Journal of Biomolecular NMR (2006) 35: 187–208
DOI 10.1007/s10858-006-9027-8

Article

Inferential backbone assignment for sparse data

Olga Vitek[a,*], Chris Bailey-Kellogg[b], Bruce Craig[c] & Jan Vitek[d]

[a]*Institute for Systems Biology, 1441 North 34th Street, Seattle, WA, 98103-8904, USA;* [b]*Department of Computer Science, Dartmouth College, Hanover, NH, 03755, USA;* [c]*Department of Statistics, Purdue University, West Lafayette, IN, 47907, USA;* [d]*Department of Computer Sciences, Purdue University, West Lafayette, IN, 47907, USA*

Received 16 December 2005; Accepted 8 May 2006

Key words: Bayesian modeling, NMR assignment, sparse data, statistical inference, stochastic search algorithm, structural genomics

Abstract

This paper develops an approach to protein backbone NMR assignment that effectively assigns large proteins while using limited sets of triple-resonance experiments. Our approach handles proteins with large fractions of missing data and many ambiguous pairs of pseudoresidues, and provides a statistical assessment of confidence in global and position-specific assignments. The approach is tested on an extensive set of experimental and synthetic data of up to 723 residues, with match tolerances of up to 0.5 ppm for C^α and C^β resonance types. The tests show that the approach is particularly helpful when data contain experimental noise and require large match tolerances. The keys to the approach are an empirical Bayesian probability model that rigorously accounts for uncertainty in the data at all stages in the analysis, and a hybrid stochastic tree-based search algorithm that effectively explores the large space of possible assignments.

Introduction

Backbone resonance assignment is required for most current procedures of NMR-based protein structure determination (Wüthrich, 2003). The basic input to an assignment procedure is the primary sequence and a set of pseudoresidues compiled from peaks in through-bond experiments such as HNCA and HN(CO)CA. As illustrated in Figure 1a, each pseudoresidue contains both within-residue chemical shifts from one residue and sequential chemical shifts from the preceding residue. A minimal set of chemical shifts includes within-residue H, N, and C^α, along with sequential C^α. When available, within-residue and sequential C^β, C', and H^α are also included. The goal of the assignment is to establish the origin of each pseudoresidue, i.e., a mapping between pseudoresidues and positions in the primary sequence. This is accomplished by matching the sequential chemical shifts of one pseudoresidue to the within-residue chemical shifts of another, and by aligning the matched pair to sequentially adjacent amino acids for which the observed chemical shifts are plausible. Each pseudoresidue can typically participate in multiple possible matches and alignments, and a one-to-one globally consistent mapping must be chosen from the combinatorial possibilities.

Significant progress has been made in the development of automated resonance assignment methods, allowing researchers to accurately and quickly assign many proteins (Moseley and Montelione, 1999). However, the desire to increase

*To whom correspondence should be addressed.
E-mail: ovitek@systemsbiology.org

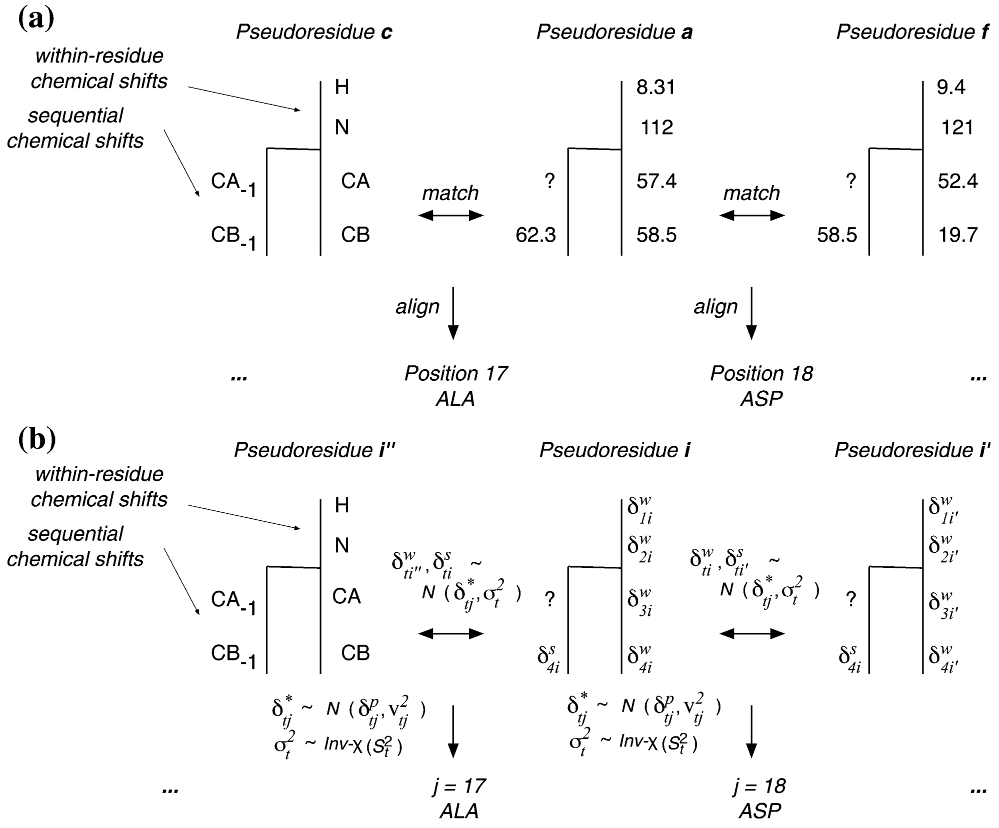

Figure 1. Matching pseudoresidues to each other and aligning them to the protein sequence (a) in text notation, (b) in mathematical notation.

throughput in NMR-based studies demands assignment algorithms that handle increasingly sparse data sets. Sparse data arise when the number, diversity, and quality of NMR experiments is low relative to the protein size. Characteristics of sparse data include limited sets of resonance types, large fractions of missing chemical shifts, entirely missing pseudoresidues for some residue positions, and spurious noise pseudoresidues. Furthermore, chemical shifts in the pseudoresidues may be obtained from only one or two peaks, making their values less certain and more subject to peak picking error. Therefore large match tolerances are required to establish sequential connectivities between chemical shifts.

Working with sparse data has important implications for scoring matches and alignments of pseudoresidues. Existing scoring functions often rely at least in part on subjectively chosen parameters and functional forms. When data are informative the choice of the scoring function does not significantly impact the conclusions of the assignment. When data are sparse, however, the choice can severely affect the accuracy of the result. It is therefore important to determine the scoring functions in the most objective manner, e.g., empirically on the basis of previously assigned data sets. In addition, noise and sparsity of the data typically result in a large number of globally consistent mappings. On one hand, stochastic variation and the approximate nature of the scores makes it dangerous to consider only the best mapping found. On the other hand, only a few positions will be reliably assigned if we make conclusions on the basis of all the detected globally consistent mappings. The existing scoring functions do not allow us to select the relevant mappings for conclusions because it is difficult to interpret the relative values of their scores. For example, if the score of the best mapping found is twice the score of an alternative mapping, it is not clear whether both mappings, or only the best one, are plausible given the data. Therefore, sparse data require a new scoring function capable of selecting a plausible subset out of all globally consistent candidate mappings.

Sparse data also have important implications for algorithms for finding the mappings. Noise and sparsity result in combinatorially large spaces of candidate mappings with essentially flat scoring landscapes. A typical approach to reducing the size of the search space is to consider "unambiguous chains" of pseudoresidues as fixed. However, this approach is not applicable to sparse data because incorrect matches will likely appear unambiguous due to missing resonance types and missing chemical shifts. New algorithms are necessary that are capable of efficiently traversing such landscapes while considering the possibility of missing pseudoresidues at any position in the primary sequence.

This paper presents an approach for resonance assignment of sparse NMR data that overcomes the above mathematical and algorithmic challenges. We develop a scoring function based on an empirical Bayesian statistical model. Bayesian modeling has recently been successfully applied to determine three-dimensional protein structures from assigned NMR data (Rieping et al. 2005), and the same conceptual approach is employed here. Bayesian modeling involves more than just an application of Bayes rule, and relies upon a joint probability distribution of all observed and unobserved quantities in the problem. In particular, one specifies the probability distribution of the observed data, and the joint prior distributions of the parameters of this distribution. The prior distributions acknowledge our uncertainty in the parameters. The qualification "empirical Bayes" applies when one takes advantage of the current and previously analyzed data sets to develop the prior distributions. Once the probability model is specified, we employ integration steps to average out the parameters that are not of our primary interest, and eventually apply Bayes rule to obtain model-based scoring functions. We then draw conclusions on the basis of a subset of all detected globally consistent mappings, selected according to their posterior probabilities.

This paper addresses the algorithmic challenges of sparse data by developing a hybrid stochastic search algorithm. The algorithm combines the advantages of systematic branch-and-bound search with desirable properties of stochastic techniques such as simulated annealing, Tabu search and population-based search, and therefore has better potential performance than any of these methods. The algorithm stochastically explores the efficient enumeration-enhanced tree-based search space we developed in Vitek et al. (2005). It considers the possibility of a missing pseudoresidue for each position, and employs adaptive learning techniques as well as multiple strategies for escaping locally optimal solutions. The output of the algorithm is a set of globally consistent mappings that have high posterior probabilities.

We test our approach on a wide range of experimental and synthetic data sets and demonstrate that it reliably and accurately assigns large portions of proteins given only sparse data. The results show that our approach is typically more accurate than the recently-published program CASA, which was demonstrated to be representative of a number of current automated assignment packages (Wang et al., 2005). In comparison to the popular program MARS (Jung and Zweckstetter 2004), our approach increases the number of reliably and accurately assigned residues when the data are noisy and require large match tolerances.

Methods

Input data

The input data to our method are the primary sequence of a protein, and a set of pseudoresidues compiled from triple-resonance experiments such as HN(CO)CA, HNCA, HN(CO)CACB or HNCB. Similarly to other automated methods of resonance assignment (e.g., Buchler et al., 1997; Coggins and Zhou, 2003; Hitchens et al., 2003; Jung and Zweckstetter, 2004) our method does not address peak picking. Thus the pseudoresidues must be compiled from spectra manually or by other means, prior to the analysis.

The pseudoresidues are illustrated in Figure 1a. They contain chemical shifts of N, H, C′, C^{α}, C^{β} and H^{α} resonance types from the same position in the protein sequence. We call these the *within-residue* chemical shifts. Pseudoresidues also contain chemical shifts of C′, C^{α}, C^{β} and H^{α} resonance types of the preceding residue, which we call the *sequential* chemical shifts. In sparse data, only a subset of C′, C^{α}, C^{β} and H^{α} resonance types will typically be present, and some of the expected chemical shifts will be missing. Some expected pseudoresidues will typically be missing, and spurious noise pseudoresidues will be present. The first position in the sequence and the Proline residues do not have experimental pseudoresidues.

An optional input to the method is the predicted secondary structure of the protein (e.g., using PSIPRED (McGuffin et al., 2000)). Predicted secondary structure is used for efficient amino acid typing and typically increases the number of reliably assigned positions. All the results discussed in the paper used secondary structure prediction as input.

Probability model for scoring candidate mappings

To establish a Bayesian inferential procedure for resonance assignment we take the following four steps (Gelman et al., 1995); Figure 1b illustrates the key terms and their mathematical notation. (1) We note that the observed chemical shifts δ_i in pseudoresidues i are random variables subject to noise and stochastic variation, and we specify their probability distribution. The distribution depends upon unknown parameters, namely the mappings $\mathcal{M}_j$ of residue positions j in the primary sequence to pseudoresidues, and the means δ_j^* and variances σ_j^2 of the chemical shifts at the residue positions. Determining the mappings is our primary goal. The means and the variances are not of particular interest, but are rather "nuisance parameters" required for a complete specification of the probability distribution. (2) We quantify our knowledge regarding plausible values of all parameters, including nuisance parameters, in terms of their prior probability distributions. These priors are determined from existing databases, previously assigned spectra, and statistical considerations. (3) We "remove" the nuisance parameters from the probability distribution by taking an average (i.e., integrating) with respect to their prior distributions. This yields an "integrated" distribution of chemical shifts that depends only on their mappings from positions in the sequence. (4) We obtain a posterior distribution of the mappings conditional on the observed chemical shifts. This allows our search algorithm to evaluate each candidate mapping, and use for inference only those that are plausible. (5) The posterior distribution of the plausible solutions allows us to assess the information content in the data, and to quantitate the confidence in individual resonance assignments in terms of their consistency across the plausible solutions. We detail these steps below.

Probability distribution of the observed data

Consider a within-residue and a sequential measurement of a resonance (say, C^α) at residue position j. Denote by δ_i^s the observed sequential chemical shift, which is in pseudoresidue i, and by $\delta_{i'}^w$ the observed within-residue chemical shift, in pseudoresidue i'. Since peak locations vary across spectra, and every peak picking procedure is subject to some errors, the observed chemical shifts are rarely the same. We assume that they are independent and Normally distributed random variables, centered at the "true" chemical shift δ_j^* with a position-specific variance σ_j^2. In our notation, the probability density function of δ_i^s and $\delta_{i'}^w$ is

$$f(\delta^s_{\mathcal{M}j}, \delta^w_{\mathcal{M}j'} \mid \delta_j^*, \sigma_j^2, \mathcal{M}_j) = \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(\delta^s_{\mathcal{M}j} - \delta_j^*)^2}{2\sigma_j^2}} \times \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(\delta^w_{\mathcal{M}j'} - \delta_j^*)^2}{2\sigma_j^2}} \tag{1}$$

The unknown parameters of the probability distribution are δ_j^* and σ_j^2, as well as the mapping $\mathcal{M}_j = i$ from position index j to pseudoresidue index i. The mapping $\mathcal{M}_j$ is of interest in our assignment procedure, and the other parameters will be integrated out with respect to their prior distributions. To simplify notation, we hereafter denote the observed chemical shifts as δ_i instead of $\delta_{\mathcal{M}j}$, with the understanding that the choice of the index i depends on the mapping $\mathcal{M}_j = i$.

Prior distributions for amino acid typing

The identification of plausible distributions for the "true" chemical shifts δ_j^* for each residue, also known as amino acid typing, has been the subject of much research. It is natural to determine this distribution from a database of previously assigned chemical shifts such as the BMRB (Seavey et al., 1991) or RefDB (Zhang et al., 2003). One often assumes that the plausible chemical shifts are independent across resonance types, and are Normally distributed with means and variances calculated from the database. One can also specify a distribution that does not rely on the Normal functional form (Wan et al., 2004; Eghbalnia et al., 2005), or study the correlation structure of resonance types within the same amino acid (Marin et al., 2004). The distribution is more informative if it uses predicted secondary structure (Jung and Zweckstetter, 2004; Wan et al., 2004; Eghbalnia et al., 2005). All these approaches are equally acceptable in Bayesian modeling.

In this paper we use the form that has been successfully used in Jung and Zweckstetter (2004). Specifically, we predict the secondary structure of the protein using PSIPRED (McGuffin et al., 2000), and calculate the predicted chemical shift δ_j^p for each resonance type and secondary structure using the algorithm described in Wan and Jardetzky (2002). We further assume that chemical shifts in the database are Normally distributed, centered at the predicted values δ_j^p with standard deviations v_j as empirically determined in Jung and Zweckstetter (2004). The specific values of v_j are 0.82 ppm for H, 4.3 ppm for N, 1.2 ppm for C^α, 1.1 ppm for C^β, 1.7 ppm for C', and 0.82 ppm for H^α. In our notation, the prior density of the "true" chemical shift is

$$f(\delta_j^*) = \frac{1}{\sqrt{2\pi v_j}} e^{-\frac{(\delta_j^* - \delta_j^p)^2}{2v_j^2}} \tag{2}$$

Prior distributions of experimental variances

In contrast to amino acid typing, empirical analysis of experimental variances σ_j^2 has received little attention. Existing assignment tools proceed by scoring matches between chemical shifts with *ad hoc* functions such as uniform scores (Atreya et al., 2000; Andrec and Levy, 2002; Coggins and Zhou, 2003; Jung and Zweckstetter, 2004), or bell-shaped scores (Bartels et al., 1997; Buchler et al., 1997; Zimmerman et al., 1997; Hitchens et al., 2003) with pre-specified parameters. Arbitrarily defined, the scores undermine the performance of the assignment algorithms. The Bayesian approach, on the other hand, empirically derives the distribution of plausible variances from previously assigned data.

We study the empirical distribution of variances σ_j^2 using 6 experimental data sets provided as a test for the AutoAssign program (Zimmerman et al., 1997). The same approach can be performed using any other training data set that is judged representative of the assignment problem at hand. For each of the 6 proteins, we consider the pseudoresidues and their reference assignments determined by AutoAssign, and compute histograms of differences in chemical shifts mapped to the same atom. Figure 2 shows the histograms combined across the proteins. One can see that AutoAssign provides a conservative view of the plausible values, allowing fairly loose matches. We truncate the distributions by desired match tolerances, translate the chemical shift differences into estimates of experimental variances by computing $(\delta_i^s - \delta_{i'}^w)^2/2$, and fit the distributions of the estimates with the inverse- χ^2 distribution with 1 degree of freedom. Inverse- χ^2 is often used to model the prior for the variance of the Normal distribution (Gelman et al., 1995). It is best defined by saying that $\sigma^2 \sim \text{Inv} - \chi_1^2$ is equivalent to $1/\sigma^2 \sim \chi_1^2$. In our notation, the prior distribution of experimental variances σ_j^2 has the probability density

$$f(\sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^3}} e^{-\frac{S^2}{2\sigma_j^2}} \tag{3}$$

The fitted parameters S^2 for large (moderate) match tolerances used in the experimental section are 0.0016 (0.0012) ppm^2 for C', 0.004 (0.002) ppm^2 for C^α, 0.005 (0.005) ppm^2 for C^β and 0.00005 (0.00005) ppm^2 for H^α (Vitek, 2005).

Integrated probability distribution of the observed data

In the Bayesian framework, the prior distributions in Equations (2) and (3) are used to integrate over the unknown mean and variance in Equation (1). If we assume that the prior distributions are independent, the integral is computed as

$$\begin{aligned} & f(\delta_i^s, \delta_{i'}^w \mid \mathcal{M}_j = i) \\ & = \int_0^\infty \left[\int_{-\infty}^{\infty} f(\delta_i^s, \delta_{i'}^w \mid \delta_j^*, \sigma_j^2, \mathcal{M}_j = i) \, f(\delta_j^*) \; d(\delta_j^*) \right] \\ & \quad \times f(\sigma_j^2) \; d(\sigma_j^2) \end{aligned} \tag{4}$$

After an analytical integration we obtain the following approximate result. The approximation relies upon the fact that the range of chemical shifts is orders of magnitude larger than the

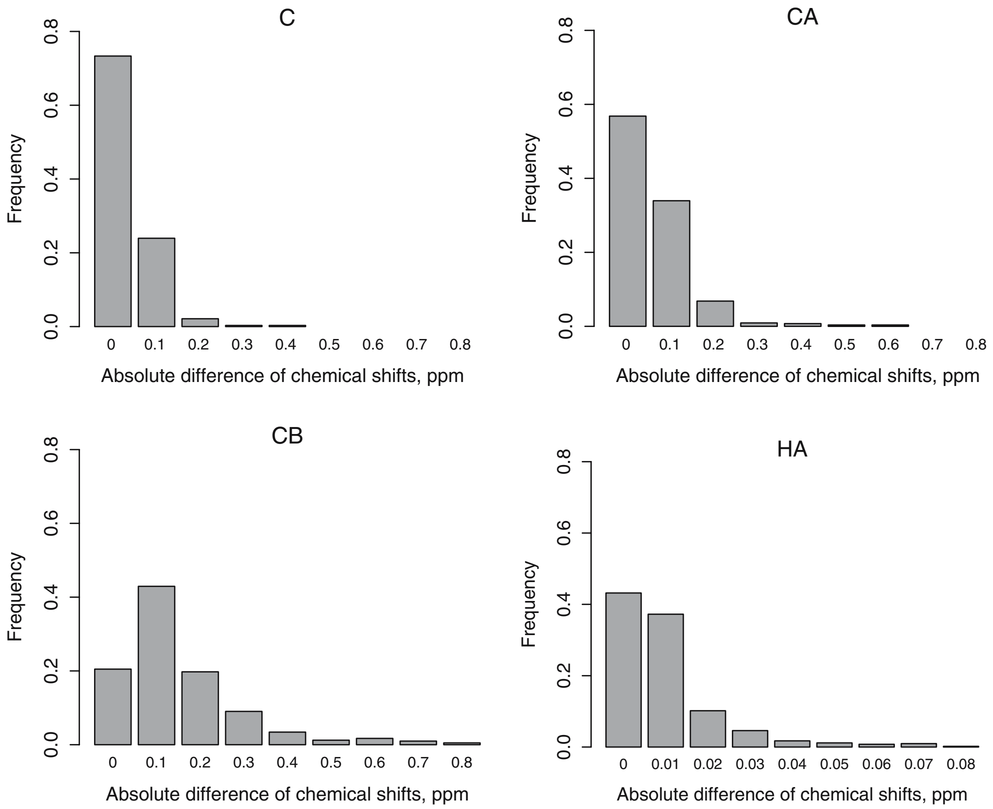

Figure 2. Distribution of absolute differences in matched chemical shifts for the data from Zimmerman et al. (1997) as assigned by AutoAssign.

variance of the signal from a particular nucleus (Vitek, 2005):

$$f(\delta_i^s, \delta_{i'}^w | \mathcal{M}_j = i) \approx \frac{1}{\sqrt{\pi} v_j} e^{-\frac{((\delta_{i'}^w + \delta_i^s)/2 - \delta_j^p)^2}{v_j^2}} \times \frac{1}{\sqrt{2\pi} S} \frac{1}{1 - \frac{(\delta_{i'}^w - \delta_i^s)^2}{2S^2}} \tag{5}$$

The first term is the probability density of a Normal distribution, and it approximately equals the prior distribution for amino acid typing. The role of this term is to quantify the plausibility of aligning the chemical shifts at position *j*. The second term appears in Equation (5) only when both $\delta_{i'}^w$ and δ_i^s are present. It is the density of the Cauchy probability distribution, and its role is to quantify the plausibility of a match between two chemical shifts.

Figure 3 compares the Cauchy distribution with the Normal distribution where the standard deviations are a third of the match tolerances, i.e., 0.25/3 for C′, 0.5/3 for C^α and C^β, and 0.05/3 for H^α. Such Normal distributions are traditionally used to score chemical shift matches. One can see that the Cauchy distribution is more concentrated around zero than the Normal distribution and thus supports better discrimination of a high quality match. At the same time it has heavier tails, reflecting the plausibility of some relatively large differences in matched chemical shifts.

Equation (5) defines the integrated probability distribution of one position and one amino acid type. The key to inference, however, is obtaining a globally consistent set of matches and alignments. Thus we consider a *full* mapping which maps each position to at most one pseudoresidue, and each pseudoresidue to at most one position. Positions not mapped to an observed pseudoresidue are mapped to placeholders for entirely missing data. If we index the distributions in Equation (5) by resonance type *t* (over *T* types) and position *j* (over

J positions) and assume that they are conditionally independent given the mapping, then the integrated probability distribution of all chemical shifts δ given a full mapping $\{\mathcal{M}_1, \ldots, \mathcal{M}_J\}$ is

$$f(\delta|\mathcal{M}_1, \ldots, \mathcal{M}_J) = \prod_{j=1}^{J} \prod_{t=1}^{T} f_{tj}(\delta_{ti}^{s}, \delta_{ti}^{w}|\mathcal{M}_j = i) \tag{6}$$

Posterior probabilities

The posterior probability of a full mapping $\{\mathcal{M}_1, \ldots, \mathcal{M}_J\}$ is obtained by specifying its prior probability $\Pr(\mathcal{M}_1, \ldots, \mathcal{M}_J)$ and by applying Bayes rule

$$\Pr(\mathcal{M}_1, \ldots, \mathcal{M}_J|\delta) \propto f(\delta|\mathcal{M}_1, \ldots, \mathcal{M}_J) \times \Pr(\mathcal{M}_1, \ldots, \mathcal{M}_J) \tag{7}$$

In our approach, the prior probability reflects our preference for assigning as many positions as possible. We quantify this using corrected Akaike Information Criterion (cAIC) prior weights (Burnham and Anderson, 2002), which penalize appropriately, but not too strictly, each observed and unassigned chemical shift. On the −log scale,

$$-\log \Pr(\mathcal{M}_1, \ldots, \mathcal{M}_J) = I' \cdot N/(N - JT) \tag{8}$$

where I' is the number of unassigned chemical shifts, N is the total number of observed chemical shifts, and J and T are respectively the protein length and the number of resonance types.

We combine Equations (5)–(8), i.e., the integrated probability distribution of chemical shifts and the Akaike Information Criterion penalty for unassigned chemical shifts, into a posterior probability of a full mapping $\mathcal{M}$. The result on the −log scale (determined up to an additive constant) is

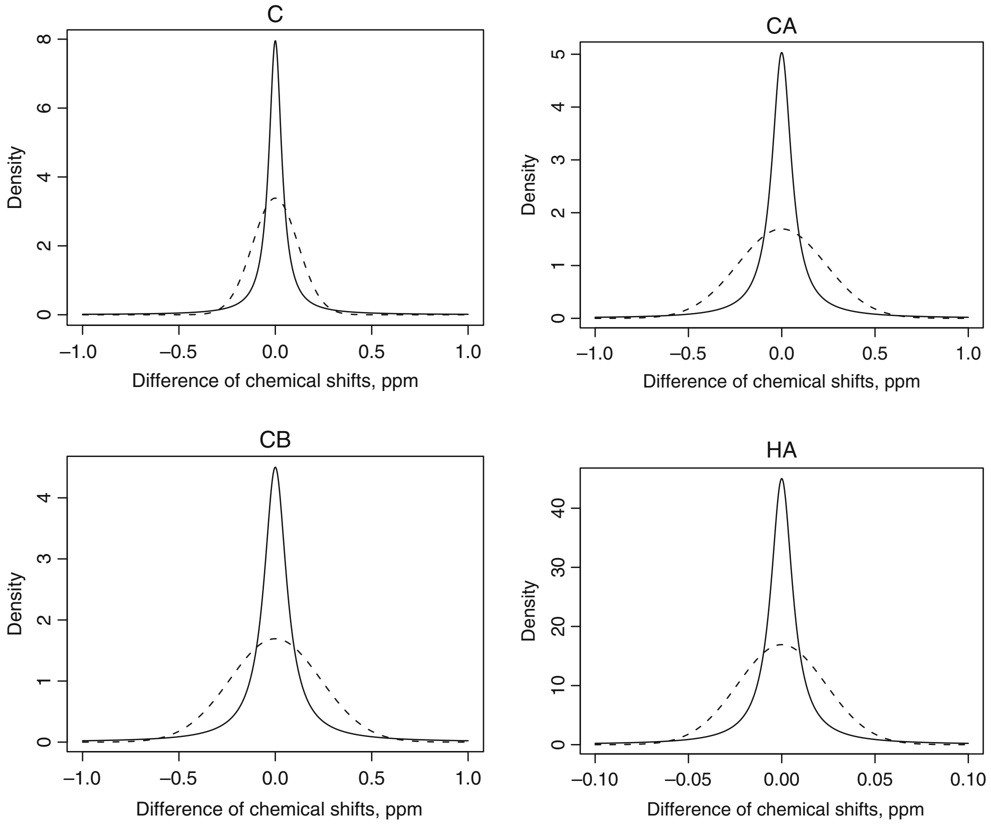

Figure 3. Distributions of chemical shift match scores. Solid lines: Cauchy distributions used in MBA, obtained from AutoAssign data (Zimmerman et al., 1997). Dashed lines: Normal distributions with standard deviations taken as a third of the corresponding match tolerances.

$$-\log \Pr(\mathcal{M}_1,\dots,\mathcal{M}_J|\delta) = \sum_{j=1}^{J}\sum_{t=1}^{T}\log(\sqrt{2}\pi S_t) + (\delta^w_{ti'} - \delta^s_{ti})^2/2S_t^2 + \sum_{j=1}^{J}\sum_{t=1}^{T}\log(\sqrt{2\pi}v_{tj}) + ((\delta^w_{ti'} + \delta^s_{ti})/2 - \delta^p_{tj})^2/2v_{tj}^2 + I' \cdot N/(N - JT) \qquad (9)$$

Recall that the choice of index i of δ_i in Equation (9) depends on the mapping $\mathcal{M}_j$ through the relationship $\mathcal{M}_j = i$. Thus Equation (9) will be used by the search algorithm to score candidate mappings.

Our goal is to determine the mapping $\{\mathcal{M}_1^*,\dots,\mathcal{M}_J^*\}$ maximizing the posterior probability or, equivalently, minimizing Equation (9). When the data are informative, the best mapping $\{\mathcal{M}_1^*,\dots,\mathcal{M}_J^*\}$ is usually the only globally consistent one that can be found. With sparse data, however, a large number of globally consistent mappings can exist. An advantage of the probability model is that it allows us to determine which mappings, out of all possible, should be considered. Specifically, we discard mappings that are 100 times less likely *a posteriori* than the best mapping $\{\mathcal{M}_1^*,\dots,\mathcal{M}_J^*\}$ found. Thus we search for all mappings $\{\mathcal{M}_1,\dots,\mathcal{M}_J\}$ satisfying

$$-\log \Pr(\mathcal{M}_1,\dots,\mathcal{M}_J|\delta) \le -\log \Pr(\mathcal{M}_1^*,\dots,\mathcal{M}_J^*|\delta) + \log 100 \qquad (10)$$

The posterior probabilities $\Pr(\mathcal{M}_1,\dots,\mathcal{M}_J|\delta)$ of the selected mappings are standardized to form a proper probability distribution as

$$\widetilde{\Pr}(\mathcal{M}_1,\dots,\mathcal{M}_J|\delta) = \frac{\Pr(\mathcal{M}_1,\dots,\mathcal{M}_J|\delta)}{\sum \Pr(\mathcal{M}_1,\dots,\mathcal{M}_J|\delta)} \qquad (11)$$

where the summation is over all mapping satisfying Equation (10). Thus our approach applies the normalization to selected mappings having high posterior probability as opposed to all possible globally consistent mappings. Making conclusions on the basis of such normalized distributions is an example of selective model averaging (Kass and Raftery, 1995; Hoeting et al., 1999). The approach is particularly useful for sparse data where the number of "interesting" mappings is relatively small compared to the number of all consistent mappings. Not only it is often impossible to consider all mappings that are globally consistent, but in doing so the denominator of Equation (11) can become large when adding a large number of small probabilities. Thus with sparse data, using selectively standardized probabilities can be necessary for obtaining reliable and numerically stable conclusions.

Inference

In the Bayesian context, statistical inference is a procedure of making conclusions on the basis of the globally consistent mappings satisfying Equation (10), and their posterior probabilities. Several types of statistical inference can be made. First, we quantify the information content in the data by considering the posterior probability distribution $\Pr(\mathcal{M}_1,\dots,\mathcal{M}_J|\delta)$ of mappings in Equation (10). If only one mapping is selected, or if the posterior probability of the best mapping $\{\mathcal{M}_1^*,\dots,\mathcal{M}_J^*\}$ is markedly higher than the posterior probabilities of the other selected mappings, we say that the data set has a high information content. It has a low information content otherwise. (Vitek et al., 2004) provide examples of posterior distributions of candidate mappings with different information content.

The second type of statistical inference quantifies the confidence in assignments of individual positions in the sequence. We say that a mapping between a position and a pseudoresidue is *reliable* if it appears in all the selected solutions, and is not surrounded by placeholders for entirely missing pseudoresidues.

Finally, one can characterize the uncertainty in an individual chemical shift by means of its posterior mean and standard deviation. While these parameters are currently averaged out in derivations of the scoring function, their posterior probability distribution is easy to obtain. This information would be used to identify positions that negatively affect the quality of the assignment, and to plan further experiments in an adaptive fashion. We plan to investigate the use of this information in our future work.

Stochastic search algorithm

The stochastic search algorithm inputs a primary sequence and experimentally determined pseudoresidues, and outputs a set of full mappings having high posterior probabilities. The algorithm relies upon three key ingredients: (1) a tree-based structure that efficiently partitions the search space into smaller subspaces, (2) a set of bounds eliminating poor quality mappings even when they are only partially specified, and (3) techniques of stochastic search that focus computational resources on promising portions of the space.

Tree-based structure of the search space

The tree recursively partitions the space of all possible globally consistent mappings. Each node in the tree represents a search subspace constrained by matches and alignments for some positions in the primary sequence and some pseudoresidues. Construction of the tree is enhanced by enumeration of *partial mappings*, which map chains of pseudoresidues to consecutive positions in the primary sequence.

Construction of the tree is illustrated in Figure 4. The root node is a set of all plausible partial mappings covering the entire protein sequence, as illustrated in the top of the figure. The root node is created by first considering all combinations of pseudoresidues mapped to the first

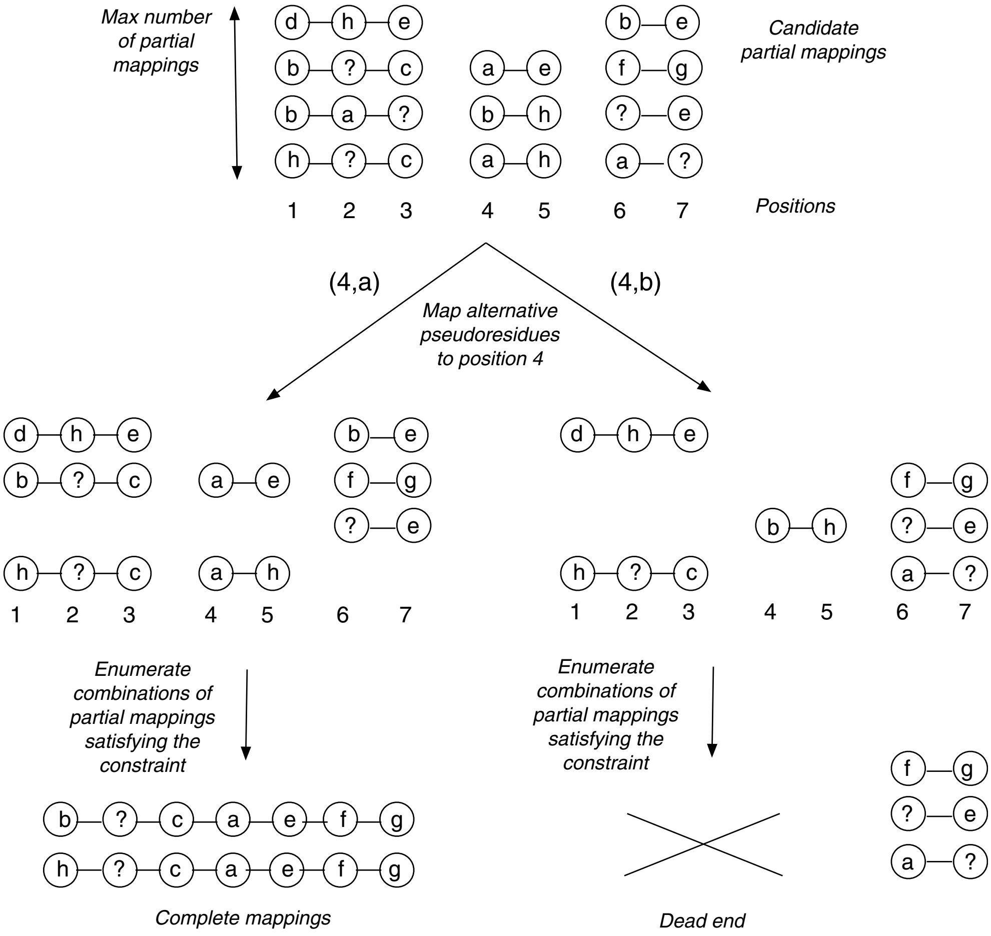
Max number of partial mappings
Candidate partial mappings
Positions
(4,a)
(4,b)
Map alternative pseudoresidues to position 4
Enumerate combinations of partial mappings satisfying the constraint
Enumerate combinations of partial mappings satisfying the constraint
Complete mappings
Dead end
1 2 3 4 5 6 7

Figure 4. Tree-based representation of the search space. Circles with characters are pseudoresidues, with "?" denoting a missing pseudoresidue. Numbers are positions in the protein sequence.

two positions in the sequence. The scores of the partial mappings are computed, and only the mappings satisfying the bounds derived from the probability model (described below) are kept. The enumeration is then extended to the following residues, and an increasingly large number of plausible partial mappings is examined. When the number of mappings exceeds a pre-specified parameter, enumeration at the current position stops, and a new independent enumeration begins at the following two positions. For example, in the case shown in the top of Figure 4, enumeration of partial mappings was stopped after the first three positions, restarted at the fourth position, stopped at the fifth and restarted again at the sixth position. The enumeration accounts for missing data by considering a placeholder for an entirely missing pseudoresidue as a potential mapping at each position.

The children of a node are obtained by branching on the choice of a position in the primary sequence to fix, and of a pseudoresidue to map to that position. The consequences of a choice are propagated to descendant nodes: all partial mappings that are inconsistent with the choice are removed, and the remaining partial mappings are combined into longer ones as illustrated in Figure 4. Note that the positions fixed at a parent node and one of its children need not be adjacent in the primary sequence. A leaf node in the tree represents a set of globally consistent mappings covering the entire sequence (bottom left of Figure 4). A dead end in the tree is a search subspace that cannot contain a full globally consistent mapping satisfying the bounds (bottom right of Figure 4).

The tree is a compact and efficient representation of the search space, and its size is determined by the information content in the data. In informative data sets the bounds are powerful, dead ends occur early in the tree, and the total number of nodes is small. In sparse data sets, the bounds become effective only after fixing a number of positions, and the tree is large.

Bounds on the search space of candidate mappings
Partial mappings in the tree must satisfy validity bounds on the elements of the score in Equation (9). Bounds specifying plausible matches and alignments of pseudoresidues are used by most assignment algorithms. In our case, these bounds are as follows.

(a) *Match*: $|\delta^{w}_{ti'} + \delta^{s}_{ti}| \leq \xi_t$ where ξ_t are match tolerances of resonance type t. We take fairly large match tolerances to accommodate sparse data. The specific values are 0.25 ppm for C′, 0.5 ppm for C^{α} and C^{β}, and 0.05 ppm for H^{α}.

(b) *Align*: $|(\delta^{w}_{ti'} + \delta^{s}_{ti})/2 - \delta^{p}_{tj}| \;/\; v_{tj} \leq 3.89$. The value 3.89 is such that the probability of rejecting a valid alignment is approximately 0.00001.

Key to the effectiveness of our search algorithm is that we also apply bounds on partial mappings. A partial mapping is evaluated according to its total score obtained as in Equation (9), but without the last term since the penalty is only defined for full mappings. Intuitively, we can eliminate a partial mapping if any extension of it is guaranteed to be much worse than the best complete mapping $\mathcal{M}^*$ that we have found. Therefore, a partial mapping covering a set of positions $\mathcal{J}$ must satisfy the following:

(c) *Score of a partial mapping*: $-\log \Pr(\mathcal{M}_{j\in\mathcal{J}}|\delta)$ $-\min\,\log \Pr(\mathcal{M}_{j\notin\mathcal{J}}|\delta) \leq -\log \Pr(\mathcal{M}^*|\delta)+$ $\log 100$. In other words, we discard partial mappings which, in combination with a guaranteed lower bound on the best score at the remaining positions, yield an assignment that has a posterior probability at least 100 times smaller than the best mapping found.

(d) *Number of unassigned pseudoresidues*: $\mathrm{Miss}(\mathcal{M}_{j\in\mathcal{J}}) + \min\,\mathrm{Miss}(\mathcal{M}_{j\notin\mathcal{J}}) \leq \mathrm{Miss}(\mathcal{M}^*)$ $+1$. The prior distribution of candidate mappings penalizes unassigned chemical shifts. Therefore, we discard mappings which, in combination with a guaranteed lower bound on the minimum number of missing pseudoresidues at the remaining positions, exceed the number of missings in $\mathcal{M}^*$ by more than 1.

Searching the space
Trees are often linked to systematic search algorithms, such as branch-and-bound, which provably find globally optimal solutions. Although attractive, a naïve tree-based representation and search algorithm are not adequate for sparse NMR data (Andrec and Levy, 2002). Better performance can be obtained by approximation algorithms which impose heuristic bounds on the branches of the tree, but only guarantee that solutions will be within some maximum distance from the optimum (Wan et al., 2004). A faster

execution time can be obtained by best-first and depth-first search algorithms, but such algorithms risk missing globally best solutions (Zimmerman et al., 1997; Wang et al., 2005). In our previous work we demonstrated that an enumeration-enhanced branch-and-bound algorithm can efficiently search surprisingly large spaces of NMR assignments (Vitek et al., 2005). Such systematic searches have not, however, been able to handle sparse data sets and large search spaces. An effective strategy for sparse data is to stochastically explore the tree by making randomized choices such that most resources are devoted to "promising" portions of the space, but some time is also spent exploring other subspaces that might be hiding good solutions. Here we design a hybrid algorithm that combines systematic exploration with desirable properties of local search methods such as simulated annealing (Buchler et al., 1997; Lukin et al., 1997), genetic algorithms (Bartels et al., 1997) and Monte Carlo optimization (Hitchens et al., 2003) that have been successfully applied to backbone resonance assignment. We also incorporate properties of other well-known methods, namely Tabu search and ant colony optimization (Hoos and Stützle, 2005) which have not yet been applied to the problem.

The search algorithm is illustrated in Figure 5. Its steps can be characterized as performing both intensification (i.e., searching for full mappings with better scores) and diversification (i.e., escaping locally optimal portions of the search space). The inner loop, *A*, of the algorithm iteratively makes *descents* into the tree, in each descent seeking to identify more plausible assignments. In a manner similar to local search methods such as simulated annealing, the algorithm preserves some mappings (according to a pre-specified probability) between iterations (Step 5 back to Step 2). The remaining portion of the space is explored by enumerating partial mappings that are consistent with the decisions at Step 2, and then systematically descending into the tree (Steps 3 and 4). The descent follows branches in an order that heuristically attempts to keep the search focused: at each step, it fixes the position to which the smallest number of alternative pseudoresidues can be mapped. To avoid being trapped in lower portions of large trees, the algorithm only visits a pre-specified number of nodes and then returns to Step 2. The systematic traversal of an ordered tree within a descent is a special case of Tabu search which prevents multiple examination of recently visited nodes.

Throughout the intensification steps, the algorithm dynamically learns promising directions of search. For example, only positions that are unambiguous according to the currently best mappings can be preserved at Step 2. The order in which the algorithm maps the possible pseudoresidues to the selected position is also determined from the best mappings found. Pseudoresidues mapped to the position in the currently best mappings will be examined first, other non-missing pseudoresidues next, and placeholders for entirely missing pseudoresidues will be tried last. Every time a better full mapping is found, the algorithm accordingly strengthens the bounds (a)–(d) derived from the probability model. As a result, more partial mappings will be eliminated during the enumeration step, and new positions will be chosen to fix at Step 2 and to branch at Step 3.

The algorithm employs multiple diversification strategies to escape locally optimal portions of the space. First, randomization is used when no full mappings are found, or in the case of a tie (e.g., when choosing which position to fix in the case where multiple positions have the same minimal number of alternative pseudoresidues). Second, with a small probability, an entirely random order is selected for visiting the branches from a node. Third, when no improvement has been made for a number of descents into the tree, the algorithm is allowed to "forget" the currently best solutions, discard the learning mechanisms, and at Steps 2 through 5 consider full mappings that are worse than the best mappings found so far. Such worsening steps are helpful to escape plateaus and basins of the search space in the neighborhood of the currently best solutions. The final diversification strategy is based on the observation that a single execution chain is more likely to be trapped in a local optimum than several independent ones. This is the basis of population-based optimization methods such as evolutionary algorithms and ant colony optimization. Our algorithm launches multiple independent search chains (Step 1) and, after a pre-specified number of descents, exchanges information regarding the best mappings found by the chains (Step 6). The chains are then re-launched using as their starting points the best of all mappings (the outer loop, *B*, in the figure).

The algorithm stops when no improvements have been made over a user-specified number of descents. As is the case with all stochastic search algorithms, it always outputs some solution, which may be either globally or locally optimal. Although finding the global optimum is the goal, in some cases it may be more important to find reasonably good mappings than to find no solution at all. This is arguably the case in sparse data situations where even the global optimum is expected to contain errors due to low information content. To validate the results obtained by the algorithm we suggest repeating the execution and comparing the solutions.

Implementation

Our method is implemented as version 3 of the software package Model-Based Assignment (MBA). MBA takes the same input format as MARS, but since we do not use structural information, those lines of the MARS parameter file are ignored. MBA supports three amino acid typing methods: prediction of chemical shifts using

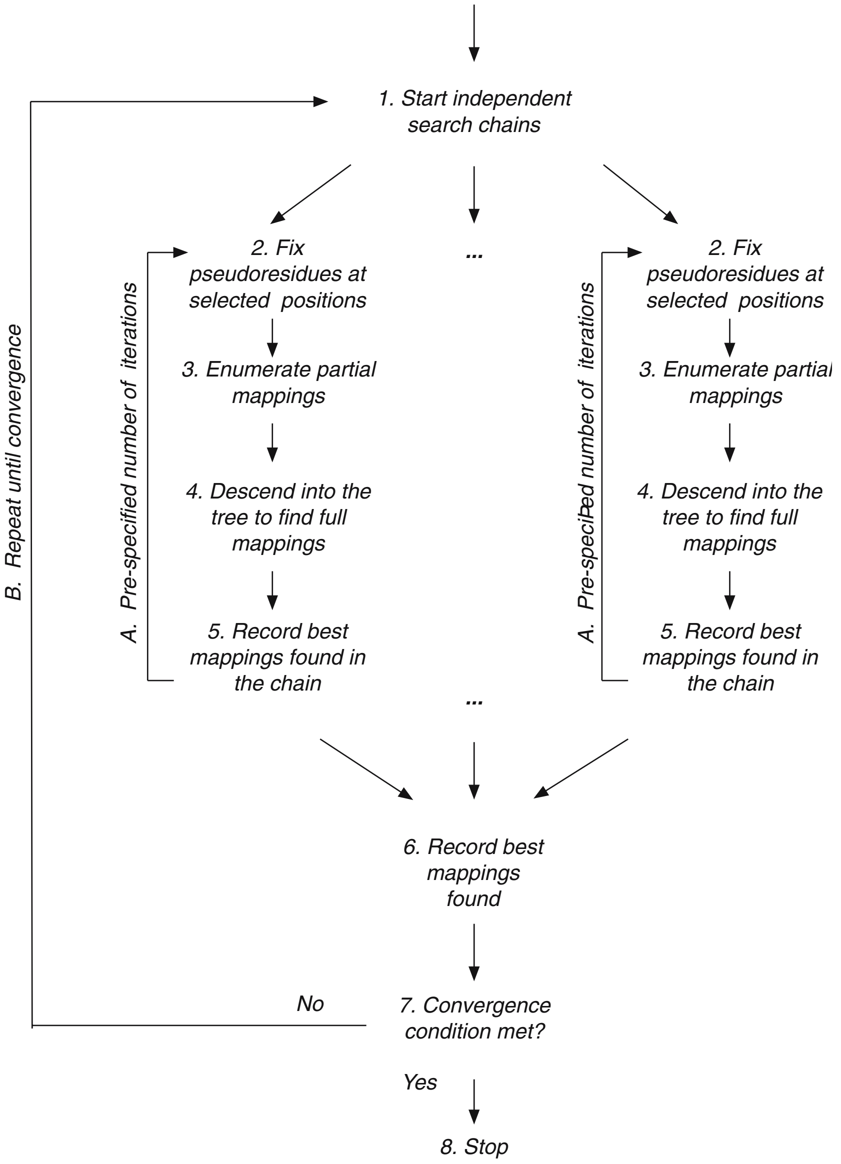
1. Start independent search chains
B. Repeat until convergence
A. Pre-specified number of iterations
2. Fix pseudoresidues at selected positions
3. Enumerate partial mappings
4. Descend into the tree to find full mappings
5. Record best mappings found in the chain
...
...
A. Pre-specified number of iterations
2. Fix pseudoresidues at selected positions
3. Enumerate partial mappings
4. Descend into the tree to find full mappings
5. Record best mappings found in the chain
6. Record best mappings found
No
7. Convergence condition met?
Yes
8. Stop

Figure 5. Schematic illustration of the stochastic search algorithm for inferential backbone resonance assignment.

(Wan and Jardetzky, 2002) as described above, the method developed by (Marin et al., 2004), and a method using raw statistics from the BMRB to generate predictive distributions of chemical shifts. The current implementation does not allow fixing pseudoresidue connectivities or restricting amino acid types, but these extensions can be easily implemented in the future.

The output of the algorithm is contained in one file. The file states the execution parameters and summarizes the data set and the reference solution if it is provided. The output is followed by a description of the assignment result: the posterior distribution of globally consistent mappings, and the posterior distributions of the pseudoresidues mapped to each position in the sequence. Finally, if a reference solution is provided, the file provides a short comparison between it and the assignment by MBA.

The program in its current form requires a significant amount of computational resources. However, the algorithm is highly parallel in that independent search chains can be run on separate processors. Synchronization of the chains is implemented through a simple shared file; each chain deposits and reads search results in this file at Step 6 in Figure 5. Besides being simple, the advantage of this implementation is that a shared filesystem is the only requirement for synchronization of multiple processors running the search. The shared file can also be used to monitor the progress of the search and diagnose convergence. In addition, a search can be stopped and subsequently restarted from the same state. Search chains in this case will be initialized by reading the last mappings deposited in the shared file. All the data sets described in the following section were assigned using a cluster of 18–30 nodes. Each execution time was under 24 h and compared favorably to the execution time required by MARS for large proteins.

The program is implemented in the Java programming language, is open source and is freely available under the JPL license at http://www.stat.purdue.edu/~ovitek/mba/mba.html.

Results

We evaluate the performance of our algorithm for six experimental and 24 simulated data sets. The simulated data are used to illustrate two types of sparsity: sparsity due to many missing peaks and few unambiguous matches of pseudoresidues (which results in large search spaces of candidate mappings), and sparsity due to experimental noise. Software tools MARS (Jung and Zweckstetter 2004) and CASA (Wang et al., 2005) serve as representative automated assignment programs for assessing the difficulty of assignment and quality of our method. Assignments with MARS were obtained using version 1.1.3 of the software[1] . We report both assignments of high confidence (marked "H"), and of moderate-to-high confidence (marked "M+H") in the MARS output. The results of assignments with CASA are as reported in (Wang et al., 2005).

Experimental data

Table 1 describes the experimental data sets used to test our approach. The data for Syk protein-tyrosine kinase SH2 domains, PLCC SH2 domains and West nile capsid (unpublished data), as well as the data for Dengue fever virus capsid (Ma et al., 2004), were provided by Dr. Post, Purdue University. The data for Human ubiquitin are publicly available from the University College London/Ludwig Institute for Cancer Research Joint NMR Laboratory (Ubiq. NMR Resource). The data for Z domain were provided as a test to the AutoAssign program (Zimmerman et al., 1997). Pseudoresidues for input to our program were manually compiled for all proteins except Z domain. For Z domain, pseudoresidues were extracted from the output of the AutoAssign program.

Table 1 illustrates the sparsity characteristics of the proteins. The lengths of the proteins cover a range between 71 and 257 residues. Pseudoresidues for the first four proteins were compiled from a minimal number of triple-resonance experiments, and only C^{α} and C^{β} resonance types are available. More resonance types are available for Human ubiquitin and Z domain, and we study the performance of our approach by sequentially reducing the number of resonance types. Overall, the data sets have between 2 and 40 entirely missing pseudoresidues, and between 2% and 20% of chemical shifts of C', C^{α}, C^{β} and H^{α} resonance

[1] http://www.mpibpc.gwdg.de/abteilungen/030/zweckstetter

Table 1. Assignment results for experimental data

Protein	# Residues		Resonance types[a]	# Miss pseudor.	% Miss cs	Unique pairs[b] (%)	Reliable/Disagree[c]			# Residues improved[d]	
	Total	Pseudo					MARS M + H[e]	MARS H[e]	MBA	M + H[e]	H[e]
Syk protein-tyrosine kinase	257	209	C^{α} C^{β}	40	6.7	1.9	180/NA	129/NA	154/6	−32	19
PLCC	106	91	$C^{\alpha}C^{\beta}$	11	7.9	25.3	82/NA	75/NA	74/0	−8	−1
West nile capsid	105	100	$C^{\alpha}C^{\beta}$	5	10.4	0	88/NA	79/NA	86/0	−2	7
Dengue fever virus capsid	100	81	$C^{\alpha}C^{\beta}$	14	14.3	0	52/NA	46/NA	57/1	5	10
Human ubiquitin	76	70	$C'_{-1}C^{\alpha}C^{\beta}$	2	6.5	20	70/NA	68/NA	70/0	0	2
			$C^{\alpha}C^{\beta}$		7.5	20	70/0	68/0	68/0	−2	0
			C^{α}		2.1	0	4/0	3/0	53/0	49	50
Z domain	71	67	$C'_{-1}C^{\alpha}C^{\beta}H^{\alpha}$	1	12.6	5.9	45/NA	26/NA	65/0	20	39
			$C'_{-1}C^{\alpha}C^{\beta}$		16.7	0	39/0	25/0	65/0	26	40
			$C^{\alpha}C^{\beta}$		20.0	0	39/0	24/0	65/0	26	41
			C^{α}		7.4	0	0/0	0/0	14/0	14	14

The true assignment for these data are unknown. We compare the results to a reference mapping, i.e., the highly reliable assignments by MARS obtained with all resonance types.

[a] Resonance types used in addition to H and N. Match tolerances are 0.25 ppm for C', and 0.5 ppm for both C^{α} and C^{β}, and 0.05 ppm for H^{α}. Subscript −1 indicates that only sequential chemical shifts of the resonance type were used.

[b] Proportion of unambiguous pairs of pseudoresidues among all pairs with valid matches and alignment.

[c] Number of reliably mapped pseudoresidues/number of reliably mapped pseudoresidues that disagree with the reference assignment.

[d] Number of reliably mapped and agreeing with the reference pseudoresidues by MBA minus number of reliably mapped and agreeing with the reference pseudoresidues by MARS.

[e] M + H: "reliable" in MARS output defined as medium or highly reliable. H: "reliable" in MARS output defined as highly reliable.

types are missing in the observed pseudoresidues. We analyze the data with match tolerances typically used in the Post lab, namely 0.25 ppm for C′, 0.5 ppm for C^{α} and C^{β}, and 0.05 ppm for H^{α}. These tolerances are likely also appropriate for Human ubiquitin and Z domain when using a small subset of available resonance types. Under these conditions, only 0–25% of the valid matched and aligned pairs of pseudoresidues are unambiguous.

Table 1 summarizes assignment results for these experimental data. The "true" assignments for these proteins are unknown. Therefore we estimate accuracy by comparing the results to a reference solution which is determined by highly reliable assignments in MARS when using all resonance types. One can see from the table that there is a small number of disagreements between the two methods. However, whenever there are no unique pairs or only a small fraction of unique pairs, MBA consistently assigns more positions than MARS. Our approach is particularly efficient at assigning small proteins with a minimal number of resonance types, such as Human ubiquitin with only C^{α} chemical shifts. In this case MBA reliably and correctly assigned 53 residues whereas MARS only provided 3 or 4 reliable assignments. MBA also typically assigns more positions than the version of MARS considering only highly reliable assignments. The numbers vary when we consider both moderate and high quality assignments in MARS. We return to the discussion of appropriate comparison between the two methods at the end of this section.

Synthetic data with large search spaces

To test our approach on data for which the correct assignment is known, we analyzed results for 9 synthetic data sets used in the publication describing MARS (Jung and Zweckstetter, 2004). The data sets were compiled and kindly provided to us by the authors of the paper. Pseudoresidues were created from entries to the BMRB database (Seavey et al., 1991) by taking the chemical shifts for each position together with those for the preceding position. All chemical shifts missing in the BMRB entries were left missing. The sparsity characteristics of the data sets are shown in Table 2. The proteins cover a wide range of

Table 2. Description of simulated data sets from (Jung and Zweckstetter, 2004)

Protein	# Residues		Resonance types[a]	# Miss pseudor.	% Miss cs	Unique pairs[b]	
	Total	W.data				Moderate match tol.[c] (%)	Large match tol.[c] (%)
Malate synthase G	723	654	$C'C^{\alpha}C^{\beta}$	37	2.7	71	51.4
Maltose bind. protein	370	335	$C'_{-1}C^{\alpha}C^{\beta}$	13	3.7	0	0
			$C^{\alpha}C^{\beta}$		4.8	0	0
N-term. dom. enzyme I	259	248	$C'_{-1}C^{\alpha}C^{\beta}$	6	8.3	0	0
			$C^{\alpha}C^{\beta}$		3.9	0	0
E-cadherin dom. II & III	227	167	$C^{\alpha}C^{\beta}$	45	13	0	0
Human prion protein	210	190	$C^{\alpha}C^{\beta}$	4	11.3	0.5	0
Superoxide dismutase	192	117	$C'C^{\alpha}C^{\beta}$	66	9.2	67.5	60.7
			$C^{\alpha}C^{\beta}$		4.9	46.2	36.8
Calmodulin /M13	148	144	$C'C^{\alpha}$	1	0	73.6	29.2
E. coli EmrE	110	74	$C'C^{\alpha}C^{\beta}$	0	30.3	0	0
			$C^{\alpha}C^{\beta}$		32.6	0	0
Human ubiquitin	76	72	$C'_{-1}C^{\alpha}C^{\beta}$	0	3	80.56	66.7
			$C^{\alpha}C^{\beta}$		3.8	80.56	66.7
			C^{α}		0	11.1	1.4

[a] Resonance types used in addition to H and N. Subscript −1 indicates that only sequential chemical shifts of the resonance type were used.

[b] Proportion of unambiguous pairs of pseudoresidues among all pairs with valid matches and alignment.

[c] Moderate match tolerances: 0.15 ppm for C′, 0.2 ppm for C^{α} and 0.4 ppm for C^{β}. Large match tolerances: 0.25 ppm for C′, 0.5 ppm for both C^{α} and C^{β}.

76–723 residues. Overall the data sets contain 0–66 entirely missing pseudoresidues, 0–32% missing chemical shifts for the C′, C^{α} and C^{β} resonance types, and at most three resonance types. To test the performance of our method under different assignment conditions, we use two sets of match tolerances: moderate (0.15 ppm for C′, 0.2 ppm for C^{α} and 0.4 ppm for C^{β}), and large (0.25 ppm for C′ and 0.5 ppm for both C^{α} and C^{β}). One can see from Table 2 that, although smaller match tolerances increase the number of unambiguous valid pairs of pseudoresidues, many data sets have no or few such pairs. Therefore, sparsity characteristics of the data will result in relatively large search spaces of candidate assignments.

Assignment results for the synthetic data sets are summarized in Table 3. On the subset of the proteins reported in Wang et al. (2005), MBA is typically more accurate than CASA: the median number of errors in these data sets is 0 for MBA and 9 for CASA for moderate match tolerances (respectively 2.5 and 32.5 for large tolerances). In comparison with MARS when using large match tolerances, MBA typically assigns more positions and has a comparably small number of errors. The difference between two methods is small when moderate match tolerances are used. We defer the discussion of the choice of match tolerances to the end of this section.

Table 3 provides interesting insights into the effects of sparsity. MARS performs comparably to or better than MBA in the cases of Human ubiquitin with C^{β} resonance type, Superoxide dismutase and Malate synthase G. These are the data sets with the largest number of unambiguous pairs of pseudoresidues (Table 2). Calmodulin, however, demonstrates the effect of sparsity (via match tolerance): increasing the match tolerance decreases the number of unambiguous pairs from 73% to 29% (Table 2) and results in a 103- or 109-residue shortfall of MARS relative to MBA (Table 3). A similar effect is observed with increased sparsity in Human ubiquitin due to fewer resonance types. In general, MBA and MARS perform comparably for proteins with a sufficient fraction of unambiguous pairs of pseudoresidues. On the other hand, MBA performs better on proteins with a small fraction of unambiguous pairs (particularly when there are none) and when using large match tolerances.

These synthetic data sets are useful as they allow us to evaluate the performance of algorithms for large proteins with many missing chemical shifts, the characteristics that result in large search spaces of candidate assignments. However, in several ways the data sets are not representative of typical noisy experimental data sets. First, the chemical shifts in the pseudoresidues have no experimental error. This results in all match differences of 0 ppm, not a realistic feature of experimental data. Second, the synthetic data sets overestimate the number of missing chemical shifts. Missings in a BMRB entry indicate that the chemical shifts could not be reliably assigned, but not necessarily that the spectra did not contain the corresponding peaks. Third, all chemical shifts from a BMRB entry are considered as both sequential and within-residue chemical shifts, but in a real-life scenario, a chemical shift may be represented by only one of the two types. Finally, the data sets contain no extra pseudoresidues. Since extra and noise peaks are common and contribute greatly to the uncertainty in assignment, it is desirable to include extra pseudoresidues in a synthetic data set. The impact of these restrictive simulation assumptions can be seen in for the case of Human ubiquitin. The experimental data set for Human ubiquitin with C'_{-1}, C^{α} and C^{β} resonance types has 20% of valid pseudoresidue pairs as unambiguous under large match tolerances, but the simulated data set with the same resonance types has 67% unambiguous pairs. This in turn results in an overly optimistic assessment of the performance of the algorithms. Note the difference in assigned positions between real and simulated data for Human ubiquitin in Tables 1 and 3. Similar differences between experimental and simulated data were observed for Calmodulin in (Wang et al., 2005).

Synthetic data with sparsity and noise estimated from real data sets

There is currently no consensus on how to appropriately generate synthetic NMR data. Here we attempt to design a simulation that is representative of a typical experiment. In the following we study statistical properties of noise of previously assigned experimental data sets, and create noisy versions of entries to a database. The

Table 3. Assignment results for simulated data from (Jung and Zweckstetter, 2004)

Protein	Resonance types[a]	Moderate match tolerances[b]						Large match tolerances[b]					
		Reliable/error[c]				# Residues improved[d]		Reliable/error[c]				# Residues improved[d]	
		CASA	MARS M + H[e]	MARS H[e]	MBA	M + H[e]	H[e]	CASA	MARS M + H[e]	MARS H[e]	MBA	M + H[e]	H[e]
Malate synthase G	$C'C^{\alpha}C^{\beta}$	653/4	648/0	637/0	638/0	−10	1	519/68	648/0	637/0	645/0	−3	8
Maltose bind. protein	$C'_{-1}C^{\alpha}C^{\beta}$	328/2	333/0	324/0	329/0	−4	5	328/2	314/0	297/0	328/0	14	31
	$C^{\alpha}C^{\beta}$	–	329/0	321/0	329/0	0	8	–	320/1	306/0	329/0	10	23
N-term. dom. enzyme I	$C'_{-1}C^{\alpha}C^{\beta}$	–	246/0	246/0	247/0	1	1	–	240/0	234/0	247/0	7	13
	$C^{\alpha}C^{\beta}$	–	246/0	246/0	247/0	1	1	–	237/0	233/0	247/0	10	14
E-cadherin dom. II & III	$C^{\alpha}C^{\beta}$	133/54	114/4	104/3	124/13	1	10	133/76	82/3	74/2	108/4	25	32
Human prion protein	$C^{\alpha}C^{\beta}$	132/19	125/0	117/0	133/0	8	16	96/31	116/0	107/0	128/3	9	18
Superoxide dismutase	$C'C^{\alpha}C^{\beta}$	117/9	107/0	106/0	106/0	−1	0	117/10	108/1	104/0	103/2	−6	−3
	$C^{\alpha}C^{\beta}$	117/17	105/0	103/0	94/1	−12	−10	117/34	105/0	103/0	94/0	−11	−9
Calmodulin/M13	$C'C^{\alpha}$	132/1	144/0	144/0	143/0	−1	−1	NA	40/0	34/0	143/0	103	109
E. coli EmrE	$C'C^{\alpha}C^{\beta}$	79/9	62/3	58/2	71/11	1	4	79/29	42/0	36/0	67/4	18	24
	$C^{\alpha}C^{\beta}$	79/31	59/6	55/6	55/3	−1	3	79/49	26/1	22/0	60/5	23	26
Human ubiquitin	$C'_{-1}C^{\alpha}C^{\beta}$	–	72/0	72/0	72/0	0	0	–	72/0	72/0	72/0	0	0
	$C^{\alpha}C^{\beta}$	–	72/0	72/0	72/0	0	0	–	72/0	72/0	72/0	0	0
	C^{α}	–	18/0	13/0	65/0	47	52	–	4/0	1/0	65/0	61	64

[a] Resonance types used in addition to H and N. Subscript −1 indicates that only sequential chemical shifts of the resonance type were used.

[b] Moderate match tolerances: 0.15 ppm for C′, 0.2 ppm for C^{α} and 0.4 ppm for C^{β}. Large match tolerances: 0.25 ppm for C′, 0.5 ppm for both C^{α} and C^{β}.

[c] Number of reliably mapped pseudoresidues/number of reliably but incorrectly mapped pseudoresidues.

[d] Number of reliably and correctly mapped pseudoresidues by MBA minus number of reliably and correctly mapped pseudoresidues by MARS.

[e] M + H: "reliable" in MARS output defined as medium or highly reliable. H: "reliable" in MARS output defined as highly reliable.

Table 4. Average fraction of missing and extra data from (Zimmerman et al., 1997) used to simulate data in Table 6

Missing pseudo residues	Extra pseudo residues	Missing chemical shifts in correct pseudoresidues				Missing chemical shifts in extra pseudoresidues			
		C^{α}_{-1}	C^{β}_{-1}	C^{α}	C^{β}	C^{α}_{-1}	C^{β}_{-1}	C^{α}	C^{β}
0.021	0.037[a]	0.056	0.118	0.003	0.187	0.468	0.328	0.362	0.497

[a]Calculated from proteins having no minor conformations.

simulated data are sparse in that they only contain C^{α} and C^{β} resonance types.

To determine realistic properties of noise we consider experimental data sets for six proteins, Fgf, Rnase, RnaseC6572S, Ns1, CspA and Z domain, provided as a test to the AutoAssign program (Zimmerman et al., 1997). Although the correct assignments for these proteins are unknown, the quality of the reference solutions provided by AutoAssign is sufficient for statistical considerations. On average the number of missing pseudoresidues is 2.1% of assignable positions, and the number of extra pseudoresidues is 3.7% of assignable positions for Ns1, CspA and Zdomain (the proteins with no minor conformations). The fractions of missing chemical shifts in the observed pseudoresidues are reported in Table 4. Note that C^{β} contains more missing chemical shifts than C^{α}, sequential chemical shifts have more missings than within-residue chemical shifts, and extra pseudoresidues have more missing chemical shifts than correct pseudoresidues. The distributions of the differences of matched chemical shifts are shown in Figure 2. One can see that AutoAssign allows fairly loose matches and on average 3.5% of “true” match differences fall outside of match tolerances of 0.5 ppm for C^{α} and C^{β}.

As a basis for simulations, we randomly selected 16 entries from RefDB (Zhang et al., 2003), ranging in length from 52 to 216 residues. Since it is unlikely that all the missing chemical shifts in the database correspond to truly missing peaks, we replaced missing data with values simulated from the expected distributions of the corresponding resonance and amino acid types. We then compiled pseudoresidues from the set of chemical shifts for each residue in conjunction with those for the preceding residue. At this point all the pseudoresidues had no missing chemical shifts.

We introduced noise by randomly deleting correct pseudoresidues and by adding extra pseudoresidues with probabilities equal to the observed frequencies reported above. We introduced missing chemical shifts independently for each resonance and amino acid type with probabilities in Table 4. Finally, we added noise to the chemical shifts by sampling from the histograms of match tolerances in Figure 2. As shown in Table 5, the procedure resulted in data sets with between 1 and 6 entirely missing pseudoresidues. Only 5 of the 18 data sets have more than 15% unambiguous pairs of pseudoresidues under moderate match tolerances, and only one does under large match tolerances.

Assignment results for these data sets are detailed in Table 6. As with previously discussed data sets, MBA performs best, relative to MARS, under large match tolerances (and particularly when only considering MARS' highly reliable assignments). The difference between the methods is less pronounced when using a moderate match tolerance. At the same time, one should note an increased number of errors in both MBA and MARS as compared to the previously discussed data, primarily due to the extra pseudoresidues. The increase is particularly apparent in the moderate-to-high reliability assignments by MARS.

Summary of assignment results

Figure 6 provides a summary overview of the assignment results for all the synthetic data presented in the paper. Panels (a) and (b) in the figure summarize the true positive rates of the assignments, i.e., the fraction of assignable positions that are correctly assigned. Panels (c) and (d) summarize the error rates, i.e., the fraction of incorrectly assigned positions among all the assigned positions. The data sets are separated by simulation type, match tolerances, and method of finding the assignments. The overview provides useful insight into the following questions.

How important is the choice of simulation procedure?

The data sets in panels (a) and (c) are generated without adding simulated experimental noise, while the data sets in panels (b) and (d) are instances of noisy data. One can see from the figure that noisy data yield smaller assigned fractions and larger error rate. This indicates that simulation procedure can have a significant impact on our conclusions. It is therefore important to generate synthetic data carefully, and accurately represent the specific experimental conditions under investigation.

Are large match tolerances necessary?

Figure 6 shows that the choice of match tolerances depends on both the noise level in the data and the assignment procedure. For data with no experimental noise, match tolerances did not dramatically affect error rates for either MARS or MBA, and MARS benefited from smaller match tolerances in terms of the number of assigned positions. The conclusions are different for noisy data. Increased match tolerances led to MARS making more M + H assignments at a higher error rate, but fewer H assignments at a lower error rate. Therefore, the choice of match tolerances for MARS with noisy data trades off between true and false positive assignments. MBA, on the other hand, benefited from large match tolerances by both increasing the number of true positive assignments, and decreasing the number of false positive assignments. We conclude that large match tolerances optimize the performance of MBA, and recommend using large match tolerances with our approach.

How can one compare two assignment methods?

When an assignment program provides multiple metrics of reliable assignments, as "M + H" and "H" in the case of MARS, it is best to compare

Table 5. Description of data simulated from randomly chosen entries from RefDB. The noise model is reported in Table 4 and Figure 2

BMRB id	# Residues		Resonance types[a]	# Miss pseudores	Unique pairs[b]	
	Total	W.data			Moderate match tol.[c] (%)	Large match tol.[c] (%)
4101	216	205	$C^{\alpha}C^{\beta}$	4	0	0
5299	197	174	$C^{\alpha}C^{\beta}$	3	0	0
5756	184	174	$C^{\alpha}C^{\beta}$	6	7.4	0
5362	175	168	$C^{\alpha}C^{\beta}$	5	9.7	2.3
4081	165	157	$C^{\alpha}C^{\beta}$	3	0	0
4053	149	138	$C^{\alpha}C^{\beta}$	5	0.7	0
4641	144	136	$C^{\alpha}C^{\beta}$	3	0	0
4046	134	127	$C^{\alpha}C^{\beta}$	5	0	0
5579	134	129	$C^{\alpha}C^{\beta}$	3	0	0
5866	129	115	$C^{\alpha}C^{\beta}$	4	0	0
5507	128	116	$C^{\alpha}C^{\beta}$	3	15.3	4.2
4162	101	92	$C^{\alpha}C^{\beta}$	3	20.2	7.5
5573	90	83	$C^{\alpha}C^{\beta}$	3	0	0
4898	86	80	$C^{\alpha}C^{\beta}$	3	24.4	11.0
4895	66	61	$C^{\alpha}C^{\beta}$	3	25.4	9.5
		63	C^{α}	1	0	0
5594	52	43	$C^{\alpha}C^{\beta}$	3	51.1	37.8
		45	C^{α}	1	0	0

[a] Resonance types used in addition to H and N. Subscript −1 indicates that only sequential chemical shifts of the resonance type were used.
[b] Proportion of unambiguous pairs of pseudoresidues among all pairs with valid matches and alignment.
[c] Moderate match tolerances: 0.15 ppm for C′, 0.2 ppm for C^{α} and 0.4 ppm for C^{β}. Large match tolerances: 0.25 ppm for C′, 0.5 ppm for both C^{α} and C^{β}.

metrics that have similar error rates. Figure 6c shows that MBA can be compared to moderate-to-high confidence assignments on data sets with no added noise. However, when assigning noisy data, the error rates of MBA are comparable to the rates of highly confident assignments, and are smaller than the rates of moderate-to-high confidence assignments by MARS (Figure 6d). Therefore, with noisy data it is most appropriate to compare the results of MBA with the assignments by MARS marked by "H".

When is MBA particularly helpful?

Figure 6 shows that MBA is particularly helpful when data are noisy and require large match tolerances. For these cases, MBA has the highest median fraction of assigned positions and the lowest median error rate.

Discussion

In this paper we presented an inferential approach and a hybrid stochastic search algorithm for backbone resonance assignment with sparse data. We tested the approach on a total of 44 experimental and synthetic data sets for proteins ranging in length from 52 to 723 residue. The accuracy of our approach is due to the scoring function based on an empirical Bayesian probability model. When data are sparse and support a large number of globally consistent mappings, the statistical model allows us to select a subset of the mappings for reliable conclusions. At the same time, our algorithm efficiently searches large spaces of candidate mappings. The efficiency is due to the hybrid nature of the algorithm that combines desirable properties of several successful methods of

Table 6. Assignment results for simulated data for randomly chosen entries from RefDB. The noise model is reported in Table 4 and Figure 2

Protein	Resonance types[a]	Moderate match tolerances[b]					Large match tolerances[b]				
		Reliable/error[c]			# Residues improved[d]		Reliable/error[c]			# Residues improved[d]	
		MARS M + H[e]	MARS H[e]	MBA	M + H[e]	H[e]	MARS M + H[e]	MARS H[e]	MBA	M + H[e]	H[e]
4101	$C^\alpha C^\beta$	127/1	105/0	142/6	10	31	121/1	98/0	163/1	42	64
5299	$C^\alpha C^\beta$	159/4	130/1	133/1	−23	3	163/2	124/0	139/0	−22	15
5756	$C^\alpha C^\beta$	106/2	93/2	140/6	30	43	133/7	89/4	148/4	18	59
5362	$C^\alpha C^\beta$	146/10	138/7	135/2	−3	2	141/7	114/1	143/2	7	28
4081	$C^\alpha C^\beta$	132/1	109/1	102/2	−31	−8	125/9	83/1	109/1	−8	26
4053	$C^\alpha C^\beta$	121/6	93/3	101/9	−23	2	130/8	100/3	134/12	0	25
4641	$C^\alpha C^\beta$	53/1	37/0	86/3	31	46	68/0	56/0	91/0	23	35
4046	$C^\alpha C^\beta$	121/16	107/4	99/4	−10	−8	123/5	116/4	117/5	−6	0
5579	$C^\alpha C^\beta$	108/2	93/1	99/0	−7	7	111/4	76/2	108/0	1	34
5866	$C^\alpha C^\beta$	55/2	51/2	78/2	23	27	65/2	54/1	90/3	24	34
5507	$C^\alpha C^\beta$	108/1	96/1	88/0	−19	−7	107/2	91/0	101/0	−4	10
4162	$C^\alpha C^\beta$	78/2	68/2	80/1	3	13	85/4	67/2	80/1	−2	14
5573	$C^\alpha C^\beta$	34/1	28/0	42/1	8	13	58/6	29/0	49/4	−7	16
4898	$C^\alpha C^\beta$	75/4	60/0	65/0	−6	5	68/1	52/0	72/0	5	20
4895	$C^\alpha C^\beta$	52/3	50/2	40/0	−9	−8	49/3	34/1	53/0	7	20
	C^α	6/2	2/1	16/0	12	15	3/1	1/1	19/3	14	16
5594	$C^\alpha C^\beta$	43/2	43/2	42/1	0	0	43/2	43/2	40/0	−1	−1
	C^α	7/0	7/0	18/1	10	10	9/0	7/0	17/0	8	10

[a] Resonance types used in addition to H and N.

[b] Moderate match tolerances: 0.15 ppm for C', 0.2 ppm for C^α and 0.4 ppm for C^β. Large match tolerances: 0.25 ppm for C', 0.5 ppm for both C^α and C^β.

[c] Number of reliably mapped pseudoresidues/number of reliably but incorrectly mapped pseudoresidues.

[d] Number of reliably and correctly mapped pseudoresidues by MBA minus number of reliably and correctly mapped pseudoresidues by MARS.

[e] M + H: "reliable" in MARS output defined as medium or highly reliable. H: "reliable" in MARS output defined as highly reliable.

stochastic search. Statistical bounds on scores of partial mappings also contribute to the efficiency of the method.

One should note that assignment errors are not necessarily due to the performance of the algorithm, but may also result from the insufficient information content in the data. For example, when we tried to assign larger proteins with only the C^{α} resonance type, the algorithm typically found globally consistent mappings covering most positions and having better scores than the reference solution. In other words, the algorithm could provide tighter matches and better quality alignments of C^{α} chemical shifts than in the reference solution. The algorithm could not make use of any other information to correct the errors. The situation where the assignment reflects the properties of the specific data set rather the true resonances that generated the data is known as overfitting. This can be diagnosed by altering the data set slightly without modifying its underlying structure. For example, one can randomly exchange sequential and within-residue chemical shifts that were generated by the same nucleus according to

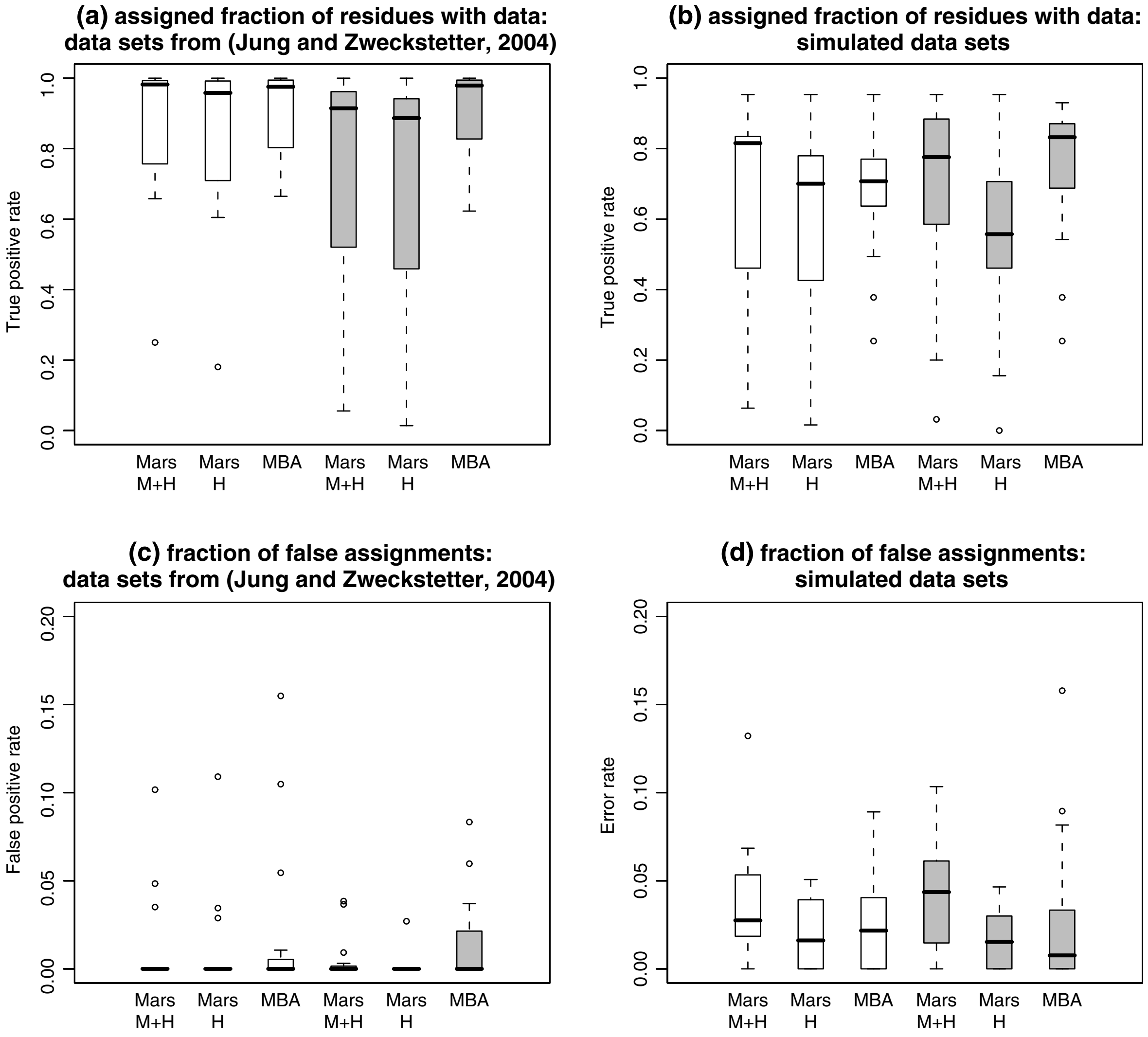

Figure 6. Boxplots of the assignment results for all the synthetic data sets. (a)–(b): number of assigned residues divided by number of assignable residues; (c)–(d): number of incorrect assignments divided by number of assigned residues. Each box covers 50% of the values, horizontal line in the center indicating the median, and circles indicating outlying values. White boxplots denote assignments with moderate match tolerances, filled boxplots denote assignments with large match tolerances.

the assignment. Alternatively, one can add a small amount of noise to the chemical shifts. Overfitting will typically result in a different set of mappings on the modified data set. However, the only way to obtain the correct assignment in this case is to collect more data.

Another error-prone situation arises when the algorithm does not reach a globally optimal solution. Sub-optimal solutions can be detected by launching independent executions multiple times. If the searches find different solutions it may be necessary to invest more computational resources (e.g., more iterations, more independent search chains, or more iterations with no improvement before convergence) for a better exploration of the space.

A limitation of the proposed approach is the significant computational resources it requires. However, the algorithm is highly parallel. The data sets discussed in this paper were analyzed on a cluster with between 18 and 30 nodes, and each execution took less then 24 h of computing time. Such clusters are becoming standard resources, and a relatively small investment for the analytical ability that they provide.

Although already efficient, the proposed approach will benefit from a number of improvements. The statistical model can be extended in a straightforward manner to incorporate additional experimental information such as from COSY spectra or selective amino acid labeling. It is also possible to extend the probability model to incorporate structural information. Finally, the search algorithm can be improved by combining its elements with resampling methods employed by MARS.

Acknowledgments

The authors would like to thanks Dr. Post, Purdue University, for providing the experimental data sets, Drs. Jung and Zweckstetter, Max Plank Institute for Biophysical Chemistry, for sharing their simulated data, and Drs. Moseley and Montelione, Rutgers University, for providing access to the AutoAssign data. This work was supported in part by a Purdue Dissertation Fellowship and by NSF grants EIA-9802068 and IIS-0502801.

References

Andrec, M. and Levy, R. (2002) *J. Biomol. NMR*, **23**, 263–270.

Atreya, H.S., Sahu, S.C., Chary, K.V.R. and Govil, G. (2000) *J. Biomol. NMR*, **17**, 125–136.

Bartels, C., Güntert, P., Billeter, M. and Wüthrich, K. (1997) *J. Comp. Chem.*, **18**, 139–149.

Buchler, N.E.G., Zuiderweg, E.P.R., Wang, H. and Goldstein, R.A. (1997) *J. Magn. Res.*, **125**, 34–42.

Burnham, K.P. and Anderson, D. (2002) *Model Selection and Multi-Model Inference*, 2nd edn., Springer.

Coggins, B.E. and Zhou, P. (2003) *J. Biomol. NMR*, **26**, 93–111.

Eghbalnia, R.H., Bahrami, A., Wang, L., Assadi, A. and Markley, J.L. (2005) *J. Biomol. NMR*, **32**, 219–233.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995) *Bayesian Data Analysis*, Chapman and Hall.

Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinsky, C.T (1999) *Stat. Sci.*, **14**, 382–417.

Hitchens, T.K., Lukin, J.A., Zhan, Y., McCallum, S.A. and Rule, G.S. (2003) *J. Biomol. NMR*, **25**, 1–9.

Hoos, H. and Stützle, T. (2005) *Stochastic Local Search: Foundations and Applications*, Elsevier, CA.

Kass, R.E. and Raftery, A.E. (1995) *J. Am. Stat. Assoc.*, **90**, 773–795.

Jung, J.-S. and Zweckstetter, M. (2004) *J. Biomol. NMR*, **30**, 11–24.

Lukin, J.A., Gove, A.P., Talukdar, S.N. and Ho, C. (1997) *J. Biomol. NMR*, **9**, 151–166.

Ma, L., Jones, C.T., Groesch, T.D., Kuhn, R.J. and Post, C.B. (2004) *Proc. Natl. Acad. Sci.*, **101**, 3414–3419.

Marin, A., Malliavin, T., Nicholas, P. and Delsuc, M.-A. (2004) *J. Biomol. NMR*, **30**, 47–60.

McGuffin, L.J., Bryson, K. and Jones, D.T. (2000) *Bioinformatics*, **16**, 404–405.

Moseley, H.N.B. and Montelione, G.T. (1999) *Curr. Opin. Struct. Biol.*, **9**, 635–642.

Rieping, W., Habeck, M. and Nilges, M. (2005) *Science*, **309**, 303–306.

Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J. (1991) *J. Biomol. NMR*, **1**, 217–236.

The Ubiquitin NMR Resource, *University College London/ Ludwig Institute for Cancer Research Joint NMR Laboratory*, http://www.biochem.ucl.ac.uk/bsm/nmr/ubq/.

Vitek, O. (2005) PhD Dissertation, Department of Statistics, Purdue University.

Vitek, O., Bailey-Kellogg, C., Craig, B., Kuliniewicz, P. and Vitek, J. (2005) *Bioinformatics*, **21**ii230–ii236.

Vitek, O., Vitek, J., Craig, B. and Bailey-Kellogg, C. (2004) *Stat. Appl. Genet. Mol. Biol.* **3**, Article 6. Available at: http://www.bepress.com/sagmb/vol3/iss1/art6.

Wan, Y. and Jardetzky, O. (2002) *J. Am. Chem. Soc.*, **124**, 14075–14084.

Wan, X., Tegos, T. and Lin, G. (2004) *J. Bioinform. Comp. Biol.*, **2**, 747–764.

Wang, J., Wang, T., Zuiderweg, E. and Crippen, G. (2005) *J. Bioinform. Comp. Biol.*, **33**, 261–279.

Wüthrich, K. (2003) *Angew. Chem.-Int. Edit.*, **42**, 3340–3363.

Zhang, H., Neal, S. and Wishart, D.S. (2003) *J. Biomol. NMR*, **25**, 173–195.

Zimmerman, D.E., Kulikowski, C.A., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S.S., Chien, C., Powers, R. and Montelione, G.T. (1997) *J. Mol. Biol.*, **269**, 592–610.