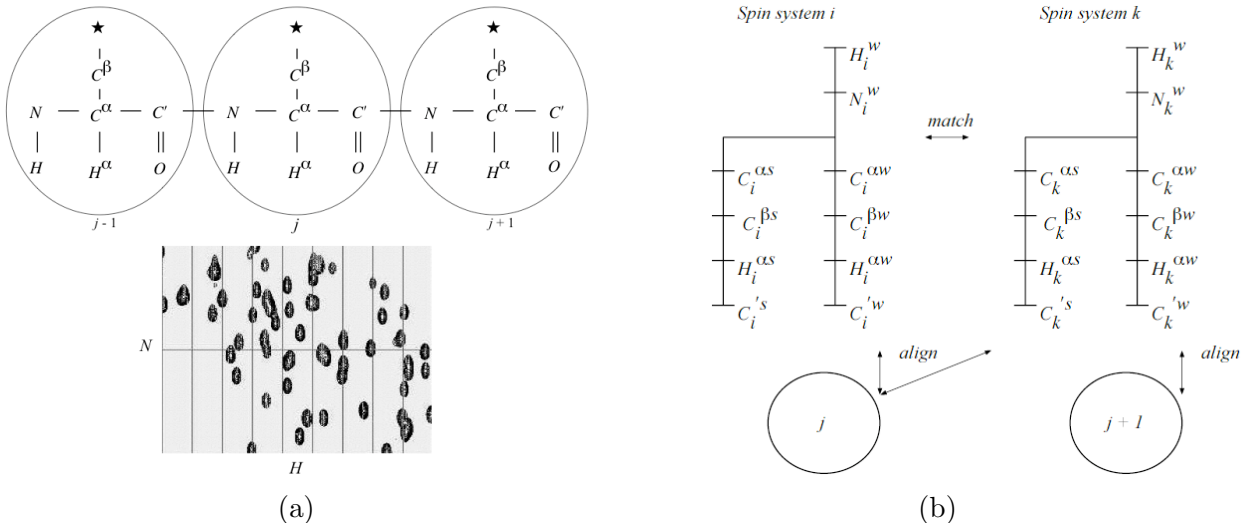
Figure 1: (a) A chain of three residues at positions $j-1$, j, and $j+1$ in the primary sequence. A star schematically denotes the side-chain atoms that vary between residue types. An HSQC spectrum is shown in the bottom of the figure; the mapping between atomic interactions and peaks is unknown and is necessary in order to determine the chemical shifts of the backbone atoms. (b) Schematic illustration of spin systems. Chemical shift values for H, N, C^α, C^β, C' and H^α are obtained by combining peaks across spectra. i and k are indexes of the spin systems, w indicates within-residue chemical shifts, and s indicates sequential chemical shifts (i.e. for the preceding residue). Values grouped into the same spin system are connected by lines. Circles indicate the (initially unknown) residues originating the spin systems and are numbered according to their position in the primary sequence.

that, generally, additional assignments beyond just the optimal one are also supported by the data. Second, it assesses the information content in the collected spectra by means of the posterior distribution of plausible assignments. Third, it quantifies the uncertainty in the individual assigned resonances in terms of posterior standard deviations, and identifies well-determined positions that can confidently be used in further NMR studies (e.g., structure determination). We believe that our approach will help avoid errors in the assignment procedure and enable researchers to determine reliable assignments at a high-throughput rate.

2 NMR data

Proteins are biological macromolecules essential for living organisms. They are polymers of amino acids, also called *residues*, that form complex three-dimensional structures in physiological conditions. The top part of Fig. 1(a) shows the organization of three neighboring residues. Each residue has a central carbon atom, denoted C^α, to which are bonded a hydrogen atom (H^α), an H–N group and a C'=O group. This chain comprises the protein's

backbone. In addition to the backbone, there are 20 different amino acid types distinguished by the *side-chain* atoms attached to the C^α atoms. The side chain typically contains a carbon atom (C^β) and other atoms schematically denoted by a star in Fig. 1(a). Two exceptions are proline, which replaces H and H^α with a cycle of atoms from the N to the C^α, and glycine, which has another H^α instead of a C^β. The number and order of amino acid types in a protein — its *primary sequence* — is determined by other experimental techniques prior to study by NMR.

NMR spectroscopy [8] takes advantage of the magnetic properties of atomic nuclei. When a protein sample is placed into a static magnetic field and exposed to another oscillating magnetic field, the individual nuclei respond at specific resonance frequencies. These resonance frequencies, or *chemical shifts*, are very sensitive to the chemical environment surrounding a nucleus and are thus in principle unique for each atom in a typical moderate-sized globular protein. Chemical shifts are central to NMR-based studies, as they provide the "IDs" of the atoms, by which the atoms can be referenced in subsequent studies of structure, dynamics, and interactions. The goal of backbone resonance assignment is to determine the chemical shift values for the nuclei in the protein backbone.

Backbone resonance assignment employs data from a set of NMR experiments. A central experiment is the HSQC, which magnetically correlates bonded H–N nuclei and yields a two-dimensional spectrum. An example fragment of an HSQC spectrum is shown in the bottom of Fig. 1(a). Peaks in the spectrum indicate H–N bonded pairs at specific chemical shifts (their coordinates). The HSQC is a core experiment since each residue except proline has an H–N pair and thus (in ideal conditions) generates such a peak. However, the correspondence between the peak and the H–N pair that generated it is unknown.

Determination of the chemical shifts of the atoms in the backbone also requires at least one, and usually several, three-dimensional NMR experiments. An example of such an experiment is the HNcoCA, which magnetically correlates bonded H–N backbone nuclei with the C^α nucleus of the preceding residue. The experiment yields a three-dimensional spectrum, in which the projection on the H–N dimensions corresponds to the HSQC, and the third dimension to the magnetically correlated C^α. Since each HNcoCA peak correlates atoms in two neighboring residues, the experiment is useful in identifying *sequential* interactions. Another three-dimensional experiment, the HNCA, correlates the bonded H–N pairs with the C^α of either the preceding residue (as in the HNcoCA) or of the same residue. It yields approximately twice as many three-dimensional peaks as the HNcoCA, gathering both sequential and *within-residue* interactions. Similar NMR experiments can be designed to involve interactions of the H–N pairs with C^β, H^α, and C'. The type of atoms involved in a spectrum is called its *resonance type*. Again, it is important to note that the residue(s) whose atoms generated each peak is initially unknown for all spectra.

The chemical shifts can be determined by finding a mapping between the peaks and the atoms. In summary, a typical procedure to find such a mapping consists of the following steps [26, 33]: (1) identify and determine the centers of peaks within the spectra; (2) group the peaks into collections called *spin systems* according to their shared H–N resonances; (3) match spin systems according to corresponding sequential and within-residue chemical

shifts; (4) align connected chains of spin systems to the primary sequence according to the expected chemical shifts for corresponding residue types. We now detail each step in this procedure.

Peak analysis. Peaks in NMR spectra do not have precise positions. As illustrated in the HSQC in Fig. 1(a), they span some volume, shown in the figure by intensity contours. The volume of a peak depends on the physical properties of the nucleus, as well as the type and sensitivity of the NMR experiment. The coordinates of the center of a peak can be determined by automated peak picking software such as NMRDraw [10]. The center can be determined fairly accurately for a well-resolved peak, but the accuracy is compromised when the peak is broad or has a low intensity, or when several peaks overlap. Peak positions are also subject to random variation between spectra, due to sample variation, differences in sample temperature and other experimental artifacts. In summary, the coordinates of the centers of the peaks can be viewed as noisy readings from the underlying chemical shifts, having a compound variation from within and between the spectra.

Peak picking software produces a list of peaks, which typically is missing some peaks and has some extra (spurious) peaks. Peaks can be missing due to physical reasons (e.g. an overly broad peak resulting from extensive dynamics), spectral degeneracy (e.g. peaks too close together to be differentiated), or noise. Spurious peaks can originate from impurities in the sample, minor conformations of the protein, or errors in the peak picking procedure.

Compilation of spin systems. The spectra we employ here all involve the "anchor" H–N pairs. Thus one can combine peaks across spectra by referencing the H–N coordinates. The third coordinate is classified according to resonance type as shown in Fig. 1(b). These collections of peaks, called *spin systems*, typically consist of within-residue and sequential chemical shifts of several resonance types, anchored at one (unknown) residue and connecting to its predecessor. Due to the noise in peaks discussed above, some spin systems can have missing chemical shift values, some spin systems can be entirely missing, and some extra, spurious spin systems can be compiled from extra, spurious peaks. Due to variability in peak positions, approximate equality of peak coordinates must be allowed when compiling spin systems. While the compilation is unambiguous for isolated peaks, ambiguities can arise in the case of close or overlapping peaks.

Matching spin systems. Spin systems can be arranged into ordered chains by matching the sequential resonances of one with the within-residue resonances of another. If two spin systems originate from the neighboring residues as shown in Fig. 1(b), the within-residue chemical shifts of the first spin system must be approximately the same as the sequential chemical shifts of the following spin system. Approximate matches must be allowed due to noise, and ambiguities arise when several spin systems have similar sequential or within-residue values. The number of plausible matches increases dramatically when the data contains a large number of missing resonance types or entirely missing spin systems — the missing chemical shifts become "wild cards" that allow matches to any spin system. Extra spin systems result in additional complications as they are often incomplete, and can sometimes form an incorrectly unambiguous match.

Aligning spin systems. Spin systems can be aligned to positions in the primary

sequence by comparing chemical shifts to the values expected for the corresponding residue types. Since an atom's chemical shift is sensitive to its local chemical environment, there is a well-characterized effect of amino acid type on chemical shift. The expected ranges can be determined from databases such as BioMagResBank [29] or RefDB [34], which contain assignments for many proteins. Unfortunately, the ranges are typically quite broad, and allow a spin system to align to any of many possible positions in the primary sequence. Chains of connected spin systems are required in order to overcome this ambiguity, where each spin system agrees that the corresponding position is consistent. However, even short chains of spin systems can typically be aligned at several places in the sequence. We note two special cases in alignment: given the H–N based spin system definition here, alignments are not possible at the first residue or proline residues; further, alignments are restricted at glycine residues due to the substitution of an extra H^{α} for the C^{β}.

This paper assumes that the data have been correctly processed and the spin systems compiled by either manual or automated methods. We focus on the problem of matching sequentially-connected spin systems and aligning them to the primary sequence. For clarity, we make a distinction between the result of matching and aligning spin systems to the sequence, which we call a "mapping," and the determination of the unknown resonance values on the basis of a mapping. We use "assignment" to denote the entire procedure.

A mapping is established by systematically matching and aligning the spin systems, much like a jigsaw puzzle. Each spin system can have at most one predecessor and can be mapped to at most one spin system, and each position can be mapped to at most one spin system. Positions with no mapped spin systems are associated with entirely missing spin systems, and spin systems not mapped to positions are considered extras. Satisfactory mappings cover the maximum number of positions. In principle, there exists one "correct" mapping that generated the data. In practice, however, the large number of ambiguities at each step of the procedure often results in several plausible mappings that need to be considered. As discussed in the introduction, our method enumerates all plausible mappings and uses their posterior distribution in a model-based approach to quantify the uncertainty.

The goal of backbone resonance assignment is to determine the chemical shifts of the protein backbone. A mapping between the spin systems and positions in the sequence is a means to this end. Given a mapping, the chemical shifts of a nucleus can be deduced from the associated peaks. However, uncertainty in the peak positions and uncertainty in the mappings result in uncertainty in chemical shifts.

3 Methods

We view backbone resonance assignment as a process of estimating the chemical shifts of the protein backbone from noisy data. A candidate mapping, combined with distributional assumptions regarding the peaks, can then be viewed as a statistical model of the data. From this perspective, the search for the "best" candidate mapping becomes a model selection problem. When several models are plausible, inference about the unknowns should incorporate both the uncertainty in the data and the uncertainty in the model selection.
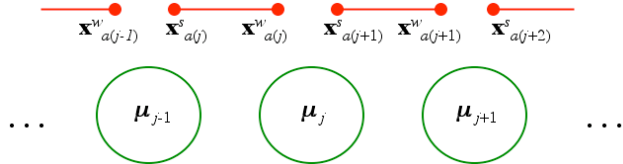
Figure 2: Quantities of interest. $\mathbf{x}^s_{a(j)}$ and $\mathbf{x}^w_{a(j)}$ are observable multivariate resonances, w indicates within-residue resonances and s indicates sequential resonances of the atom type. Red lines connect the sequential and within-residue resonances that are grouped into the same spin system. $\boldsymbol{\mu}_j$ are the unknown underlying resonances to be estimated.

3.1 Probability model

Unknowns of interest. The quantities of interest are schematically illustrated in Fig. 2. Consider a primary sequence of R residues. Let $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_R)$ denote the underlying chemical shifts of the backbone nuclei of the protein. Here each $\boldsymbol{\mu}_j$ is a vector composed of individual chemical shifts μ_{dj} for each resonance type d at position j. The $\boldsymbol{\mu}_j$ are the unknowns of interest, and the goal of the backbone resonance assignment is to estimate these values.

Input data. The input data are I observed spin systems $\mathcal{X} = \{(\mathbf{x}^s_1, \mathbf{x}^w_1), \dots, (\mathbf{x}^s_I, \mathbf{x}^w_I)\}$, where $\mathbf{x}^s_i$ is the vector of sequential chemical shifts x^s_{di}, and $\mathbf{x}^w_i$ is the vector of within-residue chemical shifts x^w_{di}, over resonance type d. We assume that the spin systems are correctly and unambiguously compiled prior to the analysis. The total number of spin systems, I, can be greater than, equal to, or less than the length of the protein R, depending on presence of extra and missing spin systems. Individual chemical shifts can also be missing in some spin systems.

Candidate mappings. Let $\boldsymbol{a} = (a_1, \dots, a_R)$ be a candidate mapping of the observed spin systems to positions in the primary sequence. In this notation, $a_j = i$ if $\mathbf{x}^w_i$ is mapped to position j (or equivalently $\mathbf{x}^s_i$ is mapped to position $j-1$). A mapping is one-to-one and gives the putative origin of the observed data. Some of the spin systems can be considered as extras by $\boldsymbol{a}$, and will be associated with sources of noise. Since we have a fixed number of positions and spin systems, considering more spin systems as extras implies that more positions have missing spin systems. Each spin system can potentially be mapped to one of several feasible positions, and therefore the space of the candidate mappings is combinatorially large.

Distributional assumptions for the observed data. We assume that, given the unknown vectors $\boldsymbol{\mu}$ and the correct assignment $\boldsymbol{a}$, the error in readings of the chemical shifts is non-systematic and Normally distributed:

$$\mathbf{x}^s_{a(j)} \mid \boldsymbol{\mu}, \boldsymbol{a} \sim \mathcal{N}(\boldsymbol{\mu}_{j-1}, W) \quad \text{and} \quad \mathbf{x}^w_{a(j)} \mid \boldsymbol{\mu}, \boldsymbol{a} \sim \mathcal{N}(\boldsymbol{\mu}_j, W).$$

The distribution of the readings is centered around $\boldsymbol{\mu}$, and the variance matrix W encompasses the variation from within and between the spectra. Intensity contours in the spectra are usually aligned with the coordinate axes, and the spectra are independently collected. It is therefore reasonable to assume that, conditional on the underlying resonance values and

assignment, the readings are independent between the resonance types. This implies that W is a diagonal matrix.

We further assume that, conditional on $\boldsymbol{\mu}$ and $\boldsymbol{a}$, the readings from the chemical shifts are independent across positions in the sequence. Note that the independence is conditional and will not hold marginally across all the candidate mappings. The assumption may oversimplify the correlation structure in the data since peaks collected within a single spectrum may co-vary. On the other hand, as discussed in Sec. 6, independence across positions is implicitly assumed by all existing automated methods for backbone resonance assignment.

Finally, we assume that the variance matrix W is known and constant for all $\mathbf{x}_{a(j)}^s$ and $\mathbf{x}_{a(j)}^w$. Weaknesses of this assumption include the possibility that peaks within a spectrum may have different resolutions, and the fact that the chemical shifts in a spin system summarize various numbers of peaks. However, estimation of the spin system-specific variances is a difficult task, requiring grouping of all peaks according to their common (unknown) source, and given that only one peak per source is available in many cases. As discussed in Sec. 6, constant and known variance is implicitly assumed by most current automated methods, and some generally accepted variance values are typically used. We follow this approach here, but plan to investigate the problem of spin system-specific variances in our future work.

Given the distributional assumptions above, the conditional likelihood of the composite of the positions in the primary sequence can be written as

$$\Pr(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{a}) \;=\; \prod_{j=1}^{R} \phi\left(W^{-\frac{1}{2}} \left(\mathbf{x}_{a(j)}^s - \boldsymbol{\mu}_{j-1}\right) \right) \; \phi\left(W^{-\frac{1}{2}} \left(\mathbf{x}_{a(j)}^w - \boldsymbol{\mu}_j\right) \right).$$

Here and in the rest of the paper ϕ denotes the density of the multivariate standard Normal distribution.

Prior distributions of the unknown $\boldsymbol{\mu}$. The distributions of chemical shifts of previously assigned proteins, as deposited in various databases, can be used to characterize the $\boldsymbol{\mu}_j$ *a priori*. We assume that the $\boldsymbol{\mu}_j$ are independent across all positions in the primary sequence and Normally distributed, i.e. $\boldsymbol{\mu}_j \sim \mathcal{N}(\boldsymbol{\theta}_j, \Sigma_j)$. The vector $\boldsymbol{\theta}_j$ contains mean chemical shifts of all resonance types, and Σ_j is the corresponding variance-covariance matrix estimated from the database. The parameters $\boldsymbol{\theta}_j$ and Σ_j are specific to the residue type at the position j.

Both $\boldsymbol{\theta}_j$ and Σ_j can be estimated from at least two databases, the BioMagResBank [29] and the RefDB [34], and in a number of ways. For example, one can assume that the resonance types are *a priori* independent, in which case Σ_j is a diagonal matrix, and take the database means and variances as the parameter estimates. Alternatively, one can assume a non-diagonal Σ_j and compute the correlation between the resonance types [24]. A more sophisticated approach would account for the redundancy in the sequences deposited into the BioMagResBank, and estimate the parameters from the subset of non-homologous sequences only [24, 31]. Or, when the information on the secondary structure of the protein is available, one can account for systematic effects of secondary structure type on chemical shift [32]. All of these methods are equally well supported by our approach, and Normal distributions with any parameters can be used provided that they are appropriately justified. In Sec. 5.2 we

conduct a sensitivity analysis and investigate the extent to which the choice of the parameters of the prior distributions affects the inference of the unknowns $\boldsymbol{\mu}$.

One may question the choice of the Normal prior distribution. Although the Normality assumption can in principle be substituted by any other distribution of choice, the Normal distribution is particularly attractive as it allows the analytical computation of posterior probabilities, and is therefore computationally efficient. The Normal distribution provides an adequate approximation of the chemical shifts in most cases and, as discussed in Sec. 6, is implicitly used by many existing methods of backbone resonance assignment. Some distributions in the databases may appear to have heavier than Normal tails. In this case, a spin system with outlying chemical shifts will not be correctly mapped to the primary sequence, and the corresponding position in the sequence will be associated with a missing spin system. An alternative solution is to increase the variance of the prior distribution in order to accommodate the tails.

Sources of noise. For completeness, we assume that the extra spin systems are readings from the underlying chemical shifts $(\boldsymbol{\mu}_{R+1}, \boldsymbol{\mu}_{R+2}, \ldots)$ of the sources of noise. We specify a non-informative Normal prior for the parameters by choosing $\boldsymbol{\theta}_j$, $j > R$, to be the mid-range of the measurements of each resonance type, and Σ_j, $j > R$, to cover the entire measurement range.

3.2   Scoring the candidate mappings

The probability model allows us to compare and evaluate the candidate mappings in terms of their posterior probabilities given the data. These can be obtained by applying Bayes theorem, and by averaging across the unknown parameters $\boldsymbol{\mu}$ [27]

$$\Pr(\boldsymbol{a}|\mathbf{x}) \propto \Pr(\mathbf{x}|\boldsymbol{a})\Pr(\boldsymbol{a}), \quad \text{where} \quad \Pr(\mathbf{x}|\boldsymbol{a}) = \int \Pr(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{a})\Pr(\boldsymbol{\mu})d\boldsymbol{\mu}. \tag{1}$$

As can be seen, the posterior probability $\Pr(\boldsymbol{a}|\mathbf{x})$ in (1) consists of two terms: the likelihood $\Pr(\mathbf{x}|\boldsymbol{a})$, and the prior distribution $\Pr(\boldsymbol{a})$ of a mapping $\boldsymbol{a}$.

Likelihood. The conjugate Normal structure of our model makes it possible to carry out the integration in (1) analytically. Upon integration, the likelihood can be written as

$$\Pr(\mathbf{x}|\boldsymbol{a}) = \prod_{j=1}^{R} \phi\left(U_1^{-\frac{1}{2}}(\mathbf{x}_{a(j+1)}^s - \mathbf{x}_{a(j)}^w) \right) \phi\left(U_2^{-\frac{1}{2}}(\bar{\mathbf{x}}_{a(j)} - \boldsymbol{\theta}_j) \right), \tag{2}$$

where $\bar{\mathbf{x}}_{a(j)}$ is the average of $\mathbf{x}_{a(j+1)}^s$ and $\mathbf{x}_{a(j)}^w$, $U_1 \triangleq 2W$ and $U_2 \triangleq \Sigma_j + W/2$. The observations mapped to a position j in the primary sequence contribute to the likelihood via two terms. The first involves the quantity $\mathbf{x}_{a(j+1)}^s - \mathbf{x}_{a(j)}^w$ and reflects the likelihood of the sequential connectivity at this position. The second involves the quantity $\bar{\mathbf{x}}_{a(j)}$ and reflects the consistency of the chemical shifts with the expectations of the mapped amino acid type. Because of the non-informative prior distribution of the sources of noise, their contribution to the likelihood is approximately 1, and can be ignored.

When comparing candidate mappings, it is helpful to consider $S(\mathbf{x}|\boldsymbol{a}) = -2 \log \Pr(\mathbf{x}|\boldsymbol{a})$, which can be viewed as a score measuring the success of $\boldsymbol{a}$ at predicting the data. In our case,

$$S(\mathbf{x}|\boldsymbol{a}) \quad = \quad \sum_{j=1}^{R} S_{match,j} + \sum_{j=1}^{R} S_{align,j} \tag{3}$$

where

$$S_{match,j} \quad = \quad (\mathbf{x}^s_{a(j+1)} - \mathbf{x}^w_{a(j)})'(2W)^{-1}(\mathbf{x}^s_{a(j+1)} - \mathbf{x}^w_{a(j)})$$
$$S_{align,j} \quad = \quad (\bar{\mathbf{x}}_{a(j)} - \boldsymbol{\theta}_j)'(\Sigma_j + W/2)^{-1}(\bar{\mathbf{x}}_{a(j)} - \boldsymbol{\theta}_j).$$

Maximizing the likelihood is equivalent to minimizing the score. Under the assumptions in Sec. 3.1 and given the correct assignment $\boldsymbol{a}$, $S(\mathbf{x}|\boldsymbol{a})$ follows the χ^2_{df} distribution. Moreover, a partial score computed over any subset of positions in the sequence also follows the χ^2_{df}. The integer df in this notation is the number of chemical shifts, x^s_{di} and x^w_{di}, used to compute the specific score. Although not part of the formal Bayesian framework, this interpretation is very useful as only those mappings that are consistent with the distribution need to be considered [28]. In practice, we can compare the scores with an α-quantile of the χ^2_{df} for a (small) pre-specified probability α, and reject the mappings with scores exceeding the quantile.

Prior distribution of the candidate mappings. A complete mapping is preferable to one with many missing observations. For example, a mapping considering most of the observed spin systems as noise is likely to be false. On the other hand, a spin system where most chemical shifts are missing is likely to be an extra. We therefore penalize the candidate mappings according to the number of missing chemical shifts associated with the positions in the sequence. Specifically, we define the prior probability of a mapping with d missing resonances as

$$\Pr(\boldsymbol{a}) \propto \exp\left\{-\tfrac{1}{2}q_{\chi^2}(\alpha, df)\right\} \tag{4}$$

where $q_{\chi^2}(\alpha, df)$ denotes the α-quantile of the χ^2 distribution with df degrees of freedom. On the $-2\log$ scale, the prior amounts to substituting a worst-case estimate for the contribution to the posterior probability of a missing observation.

The mappings with the highest posterior probabilities are used for inference about the unknown $\boldsymbol{\mu}$. Therefore, a candidate mapping can be discarded if its fit to the data is clearly worse than that of the best mapping found [16, 17]. In the following definition, we combine the interpretations on the scale of the posterior probability and on the scale of the score function.

Definition 1 (Mapping Consistency) *A mapping $\boldsymbol{a}$ is consistent with the data if, for a given (small) probability α,*

    1. $S_{match,j} < q_{\chi^2}(\alpha, df_{match,j})$ for all j.
    2. $S_{align,j} < q_{\chi^2}(\alpha, df_{align,j})$ for all j.

3. $S_{match,j} + S_{align,j} < q_{\chi^2}(\alpha, df_{match,j} + df_{align,j})$ *for all j.*
4. $S(\mathbf{x}|\boldsymbol{a}) < q_{\chi^2}(\alpha, df)$.
5. $P(\mathbf{x}|\boldsymbol{a}^*)/P(\mathbf{x}|\boldsymbol{a}) \leq 25$ *where $\boldsymbol{a}^*$ is the mapping with the highest posterior probability.*

df in the definition denotes the number of data points used to compute the corresponding score. Note that $S_{match,j}$ and $S_{align,j}$ represent the *joint* score of all resonance types of interest at position j. Therefore, a relatively loose match or alignment of one resonance type is considered as plausible if it is compensated for by a relatively strong match or alignment of the other resonance types. Although one can also define consistency separately for each resonance type, we found that in practice this definition provides more flexibility and results in more accurate assignments. The parameter α flexibly determines the search space of candidate mappings. Smaller α values result in more mappings being considered plausible, and the larger values result in rejection of more mappings.

3.3 Inference

Inference regarding $\boldsymbol{\mu}$ can be made by means of its posterior distribution given the data. If only one mapping $\boldsymbol{a}$ is consistent with the data as in Definition 1, it can be used as the basis for inference [27]

$$\mu_{dj} \mid \mathbf{x}, \boldsymbol{a} \sim \mathcal{N}(\bar{x}_{d\,a(j)}, w_d/2). \tag{5}$$

where μ_{dj} is the unknown chemical shift of resonance type d at position j, $\bar{x}_{d\,a(j)}$ is the dth element of the average of $\mathbf{x}^s_{a(j+1)}$ and $\mathbf{x}^w_{a(j)}$, and w_d is the dth diagonal element of W. Since the experimental variance is orders of magnitude smaller than the variance of the prior of $\boldsymbol{\mu}$, the contribution of the prior of $\boldsymbol{\mu}_j$ to (5) can be ignored. When K mappings are consistent with the data, inference based on the mapping with the highest posterior probability will underestimate the uncertainty in the correct mapping, as well as the differences in $\boldsymbol{\mu}_j$ under the other reasonably good alternatives. This uncertainty can be taken into account by averaging the posterior distribution of $\boldsymbol{\mu}_j$ across all candidate mappings [16, 17]

$$\mathrm{Pr}(\boldsymbol{\mu}_j|\mathbf{x}) = \sum_{k=1}^{K} \mathrm{Pr}(\boldsymbol{\mu}_j|\mathbf{x}, \boldsymbol{a}_k)\mathrm{Pr}(\boldsymbol{a}_k|\mathbf{x}) \tag{6}$$

where $\mathrm{Pr}(\boldsymbol{a}_k|\mathbf{x})$ are standardized to form a probability distribution on the set of the selected mappings. The μ_{dj} can be estimated by their posterior means, and the quality of estimation can be characterized by their posterior standard deviations:

$$\mathrm{E}(\mu_{dj}|x) = \sum_{k=1}^{K} \bar{x}_{d\,a_k(j)}\mathrm{Pr}(\boldsymbol{a}_k|\mathbf{x}) \text{ and}$$

$$\mathrm{Var}(\mu_{dj}|x) = \sum_{k=1}^{K} (w_d/2 + \bar{x}^2_{d\,a_k(j)})\mathrm{Pr}(\boldsymbol{a}_k|\mathbf{x}) - \bar{x}^2_{d\,a_k(j)} ,$$

The posterior mean of $\boldsymbol{\mu}_j$ is nothing but a weighted average of estimations according to the individual mappings. Therefore, if a spin system is mapped to the same position in all

mappings, the posterior standard deviation equals the standard deviation of the readings of the peak positions. If different spin systems are mapped to a position, the posterior deviation of $\boldsymbol{\mu}$ at this position incorporates both the uncertainty in the spin system and the experimental variance. The possibility of mapping different spin systems to a position does not necessarily imply uncertainty in the chemical shifts. Some alternative spin systems may have similar values in all or at least some resonance types, in which case the posterior standard deviation will remain close to the experimental precision. Alternatively, if the relative weight of a candidate mapping is small, the contribution of the mapping will not significantly inflate the posterior standard deviation.

In addition to the inference of $\boldsymbol{\mu}$, it is instructive to examine the posterior distribution $\Pr(\boldsymbol{a}_k|\mathbf{x})$ for the set of candidate mappings. The shape of the distribution can be used to characterize the information content in the data: the sharper the distribution, the more evidence there is in favor of the mapping with the highest posterior probability. One should not, however, confuse the inference regarding $\boldsymbol{a}$ with the inference regarding $\boldsymbol{\mu}$. As discussed in the previous paragraph, uncertainty in the candidate mappings may or may not result in uncertainty in the estimated chemical shifts.

The posterior distribution $\Pr(\boldsymbol{a}_k|\mathbf{x})$ provides a relative measure of quality for one mapping with respect to another. But it does not provide the information on how well the mapping fits the data. The scores of the candidate mappings, on the other hand, measure the goodness of fit. One can compare the scores with the α-quantiles of the corresponding χ^2_{df} distributions in order to judge the overall plausibility of the selected mappings.

4   Finding plausible candidate mappings

The proposed inferential procedure can be carried out using any algorithm which appropriately explores the probability space of candidate mappings and finds the mappings with the highest posterior probabilities. A limited number of mappings is expected to satisfy the conditions in Definition 1. Therefore, an exhaustive search is the most desirable procedure as it ensures complete examination of the search space and does not omit any mapping of interest. It has a particular advantage over greedy and best-first algorithms favoring locally optimal choices, which may not necessarily lead to overall plausible solutions and may ignore choices that do. Several exhaustive search algorithms for backbone resonance assignment have recently appeared in the literature and demonstrated the feasibility of this approach for problems of moderate size [1, 9, 20, 31]. However, the algorithms are based on scoring functions with no probability interpretation. We require an algorithm capable of exploring the probability space of candidate mappings and providing all solutions consistent with the data.

We illustrate our inferential procedure using a new algorithm of this kind. Our probability model provides information not available to the previous algorithms mentioned above. Specifically, we can evaluate the partial scores of matching and aligning any groups of spin systems with any portions of the primary sequence. In addition, the χ^2 interpretation of the scores provides an upper bound for the score of the correct mapping. Finally, the search

space can be characterized in terms of the probability of missing the mapping that generated the data. Thus we can prune the search space in a statistically sound manner, discarding from further consideration partial solutions not worth completing.

The probability model allows our algorithm to handle larger spaces of candidate mappings than what has been previously found tractable. In particular, it provides a general, consistent treatment of entirely missing spin systems. Previous algorithms only consider a missing spin system when no other matching can be found. This results in an arbitrary unequal treatment of the spin systems, as a plausible match can sometimes be successfully substituted by a match with a missing spin system. We propose an algorithm which handles all positions and all spin systems in a symmetric way. It examines the possibility of a missing spin system at any position in the sequence, and limits the search space only by a maximum allowable number of missing spin systems. A penalty (Eq. 4) discourages mappings with many missing chemical shifts, and therefore entirely missing spin systems will not appear at every position in the sequence. For most positions for which there are possible spin systems, the scores of matching and aligning are better than the corresponding penalty, and therefore a missing spin system will not appear plausible. Our current approach looks for candidate mappings with the smallest number of entirely missing spin systems; a careful choice of the maximum number of missing spin systems will be the subject of future work.

The algorithm is summarized in Fig. 3 and Fig. 4, and illustrated in Fig. 5. Rather than performing a combinatorial enumeration of individual spin systems and positions in the primary sequence as is typically done, our algorithm works at a coarser grain: it uses connected spin systems (which we call *strands*), and subsequences of the primary sequence (*windows*). It starts with an initialization step (Fig. 3(I) and Fig. 5(a)) that maps each observed spin system to each position in the sequence, subject to restrictions described in Sec. 2. A placeholder representing an entirely missing spin system is also mapped to each position. The next step (Fig. 3(II) and Fig. 5(b)) of the algorithm sets up the coarser-grained data structures for the search. It grows windows with strands sequentially starting from the first position. Specifically, the step starts by examining all possible strands that can be mapped to the first two positions, then proceeds by examining the strands that can be mapped to the first three positions, and so forth moving sequentially towards the end of the protein. Since the number of strands grows combinatorially with the number of positions examined, an execution parameter controls the maximum number of strands that can be mapped to a same set of positions. Every time the number of strands reaches the limit, the growth of the current window is stopped, and a new set of strands and a new window are started from the following position. Therefore, by design, the windows do not overlap and the number of strands in the windows is approximately equal and never exceeds the specified limit.

Not all the strands constructed during step II are kept at the end of this step. For example, a strand that is guaranteed to produce a poor score in combination with any other strand from the remaining windows (called "outer strands" in the pseudocode) can be discarded. Similarly, one can discard a strand which, in combination with other strands, will exceed the pre-specified limit on number of missing spin systems. Furthermore, the step

introduces a parameter β which provides an additional constraint on evaluation of partial mappings. Specifically, if the score of a strand exceeds the β-quantile of the corresponding χ^2 distribution, the strand is rejected. These criteria for a valid merge are summarized in Fig. 4.

The step II reduces the search space of candidate mappings. The last stage of the algorithm (Fig. 3(III)) performs an exhaustive depth-first search of the reduced space by merging the strands in the windows. This can be done sequentially starting from the first window, but a careful choice of the order of merging can greatly improve the execution time.

The parameter α in Definition 1 and β in Step II jointly limit the search space of candidate mappings. By using the conservative Bonferroni correction to multiple comparisons [15], the probability of rejecting the correct mapping for at least one position is at most $(3R+1)\alpha + 2R\beta$. Therefore, α and β can be chosen to control the familywise error rate at the desired level. Alternatively, one can undertake a more aggressive approach and choose α and β to control the False Discovery Rate (i.e. the expected proportion of positions at which the correct mapping is rejected) [5]. We believe that in this problem, the Bonferroni correction is superior to the FDR-controlling approach for two reasons. First, it is desirable to make as few incorrect rejections as possible at all positions in the sequence, and the familywise Bonferroni correction is an appropriate metric for that. Second, the Bonferroni correction is computationally inexpensive and results in a larger space of candidate mappings. Therefore, computational effort is used to explore alternative mappings rather than to calculate the FDR-based rejection thresholds.

5   Results

5.1   Data example

We illustrate our inferential procedure for resonance assignment in application to two proteins, Human Ubiquitin and the single-stranded DNA-binding cold-shock protein A (CspA) from *Escherichia coli*. Human Ubiquitin is a 76 amino acid residue protein used as a benchmark in many NMR studies. The data set containing peaks from 7 through-bond experiments and providing connectivity information for C^α, C^β and C' resonance types is publicly available from the Ubiquitin NMR Resource Web Page [30]. We manually compiled the spin systems from the observed peaks. Two expected spin systems were missing, and no extra spin systems were detected. The correct mapping of the spin systems to positions in the sequence is the one deposited in the database, and we will refer to it as the reference mapping.

The single-stranded DNA-binding cold-shock protein A (CspA) from *Escherichia coli* is a small β-sheet protein composed of 70 residues. NMR data are provided as a test for the AutoAssign program [35], and include peak lists from eight through-bond NMR experiments yielding connectivity information for C^α, C^β, and H^α. One of the experiments involves the C' resonance type. AutoAssign compiles the resonance peaks into spin systems and finds assignments for all non-proline residues except for the first two. In addition, AutoAssign

Input:

Spin systems $(\mathbf{x}_i^w, \mathbf{x}_i^s)$.

Positions j in the primary sequence.

Means $\boldsymbol{\theta}_j$ and variances Σ_j of prior distributions of resonances at each position j.

Consistency parameter α.

Maximum number of entirely missing spin systems.

Probability β limiting the search space.

Execution parameter:

Maximum number of strands mapped to a window.

Output:

Consistent mappings of spin systems to windows, as defined in Sec. 3.2.

Algorithm:

I *Initialize.*

(a) Build a table of mappings from spin systems to positions in the primary sequence.

(b) Set $maxS = q_{\chi^2}(\alpha, df)$, where df is the maximum number of data points available.

II *Construct windows.*
Let position j iterate through the primary sequence:

(a) If $j = 1$, or residue $j - 1$ is a proline, or the number of strands mapped to the current window exceeds the limit, start a new window at j. The strands mapped to the new window are simply the spin systems mapped to j in the table.

(b) Else extend the window ending at $j - 1$ by merging its strands with the spin systems mapped to j in the table and keeping only consistent extensions (Fig. 4).

III *Complete the assignment.*

(a) Depth-first merge adjacent windows, keeping consistently merged strands (Fig. 4).

(b) Whenever a complete mapping is found, set $maxS$ to $\min(maxS, -2\log(\text{posterior probability of the mapping}))$.

Figure 3: Algorithm for exhaustive search for candidate mappings.

Input:

>> Position j in the primary sequence.
>>
>> Strand ending with $\mathbf{x}_i^w$ and mapped to window ending at $j - 1$.
>>
>> Strand starting with $\mathbf{x}_i^s$ and mapped to window beginning at j.
>>
>> Parameters α, β, max missing, and $maxS$ from Fig. 3.

Output:

>> Is the merge of the input strands valid?

Algorithm:

> The merge is not valid if at least one of the following conditions holds:

1. Score of matching $\mathbf{x}_i^w$ to $\mathbf{x}_i^s$ exceeds $q_{\chi^2}(\alpha, df_{match})$.
2. Score of mapping $\mathbf{x}_i^w$ and $\mathbf{x}_i^s$ to position j exceeds $q_{\chi^2}(\alpha, df_{map})$.
3. Joint score of matching and mapping exceeds $q_{\chi^2}(\alpha, df_{match} + df_{map})$.
4. Total score of the merged strand mapped to the combined window exceeds $q_{\chi^2}(\beta, df_{total})$.
5. Number of missing spin systems in the strand plus the minimum number of missing spin systems in outer strands exceeds the limit.
6. Total score plus the minimum score in outer strands exceeds $maxS + 2\log(25)$.

Figure 4: Algorithm for testing a merge between two strands in adjacent windows. A single spin system mapped to a position in the sequence is considered as a strand of size one mapped to a window of size one.
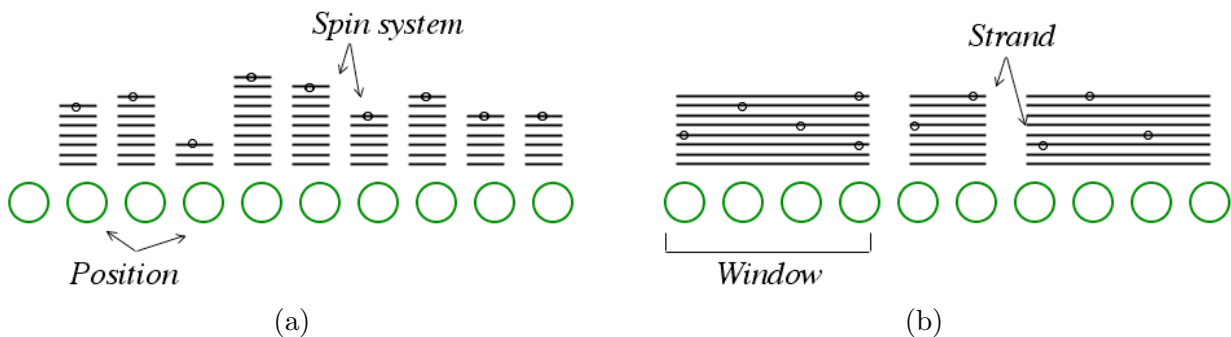
Figure 5: (a) Initial table mapping spin systems (lines) to positions in the primary sequence (circles). Circled lines are placeholders for entirely missing spin systems. (b) Strands of consecutive spin systems (lines). Circles on the lines represent placeholders for the entirely missing spin systems within the strands. Each strand covers a window in the primary sequence.

detects four extra spin systems due to noise. AutoAssign determines an assignment by matching and aligning the spin systems to positions in the sequence, and yields one complete mapping. In the following, that mapping is considered as the reference mapping. It is not possible to compare the solutions obtained by our approach for CspA to the corresponding entry in the database as our method does not handle errors in compilation of the spin systems. In this section, we analyze the uncertainty associated with the positions 3–70 assigned by AutoAssign. We investigate the impact of the two missing spin systems in the following section.

Throughout this section, we assume that the match tolerances used by AutoAssign are approximately 3 times the standard deviations of the underlying resonances for both proteins. This corresponds approximately to the standard deviation of peaks mapped to a common source by the reference mapping. We select the prior distribution for $\boldsymbol{\mu}$ by assuming their *a priori* independence across resonance types, and by using the means and variances of the protein chemical shifts in the BioMagResBank [29] as parameter estimates. The statistics provided by the database are computed on the basis of entries with no outlying observations. We ignore the H and N resonance types as they do not provide connectivity information and have little discriminatory power for alignment. Parameters α and β in the assignment algorithm were chosen to set the overall probability of rejecting the correct assignment to be at most 0.05. The assignment program was executed using a 2 GHz PowerPC G5 with 2 GB of memory.

Assignment results for Ubiquitin yield the reference mapping as the only candidate mapping consistent with the data. This demonstrates that the data set for this protein has an extremely high information content, and no uncertainty is associated with the assignment.

Assignment results for CspA yield three candidate mappings, including the reference mapping. The posterior distribution of the three selected mappings is shown in Fig. 6(a). The distribution is sharp and favors the mapping with the highest posterior probability. This
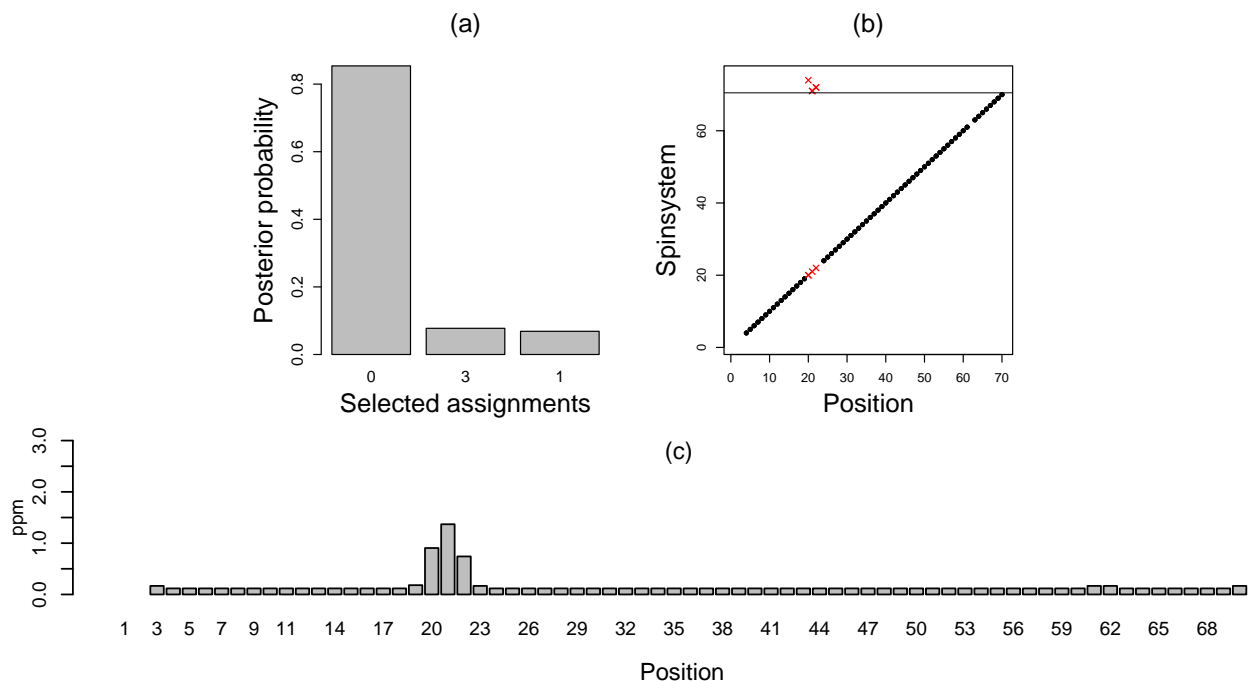
Figure 6: Assignment of CspA. (a) Posterior distribution of the selected mappings. x-axis: rank of mapping, y-axis: posterior probability. Labels under the bars show the number of positions in the sequence which distinguish the mappings from the one with the highest posterior probability. The reference solution has the second highest posterior probability. (b) Mappings of the observed spin systems. x-axis: residue position, y-axis: original mapping of the spin system with AutoAssign. Spin systems above the horizontal line are considered as extras by the original assignment. Unambiguous mappings are shown with black dots. If a spin system is unambiguously mapped to the same position as in the reference solution, the dot appears on the diagonal, at the coordinate for its position in the sequence. Unambiguous mappings are shown with red crosses. In this figure, three spin systems that were considered extras by the reference assignment can be mapped to positions 20–22. The crosses show the alternative assignments at these positions. (c) Posterior standard deviations of estimated C^α resonances. x-axis: residue position, y-axis: posterior standard deviation (in units of chemical shifts). Execution time 20 sec.

indicates high information content in the data. The candidates are overall plausible with p-values of the corresponding score functions greater than 0.9. The reference mapping has the second largest posterior probability, but the largest number of non-missing resonances. Since the original mapping was obtained using a different scoring function, we do not expect it to be the most likely *a posteriori*. Most spin systems are uniquely mapped to positions in the primary sequence, and the selected mappings differ in at most 3 positions. Fig. 6(b) details the alternative mappings. The alternatives are due to plausible mappings of extra spin systems to positions 20–22. The posterior standard deviations of the estimated resonance values for the C^α resonance type are shown in Fig. 6(c). The posterior standard deviations at the unambiguous positions are equal to the assumed experimental precision. However, the posterior standard deviations at positions 20–22 are high, indicating uncertainty in this region. This ability to uncover uncertainty is a key advantage of our approach over traditional optimization-based approaches which provide only a single best mapping.

5.2   Simulation study

In order to demonstrate the importance of inference for resonance assignment, we investigated the impact of the choices of experimental design and non-systematic sources of noise. In the following, we systematically perturb the observed spin systems of CspA and study the effect of these modifications on the assignment.

Experimental design. An experimentalist has some control over spectral resolution and experiment types. To study an effect similar to that of lower resolution, we added Gaussian noise to the observed resonances; the standard deviation of the noise is twice the assumed standard deviation of the data. As can be seen in Fig. 7, deterioration in the experimental precision results in a larger number of plausible mappings. The posterior distribution is less sharp and indicates a decrease in the information content. In addition, the overall plausibility of the selected mappings decreased. The score of the least likely of the selected mappings has a p-value of 0.53. More positions have ambiguous mappings of spin systems. However, the ambiguity in mappings does not necessarily result in ambiguity in the estimated resonances. For example, ambiguous mappings at positions 26–27 and 40 do not increase the uncertainty at these positions. Because of the reduced experimental precision, the standard deviations of unambiguous mappings are larger than in the original case.

Fig. 8 demonstrates the effect of discarding the chemical shifts of the C' and C^β resonance types. As shown in the figure, more mappings are plausible than in the original assignment. All of the selected mappings are overall plausible with p-values of their scores exceeding 0.9. Alternative spin systems are mapped to three areas in the primary sequence, namely 4–7, 19 and 70. The estimated resonances at positions 4–7 and 70 should be considered as uncertain according to the posterior standard deviations.

Non-systematic noise. The presence of extra and missing observations cannot be predicted in advance. Fig. 9 investigates the effect of extra observations by adding spin systems from a sequential segment of 20 positions of a different protein. A large number of extras can arise when, for example, a contaminated sample is used, or when the protein has a second minor conformation. As can be seen from Fig. 9, the assignment procedure selects
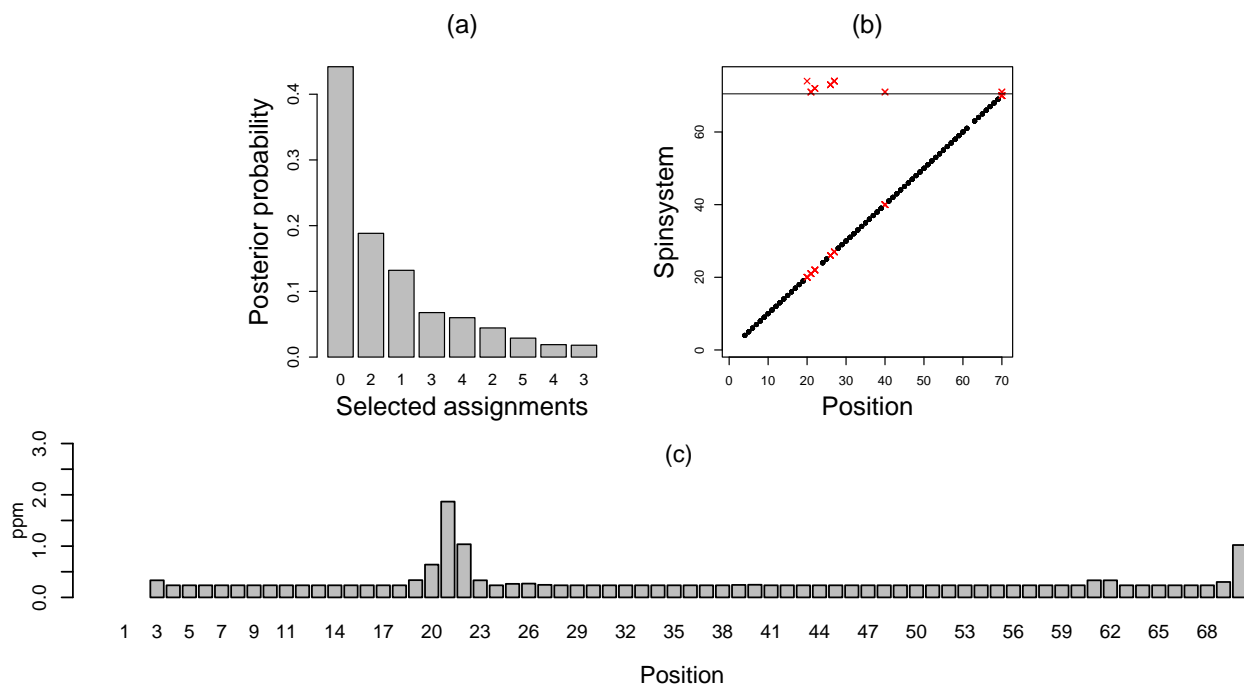
Figure 7: Assignment of peaks with reduced experimental precision. The reference solution has the second highest posterior probability. Execution time 28 sec.
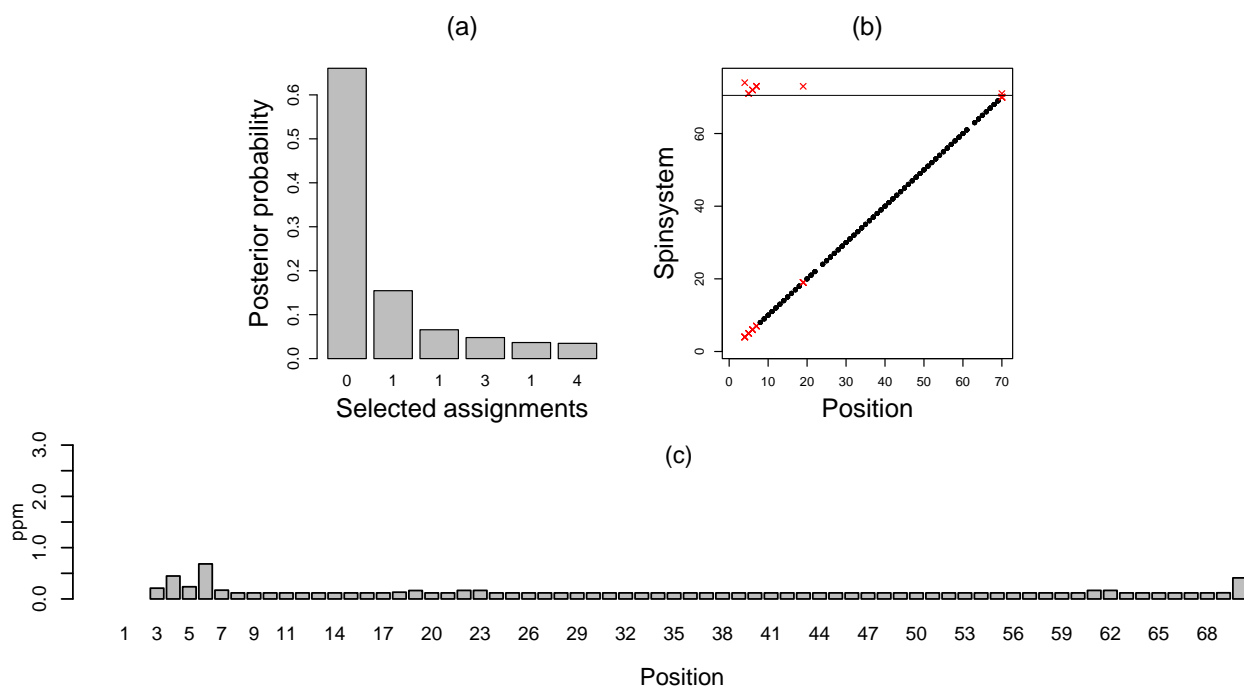

Figure 8: Assignment obtained without taking into account the C^β and C' resonance types. The reference solution has the highest posterior probability. Execution time 4 min 18 sec.
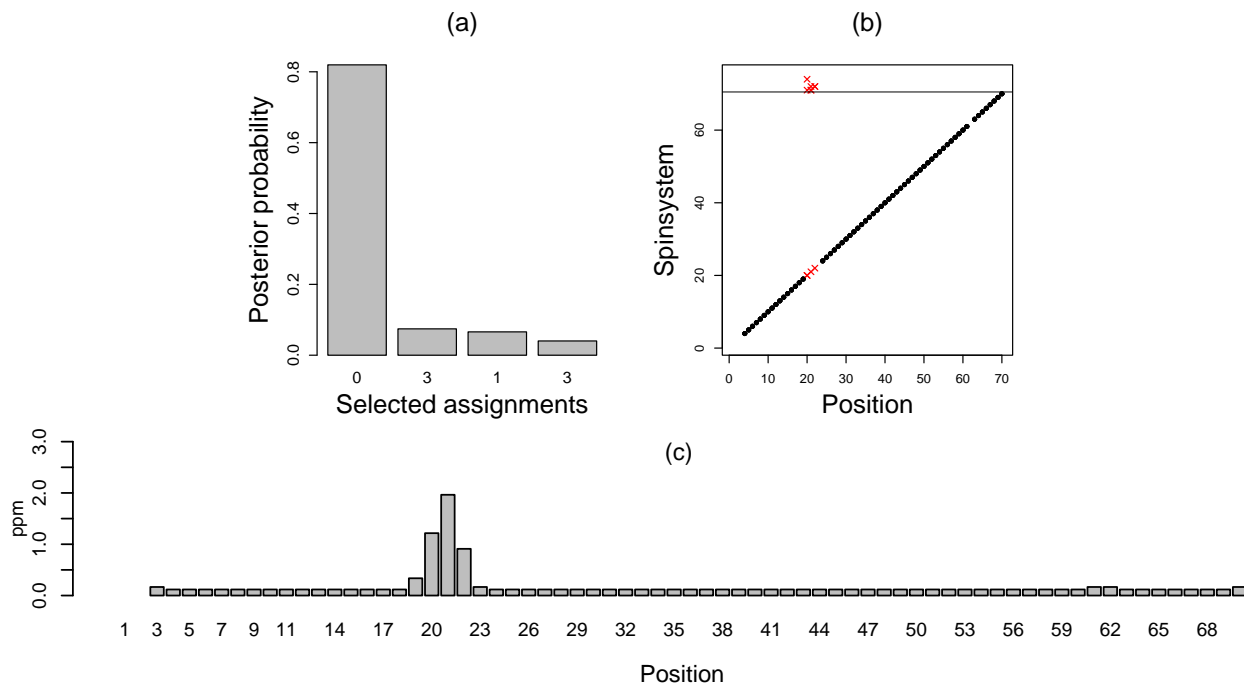
Figure 9: Assignment after introducing spin systems from a segment of a different protein. The reference solution has the second highest posterior probability. Execution time 1 min 23 sec.
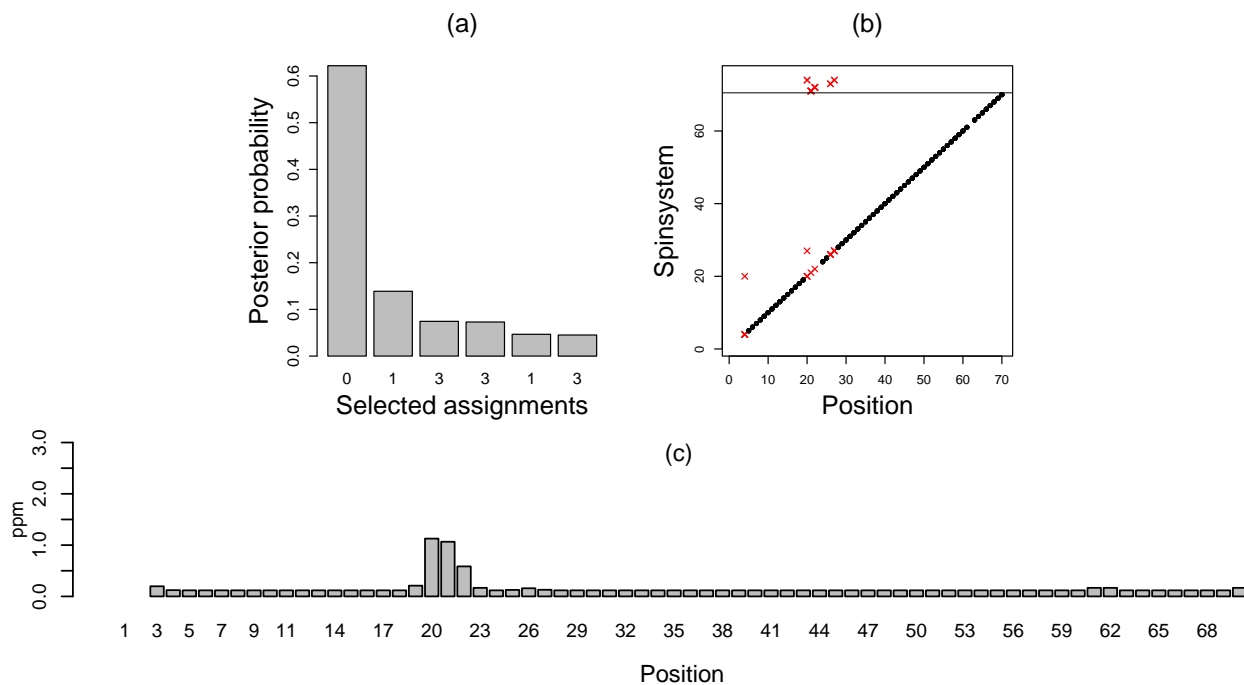

Figure 10: Assignment after removing each resonance with probability 0.1. The reference solution has the third highest posterior probability. Execution time 8 sec.
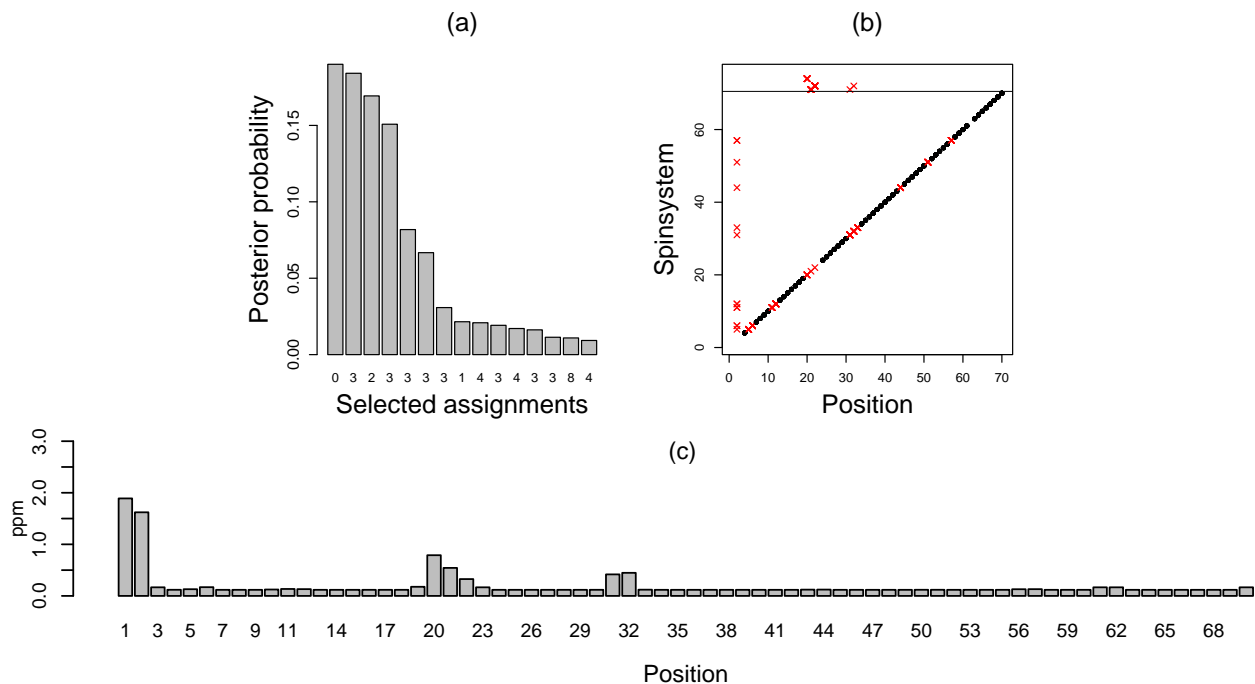
Figure 11: Assignment after adding two first positions in the sequence and introducing two missing spin systems. Red crosses at positions 5, 6, 11, 12, 33, 44, 51 and 57 indicate that a missing spin system can be successfully mapped to these positions. The reference solution is not selected. Execution time 3 min 29 sec.

	Ubiquitin			CspA, positions 3–70			CspA, positions 1–70		
Condition	Matches	Sols	Time	Matches	Sols	Time	Matches	Sols	Time
Original	7847	1	0:20	10689	3	0:20	10912	15	3:29
Prior	7147	1	0:10	9602	2	0:02	9815	20	3:41
Use N	8054	1	0:24	10509	3	0:20	10723	8	6:11
No C′	10874	2	1:14	10319	4	0:01	10539	30	3:50
Double Sd	8068	1	0:46	10594	9	0:20	10961	18	4:03
Miss 10%	15757	2	113:03	14310	6	0:08	12998	25	14:25
Extra 20	13968	4	0:56	15912	4	1:23	16180	10	106:42

Table 1: Assignments for simulated data with various complexity. The simulations were conducted under the following conditions: Original – the original data; Prior – after modifying the prior distribution of $\boldsymbol{\mu}$; Use N – including the N resonance type; No C′ – excluding C′ resonance type; Double Sd – after adding noise to the observed data; Miss 10% – after randomly removing the observed chemical shifts with probability 0.1; Extra 20 – after introducing 20 spin systems from a different protein. Matches denotes the number of consistent pairwise matches across all positions in the sequence. Sols is the number of consistent mappings found. Time is the execution time in min:sec.

one additional mapping containing an extra spin system. All mappings are plausible with p-values of their scores exceeding 0.9.

Next we study the effect of missing chemical shifts. Fig. 10 shows that removing observed resonances with a probability of 0.1 deteriorates the information content in the data and produces more plausible mappings. The alternative mappings at positions 25–26 have little impact on the estimated resonances.

Finally, we investigate the impact of entirely missing spin systems on the uncertainty in resonance assignment. To this end, we now consider the entire CspA sequence (positions 1–70), and introduce the possibility of 2 entirely missing spin systems. Our assignment algorithm treats all positions in the sequence the same, and considers a missing spin system at any position. As can be seen in Fig. 11, the presence of entirely missing spin systems can have a dramatic negative effect on the uncertainty in the assignment. We find 15 mappings consistent with the data, all with p-values greater than 0.9. This large number of candidate mappings is clearly due to the presence of the missing spin systems: an entirely missing spin system can be successfully mapped to 10 different positions in the sequence and still yield a complete mapping. The reference mapping is not selected since there exist highly plausible solutions with a non-missing spin system mapped to the second position. The alternative mappings result in uncertainty in two additional areas in the protein sequence.

Dependence of inference on the available information. Tab. 1 summarizes the simulations discussed in this section, as well as other simulations conducted using Human Ubiquitin and CspA. The first two lines in the table summarize the assignment of the same data under two prior distributions. The original assignment was completed by assuming independence of resonance types and by using the statistics from the BioMagResBank [29] as described in Sec. 3.1. The second line, denoted PRIOR, uses the prior distributions in [24] obtained by taking into account the redundancy in the database and the correlation structure across resonance types. USE N and NO C′ show the impact of respectively introducing N or removing C′. DOUBLE SD, MISS 10% and EXTRA 20 show the impact of alterations to the dataset by respectively reducing the experimental precision, randomly deleting the observed chemical shifts with probability 10%, and introducing 20 extra spin systems. As can be seen, Human Ubiquitin has a very high information content despite the presence of two entirely missing spin systems. Modifications of the assignment conditions have very little effect on the uncertainty. The original data set has only one mapping consistent with the data, and the same result would be obtained by a procedure maximizing an appropriate scoring function. Assignment of CspA, positions 3–70, represents a case of moderate information content. The uncertainty in the assignment can be affected by the assignment conditions and by artifacts in the data. CspA, positions 1–70, is a case of relatively low information content. It can be characterized by sensitivity to the assumptions and to the experimental conditions. Therefore, great care must be applied when selecting the assumptions and designing the experiments. Due to the stochastic variation in the data, an incorrect assignment may appear optimal. It is dangerous in this case to use a procedure optimizing a scoring function, and not to consider the plausible alternatives.

The information content in the data can also be characterized by the size of the search

space of candidate mappings. One such measure of the space is the number of consistent pairwise matches mapped to positions in the sequence, as shown in the column MATCHES in Tab. 1. As can be seen, the size of the space increases with the noise, and the increasing search space has a direct effect on the execution time. The assignment algorithm is fast for up to moderate-sized search spaces, but slows down precipitously for larger ones.

6 Comparison to prior work

The development of automated methods for backbone resonance assignment has recently become an active area of research [1, 2, 3, 4, 7, 9, 11, 12, 13, 14, 18, 19, 20, 22, 23, 26, 31, 35]. In this section, we briefly describe some of the existing methods and contrast them with our model-based approach. We note that, like many existing methods, we take spin systems as the input data; properly constructing the spin systems is itself an important topic with implications for assignment.

Assessment of uncertainty. The main contribution of this paper is a formal approach to assessment of uncertainty in backbone resonance assignment. Most existing methods view assignment as a deterministic optimization problem. They yield a single mapping of the observed spin systems to positions in the sequence, and provide no assessment of uncertainty associated with the result. Such optimization procedures are appropriate when the data set is extremely informative, and when no plausible alternative to the "best" candidate mapping exists. However, as demonstrated here, as well as in other cases we have studied, several candidate mappings are generally plausible. One must take this uncertainty into account in order to obtain reliable assignments and avoid erroneous conclusions.

Some other approaches also yield sets of candidate mappings. If the same spin system is always mapped to a particular position in the sequence, then they consider the chemical shifts at the position to be certain; else they consider the chemical shifts uncertain. These methods make no distinction between uncertainty in mapping spin systems, and uncertainty in determination of chemical shifts. As discussed in Sec. 3.3, this approach does not correctly represent the uncertainty because 1) alternative spin systems may have similar values, and 2) the relative weights of the mappings other than the "best" one may be small. The existing methods use score functions which do not have a probabilistic interpretation. Therefore inference of the unknown chemical shifts cannot be carried out, and the relative importance of the candidate mappings is difficult to judge.

A distinctive feature of our approach is that a probability model of sources of noise is an integral part of the assignment procedure. A global scheme of scoring, and a penalty for missing observations, allow us to directly compare assignments, and therefore perform inference regarding the unknown chemical shifts. In addition, the probabilistic interpretation of the score enables us to characterize the information content in the data and the overall plausibility of a candidate mapping.

Finding the candidate mappings. Our probability model enables a unique approach to finding candidate mappings. Specifically, one can assess not only the plausibility of a spin system at a position in the sequence, but also of any partial mapping of groups of spin

systems. The plausibility of a partial mapping (and the implications for the plausibility of a completed assignment) has not been previously used by any other assignment procedure. This results in a significant reduction of the search space, and therefore allows an exhaustive search of candidate mappings on problems which were not tractable by previous exhaustive search methods. Furthermore, the χ^2 interpretation provides an estimate for the upper bound on the total score of the correct assignment, and the penalty allows a meaningful comparison of mappings with different numbers of missing observations.

Matching spin systems. All existing methods define the total score as the sum of the individual contributions, and thus implicitly assume independence of the observed chemical shifts across positions in the sequence and across resonance types. Furthermore, by employing constant match tolerances or constant parameters for bell-shaped functions that score matches, all methods implicitly use the assumption of constant variance of chemical shifts. Some algorithms progressively increase match tolerances when no match is found under a given tolerance, resulting in an arbitrary, unequal treatment of spin systems. We follow the existing approaches by assuming constant experimental variance of the observed chemical shifts, but use a flexible scoring function which considers the plausibility of a match *jointly* for all resonance types. Therefore, a relatively loose match of one resonance type can be considered plausible if it is compensated for by very tight matches of the other resonance types. The parameter α controls the overall quality of an acceptable match, and has a probabilistic interpretation which is not available to the other methods.

Aligning spin systems. The vast majority of the existing methods use a Normal characterization of the chemical shifts deposited to a database. In particular, the program Mapper [12] uses an almost identical alignment score and its χ^2 interpretation. The score can be derived from our Bayesian standpoint under the assumption that the resonance types are *a priori* independent. However, Mapper assumes pre-compiled chains of spin systems as the input, and is not concerned with the quality of matching the spin systems. It provides no formal method for statistical inference. Our approach provides a unifying scoring system for aligning and matching, and is capable of incorporating any Normality-based characterization of the prior distributions.

7 Discussion and future work

To the best of our knowledge, our approach is the first to provide formal statistical inference for backbone resonance assignment. We quantify the uncertainty in the assigned chemical shifts in terms of their posterior standard deviations. We also characterize the information content in the data with a posterior distribution for the set of candidate mappings, and by comparing the scores of the mappings to the corresponding χ^2 distribution. The method is fully automated and does not require human intervention. It is capable of incorporating different prior characterizations of chemical shifts and different experimental variances. It requires only two tuning parameters, α and β, which control the search space of candidate mappings and are easily interpretable.

We believe that quantification of uncertainty is the key for producing reliable automated

assignments. The use of a single candidate mapping with no quantification of uncertainty can result in false optimism in the quality of reported chemical shifts. This may lead to erroneous conclusions which in turn may propagate and accumulate through subsequent stages of NMR-based analyses. Reporting posterior standard deviations of the determined chemical shifts will help prevent such errors. Specific applications of the results of our method include the following. 1) Use only the assigned chemical shifts for which the posterior standard deviations are close to the experimental precision. 2) Design additional NMR experiments based on the posterior standard deviations of chemical shifts, e.g. using isotopic labeling techniques to probe uncertain residues or residue types, or conducting experiments that focus on the resonance types containing most of the uncertainty. 3) Report posterior standard deviations of chemical shifts when depositing NMR assignments in public databases, so that entries are annotated according to their uncertainty and the underlying information content. For example, all things being equal, the distributions will help distinguish between an assignment based on, say, ten resonance experiments versus one based on just three, or between assignments obtained with different spectral resolution. 4) When a single candidate mapping must be used, assess the posterior distribution of the mappings in order to determine the relative support for the best mapping. At the same time, the score of that mapping will help judge its quality of fit with the data.

We plan to improve our method in a number of ways.

1. The space of potential mappings is combinatorially large, and the complexity of the problem grows exponentially with the protein size. Therefore, even the most efficient algorithms for exhaustive search are limited in their use. Inferential algorithms employing stochastic search are needed in order to explore large spaces of candidate mappings.

2. The probability model is based on the assumption of constant variance of resonances associated with the observed spin systems. However, the resonances are obtained using a variable number of peaks in the spectra. The quality of the observed peaks varies, and so does the uncertainty associated with their locations. The assumption may be relaxed in future work by careful modeling of the noise associated with individual peaks.

3. The current approach considers the candidate mappings with the minimum number of entirely missing spin systems. However, the penalty (Eq. 4) brings to a common scale mappings with any number of missing spin systems, and mappings with more missing spin systems can in principle have higher posterior probabilities. More research is needed in order to select an appropriate maximum allowable number of entirely missing spin systems.

4. Additional information, e.g. regarding the secondary structure of the protein, can help reduce the search space of candidate mappings, as well as the uncertainty associated with the assigned chemical shifts. Researchers have attempted to incorporate this

information by means of *prediction* of secondary structure elements [31]. However, assessment of uncertainty in the assigned chemical shifts in this case must incorporate the uncertainty in the prediction. More work is needed to correctly assess the uncertainty in this case.

The current version of our program, written in Java, can be freely obtained for academic use by request from the authors.

8   Acknowledgment

References

[1] M. Andrec and R. Levy. Protein sequential resonance assignments by combinatorial enumeration using $^{13}C^{\alpha}$, chemical shifts and their $(i,i-1)$ sequential connectivities. *Journal of Biomolecular NMR*, 23:263–270, 2002.

[2] H. S. Atreya, S. C. Sahu, K. V. R. Chary, and G. Govil. A tracked approach for automated NMR assignments in proteins (TATAPRO). *Journal of Biomolecular NMR*, 17:125–136, 2000.

[3] C. Bailey-Kellogg, A. Widge, J. J. Kelley III, M. J. Berardi, J. H. Bushweller, and B. R. Donald. The NOESY Jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *Journal of Computational Biology*, 7:537–558, 2000.

[4] C. Bartels, P. Güntert, M. Billeter, and K. Wüthrich. GARANT – a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. *Journal of Computational Chemistry*, 18(1):139–149, 1997.

[5] Y. Benjamini, Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, Ser B. 57:289-300, 1995.

[6] S. E. Brenner. A tour of structural genomics. *Nature Reviews Genetics*, 2:801–809, 2001.

[7] N. E. G. Buchler, E. P. R. Zuiderweg, H. Wang, and R. A. Goldstein. Protein heteronuclear NMR assignments using mean-field simulated annealing. *Journal of Molecular Resonance*, 125:34–42, 1997.

[8] J. Cavanagh, W. J. Fairbrother, A. G. Palmer III, and N. J. Skelton. *Protein NMR Spectroscopy*. Academic Press, 1996.

[9] B. E. Coggins and P. Zhou. PACES: Protein sequential assignment by computer-aided exhaustive search. *Journal of Biomolecular NMR*, 26:93–111, 2003.

[10] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, and A. Bax. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *Journal of Biomolecular NMR*, 6:277–293, 1995.

[11] W. Gronwald, L. Willard, T. Jellard, R. Boyko, K. Rajarathnam, D. S. Wishart, F. Sonnichsen, and B. Sykes. CAMRA: Chemical shift based computer aided protein NMR assignments. *Journal of Biomolecular NMR*, 12:395–405, 1998.

[12] P. Güntert, M. Saltzmann, D. Braun, and K. Wüthrich. Sequence-specific NMR assignment of proteins by global fragment mapping with program Mapper. *Journal of Biomolecular NMR*, 17:129–137, 2000.

[13] T. K. Hitchens, J. A. Lukin, Y. Zhan, S. A. McCallum, and G. S. Rule. MONTE: An automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins. *Journal of Biomolecular NMR*, 25:1–9, 2003.

[14] T. K. Hitchens, S. A. McCallum, and G. S. Rule. Data requirements for reliable chemical shift assignments in deuterated proteins. *Journal of Biomolecular NMR*, 25:11–23, 2003.

[15] Y. Hochberg, A. C. Tamhane. *Multiple Comparison Procedures*. John Wiley & Sons, 1987.

[16] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

[17] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[18] M. Leutner, R. M. Gschwind, J. Liermann, C. Schwarz, G. Gemmecker, and H. Kessler. Automated backbone assignment of labeled proteins using the threshold accepting algorithm. *Journal of Biomolecular NMR*, 11:31–43, 1998.

[19] K.-B. Li and B. C. Sanctuary. Automated resonance assignment of proteins using heteronuclear 3D NMR. 1. Backbone spin systems extraction and creation of polypeptides. *Journal of Chemical Informatics and Computational Sciences*, 37:359–366, 1997.

[20] G. Lin, D. Xu, Z.-Z. Chen, T. Jiang, and Y. Xu. A branch-and-bound algorithm for assignment of protein backbone NMR peaks. In *First IEEE Bioinformatics Conference*, pages 165–174, 2002.

[21] J. Liu. Bayesian modeling and computation in bioinformatics research. In T. Jiang, Y. Xu, and M. Zhang, editors, *Current topics in computational biology*, pages 11–44. MIT Press, 2002.

[22] J. A. Lukin, A. P. Gove, S. N. Talukdar, and C. Ho. Automated probabilistic method for assigning backbone resonances of (^{13}C, ^{15}N)-labeled proteins. *Journal of Biomolecular NMR*, 9:151–166, 1997.

[23] D. Malmodin, C. Papavoine, and M. Billeter. Fully automated sequence-specific resonance assignments of heteronuclear protein spectra. *Journal Biomolecular NMR*, 27:67–69, 2003.

[24] A. Marin, T. Malliavin, P. Nicholas, M.-A. Delsuc. From NMR chemical shifts to amino acid types: investigation of the predictive power carried by nuclei. Poster at the Gordon Research Conference on computational aspects of biomolecular NMR, to appear in *Journal of Biomolecular NMR*, 2004.

[25] G. T. Montelione, D. Zheng, Y. J. Huang, K. Gunsalus, and T. Szyperski. Protein NMR spectroscopy in structural genomics. *Nature America Supplement*, 2000.

[26] H. N. B. Moseley and G. T. Montelione. Automated analysis of NMR assignments and structures for proteins. *Current Opinions in Structural Biology*, 9:635–642, 1999.

[27] S. J. Press. *Subjective and Objective Bayesian Statistics : Principles, Models, and Applications*. John Wiley & Sons, 2002.

[28] D. B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172, 1984.

[29] B. R. Seavey, E. A. Farr, W. M. Westler, and J. Markley. A relational database for sequence-specific protein NMR data. *Journal Biomolecular NMR*, 1:217–236, 1991. http://www.bmrb.wisc.edu.

[30] The Ubiquitin NMR Resource Page http://www.biochem.ucl.ac.uk/bsm/nmr/ubq/

[31] X. Wan, D. Xu, C. Slupsky, and G. Lin. Automated protein NMR resonance assignments. In *Proceedings of the Computational Systems Bioinformatics*, 2003.

[32] D.S. Wishart, B.D. Sykes, and F.M. Richards. The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. *Biochemistry*, 31(6):1647–1651, 1992.

[33] K. Wüthrich. *NMR of Proteins and Nucleic Acids*. John Wiley & Sons, 1986.

[34] H. Zhang, S. Neal, D. S. Wishart. RefDB: a database of uniformely referenced protein chemical shifts. *Journal of Biomolecular NMR*, 25:173–195, 2003.

[35] D.E. Zimmerman, C.A. Kulikowski, Y. Huang, W. Feng, M. Tashiro, S. S. Shimotaka-hara, C. Chien, R. Powers, and G. T. Montelione. Automated analysis of protein NMR assignments using methods from artificial intelligence. *Journal of Molecular Biology*, 269:592–610, 1997.