Reconsidering Complete Search Algorithms for Protein Backbone NMR Assignment

Olga Vitek [a], Chris Bailey-Kellogg [b], Bruce Craig [a], Paul Kuliniewicz [c], Jan Vitek [c]

[a] Department of Statistics, Purdue University, West Lafayette, IN 47907, USA
[b] Department of Computer Science, Dartmouth College, Hanover, NH 03755, USA
[c] Department of Computer Sciences, Purdue University, West Lafayette, IN 47907, USA

ABSTRACT

Motivation: Nuclear magnetic resonance (NMR) spectroscopy is widely used to determine and analyze protein structures. An essential step in NMR studies is backbone resonance assignment, which maps individual atoms to experimentally measured resonance frequencies. Performing assignment is challenging due to the noise and ambiguity in the spectra. While automated procedures have been investigated, by-and-large they are struggling to gain acceptance due to inherent limits in scalability and unacceptable numbers of assignment errors.

To have confidence in the results, an algorithm should be *complete*, i.e. able to identify all solutions consistent with the data, including all arbitrary configurations of extra and missing peaks. The ensuing combinatorial explosion in the space of possible assignments case has led to the perception that complete search is hopelessly inefficient and cannot scale to realistic proteins.

Results: This paper presents a complete *branch-contract-and-bound* search algorithm for backbone resonance assignment. The algorithm controls the search space by hierarchically agglomerating partial assignments and employing statistically sound pruning criteria. It considers *all* solutions consistent with the data, and uniformly treats all combinations of extra and missing data. We demonstrate our approach on experimental data from 5 proteins ranging in size from 70 to 154 residues. The algorithm assigns over 95% of the positions with over 98% accuracy. We also present results on simulated data from 259 proteins from the RefDB database, ranging in size from 25 to 257 residues obtained in a fully automated manner. The median computation time is 1 minute, and the assignment accuracy is over 99%. These results demonstrate that complete search not only has the advantage of guaranteeing fair treatment of all feasible solutions, but is efficient enough to be employed effectively in practice.

Availability: The MBA$_2$ software package and data sets can be obtained from the contact author.

Contact: ovitek@stat.purdue.edu

1 INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy is an experimental method capable of determining three-dimensional structures of proteins in atomic detail under nearly physiological conditions. Some 15%-20% of new protein structures are currently determined by NMR, and the rate is likely to grow (Montelione et al., 2000). A bottlenecks in NMR studies is *backbone resonance assignment*, a computational procedure establishing values of the *chemical shifts* of the atoms of the protein backbone. Chemical shifts can be viewed as magnetic signatures of the atoms, and are extensively used in analyses of structure, dynamics, and molecular interactions.

Backbone resonance assignment typically uses a set of three-dimensional NMR experiments. An example of such an experiment is the HN(CO)CA shown in Fig. 1(a), which magnetically correlates a bonded pair of H^N–N backbone nuclei with the C^α nucleus of the preceding residue, and yields a three-dimensional spectrum. Peaks in the spectrum indicate the triples of H^N–N–C^α nuclei that exhibit magnetic interactions. The coordinates of the peaks are the chemical shifts of the nuclei. Each HN(CO)CA peak records signals from two neighboring residues and therefore captures *sequential* interactions. Another three-dimensional experiment, HNCA, magnetically correlates the bonded H^N–N pairs with the C^α either of the preceding residue (as in the HN(CO)CA), or of the same residue. It yields approximately twice as many three-dimensional peaks as the HN(CO)CA, gathering both sequential and *within-residue* magnetic interactions. Similar NMR experiments can be designed to correlate the H^N–N pairs with C^β, H^α, and C′. Coordinates of peaks from the various experiments can be combined by reference to shared coordinates of the H^N–N resonance types. The resulting *spin systems*, shown in Fig. 1(b), contain chemical shifts of the anchor H^N–N nuclei, of other backbone nuclei within the same residue, and of nuclei in the sequentially preceding residue.

While NMR studies assume that the primary sequence is known, the spectra provide no information about which position in the sequence generated a particular chemical shift. This must be inferred from the observed spin systems by the process of backbone resonance assignment. A typical resonance assignment procedure (Moseley and Montelione, 1999) searches for *mappings* between spin systems and positions in the sequence that satisfy the following constraints (Fig. 1(b)): (**1**) For two spin systems mapped to adjacent positions, the within-residue chemical shifts of the first *match* the sequential chemical shifts of the second. (**2**) Each spin system mapped to a position is *aligned* with the amino acid type, meaning that its chemical shifts are consistent with the expected values for the amino acid. (**3**) Each spin system is mapped to at most one position in the protein sequence, and each position is mapped to at most one spin system.

Noise and ambiguity in the spectra reduce the effectiveness of these constraints in the resonance assignment process. First, peak coordinates are uncertain, so approximate matches in constraint (1) must be allowed within pre-specified tolerance values. All matches must be allowed in the case of a missing chemical shift. Second, the
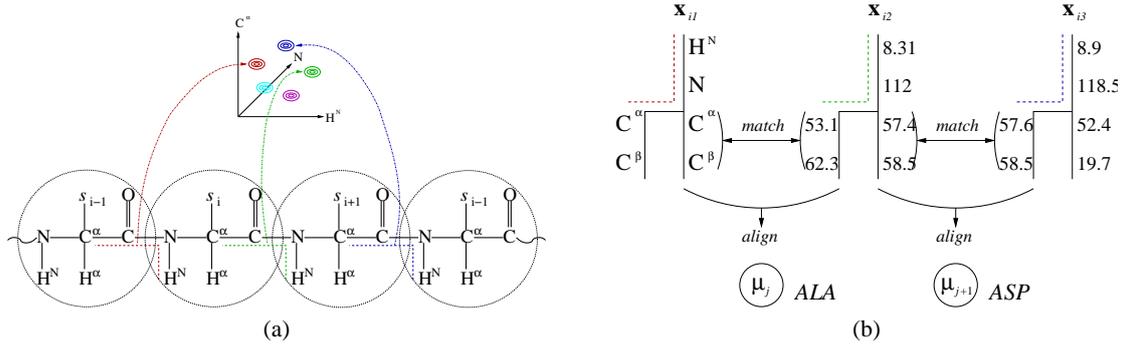
Fig. 1. NMR data. (**a**) The HN(CO)CA experiment correlates the H^N-N pair from one residue with the C^α from the preceding residue. Spectral peaks capture various such interactions. Coordinates of the peaks are the chemical shifts of the involved atoms. (**b**) Compilation of peaks from a set of experiments yields spin systems $\mathbf{x}_i$ that represent chemical shifts of atoms within a residue and in the sequentially preceding residue. Resonance assignment algorithms proceed by matching the chemical shifts of pairs of spin systems and aligning them at adjacent positions in the primary sequence.

ranges of chemical shifts for an amino acid type in constraint (2) are fairly broad, and usually allow multiple mappings of spin systems to a position. Third, positions in the sequence can have entirely missing spin systems, and the number and identity of such positions is unknown. At the same time, observed data can be extraneous.

These artifacts of noise and ambiguity result in a combinatorial explosion of the search space of candidate mappings. Consequently, exhaustive search algorithms (Andrec et al., 2001; Lin et al., 2002; Vitek et al., 2004) have been dismissed as impractical for anything but small proteins and "clean" datasets. Semi-automated procedures such as CAMRA (Gronwald et al., 1998), MAPPER (Güntert et al., 2000) and PACES (Coggins and Zhou, 2003) require human intervention to manage the search space. Fully automated approaches gain in scalability but compromise accuracy or efficiency by using, e.g., best-first search (Zimmerman et al., 1997), approximation algorithms (Chen et al., 2003), global and local stochastic optimization TATAPRO (Atreya et al., 2000), MONTE (Hitchens et al., 2003), MARS (Jung and Zweckstetter, 2004).

In order to fully characterize the confidence in an assignment, an algorithm must be *complete*. That is, it must be able to identify *all* solutions consistent with the data, including those with arbitrary configurations of matches and placements of extra and missing peaks. It is not sufficient to focus on optimization for just the "best" solution, since the ranking may be sensitive to small details in the method used to evaluate the quality of the satisfaction of the constraints. On the other hand, it is not appropriate to treat all solutions as plausible, as statistical scoring models can provide estimates of quality and indicate that some solutions are clearly inconsistent with the data. Finally, it is dangerous to fix "unambiguous" chains of matched spin systems. This does not appropriately represent our uncertainty regarding the process that generated the data as a "break" due to an entirely missing spin system can in principle appear at any position.

We present here the first efficient algorithm that performs a complete search for backbone resonance assignment. It uniformly treats all matches and combinations of extra and missing data, and returns *all* assignments that are statistically consistent with the data. Our branch-contract-and-bound algorithm explores the space of admissible solutions, not a single solution at a time, but in groups of partial solutions, so that entire sets of infeasible solutions can be ruled

out simultaneously. The algorithm branches on choices of restrictions on missing data and on selections of partial mappings, and prunes according to both local and global statistical criteria. Since missing data essentially act as wildcards, we explore all possible combinations of missing by gradually increasing their number until solutions are found. We employ a Bayesian probability model for NMR spectra (Vitek et al., 2004) which serves as a scoring function in the search for candidate mappings. This model appropriately assesses uncertainty, is amenable to formal statistical inference, and contributes greatly to the high accuracy of our algorithm.

2 METHODS

Scoring Function

Consider a primary sequence of R residues. Let $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_R)$ denote the unknown "true" chemical shifts of the backbone nuclei of the protein. Here each $\boldsymbol{\mu}_j$ is a vector composed of individual chemical shifts μ_{tj} for each resonance type $t = 1, \ldots, T$ at position j. The $\boldsymbol{\mu}_j$ are the parameters of interest, and the goal of the backbone resonance assignment is to estimate these values. The input data are I observed spin systems $\mathbf{x} = \{(\mathbf{x}_1^s, \mathbf{x}_1^w), \ldots, (\mathbf{x}_I^s, \mathbf{x}_I^w)\}$, where $\mathbf{x}_i^s$ is the vector of sequential chemical shifts x_{ti}^s, and $\mathbf{x}_i^w$ is the vector of within-residue chemical shifts x_{ti}^w, over resonance type t. 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.

Let $\boldsymbol{a} = (a_1, \ldots, a_R)$ be a candidate mapping of the observed spin systems to positions in the primary sequence. Here $a_j = i$ if $\mathbf{x}_i^w$ is mapped to position j (or equivalently $\mathbf{x}_i^s$ is mapped to position $j - 1$). We will also use the notation $\boldsymbol{a} = (j, i)$ to emphasize the relationship between positions in the sequence and spin systems. A candidate 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. According to the Bayesian paradigm, comparison between candidate mappings requires 1) specification of probability distribution of the observed data $\mathbf{x}$; 2) specification of prior distributions of the unknown parameters; 3) integrating out the unknowns with respect to the prior distributions; and 4) calculation of posterior probabilities $\Pr(\boldsymbol{a}|\mathbf{x})$
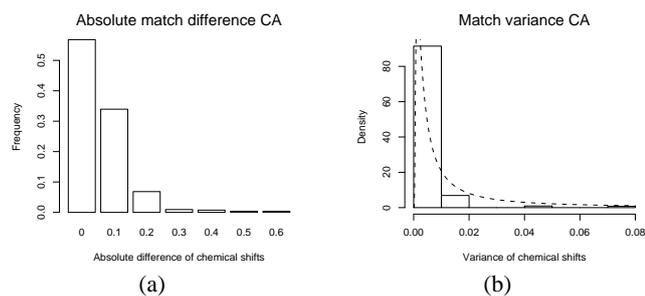
Fig. 2. (a) Histogram of absolute match differences of C^α resonance type combined from six AutoAssign data sets. (b) Histogram of estimates of experimental variance obtained as a transformation of match differences. Dashed line is a fitted Scaled Inverse χ^2 distribution with 1 degree of freedom.

Fig. 3. (a) Plausibility of a match between two C^α chemical shifts. Dashed line: scaled Cauchy density in Eq. 1. Solid line: a Normal density where the standard deviation is equal to a third of the standard match tolerance of the C^α resonance type, namely $0.5/3 = 0.17$. The latter is used by resonance assignment methods such as MARS. (b) Plausibility of aligning C^α and C^β chemical shifts to an Alanine residue. Solid lines are level curves of the multivariate Normal distribution in Eq. 1.

by applying Bayes theorem. The remainder of this section details these steps.

Likelihoods. We view the observed chemical shifts $\mathbf{x}$ as noisy readings from the unknown true chemical shifts $\boldsymbol{\mu}$. We assume that errors of the readings are independent across positions and across resonance types, and are Normally distributed. Specifically,

$$\mathbf{x}^s_{a_j} \mid \boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{a} \sim \mathcal{N}(\boldsymbol{\mu}_{j-1}, V_{j-1}) \text{ and } \mathbf{x}^w_{a_j} \mid \boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{a} \sim \mathcal{N}(\boldsymbol{\mu}_j, V_j)$$

Matrices V_j are unknown experimental variances that need not be identical for all j. The independence across resonance types implies that V_j are diagonal, and we denote the non-zero elements as v_{tj}^2.

Priors for $\boldsymbol{\mu}$, $\boldsymbol{V}$ and $\boldsymbol{a}$. A prior distribution for the chemical shifts $\boldsymbol{\mu}_j$ has been proposed by Marin *et al.* (Marin et al., 2004), and is a result of a comprehensive study of entries in the database BioMagResBank (Seavey et al., 1991). The distribution is residue-type specific and takes into account the over-representation of certain protein sequences in the database as well as the correlation of chemical shifts within a residue type. Formally, $\boldsymbol{\mu}_j \sim \mathcal{N}(\boldsymbol{\theta}_j, \Sigma_j)$ where $\boldsymbol{\theta}_j$ is a known vector, and Σ_j is a known non-diagonal matrix.

The prior distribution of V_j relaxes the stringent assumption of constant and known experimental variances made in our previous work (Vitek et al., 2004). It is obtained by examining the estimates of experimental variances for each resonance type, namely $\frac{1}{2}(x^s_{t\,a_{j+1}} - x^w_{t\,a_j})^2$, in six data sets provided as a test to the AutoAssign program (Zimmerman et al., 1997). As shown in Fig. 2, we fit the histograms with Scaled Inverse χ^2 distributions with 1 degree of freedom having densities

$$f(v_{tj}^2) = \frac{1}{\sqrt{2\pi}v_{tj}^3} \exp\left(-\frac{S_t^2}{2v_{tj}^2}\right)$$

The scale parameters S_t^2 depend on the resonance type, and the specific values are 0.0016 ppm^2 for C′, 0.004 ppm^2 for C^α, 0.005 ppm^2 for C^β and 0.00005 ppm^2 for C^α. The choice of 1 degree of freedom comes from the fact that all experimented variances are estimated on the basis of two data points.

The prior distribution of mappings $\boldsymbol{a}$ is used to determine the number of entirely missing spin systems (or, equivalently, the number of extra spin systems) in the data set. By analogy with the model
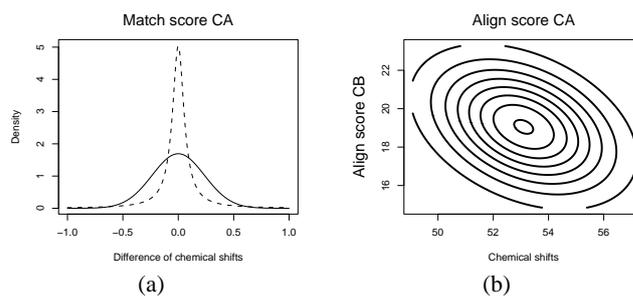
selection literature, we use Bayesian Information Criterion (BIC) weights to penalize mappings with an excessive number of extra spin systems. Specifically, $\log \Pr(\boldsymbol{a}) \propto (\log N) \cdot R'$ for mappings where N is the total number of observed chemical shifts, and R' is the number of chemical shifts considered as noise. Furthermore, we assume a uniform prior distribution of missing spin systems in the sequence conditionally on their total number. This appropriately represents our uncertainty in the physical process producing missing spin systems, but requires a search algorithm considering the possibility of a missing spin system at each position in the sequence.

Marginal likelihoods. The marginal likelihood $\Pr(\mathbf{x}|\boldsymbol{a})$ of the data given a mapping can be obtained by integrating out the unknown $\boldsymbol{\mu}_j$ and V_j with respect to their prior distributions. In our case,

$$\Pr(\mathbf{x}|\boldsymbol{a}) \approx \prod_{j=1}^{R} \prod_{t=1}^{T} \mathcal{C}\left(\frac{1}{\sqrt{2}S_t}\left(x^s_{t\,a_{j+1}} - x^w_{t\,a_j}\right)\right)$$
$$\cdot \prod_{j=1}^{R} \phi\left(\Sigma_j^{-\frac{1}{2}}(\bar{\mathbf{x}}_{a_j} - \boldsymbol{\theta}_j)\right), \quad (1)$$

where $\mathcal{C}$ denotes the density of the standard Cauchy distribution, ϕ denotes the density of the standard multivariate Normal distribution, and $\bar{\mathbf{x}}_{a_j}$ is the average of $\mathbf{x}^s_{a_{j+1}}$ and $\mathbf{x}^w_{a_j}$. The first term in Eq. 1 evaluates the plausibility of the match at a position j. As shown in in Fig. 3(a), it gives more weight to tight matches than other scoring functions, but has heavier tails. The second term in Eq. 1 evaluates the plausibility of the alignment. As shown in in Fig. 3(b), it takes into account both range and correlation of the chemical shifts.

Posterior probabilities. Through the application of the Bayes theorem, $\Pr(\boldsymbol{a}|\mathbf{x}) \propto \Pr(\mathbf{x}|\boldsymbol{a}) \cdot \Pr(\boldsymbol{a})$. The posterior probabilities are scores used to compare the candidate mappings.

Algorithm for Complete Search

We develop here a complete, fully automated, backbone resonance assignment algorithm. Our algorithm explores the space of all plausible assignments with a branch-and-bound (more precisely, branch-contract-and-bound) search technique, illustrated in Fig. 4. Nodes in a search tree compactly represent partial assignment solutions. The search recursively expands "promising" nodes, eventually
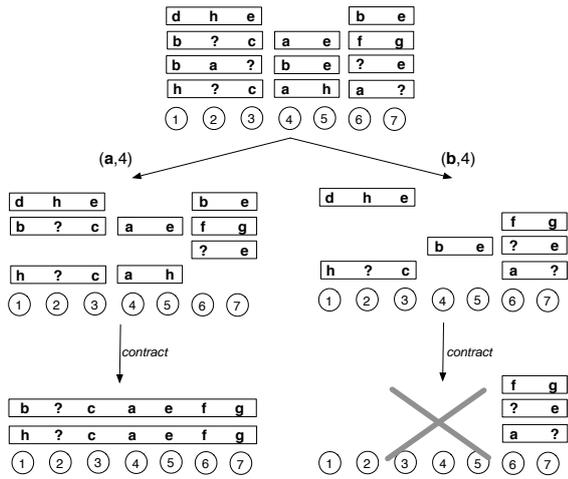
Fig. 4. Branch-contract-and-bound. here illustrate the algorithm. The topmost space consists of three windows covering, respectively, position (1-3), (4-5) and (6-7) with, respectively 4, 3 and 4 different strands. Letters **a** – **h** denote spin systems, **?** are missing observations. We show two branching steps. The left-hand side, fixes **a** at position 4. As a result, the space is pruned of strands that conflict with this assignment. The contract step yields to complete mappings (**b?caefg**) and (**h?caefg**) which differ only in one position. The right-hand side branches on (**b**, 4), and after the contraction, we find no consistent mappings. This branch is thus a dead end.

identifying entire assignments at the leaves. The expansion of a node *branches* on the possible placements for individual spin systems, as well as constraints on the number of missings. Expanded nodes are evaluated and pruned according to *bounds* that test consistency with the data according to the statistical scoring criteria, as well as plausibility relative to other solutions. In order to enhance the effectiveness of the bounds, we employ an additional *contraction* step between branching and bounding. Contraction takes advantage of the reduced combinatorics in the context of a particular branch, and generates combinations of partial solutions to be tested by the bound.

The practical utility of branch-and-bound algorithms critically depends on the effectiveness of the branching and bounding steps, and naïve approaches generally do not scale. Below we discuss the particular insights underlying our approach which result in the first complete algorithm able to handle large and noisy proteins.

Let us define a *strand* $D = \langle (j, i_1), \ldots, (j + |D|, i_{|D|}) \rangle$ as a partial assignment of spin systems i to *consecutive* positions j. Then let a *window*, $W = \{D_1, \ldots D_{|W|}\}$, be a set of alternative strands covering the same positions but with different combinations of spin systems. With these definitions, the search *space* $\mathcal{S} = \{W_1, \ldots, W_{|\mathcal{S}|}\}$ can be represented as a set of disjoint windows that cover all residues. A candidate mapping a has one strand selected for each window. We maintain a set $\mathcal{A} = \{a_1, \ldots, a_{|\mathcal{A}|}\}$ of candidate mappings discovered during the course of the search, and explore the search tree as follows. We also maintain the mapping a^* with the highest posterior probability found so far.

Initialization. For a protein with R residues and I spin systems, we initialize the search space $\mathcal{S} = \{W_1, \ldots, W_R\}$ with R windows.

Each window $W_j = \{(j, 1), (j, 2), \ldots, (j, I), (j, I + 1)\}$ contains $I + 1$ unit-length strands, one for each spin system and one for a wild-card representing a missing spin system.

Branching. The search space is split according to two types of branches: a "missing" branch and a "mapping" branch. A missing branch places an upper bound on the number of missing placeholders that can be used in a window. Only strands with at most the specified number of missings are considered in the subtree. Thus different missing branches for the same window explore solutions that leave different numbers of positions as "wild cards." A mapping branch fixes a spin system at a position in the sequence. Strands in the subtree are filtered according to the mapping, so that the same spin system is not mapped to multiple locations, and multiple spin systems are not mapped to the same location. Different mapping branches thus explore different hypotheses about individual position-spin system assignments.

Leaf nodes. The expansion of a node ends when it is identified as either a leaf node or a dead end. A leaf node has been reached when the space $\mathcal{S} = \{W\}$ has a single window W covering the entire protein. In this case, all strands in W are added to the set of candidate mappings $\mathcal{A}$, and the best mapping a^* is updated accordingly. A dead end occurs when the space $\mathcal{S} = \{W_1, \ldots, W_{|\mathcal{S}|}\}$ has some window $W_k = \emptyset$. That is, no combination of spin systems can be mapped to the positions of the window with respect to the branching constraints, and thus there is no need to explore further. In both cases the algorithm backtracks and considers the next alternative branch.

Contraction. After each branch, the algorithm contracts the search space by merging adjacent windows. A pair of windows W and W' covering positions $\{j, \ldots, k\}$ and $\{k + 1, \ldots, l\}$ is merged into a new window covering positions $\{j, \ldots, l\}$. It contains a subset of the strands $\{D \cup D' | D \in W, D' \in W'\}$ that are consistent with the bound. If the number of strands in the new window does not exceed a user-specified threshold, we substitute it for the original two windows; otherwise, we leave the original windows. We then iterate, until no adjacent windows can be merged.

Bounding. Properties of the NMR spectra, along with the probability model in the previous section, can be used to bound the search space. Specifically, we evaluate the quality of each strand at a node in the tree, and determine whether the strand should be kept. Bounds (1)–(3) arise from the constraints in Sec. 1, and from the functional form of Eq. 1.

(1) *Match*: $|x^s_{t, a_{j+1}} - x^w_{t, a_j}| \leq \xi_t$ for all t, j and a. Here ξ_t are typical match tolerances in NMR studies, namely 0.25 ppm for C', 0.5 ppm for C^α and C^β, and 0.05 ppm for H^α. ξ_t determines the valid differences of chemical shifts in Fig. 3(a).

(2) *Align*: $\Sigma_j^{-1/2}(\bar{\mathbf{x}}_{a_j} - \boldsymbol{\theta}_j) \leq \text{quantile}(\chi^2_T, 0.9999)$ for all j and a. The quantile determines a level curve such as in Fig. 3(b) which encircles the valid region of chemical shifts for the residue at position j.

(3) *Unique map*: if a_j is not a missing spin system, then $a_j \neq a(j')$ for all j and j'.

(4) and (5) are global bounds derived from the probability model, based on the solutions discovered so far.

(**4**) *Posterior probability*: $P(\mathbf{x}|\boldsymbol{a}^*)/P(\mathbf{x}|\boldsymbol{a}) \leq 100$ for all $\boldsymbol{a}$. In other words, we need not consider mappings that are more than 100 times less likely *a posteriori* than $\boldsymbol{a}^*$.

(**5**) *Number of missings*: $\text{Miss}(\boldsymbol{a}) \leq \text{Miss}(\boldsymbol{a}^*) + 1$ for all $\boldsymbol{a}$. The prior distribution of candidate mappings heavily penalizes additional missing spin systems. In the vast majority of cases it is sufficient to consider mappings where the number of missing spin systems does not exceed the number of missing spin systems in $\boldsymbol{a}^*$ by more than 1.

Heuristics. When faced with choices of branches in the tree, we must decide which branch to explore first. In practice, we partition the branches to perform all missing branches first, followed by mapping branches for the rest of the tree. The policy for missing branches is to monotonically increase the number of missings. When taking a mapping branch, it is necessary to decide which position to fix and in which order to try assigning spin systems. Experimental results suggest that an effective policy is to always select the position with the smallest set of alternative spin systems that could be mapped to it. Once a position has been selected, we order the alternative spin systems according to the the likelihood of the strands in which they occur. This policy tends to reduce the width of the tree and typically find solutions significantly faster than other policies. Finally, in choosing which windows to contract first, we prefer windows with fewer strands, as they are less likely to cause a combinatorial explosion.

Interpreting the results

The probability model above can be used to make inference regarding mappings of the observed spin systems to positions. In the following, we say that a position has a *reliable* mapping if it is mapped to a same spin system according to all solutions satisfying bounds (1)–(5).

The probability model can also be used for inference regarding the unknown chemical shifts. Given a candidate mapping $\boldsymbol{a}$, the posterior distribution of a chemical shift μ_{tj} is a scaled Cauchy distribution truncated by the match tolerances ξ_t. The overall posterior distribution of μ_{tj} can be obtained by averaging over the candidate mappings $\boldsymbol{a}_k$:

$$\Pr(\mu_{tj}|\mathbf{x}) = \sum_{k=1}^{K} \Pr(\mu_{tj}|\mathbf{x}, \boldsymbol{a}_k)\Pr(\boldsymbol{a}_k|\mathbf{x})$$

In the following, we say that a chemical shift μ_{tj} is reliably determined if its posterior variance is within the range of variances for the resonance type t. We say that it is correctly determined if the difference between the posterior mean of μ_{tj} and the reference value does not exceed ξ_t.

3 RESULTS

Experimental Data Sets

We use our algorithm to analyze several publicly available experimental datasets: Human Ubiquitin,v and Zdomain, CspA, Ns1, RnaseWt, RnaseC6572S and Fgf. (Zimmerman et al., 1997). Fig. 5 describes the data sets and the corresponding reference solutions. The length of the proteins is between small and average for modern NMR studies.

As shown in Fig. 6, multiple mappings were found for all proteins. Furthermore, the agreement between reliably mapped spin systems and the reference solution is very good, but not perfect. This is due to match differences in the reference solutions that exceed the tolerances in bound (1) of Sec. 2. In such cases the algorithm typically introduces a missing spin system that compensates for the invalid match. Therefore the number of missing spin systems is larger than the corresponding number in the reference solution.

Disagreements in the mappings of spin systems do not affect the inference regarding the chemical shifts. Fig. 6 shows a perfect agreement between the determined chemical shifts and the reference solution. All executions took less than 3 hours of running time.

Importance of missing spin systems in the search space

Many existing algorithms for resonance assignment reduce the search space by compiling chains of unambiguously connected spin systems with no breaks. In our opinion, ignoring the possibility of a "break" at each position in the sequence underestimates the impact of missing spin systems, and of match differences that exceed the tolerance values. This section illustrates that point by comparing two approaches to resonance assignment. The first, proposed in this paper and called *free* in Fig. 7, considers the possibility of a missing spin system at each position in the sequence. The second approach, proposed by (Lin et al., 2002) and (Chen et al., 2003), enforces unambiguous matches between spin systems. Specifically, if a spin system $\mathbf{x}_1$ can be uniquely followed by a spin system $\mathbf{x}_2$, and $\mathbf{x}_2$ can be uniquely preceded by $\mathbf{x}_1$, the match is considered as fixed. Therefore matches between $\mathbf{x}_1$ and a missing, or between a missing and $\mathbf{x}_2$ are not allowed in the search space. The percentage of spin systems forming such unambiguous chains is a metric of "density" of adjacency information.

Protein	Length	Miss ss	Extra ss	Resonance types	
				Sequential	Within
Ubiquitin	76	2	0	C',C^α,C^β	C',C^α,C^β
Zdomain	70	2	2	$C',C^\alpha,C^\beta,H^\alpha$	$C^\alpha,C^\beta,H^\alpha$
CspA	70	2	4	$C',C^\alpha,C^\beta,H^\alpha$	$C^\alpha,C^\beta,H^\alpha$
Ns1	73	4	2	$C',C^\alpha,C^\beta,H^\alpha$	$C^\alpha,C^\beta,H^\alpha$
RnaseW	124	0	37	$C',C^\alpha,C^\beta,H^\alpha$	$C',C^\alpha,C^\beta,H^\alpha$
RnaseC	124	0	37	$C',C^\alpha,C^\beta,H^\alpha$	$C',C^\alpha,C^\beta,H^\alpha$
Fgf	154	2	24	$C',C^\alpha,C^\beta,H^\alpha$	$C',C^\alpha,C^\beta,H^\alpha$

Fig. 5. Description of the experimental data sets.

Protein	$|\mathcal{A}|$	*Miss*	**r**	**c**	$\mathbf{r}_{C^\alpha}$	$\mathbf{c}_{C^\alpha}$	*Hours*
Ubiquitin	2	5	66	100%	71	100%	0.005
Zdomain	3	3	62	98%	68	100%	0.02
CspA	11	3	59	98%	64	100%	0.26
RnaseC	2	1	118	100%	124	100%	2.44
Fgf	4	4	140	99%	151	100%	0.74

Fig. 6. Summary of assignment results for the experimental data sets. $|\mathcal{A}|$ is the number of candidate mappings satisfying bounds (1)–(5) in Sec. 2. *Miss* is the number of missing spin systems in the mapping with the highest posterior probability. The terms in the next four columns are introduced in Sec. 2: **r** is the number of positions with unambiguous mappings of spin systems, and **c** is the fraction of correct mappings among the reliable positions. $\mathbf{r}_{C^\alpha}$ is the number of reliably determined chemical shifts, and $\mathbf{c}_{C^\alpha}$ is the fraction of correctly determine chemical shifts among the reliable ones. *Hours* is the execution time in hours on a 2.5Ghz PowerPC G5 with 3GB of memory.

The second column of Fig. 7 describes the adjacency information in Human Ubiquitin and in the AutoAssign proteins. As can be seen, it ranges between 0 and 0.81, suggesting that these data have less adjacency than the data sets in (Chen et al., 2003) where the metric varies between .5 and .9. Furthermore, a comparison of Fig. 6 and Fig. 7 suggests that adjacency is not a good measure of the difficulty of a data set. While CspA has adjacency information of 0 and RnaseC6572S has adjacency information of 0.8, it takes almost 10 times longer to complete the assignment of RnaseC6572S.

According to columns 2–4 in Fig. 7, enforcing adjacency in RnaseC6572S and Ubiquitin reduces the number of unambiguously mapped positions without compromising the accuracy of the determined C^{α} chemical shifts. In Zdomain, it increases the number of unambiguously mapped positions at the expense of accuracy. While these differences are small they can be easily exacerbated in larger and noisier data sets.

Protein	*Adjacency*	*Free*		*Enforced*	
		$\mathbf{r}_{C^{\alpha}}$	$\mathbf{c}_{C^{\alpha}}$	$\mathbf{r}_{C^{\alpha}}$	$\mathbf{c}_{C^{\alpha}}$
RnaseC	**0.81**	124	100%	121	100%
Zdomain	**0.54**	68	100%	71	95%
Ubiquitin	**0.50**	71	100%	70	100%
Fgf	**0.34**	151	100%	152	100%
CspA	**0.00**	64	100%	–	–

Fig. 7. Assignments with adjacency information. *Adjacency* is the proportion of the spin systems forming unambiguous chains. *Free* denotes the approach considering a missing spin system at each position in the sequence, and *Enforced* is the approach that enforces unambiguous matches. $\mathbf{c}_{C^{\alpha}}$ and $\mathbf{r}_{C^{\alpha}}$ are as in Fig. 6. Since CspA has 0 adjacency, one can not enforce the connectivity information for this protein.

Synthetic Data

We evaluate the large-scale performance of our approach using synthetic data sets from 259 randomly selected entries to the database RefDB (Zhang et al., 2003). In order to generate data sets of realistic quality and size, we examined noise characteristics in the AutoAssign proteins, and simulated data according to these characteristics as follows. First, since it is unlikely that all the unreported chemical shifts in the database correspond to truly missing peaks, we simulate the missing values in the RefDB entries from the prior distributions in (Marin et al., 2004), and compile spin systems on the basis of the full sets of chemical shifts. Second, we delete a random number of the correct spin systems, and add a random number of extra spin systems to the data sets. The two numbers were generated from Poisson distributions with mean 2. Third, we randomly delete chemical shifts with frequencies observed in the AutoAssign proteins for each resonance type. Finally, we consider three scenarios of experimental noise. Scenario one, called *clean*, is unrealistic but often used to evaluate assignment procedures, e.g., in (Jung and Zweckstetter, 2004). It considers spin systems with no experimental noise. Scenario two, called *consistent*, is realistic. Here noise added to the chemical shifts is sampled from the histograms of match tolerances in the AutoAssign proteins such as in Fig. 2(a). The histograms are truncated to satisfy the standard match tolerances for each resonance type. Scenario three, called *noisy*, investigates the robustness of the proposed approach. It allows invalid match tolerances in the reference solution by sampling noise from the full

(non-truncated) distribution of match differences in the AutoAssign proteins. In total, the procedure generated 777 synthetic data sets that are available from the contact author along with automated scripts for their execution.

The data sets were analyzed using the MBA_2 software with default settings and no manual intervention. The results of the executions are summarized in Fig. 8. As can be seen, the proposed procedure reliably determines most of the mappings between spin systems and positions. On average, 99.8% of assignable positions are unambiguously mapped in the *clean* data sets, and 99.7% of assignable positions are mapped in the *noisy* experiments. The accuracy of assignment is measured by $\mathbf{c}$, the proportion of reliable assignment that agree with the reference solution. It is on average 100% for *clean* data, 99.96% for *consistent* data, and 99.36% for *noisy* data. The last row in Fig. 8 gives execution times on a log scale. As expected, time grows exponentially, in particular in *noisy* data set where large numbers of missing spin systems are required to compensate for invalid matches. At the same time, the execution times remain within acceptable limits for the majority of the data sets.

4 DISCUSSION

We have presented the first approach to backbone resonance assignment that is based on a complete algorithm and scales to data sets of realistic quality and size. We have validated our claims by conducting a large-scale automated resonance assignment study. The results show that, contrary to commonly accepted wisdom, complete search algorithm can handle problems of practical interest.

Where applicable, our approach will increase the number of reliably determined chemical shifts, and yield fewer errors as compared to the existing methods of resonance assignment. The characteristics that contribute to the excellent overall performance are (1) a rigorous probability model that assesses the uncertainty in the NMR spectra and is amenable to formal statistical inference; (2) the correct definition of the search space that considers a possibility of a missing spin system at any position in the sequence, and an arbitrary number of missings; (3) a complete algorithm that relies on hierarchical association of partial mappings to control the search space.

Complete algorithms are not a panacea. In our experimental study we encountered two AutoAssign datasets, Ns1 and RnaseWt, which proved challenging to the algorithm. But this is hardly surprising considering the quality of the data. Ns1 has 4 missing spin systems (out of 73 positions) and an average of 23% of the resonances missing per spin system. RnaseWt is missing 11% of its resonances and has 37 extraneous spin systems. Moreover, in both cases the reference solutions have spin systems with scores outside the standard match tolerances (7 for RnaseWt and 5 for Ns1).

Complete search algorithms can thus be an useful tool in many situations. In cases of very noisy data, they can be used to guide heuristic or stochastic search algorithms that consider only promising portions of the space, sacrificing completeness for scalability.

REFERENCES

Andrec, M., P. Du, and R. Levy (2001). Protein structural motif recognition via NMR residual dipolar couplings. *Journal of the American Chemical Society 123*, 1222.

Atreya, H. S., S. C. Sahu, K. V. R. Chary, and G. Govil (2000). A tracked approach for automated NMR assignments in proteins (TATAPRO). *Journal of Biomolecular NMR 17*, 125–136.
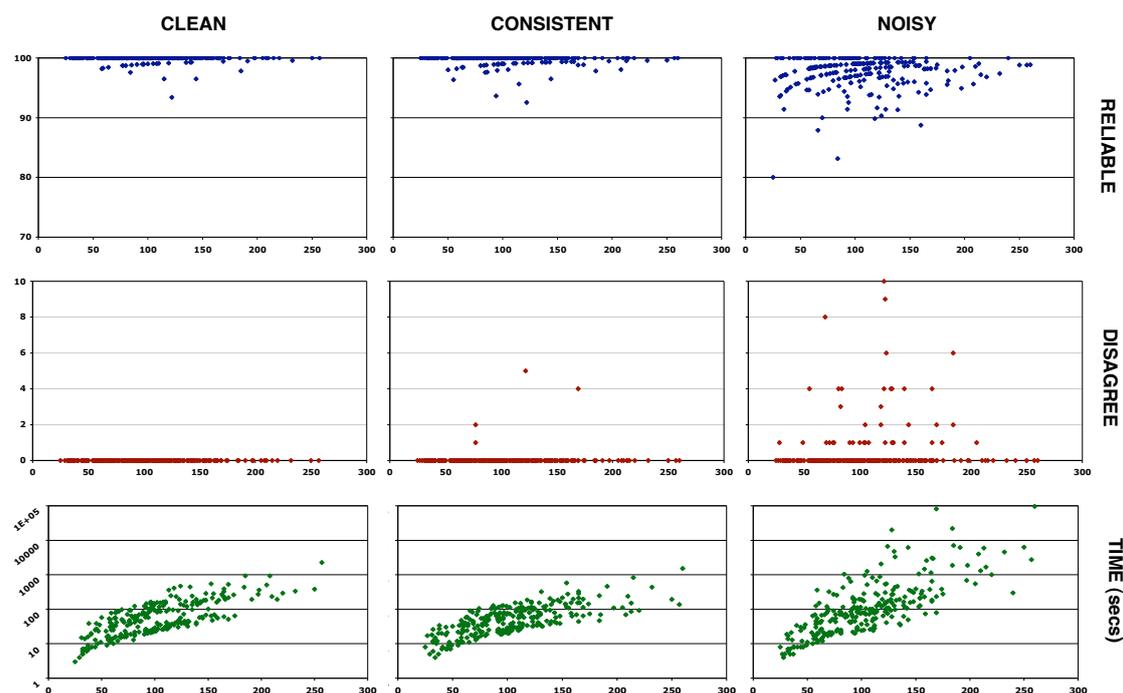
Fig. 8. Assignment Results for 259 proteins selected randomly from the RefDB. *Clean* data sets have spin systems with no experimental noise. *Consistent* data sets have distribution of experimental noise observed in the AutoAssign proteins, but within the match tolerances in the bound (1). *Noisy* data sets have non-truncated experimental noise observed in the AutoAssign data. The **x-axis** is protein length (ranging between 25 and 257). The **y-axes** have the following interpretations. For *Reliable*, it shows the percentage of reliably mapped positions (**r**/ number of assignable positions). For *Disagree*, it gives the absolute number of disagreements between reliably mapped spin systems and the reference solution. Finally, for *Time*, it shows computation time in seconds on a log scale. The median and maximum times are *Clean*: median = .7 min., max = 20 min.; *consistent*: median = .7 min., max = 25 min.; *noisy*: median = 1.4 min., max = 27 hrs.

Chen, Z.-Z., T. Jiang, G. Lin, J. Wen, D. Xu, J. Xu, and Y. Xu (2003). Approximation algorithms for NMR spectral peak assignment. *Theoretical Computer Science 299*(1-3).

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

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

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

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

Jung, J. and M. Zweckstetter (2004). MARS - robust automatic backbone assignment of proteins. *Journal of Biomolecular NMR 30*, 11–32.

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

Marin, A., T. Malliavin, P. Nicholas, and M.-A. Delsuc (2004). From NMR chemical shifts to amino acid types: Investigation of the predictive power carried by nuclei. *Journal of Biomolecular NMR 30*, 47–60.

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

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

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

Vitek, O., J. Vitek, B. Craig, and C. Bailey-Kellogg (2004). Model-based assignment and inference of protein backbone nuclear magnetic resonances. *Statistical Applications in Genetics and Molecular Biology 3*, 1–33. http://www.bepress.com/sagmb/vol3/iss1/art6.

Zhang, H., S. Neal, and D. S. Wishart (2003). A database of uniformly referenced protein chemical shifts. *Journal of Biomolecular NMR 25*(3), 173–195.

Zimmerman, D., C. Kulikowski, Y. Huang, W. Feng, M. Tashiro, S. S. Shimotakahara, C. Chien, R. Powers, and G. T. Montelione (1997). Automated analysis of protein NMR assignments using methods from artificial intelligence. *Journal of Molecular Biology 269*, 592–610.