A Bayesian Approach for Determining Protein Side-Chain Rotamer Conformations Using Unassigned NOE Data*

Jianyang Zeng[1], Kyle E. Roberts[3], Pei Zhou[2], and Bruce R. Donald[1,2,3,**]

[1] Department of Computer Science, Duke University, Durham, NC 27708, USA
[2] Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, USA
[3] Program in Computational Biology and Bioinformatics, Duke University, Durham NC 27708, USA
Tel.: 919-660-6583; Fax: 919-660-6519
brd+recomb11@cs.duke.edu

Abstract. A major bottleneck in protein structure determination via nuclear magnetic resonance (NMR) is the lengthy and laborious process of assigning resonances and nuclear Overhauser effect (NOE) cross peaks. Recent studies have shown that accurate backbone folds can be determined using sparse NMR data, such as residual dipolar couplings (RDCs) or backbone chemical shifts. This opens a question of whether we can also determine the accurate protein side-chain conformations using sparse or unassigned NMR data. We attack this question by using unassigned nuclear Overhauser effect spectroscopy (NOESY) data, which record the through-space dipolar interactions between protons nearby in 3D space. We propose a Bayesian approach with a Markov random field (MRF) model to integrate the likelihood function derived from observed experimental data, with prior information (i.e., empirical molecular mechanics energies) about the protein structures. We unify the side-chain structure prediction problem with the side-chain structure determination problem using unassigned NMR data, and apply the deterministic *dead-end elimination* (DEE) and A* search algorithms to provably find the global optimum solution that maximizes the posterior probability. We employ a Hausdorff-based measure to derive the likelihood of a rotamer or a pairwise rotamer interaction from unassigned NOESY data. In addition, we apply a systematic and rigorous approach to estimate the experimental noise in NMR data, which also determines the weighting factor of the data term in the scoring function that is derived from the Bayesian framework. We tested our approach on real NMR data of three proteins, including the FF Domain 2 of human transcription elongation factor CA150 (FF2), the B1 domain of Protein G (GB1), and human ubiquitin. The promising results indicate that our approach can be applied in high-resolution protein structure determination. Since our approach does not require any NOE assignment, it can accelerate the NMR structure determination process.

* This work is supported by the following grants from National Institutes of Health: R01 GM-65982 and R01 GM-78031 to B.R.D. and R01 GM-079376 to P.Z.
** Corresponding author.

V. Bafna and S.C. Sahinalp (Eds.): RECOMB 2011, LNBI 6577, pp. 563–578, 2011.

1 Introduction

Nuclear magnetic resonance (NMR) is an important tool for determining high-resolution protein structures in the solution state. Traditional NMR structure determination approaches [19,21,40,23,35] typically use a dense set of nuclear Overhauser effect (NOE) distance restraints to calculate the 3D coordinates of the protein structure. This process requires nearly complete assignment of both resonances (which serve as IDs of atoms in NMR spectra) and NOE data. Unfortunately, assigning resonances and NOEs is a time-consuming and laborious process, which is a major bottleneck in NMR structure determination. To address this problem, several approaches have been proposed to determine protein structures using sparse experimental data [24, 53, 54, 12, 4, 7, 51, 47] or unassigned NMR data [43,46,61,59,62]. These new approaches have shown promising results. In particular, it has been shown that accurate backbone folds can be determined using sparse NMR data, such as residual dipolar couplings (RDCs) [53, 54, 12, 24] or backbone chemical shifts [51, 47]. The question remains: After the backbone structure has been solved, can we also determine accurate side-chain conformations using sparse or unassigned NMR data? In this paper, we address this question by using unassigned nuclear Overhauser effect spectroscopy (NOESY) data, which record the through-space dipolar interactions between protons nearby in 3D space. While protein backbones have previously been determined to low resolution [43, 44] or even moderate resolution [33, 17, 18, 46] using unassigned NOESY data, it has never been shown, prior to our paper, that high-resolution side-chain conformations can be computed using only unassigned NOESY data. Since our algorithm does not require any NOE assignment, it can shorten the time required in the NMR data analysis, and hence accelerate the NMR structure determination process.

Protein side-chains have been observed to exist in a number of energetically favored conformations, called *rotamers* [42]. Based on this observation, the side-chain structure determination problem can be formulated as a discrete combinatorial optimization problem, in which a set of side-chain conformations are searched over a given rotamer library to optimize a scoring function that represents both empirical molecular mechanics and data restraints. Substantial work has been developed for predicting protein side-chain conformations without using experimental data [52, 11, 22, 31, 16, 27, 3, 56, 49, 30, 57, 34]. These side-chain structure prediction approaches might be limited by the approximate nature of the employed empirical molecular mechanics energy function, which might not be sufficient to accurately capture the real energetic interactions among atoms in the protein.

Integration of NMR data with the empirical molecular mechanics energy is a challenging problem. Most frameworks for NMR protein structure determination use heuristic models with *ad hoc* parameter settings to incorporate experimental data (which are usually *assigned* NOE data in these approaches) and integrate them with the empirical molecular mechanics energy in a scoring function to compute protein structures. These approaches suffer from the subjective choices in the data treatment, which makes it difficult to objectively calculate high-quality structures. To overcome this drawback, we use a Bayesian approach [48, 20, 50] and cast the protein side-chain structure determination problem using unassigned NOESY data into a Markov random field (MRF) framework. We treat NMR data as an experimental observation on side-chain rotamer

states, and use the MRF to encode prior information about the protein structures, such as empirical molecular mechanics energies. The priors in our framework are in essence parameterized by the random variables representing the side-chain rotamer conformations. The MRF modelling captures atomic interactions among residues both from empirical molecular mechanics energies and geometric restraints from unassigned NOESY data. The derived posterior probability combines prior information and the likelihood model constructed from observed experimental data. Unlike previous *ad hoc* models, our Bayesian framework provides a rational basis to incorporate both experimental data and modelling information, which enables us to develop systematic techniques for computing accurate side-chain conformations.

The side-chain structure determination problem is NP-hard [45,8]. Therefore, a number of algorithms have been developed to address the complexity. Stochastic techniques [52, 22, 27, 49] randomly sample conformation space to generate a set of side-chain rotamer conformations. In contrast, our approach applies deterministic algorithms with provable guarantees [11, 41, 16, 15, 9, 13] to determine the optimal side-chain rotamer conformations that satisfy both experimental restraints and prior information on the protein structures. We first apply a *dead-end elimination* (DEE) algorithm [11, 41, 16] to prune side-chain conformations that are *provably* not part of the optimal solution. After that, an A* search algorithm is employed to find the global optimum solution that best interprets our MRF model.

The guarantee to provably find the global optimum using the DEE/A* algorithms enables us to rigorously and objectively estimate the experimental noise in NMR data and the weighting factor between the empirical molecular mechanics energy and experimental data in the scoring function derived in our Bayesian framework. Specifically, we employ a grid search approach to systematically search over all possible grid point values of the noise parameter, and use the DEE/A* search algorithms to compute the optimal solution that minimizes the scoring function for each grid point. We then compare the best solutions over all grid points and find the globally optimal estimation of the weight parameter. The following contributions are made in this paper:

1. A novel framework to unify the side-chain structure prediction problem with the side-chain structure determination problem using unassigned NOESY data, by applying the provable *dead-end elimination* (DEE) and A* search algorithms to find the global optimum solution;
2. A Bayesian approach with an MRF model to derive the posterior probability of side-chain conformations by combining the likelihood function from observed experimental data with prior information (i.e., empirical molecular mechanics energies) about the protein structures;
3. A systematic and rigorous approach to estimate the experimental noise in NMR data, which determines the weighting factor of the data term in the derived scoring function, by combining grid search and DEE/A* search algorithms;
4. Introduction of a Hausdorff-based measure to derive the likelihood function from unassigned NMR data;
5. Promising test results on real NMR data recorded at Duke University.

1.1 Related Work

In [61, 59], we developed an algorithm, called HANA, that employs a Hausdorff-based pattern matching technique to place the side-chain rotamer conformations on the backbone structures determined mainly using RDC data [53,54,12]. In [62], we proposed an MRF based algorithm, called NASCA, to assign side-chain resonances and compute the side-chain rotamer conformations from unassigned NOESY data without using TOCSY experiments. Neither HANA nor NASCA completely exploits prior information or all the available information from experimental data. For example, HANA only uses the back-computed NOE pattern from side-chain rotamers to backbone to calculate the likelihood of a rotamer. In addition, HANA and NASCA do not take into account the empirical molecular mechanics energy when determining the side-chain rotamer conformations. Thus, the side-chain conformations determined by these two approaches may embrace some bad local geometry such as serious steric clashes. Our current Bayesian approach improves over HANA and NASCA by eliminating all serious steric clashes (Table 3). It is a significant extension of the HANA and NASCA modules, and can be combined with our previously-developed backbone structure techniques [53,54,12,59,62] to determine high-resolution structures, using a protocol similar to [59, 62].

Several approaches have been proposed to use backbone chemical shift data [4, 7, 51, 47] or unassigned NOESY data [33, 17, 18, 43, 44, 46] in protein structure determination at different resolutions. These frameworks use a generate-and-test strategy or stochastic techniques such as Monte Carlo (MC), simulated annealing (SA), or highly-simplified molecular dynamics (HSMD) to randomly sample conformation space and compute a set of structures that satisfy the data restraints. These approaches suffer from the problem of undersampling conformation space and overfitting to the data. They cannot provide any guarantee on the convergence to the global optimum. In addition, integration of experimental data with the empirical molecular mechanics energy and the parameter settings in these frameworks are usually performed on an *ad hoc* basis.

Unlike a previous Bayesian approach in NMR structure determination [48,20], which requires *assigned* NOE data, our approach works on unassigned NOESY data. Moreover, the Bayesian approach in [48, 20] mainly relies on heuristic techniques, such as Monte Carlo or Gibbs sampling, to randomly sample both conformation space and joint posterior distribution, while our approach employs a systematic and rigorous search method (i.e., a combination of grid search and DEE/A* algorithms) to compute the optimal parameter estimation that is only subject to the resolution used in the grid search.

MRFs offer a mathematically sound framework for describing the dependencies between random variables, and have been widely applied in computer vision [14, 39] and computational structural biology [58,29]. In [29], an MRF was used to estimate the free energy of protein structures, while in [58], a graphical model similar to an MRF was used to predict side-chain conformations. Although both graphical models in [58, 29] provide a reasonable model to describe the protein side-chain rotamer interactions, they do not use any experimental data. In addition, the belief propagation approach used in [58, 29] to search for the low-energy conformations can be trapped into local minima, while our approach computes the global optimum solution.

2 Methods

2.1 Backbone Structure Determination from Residual Dipolar Couplings

In our high-resolution structure determination protocol, we apply our recently-developed algorithms [53, 54, 59, 12] 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 (SM) [60] **Section 1** and [53, 54, 12].

2.2 Using Markov Random Fields for Rotamer Assignment

We first use a Markov random field to formulate our side-chain structure determination problem. A Markov random field is a set of random variables defined on an undirected graph, which describes the conditional dependencies among random variables. In our problem, each random variable represents the rotamer state of a residue. Formally, let X_i be a random variable representing the rotamer state at residue i, where $1 \leq i \leq n$, and n is the total number of residues in the protein sequence. Let t_i be the maximum number of rotamer states at residue i. Then each random variable X_i can take a value from set $\{1, \cdots, t_i\}$. We use x_i to represent a specific value taken by random variable X_i. We also call x_i the *rotamer assignment* or *conformation* of residue i. Let $X = \{X_1, \cdots, X_n\}$ be the set of random variables representing the rotamer assignments for all residues $1, \cdots, n$ in the protein sequence. A joint event $\{X_1 = x_1, \cdots, X_n = x_n\}$, abbreviated as $X = x$, is called a *rotamer assignment* or *conformation* for all residues in the protein sequence, where $x = \{x_1, \cdots, x_n\}$.

In our side-chain structure determination problem, we assume that the backbone is rigid. Based on this assumption, it is generally safe to argue that each residue only interacts with other residues within a certain distance threshold or energy cutoff, when considering the pairwise interactions between side-chains. We use a graph $G = (V, E)$ to represent such residue-residue interactions, where each vertex in V represents a residue, and each edge in E represents a possible interaction between two residues (i.e., the minimum distance between atoms from these two residues is within a distance threshold). Such a graph $G = (V, E)$ is called the *residue interaction graph*. Given a residue interaction graph $G = (V, E)$, the *neighborhood* of residue i, denoted by N_i, is defined as $N_i = \{j \mid j \in V, i \neq j, (i, j) \in E\}$. The neighborhood system describes the dependencies between rotamer assignments for all residues in the protein sequence. A Markov random field (MRF), defined based on the neighborhood system of an underlying graph $G = (V, E)$, encodes the following conditional independencies for each variable X_i:

$$\Pr(X_i | X_j, j \neq i) = \Pr(X_i | X_j, j \in N_i). \tag{1}$$

This condition states that each random variable X_i is only dependent on the random variables in its neighborhood.

We use $\Pr(x)$ to represent the *prior* probability for a rotamer assignment $x = \{x_1, \cdots, x_n\}$ of a protein sequence, which is derived from prior information about the protein structures, such as empirical molecular mechanics. Let D be the observation data, which in this case are the unassigned NOESY data. Let σ be the experimental noise in the

unassigned NOESY data. The parameter σ is unknown and needs to be estimated. We use $\Pr(D|x,\sigma)$ to represent the *likelihood* function of a rotamer assignment x and a parameter σ given the observation D. We use $\Pr(x,\sigma|D)$ to represent the *a posteriori* probability. Our goal is to find a combination of rotamer assignment x and parameter σ, denoted by (x,σ), that maximizes the *a posteriori* probability (MAP). By Bayes's theorem, the posterior probability can be computed by

$$\Pr(x,\sigma|D) \propto \Pr(D|x,\sigma) \cdot \Pr(x) \cdot \Pr(\sigma). \tag{2}$$

2.3 Deriving the Prior Probability

According to the Hammersley-Clifford theorem [2] on the Markov-Gibbs equivalence, the distribution of an MRF with respect to an underlying graph $G = (V,E)$ can be written in the following Gibbs form:

$$\Pr(x) \propto \exp(-U(x)/\beta), \tag{3}$$

where β is a *global control parameter*, and $U(x)$ is the *prior energy* that encodes prior information about the rotamer interactions in the protein structure. The prior energy can be defined by $U(x) = \sum_{C\in\mathcal{C}} V_C(x)$, where $V_C(\cdot)$ is a *clique potential* and $\mathcal{C}$ is the set of cliques in the neighborhood system of the underlying graph $G = (V,E)$. In our problem, we only focus on one-site and two-site interactions (i.e., with cliques of size 2) in a residue interaction graph $G = (V,E)$. Given an assignment $x = \{x_1, \cdots, x_n\}$ for a residue interaction graph $G = (V,E)$, we use the following empirical molecular mechanics energy function to define the prior energy $U(x)$:

$$U(x) = \sum_{i\in V} E'(x_i) + \sum_{i\in V}\sum_{j\in N_i} E'(x_i,x_j), \tag{4}$$

where $E'(x_i)$ is the *self energy* term for rotamer assignment x_i at residue i, and $E'(x_i,x_j)$ is the *pairwise energy* term for rotamer assignments x_i and x_j at residues i and j respectively. We can use the Boltzmann distribution to further specify the prior probability in Eq. (3) by setting $\beta = k_b T$, where k_b is the Boltzmann constant, and T is the temperature.

2.4 Deriving the Likelihood Function and the Scoring Function

An accurate likelihood function should effectively interpret the observation data, and incorporate experimental uncertainty into the model. In our framework, the likelihood $\Pr(D|x,\sigma)$ is defined as

$$\Pr(D|x,\sigma) = Z(\sigma) \cdot \exp\left(-U(D|x,\sigma)\right), \tag{5}$$

where $Z(\sigma)$ is the *normalizing factor*, and $U(D|x,\sigma)$ is called the *likelihood energy*, which evaluates the likelihood of observed NOESY data given rotamer assignment x and parameter σ.

The likelihood energy $U(D|x,\sigma)$ can be measured by matching the back-computed NOE patterns with experimental cross peaks in unassigned NOESY data D. Given a rotamer assignment x_i at residue i, we can back-compute its NOE pattern between backbone and intra-residue atoms. This NOE pattern is called the *self back-computed NOE pattern*. Similarly, we can back-compute the NOE pattern between a pair of rotamer assignments x_i and x_j at residues i and j respectively. This NOE pattern is called the *pairwise back-computed NOE pattern*. We use a criterion derived from the Hausdorff distance [25, 26], called the *Hausdorff fraction*, to measure the matching score between a back-computed NOE pattern and unassigned NOESY data. Details of deriving the Hausdorff fraction for a back-computed NOE pattern are in SM [60] **Section 2** and [61, 59]. Let $F(x_i)$ and $F(x_i, x_j)$ be the Hausdorff fractions for the self and pairwise back-computed NOE patterns respectively. Then the likelihood energy $U(D|x,\sigma)$ is defined as:

$$U(D|x,\sigma) = \sum_{i\in V} \frac{\left(1 - F(x_i)/F_0(x_i)\right)^2}{2\sigma^2} + \sum_{i\in V}\sum_{j\in N_i} \frac{\left(1 - F(x_i,x_j)/F_0(x_i,x_j)\right)^2}{2\sigma^2}, \tag{6}$$

where σ is the experimental noise in unassigned NOESY data, and $F_0(x_i)$ and $F_0(x_i, x_j)$ are the *expected values* of $F(x_i)$ and $F(x_i, x_j)$ respectively. Here we assume that the experimental noise of unassigned NOESY cross peaks follows an independent Gaussian distribution. Thus, σ represents the standard deviation of the Gaussian noise. Such an independent Gaussian distribution provides a good approximation [36, 39] when the accurate noise model to describe the uncertainty in experimental data is not available. In general, it is difficult to obtain the accurate values of the expected Hausdorff fractions $F_0(x_i)$ and $F_0(x_i, x_j)$. In principle, a rotamer conformation should be closer to the native side-chain conformation if its back-computed NOE pattern has a higher Hausdorff fraction (i.e., with higher data satisfaction score). In practice, we use the maximum value of the Hausdorff fraction among the back-computed NOE patterns of all rotamers as the expected value of $F(x_i)$ and $F(x_i, x_j)$.

The function $U(x,\sigma|D) = -\log \Pr(x,\sigma|D)$ is called the *posterior energy* for a rotamer assignment x and parameter σ, given the observed data D. Then maximizing the posterior probability is equivalent to minimizing the posterior energy function. Substituting Eqs. (3), (4) and (6) into Eq. (2), and taking the negative logarithm on both sides of the equation, we have the following form of the posterior energy function:

$$U(x,\sigma|D) \propto \frac{1}{\beta}\left(\sum_{i\in V} E'(x_i) + \sum_{i\in V}\sum_{j\in N_i} E'(x_i,x_j)\right) +$$
$$\left(\sum_{i\in V} \frac{\left(1 - F(x_i)/F_0(x_i)\right)^2}{2\sigma^2} + \sum_{i\in V}\sum_{j\in N_i} \frac{\left(1 - F(x_i,x_j)/F_0(x_i,x_j)\right)^2}{2\sigma^2}\right) + \log\frac{Z(\sigma)}{\Pr(\sigma)}. \tag{7}$$

In Sec. 2.5, we will show how to estimate parameter σ. After σ has been estimated, we have the following form of the posterior energy function:

$$U(x|D) \propto \frac{1}{\beta}\left(\sum_{i\in V} E'(x_i) + \sum_{i\in V}\sum_{j\in N_i} E'(x_i,x_j)\right)$$

$$+\left(\sum_{i\in V}\frac{\left(1-F(x_i)/F_0(x_i)\right)^2}{2\sigma^2}+\sum_{i\in V}\sum_{j\in N_i}\frac{\left(1-F(x_i,x_j)/F_0(x_i,x_j)\right)^2}{2\sigma^2}\right). \quad (8)$$

The function $U(x|D)$ is also called the *pseudo energy*. We rewrite the pseudo energy function in Eq. (8). Let $E(x_i) = E'(x_i)/\beta + \left(1-F(x_i)/F_0(x_i)\right)^2/2\sigma^2$ and $E(x_i,x_j) = E'(x_i,x_j)/\beta + \left(1-F(x_i,x_j)/F_0(x_i,x_j)\right)^2/2\sigma^2$. Then we have

$$U(x|D) = \sum_{i\in V} E(x_i) + \sum_{i\in V}\sum_{j\in N_i} E(x_i,x_j). \quad (9)$$

The pseudo energy function defined in Eq. (9) has the same form as in protein side-chain structure prediction [31, 38, 49, 30, 57, 34] or protein design [11, 41, 16, 15, 9, 13]. Thus, we can apply similar algorithms, including the dead-end elimination (DEE) and A* search algorithms, to solve this problem. A brief overview of the DEE/A* algorithms is in SM [60] **Section 3** and [11, 41, 16, 15, 9]. Similar to protein side-chain prediction and protein design, the optimal rotamer assignment x^* that minimizes the pseudo energy function in Eq. (9) is called the *global minimum energy conformation (GMEC)*. The DEE/A* algorithms employed in our framework guarantee to find the GMEC with respect to our pseudo energy function. Similar to [15,9,13], we can also extend the original A* search algorithm to compute a gap-free ensemble of conformations such that their energies are all within a user-specified window from the lowest pseudo energy.

2.5 Estimation of Experimental Noise in the NOESY Data

In practice, parameter σ in Eq. (6) is generally unknown, and needs to be estimated for each set of experimental data used in structure calculation. In the likelihood function Eq. (5), the normalizing factor $Z(\sigma)$ is related to the unknown parameter σ. Based on the independent Gaussian distribution assumption on experimental noise in unassigned NOESY data, we have $Z(\sigma) = (2\pi\sigma^2)^{m/2}$, where m is the total number of self and pairwise back-computed NOE patterns. In our problem, m is equal to the size of the residue interaction graph $G = (V, E)$, that is, $m = |V| + |E|$.

Similar to [48, 20], we use the Jeffrey prior [28] to represent the prior probability of parameter σ, that is, $\Pr(\sigma) = \sigma^{-1}$. Substituting $Z(\sigma) = (2\pi\sigma^2)^{m/2}$ and $\Pr(\sigma) = \sigma^{-1}$ into Eq. (7), we have

$$U(x,\sigma|D) \propto (m+1)\log\sigma + \frac{1}{\beta}\left(\sum_{i\in V} E'(x_i) + \sum_{i\in V}\sum_{j\in N_i} E'(x_i,x_j)\right) +$$

$$\left(\sum_{i\in V}\frac{\left(1-F(x_i)/F_0(x_i)\right)^2}{2\sigma^2}+\sum_{i\in V}\sum_{j\in N_i}\frac{\left(1-F(x_i,x_j)/F_0(x_i,x_j)\right)^2}{2\sigma^2}\right). \quad (10)$$

Now our goal is to find a value of (x,σ) that minimizes the posterior energy in Eq. (10). Here we combine a grid search approach with the DEE/A* search algorithms to compute the optimal estimation of $w = \sigma^{-2}$. Once w is determined, parameter σ can be

computed using equation $\sigma = \sqrt{1/w}$. Our parameter estimation approach first incrementally searches the grid points of weighting factor w. For each grid point of w, it uses the DEE and A* search algorithms to find the GMEC that minimizes the pseudo energy function. Finally, it compares all GMEC solutions over all searched grid points, and chooses the optimal value of parameter w that minimizes the posterior energy function in Eq. (10).

In Eq. (10), as the weighting factor w increases (i.e., the data term is weighted more), the first term $(m+1)\log\sigma$ in Eq. (10) decreases, while the third term representing the data restraints increases. Fig. 1A shows a typical plot of the posterior energy $U(x, \sigma|D)$ vs. the weighting factor w, in which a minimum is usually observed. The performance of our parameter estimation approach is only subject to the resolution used in the grid search. In practice, our approach is sufficient to find the optimal parameter estimation (Fig. 1), as we will show in the Results section.

3 Results

We implemented our Bayesian approach for side-chain structure determination and tested it on NMR data of three proteins: the FF Domain 2 of human transcription elongation factor CA150 (FF2), the B1 domain of Protein G (GB1), and human ubiquitin. The numbers of amino acid residues in these three proteins are 62, 56 and 76 for FF2, GB1 and ubiquitin respectively. The PDB IDs of the NMR reference structures are 2KIQ, 3GB1 and 1D3Z for FF2, GB1, and ubiquitin respectively. The PDB IDs of the X-ray reference structures are 3HFH, 1PGA and 1UBQ for FF2, GB1, and ubiquitin respectively.

Our algorithm uses the following input data: (1) the protein primary sequence; (2) the protein backbone; (3) the 2D or 3D NOESY peak list from both ^{15}N- and ^{13}C-edited spectra; (4) the resonance assignment list, including both backbone and side-chain resonance assignments; (5) the rotamer library [42]. The empirical molecular mechanics energy function that we used in Eq. (4) consists of the Amber electrostatic, van der Waals (vdW), and dihedral terms, the EEF1 implicit solvation energy term [37], and the rotamer energy term, which represents the frequency of a rotamer that is estimated from high-quality protein structures [42]. All NMR data, except RDCs of GB1 and ubiquitin, 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. Details on the NMR experimental procedures are provided in Supplementary Material [60] **Section 4**. Our tests were performed on a 2.20 GHz Intel core 2 Duo processor with 4 GB memory. The total running time of computing the GMEC solution for a typical medium-size protein, such as GB1, is less than an hour after parameter $w = \sigma^{-2}$ has been estimated.

We used the same rules as in [42] to classify and identify the rotamer conformations, that is, we used a window of $\pm 30^\circ$ to determine most χ angles, except that a few specific values (see Table 1 in [42]) were used in determining the terminal χ angle boundaries for glutamate, glutamine, aspartate, asparagine, leucine, histidine, tryptophan, tyrosine and phenylalanine. Since most rotamer conformations are short, the RMSD is not sufficient to measure the structural dissimilarity between two rotamers. Thus, we did not

use the RMSD to compare different rotamers. We used two measurements to evaluate the accuracies of the determined side-chain rotamer conformations. The first one is called the *accuracy of all* χ *angles*, measuring the percentage of side-chain rotamer conformations in which all χ angles agree with the NMR or X-ray reference structure. The second measurement is called the *accuracy of* (χ_1, χ_2) *angles*, which measures the percentage of side-chain rotamer conformations whose first two χ angles (i.e., both χ_1 and χ_2) agree with the NMR or X-ray reference structure. We say a determined side-chain conformation is *correct* if all its χ angles agree with the NMR or X-ray reference structure.

3.1 Parameter Estimation

We estimated the weighting factor parameter $w = \sigma^{-2}$ in the posterior energy function using the approach described in Sec. 2.5. Here we used the test on GB1 (Fig. 1) as an example to demonstrate our parameter estimation approach. The parameters for the other two proteins were estimated similarly. For GB1, the optimal weighting factor was 32, where the posterior energy $U(x, \sigma|D)$ met the minimum (Fig. 1A). This optimal weight value corresponded to the best accuracies 77.8% and 87.0% for all χ angles and (χ_1, χ_2) angles respectively (Fig. 1E and Fig. 1F).

Fig. 1C and Fig. 1D show the influence of the weight w on the empirical molecular mechanics energy and the NOE pattern matching score of the GMEC. As expected, as the data restraints were weighted more, the empirical molecular mechanics energy declined while the data satisfaction score was improved for the GMEC solution. At the optimal weight value $w = 32$, the GMEC yielded decent scores for both empirical molecular mechanics energy and NOE pattern matching score. Although the NOE pattern matching score of the GMEC jumped to a higher plateau when $w \geq 110$ (Fig. 1C), the accuracies of all χ angles and (χ_1, χ_2) angles did not increase correspondingly (Fig. 1E and Fig. 1F). Probably this high NOE satisfaction score was caused to some extent by

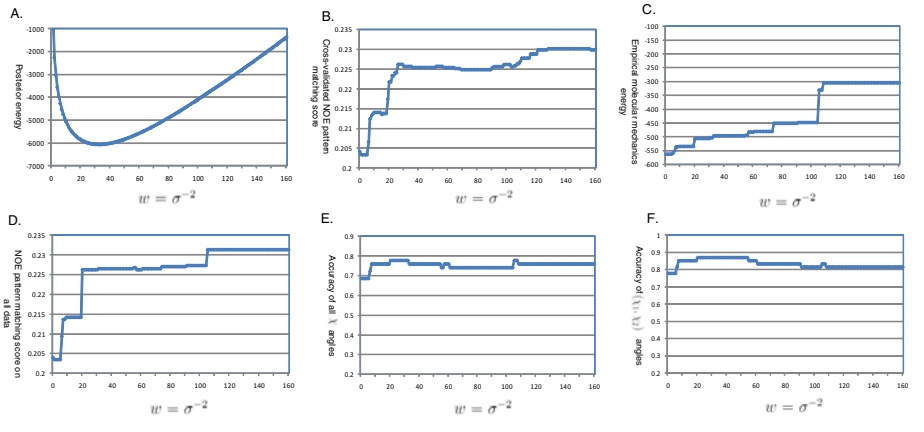

Fig. 1. Estimation of the weighting factor parameter $w = \sigma^{-2}$ for the data term in the posterior energy function for GB1. In plots (B) and (D), the Hausdorff fraction was used to measure the matching score between the back-computed NOE pattern of the GMEC and experimental spectra.

overfitting the side-chain rotamer conformations to experimental data. We also demonstrated that our approach performed better than the cross validation approach [5,6] used in estimating the weighting factor parameter $w = \sigma^{-2}$ (Fig. 1B). Details on the cross validation approach and the comparison results are provided in Supplementary Material [60] **Section 5**.

3.2 Accuracy of Determined Side-Chain Rotamer Conformations

We first tested our side-chain structure determination approach on the backbones from the NMR reference structures (Table 1). To check whether our current side-chain structure determination approach can be combined with our previously-developed backbone structure determination techniques [12,53,54,59] for high-resolution structure determination, we also tested it on the backbones computed mainly using RDC data (Table 2). The RMSD between the input RDC-defined backbone and the NMR reference structure is 0.96 Å, 0.87 Å and 0.97 Å for FF2, GB1 and ubiquitin respectively. In addition to the GMEC, we also computed the top ensemble of 50 conformations with the lowest pseudo energies (Tables 1 and 2), using an extension to the original A* algorithm [15, 9, 13]. An ensemble of computed structures is important when multiple models may agree with the experimental data [10]. In addition, an ensemble of structures can reflect the conformational difference resulting from different experimental conditions, lack of data, or protein motion in solution [10, 1].

In addition to examining the accuracies of the determined side-chain conformations in all residues, we also checked the performance of our approach in *core* residues, which are defined as those residues with solvent accessibility $\leq 10\%$. We used the software MOLMOL [32] with a solvent radius of 2.0 Å to compute solvent accessibility for each residue. Note that in the side-chain structure determination problem using experimental data, we were particularly interested in the accuracies of side-chain conformation determination in core residues because: (1) Biologically the side-chains on the interior and buried regions of the protein play more important roles in studying protein dynamics and determining the accurate structures than other residues on the protein surface; (2) In the X-ray or NMR reference structure, the data for the solvent-exposed side-chains are often missing. Thus, modeling information is often used to compute the side-chain conformations of the residues on the protein surface.

Table 1. Accuracies of the side-chain rotamer conformations determined by our approach on the backbones from the NMR reference structures

Proteins	**All residues**				**Core residues**			
	Accuracy of all χ angles (%)		**Accuracy of (χ_1, χ_2) angles (%)**		**Accuracy of all χ angles (%)**		**Accuracy of (χ_1, χ_2) angles (%)**	
	GMEC	**Top 50**	**GMEC**	**Top 50**	**GMEC**	**Top 50**	**GMEC**	**Top 50**
GB1	77.8	77.8	87.0	87.0	100.0	100.0	100.0	100.0
ubiquitin	75.4	78.3	84.1	85.5	84.0	88.0	88.0	92.0
FF2	71.9	71.9	82.5	86.0	100.0	100.0	100.0	100.0

Table 2. Accuracies of the side-chain rotamer conformations determined by our approach on the RDC-defined backbones computed using the algorithms in [12, 53, 54, 59]

Proteins	All residues				Core residues			
	Accuracy of all χ angles (%)		Accuracy of (χ_1, χ_2) angles (%)		Accuracy of all χ angles (%)		Accuracy of (χ_1, χ_2) angles (%)	
	GMEC	Top 50	GMEC	Top 50	GMEC	Top 50	GMEC	Top 50
GB1	75.9	79.6	81.5	88.9	92.9	100.0	92.9	100.0
ubiquitin	72.5	76.8	79.7	82.6	80.0	84.0	80.0	84.0
FF2	71.9	75.4	80.7	84.2	100.0	100.0	100.0	100.0

Overall, our approach determined more than 70% correct rotamer conformations, and achieved over 80% accuracy for (χ_1, χ_2) angles for all residues (Tables 1 and 2). Our results also show that computing the ensemble of top 50 conformations with the lowest pseudo energies can slightly improve the results (Tables 1 and 2), which indicates that it is necessary to compute an ensemble of conformations rather than a single GMEC solution. In core residues, our approach achieved a high percentage of accurate side-chain conformations. Our approach computed all the correct side-chain conformations in core residues for GB1 and FF2, and had accuracies $\geq$ 84% for ubiquitin, given the backbone structures from the NMR reference structures (Table 1). The tests on the RDC-defined backbones exhibited similar results (Table 2), which indicates that our current Bayesian approach can be combined with our previously-developed backbone structure determination techniques [12, 53, 54, 59] to determine high-resolution protein structures mainly using RDC and unassigned NOESY data.

We also examined the accuracies of the determined side-chain conformations for residues of different lengths (Fig. 2). In general, more short side-chain conformations (i.e., 1-χ and 2-χ side-chains) were determined correctly than the long side-chain conformations (i.e., 3-χ and 4-χ side-chains). On the other hand, although our program assigned a very low percentage of correct 4-χ rotamers (i.e., arginine and lysine), it was able to compute the first two χ angles correctly for most 4-χ side-chains (Fig. 2). In addition to their side-chain flexibility, arginine and lysine are usually exposed to the solvent and undergo many conformational changes. Also, their NOE data are often missing. Therefore, it is generally difficult to compute all the χ angles correctly for these two long side-chains. We further investigated the accuracies of the determined side-chain rotamer conformations for residues with different numbers of available data restraints. We first define the *number of matched NOE peaks* for residue i, denoted by D_i, as follows:

$$D_i = \frac{1}{t_i} \sum_{x_i} \left(f(x_i) + \sum_{j \in N_i} \max_{x_j} f(x_i, x_j) \right), \tag{11}$$

where t_i is the maximum number of rotamer states at residue i, and $f(x_i)$ and $f(x_i, x_j)$ are the numbers of experimental NOE cross peaks that are close to a back-computed NOE peak in the self and pairwise back-computed NOE patterns respectively. Basically D_i measures the degree of available data restraints for residue i averaged over all possible rotamer conformations. We define the value of D_i divided by the number of

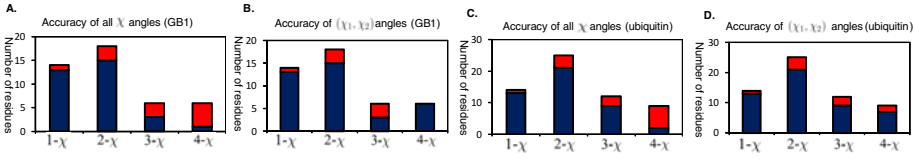

Fig. 2. Accuracies of the determined side-chain rotamer conformations for residues with different lengths (i.e., with different numbers of rotatable χ angles) for GB1 and ubiquitin. The bars represent the number of residues of the indicated type in the protein. The portions marked in blue represent the percentage of rotamers with all χ angles or (χ_1, χ_2) angles that agree with the NMR or X-ray reference structure.

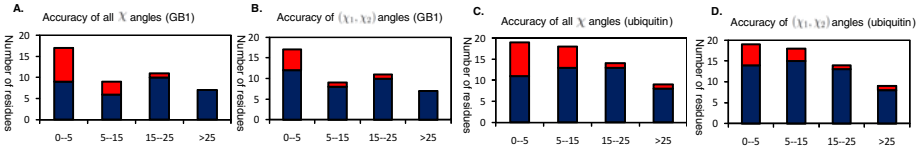

Fig. 3. Accuracies of the determined side-chain rotamer conformations for residue with different numbers of matched NOE peaks per χ angle for GB1 and ubiquitin. Diagrams are shown in the same format as in Fig. 2.

rotatable χ angles in the side-chain as the *number of matched NOE peaks per χ angle* for residue i. As shown in Fig. 3, our approach performed much better on those residues with relatively dense data restraints (i.e., with the number of matched NOE peaks per χ angle ≥ 15) than other residues.

3.3 Improvement on Our Previous Approaches HANA and NASCA

In our previous approaches, HANA [61, 59] and NASCA [62], only experimental data were used in determining side-chain conformations. Thus, they did not consider the empirical molecular mechanics energy when packing side-chain conformations. Thus, the side-chain structures computed by HANA and NASCA can contain steric clashes. Our new approach solves this problem by taking into account a molecular mechanics potential, which sharply penalizes physically unrealistic conformations. As shown in Table 3, our new approach eliminated all the serious steric clash overlaps (> 0.9 Å), which appeared previously in the side-chain conformations computed by HANA and NASCA.

3.4 Comparisons with SCWRL4

SCWRL4 [34] is one of the most popular programs for predicting side-chain rotamer conformations given a backbone structure. Note that our approach uses unassigned NOESY data, while SCWRL4 does not use any experimental data. We compared the performance of our approach with that of SCWRL4 on GB1 using different input backbone structures (Table 4). The comparison showed that our approach outperformed SCWRL4

Table 3. Comparison between our current Bayesian approach and HANA and NASCA on the number of serious steric clash overlaps (> 0.9 Å) in the determined side-chain conformations

Proteins	Current Bayesian approach	HANA	NASCA
GB1	0	10	14
ubiquitin	0	16	21
FF2	0	2	14

Table 4. Comparison with the side-chain structure prediction program SCWRL4 on GB1 using different input backbone structures. The backbone RMSD from 2GB1, 1GB1 , 1P7E, 1PGA and 1PGB to 3GB1 is 1.01 Å, 1.00 Å, 0.44 Å, 0.54 Å and 0.56 Å respectively. The program REDUCE [55] was used to add hydrogens to the X-ray backbone structures 1PGA and 1PGB. In our approach, the GMEC was computed for this comparison.

Backbones	All residues				Core residues			
	Accuracy of all χ angles (%)		Accuracy of (χ_1, χ_2) angles (%)		Accuracy of all χ angles (%)		Accuracy of (χ_1, χ_2) angles (%)	
	Our approach	SCWRL4	Our approach	SCWRL4	Our approach	SCWRL4	Our approach	SCWRL4
3GB1	77.8	72.2	85.2	79.4	100.0	85.7	100.0	85.7
2GB1	72.1	68.5	81.3	74.5	92.9	78.6	92.9	78.6
1GB1	74.1	70.4	83.3	77.8	92.9	78.6	92.9	78.6
1P7E	74.1	70.4	83.3	75.9	92.9	78.6	92.9	78.6
1PGA	70.4	64.8	79.6	70.4	92.9	71.4	92.9	71.4
1PGB	75.9	74.1	83.3	77.8	100.0	85.7	100.0	85.7

for all input backbone structures, especially on the core regions (Table 4). For core residues, our approach achieved accuracies between 92.9-100.0%, while SCWRL4 only achieved accuracies up to 85.7%. As we discussed previously, the correctness of the side-chain conformations on the core regions is crucial for determining the accurate global fold of a protein. Thus, in order to meet the requirement of high-resolution structure determination, the data restraints must be incorporated for packing the side-chain conformations in core residues.

4 Conclusions

In this paper, we unified the side-chain structure prediction problem with the side-chain structure determination problem using unassigned NOESY data. We proposed a Bayesian approach to integrate experimental data with modeling information, and used the provable algorithms to find the optimal solution. Tests on real NMR data demonstrated that our approach can determine a high percentage of accurate side-chain conformations. Since our approach does not require any NOE assignment, it can accelerate NMR structure determination.

Availability

The source code of our program is available by contacting the authors, and is distributed open-source under the GNU Lesser General Public License (Gnu, 2002). The source code can be freely downloaded after publication of this paper.

Acknowledgements

We thank Mr. Pablo Gainza and Ms. Swati Jain for helping us set up the DEE/A* code. We thank all members of the Donald and Zhou Labs for helpful discussions and comments.

References

1. Andrec, M., et al.: Proteins 69(3), 449–465 (2007)
2. Besag, J.: J. Royal Stat. Soc. B 36 (1974)
3. Bower, M.J., et al.: J. Mol. Biol. 267(5), 1268–1282 (1997)
4. Bowers, P.M., et al.: J. Biomol. NMR 18(4), 311–318 (2000)
5. Brünger, A.T.: Nature 355(6359), 472–475 (1992)
6. Brünger, A.T., et al.: Science 261(5119), 328–331 (1993)
7. Cavalli, A., et al.: Proc. Natl. Acad. Sci. USA 104(23), 9615–9620 (2007)
8. Chazelle, B., et al.: INFORMS J. on Computing 16(4), 380–392 (2004)
9. Chen, C.Y., et al.: Proc. Natl. Acad. Sci. USA 106, 3764–3769 (2009)
10. De Pristo, M.A., et al.: Structure 12(5), 831–838 (2004)
11. Desmet, J., et al.: Nature 356, 539–542 (1992)
12. Donald, B.R., Martin, J.: Progress in NMR Spectroscopy 55, 101–127 (2009)
13. Frey, K.M., et al.: Proc. Natl. Acad. Sci. USA 107(31), 13707–13712 (2010)
14. Geman, S., Geman, D.: IEEE Trans. Pattern Anal. Mach. Intell., 721–741 (1984)
15. Georgiev, I., et al.: Journal of Computational Chemistry 29, 1527–1542 (2008)
16. Goldstein, R.F.: Biophysical Journal 66, 1335–1340 (1994)
17. Grishaev, A., Llinás, M.: Proc. Natl. Acad. Sci. USA 99, 6707–6712 (2002)
18. Grishaev, A., Llinás, M.: Proc. Natl. Acad. Sci. USA 99, 6713–6718 (2002)
19. Güntert, P.: Progress in Nuclear Magnetic Resonance Spectroscopy 43, 105–125 (2003)
20. Habeck, M., et al.: Proc. Natl. Acad. Sci. USA 103(6), 1756–1761 (2006)
21. Herrmann, T., et al.: Journal of Molecular Biology 319(1), 209–227 (2002)
22. Holm, L., Sander, C.: Proteins 14(2), 213–223 (1992)
23. Huang, Y.J., et al.: Proteins 62(3), 587–603 (2006)
24. Hus, J.C., et al.: J. Am. Chem. Soc. 123(7), 1541–1542 (2001)
25. Huttenlocher, D.P., Kedem, K.: Distance Metrics for Comparing Shapes in the Plane. In: Donald, B.R., et al. (eds.) Symbolic and Numerical Computation for Artificial Intelligence, pp. 201–219. Academic press, London (1992)
26. Huttenlocher, D.P., et al.: IEEE Trans. Pattern Anal. Mach. Intell. 15(9), 850–863 (1993)
27. Hwang, J.K., Liao, W.F.: Protein Eng. 8(4), 363–370 (1995)
28. Jeffreys, H.: Proceedings of the Royal Society of London (Series A) 186, 453–461 (1946)
29. Kamisetty, H., et al.: Journal of Computational Biology 15, 755–766 (2008)
30. Kingsford, C.L., et al.: Bioinformatics 21(7), 1028–1036 (2005)
31. Koehl, P., Delarue, M.: J. Mol. Biol. 239(2), 249–275 (1994)
32. Koradi, R., et al.: J. Mol. Graph. 14(1) (1996)
33. Kraulis, P.J.: J. Mol. Biol. 243(4), 696–718 (1994)
34. Krivov, G.G., et al.: Proteins 77(4), 778–795 (2009)
35. Kuszewski, J., et al.: J. Am. Chem. Soc. 126(20), 6258–6273 (2004)
36. Langmead, C.J., Donald, B.R.: J. Biomol. NMR 29(2), 111–138 (2004)
37. Lazaridis, T., Karplus, M.: Proteins 35(2), 133–152 (1999)
38. Leach, A.R., Lemon, A.P.: Proteins 33(2), 227–239 (1998)
39. Li, S.Z.: Markov random field modeling in computer vision. Springer, London (1995)

40. Linge, J.P., et al.: Bioinformatics 19(2), 315–316 (2003)
41. Looger, L.L., Hellinga, H.W.: J. Mol. Biol. 3007(1), 429–445 (2001)
42. Lovell, S.C., et al.: Proteins: Structure Function and Genetics 40, 389–408 (2000)
43. Meiler, J., Baker, D.: Proc. Natl. Acad. Sci. USA 100(26), 15404–15409 (2003)
44. Meiler, J., Baker, D.: J. Magn. Reson. 173(2), 310–316 (2005)
45. Pierce, N.A., Winfree, E.: Protein Eng. 15(10), 779–782 (2002)
46. Raman, S., et al.: J. Am. Chem. Soc. 132(1), 202–207 (2010)
47. Raman, S., et al.: Science 327(5968), 1014–1018 (2010)
48. Rieping, W., et al.: Science 309, 303–306 (2005)
49. Rohl, C.A., et al.: Proteins 55(3), 656–677 (2004)
50. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice Hall, Englewood Cliffs (2002)
51. Shen, Y., et al.: Proc. Natl. Acad. Sci. USA 105(12), 4685–4690 (2008)
52. Tuffery, P., et al.: J. Biomol. Struct. Dyn. 8(6), 1267–1289 (1991)
53. Wang, L., Donald, B.R.: Jour. Biomolecular NMR 29(3), 223–242 (2004)
54. Wang, L., et al.: Journal of Computational Biology 13(7), 1276–1288 (2006)
55. Word, J.M., et al.: J. Mol. Biol. 285(4), 1735–1747 (1999)
56. Xiang, Z., Honig, B.: J. Mol. Biol. 311(2), 421–430 (2001)
57. Xu, J., Berger, B.: Journal of the ACM 53(4), 533–557 (2006)
58. Yanover, C., Weiss, Y.: In: NIPS (2002)
59. Zeng, J., et al.: Journal of Biomolecular NMR 45(3), 265–281 (2009)
60. Zeng, J., et al. A Bayesian Approach for Determining Protein Side-Chain Rotamer Conformations Using Unassigned NOE Data–Supplementary Material (2011), `http://www.cs.duke.edu/donaldlab/Supplementary/recomb11/bayesian/`
61. Zeng, J., et al. In: Proceedings of CSB 2008, Stanford CA (2008) PMID: 19122773
62. Zeng, J., et al.: A markov random field framework for protein side-chain resonance assignment. In: Berger, B. (ed.) RECOMB 2010. LNCS, vol. 6044, pp. 550–570. Springer, Heidelberg (2010)