A Markov Random Field Framework for Protein Side-Chain Resonance Assignment[⋆]

Jianyang Zeng[1], Pei Zhou[2], and Bruce R. Donald[1,2,⋆⋆]

[1] Department of Computer Science, Duke University, Durham, NC 27708, USA.
[2] Department of Biochemistry, Duke University Medical Center, Durham, NC 27708, USA.

Abstract. Nuclear magnetic resonance (NMR) spectroscopy plays a critical role in structural genomics, and serves as a primary tool for determining protein structures, dynamics and interactions in physiologically-relevant solution conditions. The current speed of protein structure determination via NMR is limited by the lengthy time required in resonance assignment, which maps spectral peaks to specific atoms and residues in the primary sequence. Although numerous algorithms have been developed to address the *backbone* resonance assignment problem [68, 2, 10, 37, 14, 64, 1, 31, 60], little work has been done to automate *side-chain* resonance assignment [43, 48, 5]. Most previous attempts in assigning side-chain resonances depend on a set of NMR experiments that record through-bond interactions with side-chain protons for each residue. Unfortunately, these NMR experiments have low sensitivity and limited performance on large proteins, which makes it difficult to obtain enough side-chain resonance assignments. On the other hand, it is essential to obtain almost all of the side-chain resonance assignments as a prerequisite for high-resolution structure determination. To overcome this deficiency, we present a novel side-chain resonance assignment algorithm based on alternative NMR experiments measuring through-space interactions between protons in the protein, which also provide crucial distance restraints and are normally required in high-resolution structure determination. We cast the side-chain resonance assignment problem into a Markov Random Field (MRF) framework, and extend and apply combinatorial protein design algorithms to compute the optimal solution that best interprets the NMR data. Our MRF framework captures the contact map information of the protein derived from NMR spectra, and exploits the structural information available from the backbone conformations determined by orientational restraints and a set of discretized side-chain conformations (i.e., rotamers). A Hausdorff-based computation is employed in the scoring function to evaluate the probability of side-chain resonance assignments to generate the observed NMR spectra. The complexity of the assignment problem is first reduced by using a *dead-end elimination* (DEE) algorithm, which prunes side-chain resonance assignments that are *provably* not part of the optimal solution. Then an A* search algorithm is used to find a set of optimal side-chain resonance assignments that best fit the NMR data. We have tested our algorithm on NMR data for five proteins, including the FF Domain 2 of human transcription elongation factor CA150 (FF2), the B1 domain of Protein G (GB1), human ubiquitin, the ubiquitin-binding zinc finger domain of the human Y-family DNA polymerase Eta (pol η UBZ), and the human Set2-Rpb1 interacting domain (hSRI). Our algorithm assigns resonances for more than 90% of the protons in the proteins, and achieves about 80% correct side-chain resonance assignments. The final structures computed using distance restraints resulting from the set of assigned side-chain resonances have backbone RMSD $0.5 - 1.4$ Å and all-heavy-atom RMSD $1.0 - 2.2$ Å from the reference structures that were determined by X-ray crystallography or traditional NMR approaches. These results demonstrate that our algorithm can be successfully applied to automate side-chain resonance assignment and high-quality protein structure determination. Since our algorithm does not require any specific NMR experiments for measuring the through-bond interactions with side-chain protons, it can save a significant amount of both experimental cost and spectrometer time, and hence accelerate the NMR structure determination process.

Abbreviations used: NMR, nuclear magnetic resonance; ppm, parts per million; RMSD, root mean square deviation; HSQC, heteronuclear single quantum coherence spectroscopy; NOE, nuclear Overhauser effect; NOESY, nuclear Overhauser and exchange spectroscopy; TOCSY, total correlation spectroscopy; TROSY, transverse relaxation-optimized spectroscopy; RDC, residual dipolar coupling; PDB, Protein Data Bank; BMRB, Biological Magnetic Resonance Bank; pol η UBZ, ubiquitin-binding zinc finger domain of the human Y-family DNA polymerase Eta; hSRI, human Set2-Rpb1 interacting domain; FF2, FF Domain 2 of human transcription elongation factor CA150; GB1, B1 domain of Protein G; CH, C^{α}-H^{α}; SSE, secondary structure element; C′, carbonyl carbon; MRF, Markov Random Field; DEE, dead-end elimination; GMEC, global minimum energy conformation.

[⋆] This work is supported by the following grants from National Institutes of Health: R01 GM-65982 to B.R.D. and R01 GM-079376 to P.Z.

[⋆⋆] Corresponding author: Bruce Randall Donald, `brd+recomb10@cs.duke.edu`, Tel: 919-660-6583, Fax: 919-660-6519.

1 Introduction

The knowledge of the 3D structures of proteins plays an important role in understanding protein functions and discovering new drugs. Although high-throughput DNA sequencing technologies have been able to identify nearly the complete sequence of the human genome, studies of the 3D structures of proteins on a genome-wide scale (i.e., structural proteomics) are still limited by current slow speed of protein structure determination. X-ray crystallography and nuclear magnetic resonance (NMR) are two primary experimental methods for high-resolution protein structure determination. Unfortunately, structure determination by either method is laborious and time-consuming. In X-ray crystallography, growing a good quality crystal is in general a difficult task. NMR structure determination does not require crystals, thus it can be used to determine protein structures in the physiologically-relevant solution state, and has become a premier tool for studying protein dynamics. However, current NMR structure determination is still limited by the lengthy time required to process and analyze the experimental data. The development of automated and efficient procedures for analyzing NMR data and acquiring experimental restraints will thereby speed up protein structure determination and advance structural proteomics research. In practice, *side-chain resonance assignments* (the focus of this paper) are required for both side-chain dynamics studies and high-resolution structure determination.

In NMR terminology, each atom in the known primary sequence of a target protein is represented by a unique *chemical shift* (or *resonance*) in NMR spectra, that is, chemical shift serves as a scalar "ID" for an atom in the primary sequence. The magnetic interactions captured by an NMR spectrum can be described as a graph, in which each node represents the resonance of an atom in the primary sequence, and each edge represents a possible atomic interaction either through bond or through space. We call such a graph the *NMR interaction graph* [2]. For example, in an NMR interaction graph derived from a *heteronuclear single quantum coherence spectroscopy* (HSQC) spectrum, each edge represents an amide bond (i.e., $\mathrm{H^N}-\mathrm{N}$) interaction, while in an NMR interaction graph derived from a *nuclear Overhauser and exchange spectroscopy* (NOESY) spectrum, each edge represents a through-space interaction between a pair of protons closer than 6 Å, measured via the *nuclear Overhauser effect* (NOE).

In general, NMR structure determination is accomplished through the following procedure. The first step is to identify the correspondence between chemical shifts (i.e., nodes in the NMR interaction graph) and atoms in the primary sequence. Such a process is called *resonance assignment*, which is a crucial step in NMR data analysis and structure calculation. The resonance assignment can be classified into two categories: *backbone resonance assignment* and *side-chain resonance assignment*, which refers to resonance assignment for backbone or side-chain atoms. A typical approach for backbone resonance assignment is to exploit the connectivity information in an NMR spectrum that measures the bond interactions between backbone atoms in the main-chain of the primary sequence. For instance, in [1] a globally-consistent Hamiltonian path from an NMR interaction graph is found to align to the primary sequence and obtain backbone resonance assignments. On the other hand, side-chain resonances are normally assigned by exploiting the chemical shift pattern and the through-bond connectivity information in side-chains from an HCCH *total correlation spectroscopy* (HCCH-TOCSY) spectrum, which links up the side-chain resonances with the pre-determined backbone resonances using sequential connectivities. The Biological Magnetic Resonance Bank (BMRB) [59] has collected statistics on observed chemical shifts of all amino acids from a large database of solved protein structures. We call this information the *BMRB statistical information*. This information is often used to assist both backbone and side-chain resonance assignment. Once the correspondence between chemical shifts and atoms in the primary sequence has been identified after resonance assignment, each NOESY cross peak can be assigned to a pair of protons that are potentially correlated via a through-space NOE interaction. This process is called *NOE assignment*. In practice, neither resonance assignment nor NOE assignment is an easy task, since NMR spectra are often complicated by spectral artifacts, missing peaks, experimental noise and peak overlap. The completion of the NOE assignment process immediately provides a set of NOE distance restraints between spatially-neighboring protons, and enables structure calculation software, such as XPLOR-NIH [55] and CYANA [23], to compute the 3D structure of the protein. Besides NOE distance restraints, other NMR geometric constraints can be also used in structure determination. For example, residual dipolar couplings (RDCs) provide global orientational restraints on the internuclear bond vectors [58, 57], and can be also used in structure determination [58, 17, 53, 51, 13, 61, 62, 66].

Although substantial progress has been made in automated backbone resonance assignment [68, 2, 10, 37, 14, 64, 1, 31, 60], general approaches for automated side-chain resonance assignment are still not well developed [43, 48, 5]. Generally the side-chain resonance assignment problem is much more challenging than the backbone resonance assignment problem [48, 5, 47]. Traditional approaches for side-chain resonance assignment [40, 41, 46, 47] usually require a combination of several insensitive side-chain NMR experiments, including HCCH-TOCSY experiments, to obtain enough side-chain resonance assignments. Unfortunately, the performance of HCCH-TOCSY experiments is limited on large proteins due to the fast transverse relaxation of protonated carbons, which causes severe signal loss in NMR spectra. In addition, most large proteins must be deuterated (i.e., most aliphatic protons are replaced with deuterium isotope, and NMR signals from these atoms are muted), to reduce peak overlap and congestion in NMR spectra. The deuteration is also required to increase the efficiency of the *transverse relaxation-optimized spectroscopy* (TROSY) experiments that are generally used to enhance the sensitivity of NMR spectra. The deuteration for large proteins also drastically reduces the number of the NMR-active protons attached to side-chain carbons, which further limits the utility of TOCSY experiments, and thus makes it difficult to attain complete side-chain resonance assignments. On the other hand, it is essential to obtain almost all of the side-chain resonance assignments as a prerequisite for high-resolution structure determination, since they enable the NOE assignment, which constrains side-chain conformations geometrically, thereby enabling high-resolution structure determination. Although new techniques based on high-dimensional NMR experiments have been proposed to overcome the peak overlap issue in side-chain resonance assignment [25, 16], they still incur a penalty in absolute sensitivity. In general, it takes weeks or even months for traditional NMR approaches to collect all these required experimental data, and obtain a nearly complete set of side-chain resonance assignments.

In this paper, we describe a novel algorithm that assigns side-chain resonances from NOESY, backbone chemical shift and RDC data rather than from TOCSY spectra. We cast the side-chain resonance assignment problem into a Markov Random Field (MRF) framework, and apply combinatorial protein design algorithms to compute the optimal solution. Our MRF captures the contact map information in the backbone conformations determined from RDCs using our recently-developed techniques [61, 62, 13, 66], and a set of discretized side-chain conformations (i.e., rotamers) obtained from a high-resolution structure database. A Hausdorff-based computation is incorporated in the scoring function to compute the probability of side-chain resonance assignments to generate the observed NOESY spectra. The optimal side-chain resonance assignments are computed using the following protein design algorithms [12, 45, 19, 18, 9]. First, a *dead-end elimination* (DEE) algorithm [12, 45, 19] is applied to prune side-chain resonance assignments that are *provably* not part of the optimal solution. Second, an A* search algorithm is employed to find a set of optimal side-chain resonance assignments that best interpret the NMR data. Note that MRF and other graphical models have been used in structural and computational biology. Often they are used with techniques such as belief propagation, which can only be proven to compute a local optimum for a general graph. In contrast, we use DEE and A* algorithms to provably compute the global optimal solution to the MRF.

In [66], we proposed a high-resolution structure determination approach using an RDC-defined backbone conformation and a pattern-matching technique. Unlike the algorithm in [66] and other previous structure calculation approaches [22, 24, 44, 26, 34], all of which require a nearly complete set of both side-chain and backbone resonance assignments, in this paper the high-resolution structure determination strategy encoded by our algorithm only needs backbone resonance assignments, and does not require any TOCSY-like experiments. Such an advantage can help structural biologists reduce both experimental cost and NMR instrument time, and hence speed up the NMR structure determination process. The following contributions are made in this paper:

1. Introduction of a novel side-chain resonance algorithm that only requires NOESY spectra, backbone chemical shifts, and RDCs, and does not require any TOCSY-like experiments;
2. Development of an MRF framework for side-chain resonance assignment, which captures the contact map information of the protein derived from NOESY spectra, and exploits the structural information inferred from orientational restraints and side-chain rotamers;
3. Introduction of a Hausdorff-based measure to compute a probability distribution of side-chain resonance assignments in the MRF framework;
4. Application of protein design algorithms, including the DEE and A* search algorithms, to solve the side-chain resonance assignment problem; and
5. Testing and excellent results on real NMR spectra for five proteins recorded at Duke University.

2 Methods

2.1 Backbone Structure Determination from Residual Dipolar Couplings

We apply our recently-developed algorithms [61, 62, 66, 13] to compute the protein backbone structures using two RDCs per residue (either NH RDCs measured in two media, or NH and CH RDCs measured in a single medium). Details on backbone structure determination from RDCs are available in Supplementary Material [67] **Section 1** and [61, 62, 13]. Alternatively, the global fold (i.e., backbone) could in principle be computed by other approaches, such as protein structure prediction [3], protein threading [65] or homology modeling [35, 36].

2.2 Markov Random Field for Side-Chain Resonance Assignment

We introduce notation to describe our side-chain resonance assignment problem. Let $U = \{r_1, \cdots, r_n\}$ be the set of all resonances, including both backbone and side-chain resonances. Here backbone resonances are assigned and taken as input to our algorithm. Side-chain resonances are, of course, unassigned. Let t be the number of unassigned side-chain resonances, so the number of assigned backbone resonances is $n - t$. Without loss of generality, let $V = \{r_1, \cdot\cdot\cdot, r_t\}$ be the set of unassigned side-chain resonances, and let $U - V = \{r_{t+1}, \cdot\cdot\cdot, r_n\}$ be the set of assigned backbone resonances.

A graph $G = (U, E)$, called the *NOESY graph* [2, 1], represents the *contact map* information of resonances from NOESY spectra. In a NOESY graph $G = (U, E)$, U is the set of proton resonances (including both assigned backbone and unassigned side-chain proton resonances). Two resonances in U are connected by an edge in E, when a NOESY cross peak is observed at the coordinates (within a parameterized error window) of these two resonances (Fig. 1A). Nodes in U are called the *resonance nodes* (or *resonances*). Given a resonance node u in a NOESY graph $G = (U, E)$, $N(u) = \{v \mid (u, v) \in E$ and $u, v \in U,\ u \neq v\}$ is called the *neighborhood* of u. A *proton label* is defined as a 3-tuple that consists of the *proton name* (e.g., Arg16-H$_{\gamma_2}$), the *rotamer index* (e.g., the 3rd rotamer in the rotamer library) and the *proton coordinates* in $\mathbb{R}^3$. The set of all proton labels is called the *label set* L of the NOESY graph. We obtain a discrete and finite label set by considering all possible side-chain rotamer conformations on the RDC-defined backbone (Fig. 1B). Since the backbone has been solved and each side-chain rotamer conformation is rigid, each proton label corresponds to a proton on a particular rotamer after being placed on the backbone (with fixed positions in $\mathbb{R}^3$ with respect to backbone conformation). In our assignment problem, we aim to find a *map* $\pi : V \to L$, such that the contact map information through the mapped resonance nodes in a NOESY graph optimally interprets NOESY spectra. Given a resonance node $r_i \in V$ and a map π, we call $\pi(r_i) \in L$ a *proton label assignment* (or *assignment*) of r_i. Given a sequence of resonances $W = (r_1, \cdots, r_m)$, we call the sequence $(\pi(r_1), \cdots, \pi(r_m))$ an *assignment* of W, where $\pi(r_i)$ is the assignment of resonance node r_i.

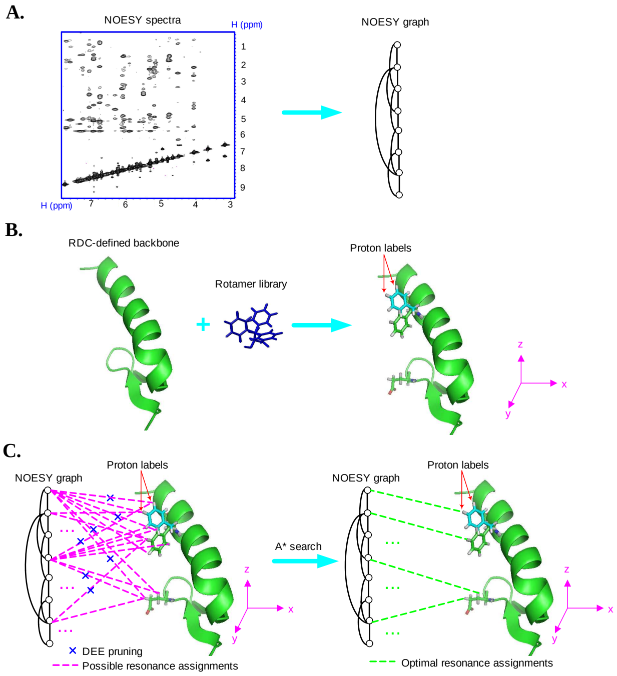

Fig. 1. Schematic illustration of our side-chain resonance assignment algorithm. (A): Construction of the NOESY graph. (B): Construction of proton labels. (C): The side-chain resonance assignment process.

Unlike previous side-chain resonance assignment algorithms [40, 41, 46, 47, 15], which only assign proton names to resonances, our algorithm computes not only the resonance assignments but also the rotamer assignments, since each proton label contains both the proton name and the rotamer index of this proton. The rotamer assignments included in the proton label assignments yield an ensemble of side-chain rotamer conformations for each residue, which are unified by the logical "OR" operation. In our algorithm, proton labels are treated as a cloud of unconnected points in $\mathbb{R}^3$. This formulation is similar to [20, 21] which uses a spatial proton distribution to represent a gas of unbound and unassigned hydrogen atoms. Unlike in [20, 21], which depends on molecular dynamics to embed the structure from the unassigned proton density, here we exploit the RDC-defined backbone conformations and apply an MRF to compute the correspondence between side-chain resonances and protons. Although the absence of the covalent structure in proton labels may allow resonances to map to the protons on the same side-chain in different rotameric states, we take into account the distance information of the covalent structure when computing the probability of side-chain resonance assignments (Sec. 2.3). In practice, as we will show in Sec. 3, our MRF can compute a high percentage of correct side-chain resonance assignments for accurate structure determination.

Given a NOESY graph, the assignment of each unassigned resonance r_i only depends on the resonance assignments of its neighborhood $N(r_i)$ in G. We can use a Markov Random Field (MRF) model [33] to encode this assignment problem. The assignment of a resonance node r_i satisfies the following property:

$$\Pr\big(\pi(r_i) \mid \pi(r_j), i \neq j\big) = \Pr\big(\pi(r_i) \mid \pi(r_j),\ r_j \in N(r_i)\big), \tag{1}$$

where $\Pr(\cdot)$ is the probability of an event, and $N(r_i)$ is the set of resonance nodes adjacent to r_i in the graph.

According to the Hammersley-Clifford theorem [6], the distribution of an MRF can be written in a closed form. Let C be a clique in the underlying graph G, and let $T_C(\cdot)$ be a *clique potential* [6] that represents the probability of a particular assignment of all resonance nodes in clique C. Let $V' = (r_1, \cdots, r_t)$ be an ordered sequence of resonances from set $V = \{r_1, \cdots, r_t\}$. Let $F = (\pi(r_1), \cdots, \pi(r_t))$ be an assignment for the sequence of resonances V'. By the Hammersley-Clifford theorem, the probability of an assignment F is defined by $\Pr(F) \propto \exp(-\sum_C T_C(F))$. We consider the potential function T_C for cliques of size 2, that is, the clique potential involves pairs of neighboring resonance nodes in G. Note that MRFs with cliques of size of 2 have been widely applied in several areas such as computer vision [8] and computational biology [32, 63]. In our MRF, $\Pr(F)$ measures the distribution of side-chain resonance assignments by capturing the pairwise resonance interactions in NOESY spectra and exploiting the structural information available from the RDC-defined backbone conformations and the discretized side-chain rotamer conformations.

Given two proton labels with the distance between their coordinates less than 6 Å, we expect to observe an NOE peak in NMR spectra. Such an expected peak is called a *back-computed NOE peak*. In contrast, an NOE peak that has been observed in experimental (NOESY) spectra is called the *experimental NOE peak*. A *back-computed NOE pattern* is defined as a set of back-computed NOE peaks. Since each proton label consists of the proton name, the rotamer index and the discrete coordinates of the rotamer's side-chain proton, the assignments of a resonance r_i and its neighborhood $N(r_i)$ determine a back-computed NOE pattern. A back-computed NOE pattern is constructed as follows. Let $d(\pi(r_i), \pi(r_j))$ be the Euclidean distance between two proton labels $\pi(r_i)$ and $\pi(r_j)$. Let $I_{ij} = c \cdot (d(\pi(r_i), \pi(r_j)))^{-6}$ be the back-computed peak intensity using distance $d(\pi(r_i), \pi(r_j))$, where c is the calibration constant that can be computed using the same strategy as in [49, 34]. Let $\lambda(r_i)$ be the resonance of the heavy atom that is covalently bound to the proton corresponding to resonance r_i. Given a pair of assignments $\pi(r_i)$ and $\pi(r_j)$, we call $b_{ij}(\pi(r_i), \pi(r_j)) = (r_i, \lambda(r_i), r_j, I_{ij})$ the *back-computed NOE peak* of $\pi(r_i)$ and $\pi(r_j)$. The definitions of back-computed NOE peaks here and experimental NOE peaks in Sec. 2.3 are presented for 3D NOESY spectra. They can be easily extended to other dimensional cases (e.g., 4D). When $d(\pi(r_i), \pi(r_j))$ is larger than the NOE cutoff 6 Å or two proton labels represent the same proton name, the back-computed NOE peak is a null point. Given a set of resonances $W \subset U$ and the assignment π, let $B_\pi(W) = \{b_{ij}(\pi(r_i), \pi(r_j)) | r_i, r_j \in W, r_i \neq r_j\}$ be the *back-computed NOE pattern* of W.

In our MRF formulation, the clique potential for node r_i and its neighborhood $N(r_i)$ can be measured by the matching score of their back-computed NOE pattern. Specifically, let $V_i = \{r_i\} \cup N(r_i)$, and let $B_\pi(V_i)$ be the back-computed NOE pattern of V_i under the assignment π. Without ambiguity, we will use B_i to represent $B_\pi(V_i)$. Let $s(B_i)$ be the matching score of the back-computed NOE pattern B_i, where the function $s(\cdot)$ will be defined in Sec. 2.3. We use $T_\pi(r_i, N(r_i)) = -s(B_i)$ to represent the clique potential of the pairwise interactions between r_i and its neighborhood $N(r_i)$. Thus, we have the following function for the probability of an MRF $F = (\pi(r_1), \cdots, \pi(r_t))$:

$$\Pr(F) \propto \exp\left(-\sum_{r_i \in V} T_\pi(r_i, N(r_i))\right) = \exp\left(\sum_{r_i \in V} s(B_i)\right). \tag{2}$$

We use Q to represent the BMRB statistical information (see Sec. 1). To estimate the probability of an MRF F based on the BMRB statistical information Q, we first relate them using the probability function $\Pr(Q|F)$. Recall that $\lambda(r_i)$ represents the frequency of the heavy atom covalently bound to the proton corresponding to r_i. The probability function $\Pr(Q|F)$ is defined by

$$\Pr(Q|F) = \prod_{r_i \in V} P(|r_i - \mu_i|, \sigma_i) \cdot P(|\lambda(r_i) - \mu'_i|, \sigma'_i), \tag{3}$$

where $P(|x - \mu|, \sigma)$ is the probability of observing the difference $|x - \mu|$ in a normal distribution with mean μ and standard deviation σ, and μ_i, σ_i, μ'_i, σ'_i are average values and standard deviations of chemical shifts derived from BMRB given the assignment $\pi(r_i)$. We note that the normal distribution and other similar distribution families have been widely used to model the noise in the NMR data, e.g., see [52] and [38].

By Bayes' Rule, $\Pr(F|Q)$, the probability of the assignment F conditioned on the BMRB statistical information Q (namely the *posterior probability*), can be computed as follows:

$$\begin{aligned}\Pr(F|Q) &\propto \Pr(F)\cdot\Pr(Q|F) && (4)\\ &\propto \exp\Big(-\sum_{r_i\in V} T\big(\pi(r_i),\pi(N(r_i))\big)\Big)\cdot \prod_{r_i\in V} P(|r_i-\mu_i|,\sigma_i)\cdot P(|\lambda(r_i)-\mu'_i|,\sigma'_i) && (5)\\ &= \exp\Big(\sum_{r_i\in V} s(B_i)\Big)\cdot \prod_{r_i\in V} P(|r_i-\mu_i|,\sigma_i)\cdot P(|\lambda(r_i)-\mu'_i|,\sigma'_i). && (6)\end{aligned}$$

Our goal is to compute an assignment $F^* = (\pi^*(r_1),\cdot\cdot\cdot,\pi^*(r_t))$ that maximizes the posterior probability $\Pr(F|Q)$. Taking the negative logarithm on both sides of Eq. (6), we have the following *pseudo-energy* function for an assignment $F = (\pi(r_1),\cdot\cdot\cdot,\pi(r_t))$:

$$E_F = -\sum_{r_i\in V} \ln P(|r_i-\mu_i|,\sigma_i)\cdot P(|\lambda(r_i)-\mu'_i|,\sigma'_i) - \sum_{r_i\in V} s(B_i). \qquad (7)$$

The pseudo-energy function in Eq. (7) measures how well an assignment $F = (\pi(r_1),\cdot\cdot\cdot,\pi(r_t))$ satisfies both the BMRB statistical information and the experimental NMR data. Maximizing the posterior probability $\Pr(F|Q)$ in Eq. (6) is equivalent to minimizing the pseudo-energy function in Eq. (7). We call the assignment $F^* = (\pi^*(r_1),\cdot\cdot\cdot,\pi^*(r_t))$, that minimizes the scoring function E_F and thus best interprets the NMR data restraints, the *optimal assignment* or *optimal solution* to our MRF. Since our proton label assignments contain both resonance assignments and molecular side-chain coordinates, the optimal assignment is analogous to the *global minimum energy conformation* (GMEC) in the protein design literature.

2.3 The Matching Score of a Back-Computed NOE Pattern

The *matching score* of a back-computed NOE pattern can be measured by comparing the back-computed peaks with NOESY spectra. Given a set of resonance nodes $W \subset U$ and an assignment π, let $B_\pi(W)$ denote their back-computed NOE pattern. Without ambiguity, we will use B to stand for $B_\pi(W)$. Let Y be the set of experimental peaks. The matching score between the back-computed NOE pattern B and experimental spectrum Y can be measured by the conventional Hausdorff distance $H(B,Y) = \max(h(B,Y),h(Y,B))$, where $h(B,Y) = \max_{b\in B}\min_{y\in Y}\|b-y\|$ and $\|\cdot\|$ is the normed distance. This conventional Hausdorff distance is sensitive to a single outlying point of B or Y [28, 29]. For example, suppose that an NOE peak is missing in Y (which is quite common in NMR data), and its corresponding back-computed peak in B has a large distance from any peak in Y. In such a case, the Hausdorff distance between B or Y is dominated by this missing NOE peak. To take into account the missing NOE peaks, we employ a generalized Hausdorff distance measure, called the *Hausdorff fraction (fractional Hausdorff distance)*, which is derived from the k^{th} *Hausdorff distance* h_k from B to Y [29, 27]:

$$h_k(B,Y) = \underset{b\in B}{k\mathrm{th}}\min_{y\in Y}\|b-y\|,$$

where k^{th} is the k^{th} largest value. Now, let δ be the error window in chemical shift. Then the probability of the back-computed NOE pattern B under $h_k(B,Y)\le\delta$, is computed by the following *Hausdorff fraction* equation [27]:

$$s(B) = \frac{\tau(B\cap Y_\delta)}{\tau(B)}, \qquad (8)$$

where Y_δ denotes the union of all balls obtained by replacing each point in Y with a ball of radius δ, and $\tau(\cdot)$ denotes the size of a set.

Next, we will show how to compute the matching score of a back-computed NOE pattern in Eq. (8). Let $b_{ij}(\pi(r_i),\pi(r_j)) = (r_i,\lambda(r_i),r_j,I_{ij})$ be a back-computed NOE peak in B based on assignments

$\pi(r_i)$ and $\pi(r_j)$, where $\lambda(r_i)$ is the frequency of the heavy atom covalently bound to the proton corresponding to r_i, and I_{ij} is the back-computed peak intensity. Without ambiguity, we will use b_{ij} to represent $b_{ij}(\pi(r_i),\pi(r_j))$. Note that the distance information of the covalent structure is also included when computing a back-computed NOE pattern, since the distances between protons within a residue or in consecutive residues are generally < 6 Å. Let (x, y, z, I') be the experimental NOESY cross peak that is closest to the back-computed NOE peak b_{ij} under the Euclidean distance measure, where x and z are frequencies of NOE interacting protons, y is the frequency of the heavy atom covalently bound to the first proton, and I' is the peak intensity. When computing the geometric count $\tau(B \cap Y_\delta)$, we must take into account the uncertainty in chemical shift. For example, suppose that the back-computed NOE peak b_{ij} is within the Euclidean distance δ from an experimental NOESY cross peak. When b_{ij} is closer to this experimental peak, it should contribute more to counting $\tau(B \cap Y_\delta)$. To measure the probability of a back-computed NOE peak to intersect with Y_δ, we model the uncertainty of chemical shifts in individual dimensions as independent normal distributions. Formally, the following equation is employed to compute $\tau(B \cap Y_\delta)$:

$$\tau(B \cap Y_\delta) = \sum_{b_{ij}\in B} P(|I' - I_{ij}|, \sigma_{I\delta}) \cdot P(|x - r_i|, \sigma_{x\delta}) \cdot P(|y - \lambda(r_i)|, \sigma_{y\delta}) \cdot P(|z - r_j|, \sigma_{z\delta}), \tag{9}$$

where $P(|x - \mu|, \sigma)$ is the probability of observing the difference $|x - \mu|$ in a normal distribution with mean μ and standard deviation σ. We define the standard deviations in Eq. (9) as a function of the error window δ. We choose $\sigma = \delta/3$ for each dimension such that the probability for a back-computed NOE peak outside Y_δ to contribute to $\tau(B \cap Y_\delta)$ is almost 0.

2.4 A DEE Pruning Algorithm

The chemical shift of each proton in a particular residue usually lies within an interval derived from the BMRB statistical information [59]. Therefore, each resonance node r_i in the NOESY graph is only allowed to map to a subset of proton labels, in which the BMRB-derived chemical shift intervals contain the frequency of r_i. Given a resonance r_i, we call the subset of proton labels in L, that r_i is allowed to map to, the *candidate mapping set* of r_i, denoted by $A(r_i)$. When we know the backbone resonance assignments, we have $|A(r_i)| = 1$ for all backbone resonance nodes r_i. Given a sequence of resonances $W = (r_1, \cdots, r_m)$, we call $A(W) = (A(r_1), \cdots, A(r_m))$ the *candidate mapping set* of W. Let $D = (\pi(r_1), \cdots, \pi(r_m))$, where $\pi(r_i) \in A(r_i)$ is the assignment of r_i. We write $D \dot{\in} A(W)$ when $\pi(r_i) \in A(r_i)$ for every $i = 1, \cdots, m$, i.e., the assignment of r_i lies in the candidate mapping sets for all resonances.

We use $\gamma(r_i, u)$ to mean that proton label $u \in L$ is assigned to resonance node r_i, where $u \in A(r_i)$. Initially, we prune any proton label assignment $\gamma(r_i, u)$ in which the frequency of r_i falls outside the BMRB-derived chemical shift interval. Let $N(r_i) = \{r'_{i1}, \cdots, r'_{im}\}$ be the set of resonance nodes in the neighborhood of r_i, and let $N'(r_i) = (r'_{i1}, \cdots, r'_{im})$ be a sequence of resonance nodes in $N(r_i)$, where m is total number of resonance nodes in the neighborhood. Then the candidate mapping set of $N'(r_i) = (r'_{i1}, \cdots, r'_{im})$ is $A(N'(r_i)) = (A(r'_{i1}), \cdots, A(r'_{im}))$. Let $D_i = (\pi(r'_{i1}), \cdots, \pi(r'_{im})) \dot{\in} A(N'(r_i))$ be an assignment of $N'(r_i)$, where $\pi(r'_{ij}) \in A(r'_{ij})$, and we use $\gamma(N'(r_i), D_i)$ to mean that D_i is assigned to $N'(r_i)$.

Given an assignment $F = (\pi(r_1), \cdots, \pi(r_t))$ for the sequence of resonances $V' = (r_1, \cdots, r_t)$, we use $E(\gamma(r_i, \pi(r_i))$ to represent the first energy term in Eq. (7) under the assignment π. We use $E(\gamma(r_i, \pi(r_i)), \gamma(N'(r_i), D_i))$ to represent the second energy term in Eq. (7) when assigning $\pi(r_i)$ to resonance node r_i and D_i to $N'(r_i)$, where $\pi(r_i) \in A(r_i)$ and $D_i \dot{\in} A(N'(r_i))$. Then the pseudo-energy scoring function in Eq. (7) for an assignment $F = (\pi(r_1), \cdots, \pi(r_t))$ can be rewritten as

$$E_F = \sum_{r_i \in V} E\big(\gamma(r_i, \pi(r_i))\big) + \sum_{r_i \in V} E\big(\gamma(r_i, \pi(r_i)), \gamma(N'(r_i), D_i)\big), \tag{10}$$

where $\pi(r_i) \in A(r_i)$ and $D_i \dot{\in} A(N'(r_i))$.

An algorithm that is similar to the GMEC calculation method in protein design [12, 45, 19, 18, 9] can be applied here to compute the optimal proton label assignments. The *dead-end elimination* (DEE) algorithm has been effectively applied to prune rotamers when their contribution to the total energy is always less than another (competing) rotamer [12, 45, 19, 18, 9]. We use a similar idea here to prune proton label assignments that are *provably* not part of the optimal solution. Given an unassigned side-chain resonance node $r_i \in V$, a proton label assignment $v \in A(r_i)$ is eliminated if an alternative proton label assignment $u \in A(r_i)$ satisfies the following Goldstein criterion [19]:

$$E\big(\gamma(r_i, v)\big) - E\big(\gamma(r_i, u)\big) + \min_{D_i \dot{\in} A(N'(r_i))} \Big(E\big(\gamma(r_i, v), \gamma(N'(r_i), D_i)\big) - E\big(\gamma(r_i, u), \gamma(N'(r_i), D_i)\big) \Big) > 0. \tag{11}$$

Any assignment $\gamma(r_i, v)$ satisfying Eq. (11) is *provably* not part of the optimal solution, and thus can be safely pruned. The complexity of computing the Goldstein criterion in Eq. (11) is $O(na^2w)$, where n is the total number of resonances, a is the maximum number of proton labels in the candidate mapping set of a resonance, and w is the maximum number of proton labels that can be assigned to a resonance node's neighborhood. DEE reduces the conformation search space by pruning proton label assignments that can not be in the optimal solution, and provides a combinatorial factor reduction in computational complexity.

2.5 Computing Optimal Side-Chain Resonance Assignments

To compute the optimal solution to our MRF, we apply an A* algorithm [39, 54, 56] to search over all possible combinations of the remaining proton label assignments surviving from DEE. An A* algorithm provably finds the optimal (i.e., least-cost) path from a given starting node to the goal node in a search tree or graph. It uses a heuristic cost function to determine the order of visiting nodes during the search. The heuristic cost function consists of two parts: the *actual* cost from the starting node to the current node, and the *estimated* cost from the current node to the goal node. Next, we will define both actual and estimated cost functions that are used to determine the order of searching nodes in our A* algorithm.

Recall that $V' = (r_1, \cdots, r_t)$ denotes the sequence of unassigned side-chain resonances, and $(r_{t+1}, \cdots, r_n)$ denotes the sequence of assigned backbone resonances. Let X_i be the variable representing the assignment of resonance node r_i. Similar to the protein design problem [39, 18], our search configuration space can be also formulated as a tree, in which the root represents an empty assignment, a leaf node represents a full assignment of V', and an internal node represents a partial assignment of V' (i.e., only a subsequence of resonances in V' are assigned). Let $H = (X_{t+1}, \cdots, X_n)$ be the sequence of known assignments for backbone resonances $(r_{t+1}, \cdots, r_n)$. Let $S = (X_1, \cdots, X_t)$ be a sequence of assignments for side-chain resonances in V'. Given the BMRB statistical information Q and the known backbone chemical shifts H, the probability for a sequence of side-chain resonance assignments S is

$$\Pr(S|H,Q) = \Pr(X_t, X_{t-1}, \cdots, X_1|H,Q) = \Pr(X_t|X_{t-1}, \cdots, X_1, H, Q) \cdots \Pr(X_2|X_1, H, Q) \cdot \Pr(X_1|H,Q). \tag{12}$$

Suppose that the A* algorithm has assigned resonances $r_1, \cdots, r_{i-1}$. We rewrite Eq. (12) as

$$\begin{aligned} \Pr(S|H,Q) = {} & \Pr(X_t|X_{t-1}, \cdots, X_1, H, Q) \cdots \Pr(X_{i+1}|X_i, \cdots, X_1, H, Q) \\ & \cdot \Pr(X_i|X_{i-1}, \cdots, X_1, H, Q) \cdots \Pr(X_1|H,Q). \end{aligned} \tag{13}$$

Taking the negative logarithm on both sides of Eq. (13), we have

$$\begin{aligned} -\ln \Pr(S|H,Q) = {} & -\ln\big(\Pr(X_t|X_{t-1}, \cdots, X_1, H, Q) \cdots \Pr(X_{i+1}|X_i, \cdots, X_1, H, Q)\big) \\ & -\ln\big(\Pr(X_i|X_{i-1}, \cdots, X_1, H, Q) \cdots \Pr(X_1|H,Q)\big). \end{aligned} \tag{14}$$

Eq. (14) measures the *cost* of a path from the root (i.e., empty assignment) to one of leaf nodes (i.e., full assignments) in our A* search tree.

Let
$$g = -\ln\big(\Pr(X_i|X_{i-1}, \cdots, X_1, H, Q) \cdots \Pr(X_1|H,Q)\big), \tag{15}$$

which measures the probability of the set of the first i assignments $X_1, \cdots, X_i$, and leads to the actual cost of the path from the root to the current node in the A* search tree.

Let
$$h = -\ln\big(\Pr(X_t|X_{t-1}, \cdots, X_1, H, Q) \cdots \Pr(X_{i+1}|X_i, \cdots, X_1, H, Q)\big), \tag{16}$$
which estimates the cost of assigning the remaining resonance nodes (i.e., the cost of the path from current node to the leaf nodes in the A* search tree).

Then the cost function in our A* search is defined by
$$f = g + h, \tag{17}$$
where g is the *actual cost* from the root to the current node in the search tree, and h is the *estimated cost* from the current node to one of leaf nodes, in which all side-chain resonances are assigned.

In Eq. (16), $\Pr(X_j|X_{j-1}, \cdots, X_i, \cdots, X_1, H, Q)$, $j > i$, is estimated as follows:
$$\Pr(X_j|X_{j-1}, \cdots, X_i, \cdots, X_1, H, Q) = \max_{\substack{u_j \in A(r_j) \\ \cdots \\ u_{i+1} \in A(r_{i+1})}} \Pr(\gamma(r_j, u_j)|\gamma(r_{j-1}, u_{j-1}), \cdots, \gamma(r_{i+1}, u_{i+1}), X_i, \cdots, X_1, H, Q), \tag{18}$$
where $\gamma(r_j, u_j)$ denotes the assignment of u_j to resonance node r_j.

The A* algorithm maintains a list of search nodes, which are ranked according to the cost function (Eq. (17)). Similar to the protein design work in [18], here the A* search algorithm expands the nodes in order of the cost function f. In each iteration, the node with the smallest f value is visited, and the values of f in the remaining nodes are updated. All remaining nodes in the list are re-ordered according to the new f values, and form the children of current visited node. Such a process is repeated until all side-chain resonances are assigned (i.e., when a leaf node in the search tree is reached).

An estimated cost function is *admissible*, if it does not overestimate the cost from any node to the goal node. The admissibility of the estimated cost function ensures that an A* search algorithm will find the optimal solution. The following claim provides the soundness of our A* algorithm in computing the optimal assignment. The proof of this claim is provided in Supplementary Material **Section 2** available online in Ref. [67].

***Claim* 1** *The estimated cost function defined in Eq. (18) is* admissible*, which guarantees that our A* search algorithm will find the optimal solution.*

The A* algorithm is proven to be complete and optimal in searching for the least-cost path [39, 54, 56]. Although the time complexity of the A* algorithm is exponential in the number of side-chain resonances in the worst case, in practice, our algorithm, including both DEE and A* modules, runs only in hours for a medium-size protein. For instance, it takes about 7 hours to compute the set of side-chain resonance assignments on a single-processor machine for the human ubiquitin protein without human intervention.

3 Results

We have tested our algorithm on NMR data of five proteins: the FF Domain 2 of human transcription elongation factor CA150 (FF2), the B1 domain of Protein G (GB1), human ubiquitin, the ubiquitin-binding zinc finger domain of the human Y-family DNA polymerase Eta (pol η UBZ), and the human Set2-Rpb1 interacting domain (hSRI). The numbers of amino acid residues in these proteins are 62 for FF2, 39 for pol η UBZ, 56 for GB1, 76 for ubiquitin, and 112 for hSRI. Note that by the standards of the NMR community [22, 5, 24, 55, 23, 64, 52, 60], tests on real experimental data of five proteins are sufficient to demonstrate the feasibility of an algorithm in NMR data analysis and structure determination. All NMR data except RDCs of ubiquitin and GB1 were recorded and collected using Varian 600 and 800 MHz spectrometers at Duke University. The NOE cross peaks were picked from 3D ^{15}N- and ^{13}C-edited NOESY-HSQC spectra. The NH and CH RDC data of FF2, pol η UBZ and hSRI were measured from a 2D ^{1}H-^{15}N IPAP experiment [50] and a modified (HACACO)NH experimental [4] respectively. Details on the NMR experimental procedures and results on the backbone structure calculation from RDCs are provided in Supplementary Material **Section 3** available online in Ref. [67].

3.1 Accuracy of Side-Chain Resonance Assignments

We evaluated the accuracy of the side-chain resonances assigned by our algorithm by comparing them with the chemical shifts of the proteins that were assigned manually using other additional side-chain NMR experiments. Our algorithm achieved the completeness of over 90% for resonance assignments, that is, it assigned the resonances of over 90% of protons (Table 1). Note that the manual assignments are usually obtained from TOCSY experiments, while frequencies in our resonance list are extracted from NOESY spectra. Due to the experimental uncertainty, frequencies of our assigned resonances are not exactly equal to the manually-assigned chemical shifts. We used an error window 0.04 ppm for ^{1}H, and 0.4 ppm for heavy atoms (i.e., ^{13}C and ^{15}N) to check whether two resonance assignments agree with each other. We say a resonance assignment is *correct* if its frequency is within the error window from the reference assignment, which was assigned manually using other additional experiments. Our tests show that our algorithm computes about 80% of the correct resonance assignments (Table 1).

Proteins	**GB1**	**ubiquitin**	**hSRI**	**pol η UBZ**	**FF2**
Completeness (%)	97.7	94.9	90.2	97.6	92.7
Correctness (%)	81.9	81.8	83.6	92.2	78.0

Table 1. Summary of side-chain resonance assignment results.

In a hypothetical ideal case without any experimental error and noise, the goal of an NMR assignment problem is to find a one-to-one correspondence (i.e., bijection) between resonances and proton names in the protein sequence. In practice, a proton can be mapped to 2-3 different resonances due to the ambiguity arising from chemical shift degeneracy, that is, chemical shifts of two different protons may be so close that the probabilities measuring their assignments are not sufficient to distinguish them. In practice, the optimal solution to our MRF finds the one-to-one mapping for most resonance assignments (Table 1), because the local neighborhood structure of our MRF has enforced these correct assignments. Most of the *inconsistent* assignments (i.e., two resonances are assigned to the same proton label) occur in the methylene protons bound to the same carbon, or neighboring ring protons in aromatic residues. These protons often have both similar chemical shifts and close coordinates in $\mathbb{R}^3$, which makes it difficult to distinguish them using the probability functions derived from our MRF framework. We use the Boolean operation "XOR" to unify these inconsistent assignments. As we will show in Sec. 3.2, the NOE assignment ambiguity arising from these inconsistent resonance assignments does not degrade high-resolution structure determination, probably because these protons are adjacent in $\mathbb{R}^3$ (with distance $< 1.8 - 2.5$ Å).

3.2 Effectiveness for High-Resolution Structure Determination

To investigate the effect of assigned side-chain resonances on high-resolution structure determination, we first computed a set of NOE assignments using the side-chain resonance assignments computed by our algorithm. We then examined the quality of the structures calculated using these NOE distance restraints. Details on computing NOE distance restraints using assigned side-chain resonances are provided in Supplementary Material **Section 4** available online in Ref. [67].

To examine the accuracy of the NOE assignments computed by our algorithm, we compared them with the reference structures. We say an NOE assignment is *correct* if it agrees with the reference structure, that is, the distance between the assigned pair of NOE protons in the reference structure satisfies the NOE restraint whose distance is calibrated from the experimental peak intensity. As shown in Table 2, our algorithm computes over 80% correct NOE restraints. To further investigate these NOE distance restraints, we fed them into XPLOR-NIH [55] for the structure calculation. To fairly compare the accuracy of our NOE restraints, we fed the same hydrogen bond and dihedral angle constraints into XPLOR-NIH, as in computing the NMR reference structures. In addition, the structures were refined with RDC data using XPLOR-NIH with a water-refinement protocol [55]. We chose the ensemble of top 20 structures with the lowest energies out of 50 structures computed by XPLOR-NIH as the ensemble of final structures. For all five proteins, the ensemble of top 20 structures with the lowest energies converge into a compact cluster (Table 3 and Fig. 2). The average RMSD to the mean coordinates is $\leq$ 0.6 Å for backbone atoms and $\leq$ 1.0 Å for all-heavy atoms. We superimposed the mean structure of the ensemble with the reference structure for each protein. The RMSD between the mean structure and the reference structure (ordered region) is $0.5-1.4$ Å for backbone atoms and $1.0-2.2$ Å for all-heavy atoms (Table 3 and Fig. 2). These results indicate that the NOE assignments computed by our algorithm are sufficient for high-resolution structure determination.

Proteins	**GB1**	**ubiquitin**	**hSRI**	**pol η UBZ**	**FF2**
Total # of assigned NOEs	1421	1531	3540	960	1354
Intraresidue	597	648	1326	419	619
Sequential ($\lvert i-j \rvert = 1$)	295	321	777	254	282
Medium-range ($\lvert i-j \rvert \leq 4$)	185	202	984	177	281
Long-range ($\lvert i-j \rvert \geq 5$)	344	360	453	110	172
Percentage of correct NOE assignments (%)	87.0	81.7	83.3	89.4	85.5

Table 2. Summary of NOE assignment results.

Proteins	**GB1**	**ubiquitin**	**hSRI**	**pol η UBZ**	**FF2**
Average RMSD to mean coordinates					
SSE region (backbone, heavy) (Å)	0.18, 0.38	0.36, 0.71	0.29, 0.75	0.12, 0.43	0.25, 0.67
Ordered region (backbone, heavy) (Å)	0.20, 0.41	0.58, 0.95	0.35, 0.81	0.15, 0.67	0.34, 0.89
RMSD to reference structure					
SSE region (backbone, heavy) (Å)	0.56, 1.14	0.63, 1.40	1.25, 1.93	0.62, 1.39	0.58,1.53
Ordered region (backbone, heavy) (Å)	0.54, 1.08	0.93, 1.51	1.37, 2.09	0.97, 1.73	1.06, 2.17

Table 3. Summary of final calculated structures.

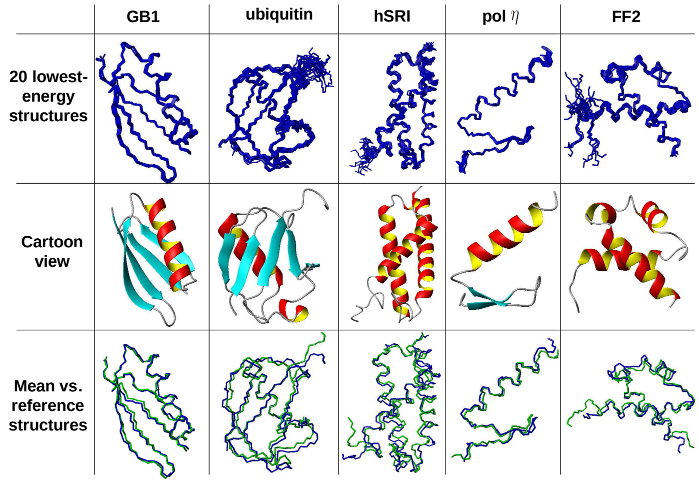

Fig. 2. Final NMR structures computed using our automatically-assigned NOEs. Row 1: the ensemble of 20 lowest-energy NMR structures. Row 2: ribbon view of one structure in the ensemble. Row 3: backbone overlay of the mean structures (blue) vs. corresponding NMR reference structures (green) (PDB ID of GB1 [30]: 3GB1; PDB ID of ubiquitin [11]: 1D3Z; PDB ID of FF2: 2E71; PDB ID of hSRI [42]: 2A7O; PDB ID of pol η UBZ [7]: 2I5O).

4 Conclusions

Side-chain resonance assignments are essential for high-resolution structure determination and side-chain dynamics studies. In this paper we proposed an MRF with protein design algorithms to compute the set of optimal side-chain resonance assignments that best interpret the NMR data. Tests on real NMR data demonstrated that our algorithm computes a high percentage of accurate side-chain resonance assignments for high-resolution structure determination. Since our algorithm does not require any TOCSY-like experiments, it can advance NMR structure determination by saving a significant amount of both experimental cost and NMR instrument time.

In [15], the authors proposed an algorithm that uses the knowledge of local covalent polypeptide structures to iteratively assign side-chain resonances from previously-assigned resonances (initially backbone resonances were assigned) using NOESY or TOCSY spectra. Compared to [15], in which only the conformation-independent bounds on intra-residue and sequential inter-proton distances are used to iteratively assign side-chain resonances, our algorithm applies an MRF that effectively exploits the RDC-defined backbone conformations to derive side-chain resonance assignments.

Although our algorithm is only implemented for 3D NOESY spectra, it is general and can be easily extended to higher-dimensional NOESY spectra. In addition, it would be interesting to extend our algorithm to perform side-chain resonance assignment without requiring backbone resonance assignments. Because RDCs are mapped to backbone resonances, in this case, we might have to resort to other approaches such as protein structure prediction, protein threading or homology modeling to obtain the initial global fold.

Availability

The source code of our algorithm is available by contacting the authors, and is distributed open-source under the GNU Lesser General Public License (Gnu, 2002).

Acknowledgements

We thank Dr. C. Bailey-Kellogg, Dr. M.S. Apaydin and Mr. J. Martin for reading our draft and providing us with valuable comments. We thank all members of the Donald and Zhou Labs for helpful discussions and comments. We are grateful to Ms. M. Bomar for helping us with pol η UBZ NMR data.

References

1. C. Bailey-Kellogg, S. Chainraj, and G. Pandurangan. A Random Graph Approach to NMR Sequential Assignment. *Journal of Computational Biology*, 12(6):569–583, 2005.
2. C. Bailey-Kellogg, A. Widge, J. J. Kelley, M. J. Berardi, J. H. Bushweller, and B. R. Donald. The NOESY jigsaw: automated protein secondary structure and main-chain assignment from sparse, unassigned NMR data. *Journal of Computational Biology*, 7(3-4):537–558, 2000.
3. D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294:93–96, 2001.
4. G. Ball, N. Meenan, K. Bromek, B. O. Smith, J. Bella, and D. Uhrín. Measurement of one-bond $^{13}C^{\alpha}$-$^{1}H^{\alpha}$ residual dipolar coupling constants in proteins by selective manipulation of $C^{\alpha}H^{\alpha}$ spins. *Journal of Magnetic Resonance*, 180:127–136, 2006.
5. M. C. Baran, Y. J. Huang, H. N. Moseley, and G. T. Montelione. Automated analysis of protein NMR assignments and structures. *Chem Rev.*, 104:3541–3456, 2004.
6. J. Besag. Spatial interaction and the statistical analysis of lattice systems. *J. Royal Stat. Soc. B*, 36, 1974.
7. M. G. Bomar, M. Pai, S. Tzeng, S. Li, and P. Zhou. Structure of the ubiquitin-binding zinc finger domain of human DNA Y-polymerase η. *EMBO reports*, 8:247–251, 2007.
8. Y. Boykov, O. Veksler, and R. Zabih. Markov random fields with efficient approximations. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, page 648, 1998.
9. C.Y. Chen, I. Georgiev, A.C. Anderson, and B.R. Donald. Computational structure-based redesign of enzyme activity. *Proc Natl Acad Sci U S A*, 106:3764–3769, 2009.
10. B. E. Coggins and P. Zhou. PACES: Protein sequential assignment by computer-assisted exhaustive search. *Journal of Biomolecular NMR*, 26:93–111, 2003.
11. G. Cornilescu, J. L. Marquardt, M. Ottiger, and A. Bax. Validation of Protein Structure from Anisotropic Carbonyl Chemical Shifts in a Dilute Liquid Crystalline Phase. *Journal of the American Chemical Society*, 120:6836–6837, 1998.
12. J. Desmet, M.D. Maeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356:539 – 542, 1992.
13. B. R. Donald and J. Martin. Automated NMR assignment and protein structure determination using sparse dipolar coupling constraints. *Progress in NMR Spectroscopy*, 55:101–127, 2009.

14. H.R. Eghbalnia, A. Bahrami, L.Y. Wang, A. Assadi, and J.L. Markley. Probabilistic identification of spin systems and their assignments including coil-helix inference as output (PISTACHIO). *J. Biomol. NMR*, 32:219–33, 2005.
15. F. Fiorito, T. Herrmann, F.F. Damberger, and K. Wüthrich. Automated amino acid side-chain NMR assignment of proteins using (13)C- and (15)N-resolved 3D [(1)H, (1)H]-NOESY. *J Biomol NMR*, 42:23–33, 2008.
16. F. Fiorito, S. Hiller, G. Wider, and K. Wüthrich. Automated resonance assignment of proteins: 6D APSY-NMR. *J Biomol NMR*, 35:27–37., 2006.
17. C. A. Fowler, F. Tian, H. M. Al-Hashimi, and J. H. Prestegard. Rapid determination of protein folds using residual dipolar couplings. *Journal of Molecular Biology*, 304:447–460, 2000.
18. I. Georgiev, R. H. Lilien, and B. R. Donald. The minimized dead-end elimination criterion and its application to protein redesign in a hybrid scoring and search algorithm for computing partition functions over molecular ensembles. *Journal of Computational Chemistry*, 29:1527–1542, 2008.
19. R. F. Goldstein. Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophysical Journal*, 66:1335–1340, 1994.
20. A. Grishaev and M. Llinás. CLOUDS, a protocol for deriving a molecular proton density via NMR. *Proc Natl Acad Sci U S A*, 99:6707–6712, 2002.
21. A. Grishaev and M. Llinás. Protein structure elucidation from NMR proton densities. *Proc Natl Acad Sci U S A*, 99:6713–6718, 2002.
22. P. Güntert. Automated NMR Protein Structure Determination. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 43:105–125, 2003.
23. P. Güntert. Automated NMR protein structure calculation with CYANA. *Meth. Mol. Biol.*, 278:353–378, 2004.
24. T. Herrmann, P. Güntert, and K. Wüthrich. Protein NMR Structure Determination with Automated NOE Assignment Using the New Software CANDID and the Torsion Angle Dynamics Algorithm DYANA. *Journal of Molecular Biology*, 319(1):209–227, 2002.
25. S. Hiller, R. Joss, and G. Wider. Automated NMR assignment of protein side chain resonances using automated projection spectroscopy (APSY). *J Am Chem Soc.*, 130(36):12073–12079, 2008.
26. Y. J. Huang, R. Tejero, R. Powers, and G. T. Montelione. A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins: Structure Function and Bioinformatics*, 62(3):587–603, 2006.
27. D. P. Huttenlocher and E. W. Jaquith. Computing visual correspondence: Incorporating the probability of a false match. In *Procedings of the Fifth International Conference on Computer Vision (ICCV 95)*, pages 515–522, 1995.
28. D. P. Huttenlocher and K. Kedem. *Distance Metrics for Comparing Shapes in the Plane*. In B. R. Donald and D. Kapur and J. Mundy, editors, Symbolic and Numerical Computation for Artificial Intelligence, pages 201-219, Academic press, 1992.
29. D. P. Huttenlocher, G. A. Klanderman, and W. Rucklidge. Comparing Images Using the Hausdorff Distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(9):850–863, 1993.
30. K. Juszewski, A. M. Gronenborn, and G. M. Clore. Improving the Packing and Accuracy of NMR Structures with a Pseudopotential for the Radius of Gyration. *Journal of the American Chemical Society*, 121:2337–2338, 1999.
31. H. Kamisetty, C. Bailey-Kellogg, and G. Pandurangan. An efficient randomized algorithm for contact-based NMR backbone resonance assignment. *Bioinformatics*, 22(2):172–180, 2006.
32. H. Kamisetty, E.P. Xing, and C.J. Langmead. Free Energy Estimates of All-atom Protein Structures Using Generalized Belief Propagation. *Journal of Computational Biology*, 15:755–766, 2008.
33. R. Kindermann and J.L. Snell. *Markov Random Fields and Their Applications*. American Mathematical Society, 1980.
34. J. Kuszewski, C. D. Schwieters, D. S. Garrett, R. A. Byrd, N. Tjandra, and G. M. Clore. Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear overhauser enhancement spectra and chemical shift assignments. *J. Am. Chem. Soc.*, 126(20):6258–6273, 2004.
35. C. J. Langmead and B. R. Donald. 3D structural homology detection via unassigned residual dipolar couplings. In *Procedings of 2003 IEEE Comput Syst Bioinform Conf*, pages 209–217, 2003.
36. C. J. Langmead and B. R. Donald. High-throughput 3D structural homology detection via NMR resonance assignment. In *Procedings of 2004 IEEE Comput Syst Bioinform Conf*, pages 278–289, 2004.
37. C. J. Langmead, A. K. Yan, R. H. Lilien, L. Wang, and B. R. Donald. A polynomial-time nuclear vector replacement algorithm for automated NMR resonance assignments. In *Proceedings of the seventh annual international conference on Research in computational molecular biology*, pages 176–187, 2003.
38. C.J. Langmead and B.R. Donald. An expectation/maximization nuclear vector replacement algorithm for automated NMR resonance assignments. *J. Biomol. NMR*, 29(2):111–138, 2004.
39. A.R. Leach and A.P. Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins*, 33(2):227–239, 1998.
40. K.B. Li and B.C. Sanctuary. Automated extracting of amino acid spin systems in proteins using 3D HCCH-COSY/TOCSY spectroscopy and constrained partitioning algorithm (CPA). *J. Chem. Inf. Comput. Sci.*, 36:585–593, 1996.
41. K.B. Li and B.C. Sanctuary. Automated resonance assignment of proteins using heteronuclear 3D NMR. 2. Side chain and sequence-specific assignment. *J. Chem. Inf. Comput. Sci.*, 37:467–477, 1997.

42. M. Li, H. P. Phatnani, Z. Guan, H. Sage, A. L. Greenleaf, and P. Zhou. Solution structure of the Set2-Rpb1 interacting domain of human Set2 and its interaction with the hyperphosphorylated C-terminal domain of Rpb1. *Proceedings of the National Academy of Sciences*, 102:17636–17641, 2005.
43. Y. Lin and G. Wagner. Efficient side-chain and backbone assignment in large proteins: Application to tGCN5. *J. Biomol. NMR*, 15:227–239, 1999.
44. J. P. Linge, M. Habeck, W. Rieping, and M. Nilges. ARIA: Automated NOE assignment and NMR structure calculation. *Bioinformatics*, 19(2):315–316, 2003.
45. L.L. Looger and H.W. Hellinga. Generalized dead-end elimination algorithms make large-scale protein side-chain structure prediction tractable: implications for protein design and structural genomics. *J Mol Biol.*, 3007(1):429–445, 2001.
46. A. Marin, T.E. Malliavin, P. Nicolas, and M.A. Delsuc. From NMR chemical shifts to amino acid types: investigation of the predictive power carried by nuclei. *Journal of Biomolecular NMR*, 30:47–0, 2004.
47. J.E. Masse, R. Keller, and K. Pervushin. SideLink: automated side-chain assignment of biopolymers from NMR data by relative-hypothesis-prioritization-based simulated logic. *Journal of Magnetic Resonance*, 181(1):45–67, 2006.
48. G. T. Montelione and H. N. B. Moseley. Automated analysis of NMR assignments and structures for proteins. *Curr. Opin. Struct. Biol.*, 9:635–642, 1999.
49. C. Mumenthaler, P. Güntert, W. Braun, and K. Wüthrich. Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. *Journal of Biomolecular NMR*, 10(4):351–362, 1997.
50. M. Ottiger, F. Delaglio, and A. Bax. Measurement of J and dipolar couplings from simplified two-dimensional NMR spectra. *Journal of Magnetic Resonance*, 138:373–378, 1998.
51. J. H. Prestegard, C. M. Bougault, and A. I. Kishore. Residual Dipolar Couplings in Structure Determination of Biomolecules. *Chemical Reviews*, 104:3519–3540, 2004.
52. W. Rieping, M. Habeck, and M. Nilges. Inferential Structure Determination. *Science*, 309:303 – 306, 2005.
53. K. Ruan, K. B. Briggman, and J. R. Tolman. De novo determination of internuclear vector orientations from residual dipolar couplings measured in three independent alignment media. *Journal of Biomolecular NMR*, 41:61–76, 2008.
54. S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 2002.
55. C. D. Schwieters, J. J. Kuszewski, N. Tjandra, and G. M. Clore. The Xplor-NIH NMR molecular structure determination package. *J Magn Reson*, 160:65–73, 2003.
56. X. Sun, M. J. Druzdzel, and C. Yuan. Dynamic Weighting A* Search-Based MAP Algorithm for Bayesian Networks. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2385–2390, 2007.
57. N. Tjandra and A. Bax. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science*, 278:1111–1114, 1997.
58. J. R. Tolman, J. M. Flanagan, M. A. Kennedy, and J. H. Prestegard. Nuclear magnetic dipole interactions in field-oriented proteins: Information for structure determination in solution. *Proc. Natl. Acad. Sci. USA*, 92:9279–9283, 1995.
59. E.L. Ulrich, H. Akutsu, J.F. Doreleijers, Y. Harano, Y.E. Ioannidis, J. Lin, M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C.F. Schulte, D.E. Tolmie, R.K. Wenger, H. Yao, and J.L. Markley. BioMagResBank. *Nucleic Acids Research*, 36:D402–D408, 2007. doi:10.1093/nar/gkm957.
60. O. Vitek, C. Bailey-Kellogg, B. Craig, and J. Vitek. Inferential backbone assignment for sparse data. *J. Biomolecular NMR*, 35:187–208, 2006.
61. L. Wang and B. R. Donald. Exact solutions for internuclear vectors and backbone dihedral angles from NH residual dipolar couplings in two media, and their application in a systematic search algorithm for determining protein backbone structure. *Jour. Biomolecular NMR*, 29(3):223–242, 2004.
62. L. Wang, R. Mettu, and B. R. Donald. A Polynomial-Time Algorithm for De Novo Protein Backbone Structure Determination from NMR Data. *Journal of Computational Biology*, 13(7):1276–1288, 2006.
63. Z. Wei and H. Li. A Markov random field model for network-based analysis of genomic data. *Bioinformatics*, 23:1537–44, 2007.
64. K.-P. Wu, J.-M. Chang, J.-B. Chen, C.-F. Chang, W.-J. Wu, T.-H. Huang, T.-Y. Sung, and W.-L. Hsu. RIBRA-an Error-Tolerant Algorithm for the NMR Backbone Assignment Problem. In *Proceedings of the International conference on Research in Computational Molecular Biology (RECOMB'05)*, pages 229–244, 2005.
65. Y. Xu, D. Xu, and E. C. Uberbacher. An efficient computational method for globally optimal threading. *J Comput Biol.*, 5(3):597–614, 1998.
66. J. Zeng, J. Boyles, C. Tripathy, L. Wang, A. Yan, P. Zhou, and B. R. Donald. High-Resolution Protein Structure Determination Starting with a Global Fold Calculated from Exact Solutions to the RDC Equations. *Journal of Biomolecular NMR.*, 45:265–281, 2009. PMID: 19711185.
67. J. Zeng, P. Zhou, and B. R. Donald. A Markov Random Field Framework for Protein Side-Chain Resonance Assignment – Supplementary Material. Department of Computer Science, Duke University, [online]. Available: http://www.cs.duke.edu/donaldlab/Supplementary/recomb10/. Jan, 2010.
68. D.E. Zimmerman, C.A. Kulikowski, W. Feng, M. Tashiro, C.-Y. Chien, C.B. Ríos, F.J. Moy, R. Powers, and G.T. Montelione. Automated analysis of protein NMR assignments using methods from artificial intelligence. *J. Mol. Biol.*, 269:592 – 610, 1997.