A Hausdorff-Based NOE Assignment Algorithm Using Protein Backbone Determined from Residual Dipolar Couplings and Rotamer Patterns (Supporting Material)

Jianyang (Michael) Zeng and Chittaranjan Tripathy
Department of Computer Science, Duke University, Durham, NC 27708, USA

Pei Zhou
Department of Biochemistry, Duke University Medical Center, Durham, NC 27708, USA

Bruce R. Donald*†
Department of Computer Science, Duke University, Department of Biochemistry, Duke University Medical Center, Durham, NC 27708, USA
**Email: brd+csb08@cs.duke.edu*

Below is supplementary material for the following paper:

- J. Zeng, C. Tripathy, P. Zhou, and B. R. Donald. "A Hausdorff-Based NOE Assignment Algorithm Using Protein Backbone Determined from Residual Dipolar Couplings and Rotamer Patterns." *Proceedings of the 7th Annual International Conference on Computational Systems Bioinformatics (CSB)*, Stanford CA. In Press. (2008).

Appendix

The following is an appendix which provides additional information to substantiate the claims of the paper.[15] **Appendix 1** describes details of the high-resolution protein backbone computation from residual dipolar coupling data. In **Appendix 2**, we state Chernoff bounds. **Appendices 3** and **4** describe the pseudocode for the Hausdorff-based similarity measure and the NOE assignment algorithm HANA, respectively. **Appendix 5** provides a detailed proof our main theorem (Theorem 5.1 in our paper[15]). Finally, in **Appendix 6** we present an analysis of the running time of HANA.

1. Details of Protein Backbone Structure Determination from Residual Dipolar Couplings

RDC-EXACT refers to the first polynomial-time *de novo* algorithm (in fact, linear-time in the number of residues of a protein) for high-resolution protein backbone structure determination developed in Refs. 13 and 11. It takes as input (a) two RDCs per residue (e.g., assigned NH RDCs in two media or NH and CH RDCs in a single medium), (b) delimited α-helices and β-sheets with known hydrogen bond information between paired strands, and a few unambiguous NOEs (used to pack the helices and strands). RDC-EXACT differs from previous approaches for computing backbone conformation in native state from experimental data in (a) the number of restraints used, (b) how backbone dihedral angles are computed, and (c) how the conformational space is searched. RDC-EXACT does not randomly search the entire conformation space to find solutions consistent with the RDC data. Rather, it formulates the problem such that the structures computed are *exact solutions* of a system of quartic monomial equations derived from the RDC equation

$$r = D_{\mathtt{max}}\mathbf{v}^T\mathbf{S}\mathbf{v}, \tag{1}$$

*Corresponding author.
†This work is supported by the following grant to B.R.D.: National Institutes of Health (R01 GM-65982).

where r is the experimentally observed RDC, $D_{\texttt{max}}$ is the dipolar interaction constant, $\mathbf{S}$ is the 3×3 *Saupe order matrix*,[6] or *alignment tensor* that specifies the ensemble-averaged anisotropic orientation of the protein in the laboratory frame, and $\mathbf{v}$ represents the internuclear bond vector. Letting $D_{\texttt{max}} = 1$ for simplicity of exposition, and considering a global coordinate frame that diagonalizes the alignment tensor $\mathbf{S}$ (such a coordinate frame is called *principal order frame (POF)*), Equation (1) can be rewritten as

$$r = S_{xx}x^2 + S_{yy}y^2 + S_{zz}z^2, \tag{2}$$

where S_{xx}, S_{yy} and S_{zz} are the three diagonal elements of $\mathbf{S}$, and x, y and z are, respectively, the x, y and z components of the unit vector $\mathbf{v}$ in a POF which diagonalizes $\mathbf{S}$, which is a 3×3 symmetric, traceless matrix with five independent elements.[9, 10] Given NH RDCs in two aligning media (or NH and CH RDCs in single medium), the associated NH vector $\mathbf{v}$ must lie on the intersection of two conic curves.[8, 14] We state the following two propositions without proof (detailed proofs an be found in Refs. 11 and 13), which form the corner-stone of the exact-solution based polynomial time algorithm for backbone computation.

Proposition 1.1. ***(Ref. 11)***. *Given the diagonal Saupe elements S_{xx} and S_{yy} for medium 1, S'_{xx} and S'_{yy} for medium 2 and a relative rotation matrix $\mathbf{R}$ between the POFs of medium 1 and 2, the square of the x-component of the unit vector $\mathbf{v}$ satisfies a monomial quartic equation.*

Proposition 1.2. ***(Ref. 11)***. *Given the NH unit vectors $\mathbf{v}_i$ and $\mathbf{v}_{i+1}$ of residues i and $i+1$ and the $\overrightarrow{NC_\alpha}$ vector of residue i the sines and cosines of the intervening backbone dihedral angles (ϕ, ψ) satisfy the trigonometric equations $\sin(\phi + g_1) = h_1$ and $\sin(\psi + g_2) = h_2$, where g_1 and h_1 are constants depending on $\mathbf{v}_i$ and $\mathbf{v}_{i+1}$, and g_2 and h_2 depend on $\mathbf{v}_i$, $\mathbf{v}_{i+1}$, $\sin\phi$ and $\cos\phi$. Furthermore, exact solutions for $\sin\phi$, $\cos\phi$, $\sin\psi$, and $\cos\phi$ can be computed from a quadratic equation by tangent half-angle substitution.*

RDC-EXACT reduces the problem of searching the conformational space to finding the roots of a system of low-degree (quartic) monomial equations, which are discrete, finite and algebraic. A depth-first systematic search over all possible conformations (solutions) that employs a provable pruning strategy (which guarantees pruned conformations need not be considered further) based on a real solution filter and a Ramachandran Filter is used to output the conformations of a given secondary structure element that agrees the best with both the experimental RDCs and the geometry of the corresponding secondary structure type.

Given a set of computed secondary structure elements, the backbone fold is computed by computing the relative translations between these oriented secondary structure elements. This is done by using a sparse set of NOE distance restraints. At least three NOEs are needed to pack a pair of secondary structure elements, and to resolve the 4-fold orientational degeneracy in the relative pose between the secondary structure elements arising due to the symmetry of the dipolar operator (when RDCs are measured in one medium).

A loop connects two consecutive secondary structure elements. Unlike secondary structure elements, the geometry of a loop does not follow any specific pattern, and can be less ordered. Given the orientation of the end peptide plane of one and the beginning peptide plane of another secondary structure element, the loop computation (a.k.a., *loop closure*) problem involves computing an ensemble of loops that fit the missing portion of the backbone chain without violating the backbone geometry, and simultaneously satisfying the experimental data recorded for the loop. The loop closure problem is an instance of an inverse-kinematics problem, which can be solved exactly to enumerate all solutions (16 of them at most) for 3 residue-long loops, i.e., for 6 degrees of freedom (DOF) in the absence of experimental data. For loops with more than 3 residues (i.e., with more than 6 DOFs) this problem is underconstrained, thus a continuous family of solutions are possible in the absence of additional constraints. RDCs provide algebraic restraints on the global orientation of many bond vectors in the loops. Whenever two RDCs per residue are available for each residue in a loop, we use RDC-EXACT with real solution and steric filters to compute the loops.[13] In case of missing RDC data for the loops, we used an enhanced version of robotics-based cyclic coordinate descent (CCD) algorithm[3, 7] in conjunction with a steric filter to compute loops that also minimize the deviation between the experimental RDCs and the back-computed

RDCs (measured on the respective bond vectors in each residue), and without any steric clash with the remainder of the protein structure.

2. Chernoff Tail Bounds

The Chernoff bound provides a bound for the success of majority agreement for a sequence of independent events. The following lemma (Lemma 2.1) gives common formats of the Chernoff bound that determines the bound on the number of trials in order to obtain the majority agreement up to a specified probability.

Lemma 2.1. ***(Chernoff Tail Bounds[5, 1]):*** *Let $X_1, \cdots, X_n$ be a sequence of independent Poisson trials such that $\Pr(X_i) = p_i$, where $0 < p_i < 1$ and $1 \leq i \leq n$. Suppose that $X = \sum_{i=1}^{n} X_i$ and $\mu = E(X) = \sum_{i=1}^{n} p_i$. Then for any $\epsilon > 0$, we have*

$$\Pr\big(X \leq (1-\epsilon)\mu\big) \leq \exp(-\mu\epsilon^2/2), \text{ for any } 0 < \epsilon < 1; \tag{3}$$

$$\Pr\big(X \geq (1+\epsilon)\mu\big) \leq \exp(-\mu\epsilon^2/(2+\epsilon)), \text{ for any } \epsilon > 0. \tag{4}$$

Based on Lemma 2.1, we can easily derive the following extended lemma of tail bounds that is useful for the proof of Theorem 5.1 (Section 5 in our paper[15] and **Appendix 5**).

Lemma 2.2. ***(Extended Tail Bounds):*** *Let $X_1, \cdots, X_n$ be a sequence of independent Poisson trials such that $\Pr(X_i) = p_i$, where $0 < p_i < 1$ and $1 \leq i \leq n$. Suppose that $X = \sum_{i=1}^{n} X_i$ and $\mu = E(X) = \sum_{i=1}^{n} p_i$. Then we have*

$$\Pr(X \leq \gamma) \leq \exp\big(-\frac{(\mu-\gamma)^2}{2\mu}\big), \text{ for any } 0 < \gamma < \mu. \tag{5}$$

$$\Pr(X \geq \gamma) \leq \exp\big(-\frac{(\mu-\gamma)^2}{\mu+\gamma}\big), \text{ for any } \gamma > \mu; \tag{6}$$

3. Pseudo Code for Computing the Similarity Score for an NOE Pattern

Let B be the back-computed NOE pattern, and let Y be the experimental NOESY spectrum. Let δ_j be the error tolerance in the NOESY spectrum in the jth dimension, and let σ_j be the uncertainty of the NOE peak position in the jth dimension, where $j = 1, 2, 3$. Let $(\omega(a_1), \omega(a_2), \omega(a_3)) \in B$ be the back-computed NOE peak for an expected NOE $\big((a_1), (a_2), (a_3)\big)$. The pseudocode for calculating the similarity score between the back-computed NOE pattern B and the experimental NOESY spectrum Y is given in Algorithm 3.1. For each rotamer, the computation of its similarity score based on the Hausdorff distance using Algorithm 3.1 takes $O(mw)$ time, where m is the number of back-computed NOE peaks, and w is the total number of cross peaks in the experimental NOESY spectrum.

4. Pseudo Code for NOE Assignment Algorithm HANA

Figure 1 shows the flow chart of our NOESY data interpretation approach for the structure determination. The NOE assignment process is divided into three phases: initial NOE assignment (phase 1), rotamer selection (phase 2) and filtration of ambiguous NOE assignments (phase 3). In the initial NOE assignment phase, all possible pairs of ambiguous NOEs are assigned to a NOESY cross peak if the resonances of corresponding atoms fall within a tolerance window around the NOE peak. In the rotamer selection phase, an extended model of the Hausdorff distance (Section 4.3) is used to measure the match between the back-computed NOE pattern and the experimental spectrum, and thus choose the ensemble of best rotamers with top match scores. Here our rotamer selection is different from that in Ref. 12:

Algorithm 3.1 Similarity Score Calculation Based on the Hausdorff Distance

Function **Hausdorff_Score** (B, Y) /* B is the back-computed NOE pattern, and Y is the NOESY spectrum. */
$x_0, x_{\max}, x, s, \theta \leftarrow 0$;
$m \leftarrow |B|$; /* m is the number of back-computed NOE peaks. */
for each $(\omega(a_1), \omega(a_2), \omega(a_3)) \in B$ **do**
  for each $(p_1, p_2, p_3) \in Y$ **do**
    if $|p_1 - \omega(a_1)| < \delta_1$ **and** $|p_2 - \omega(a_2)| < \delta_2$ **and** $|p_3 - \omega(a_3)| < \delta_3$ **then**
      /* δ_j is the error tolerance in the NOESY spectrum in the jth dimension, $j = 1, 2, 3$. */
      $x_0 \leftarrow \prod_{j=1}^{3} \mathcal{N}(|\omega(a_j) - p_j|, \sigma_j)$;
      /* $\mathcal{N}(|x - \mu|, \sigma)$ is the probability of observing the difference $|x - \mu|$ with mean μ and deviation σ. */
      if $x_0 > x_{\max}$ **then**
        $x_{\max} \leftarrow x_0$;
      end if
    end if
  end for
  $x \leftarrow x + x_{\max}$;
end for
$s \leftarrow x/m$;
$\rho \leftarrow \sqrt{2mp(1-p)}$; /* p is the probability for a back-computed NOE peak to randomly match an experimental peak. */
$\theta \leftarrow \frac{1}{2}(\Phi((1-p)m\rho^{-1}) - \Phi((s-p)m\rho^{-1}))$; /* probability of a false random match. */
return $(1 - \theta)x/m$;

rotamers in Ref. 12 were chosen statistically from a high-resolution protein structure database, while our rotamer selections are driven directly from the pattern match score between back-computed and experimental NOE peaks. In the last phase, ambiguous NOE assignments are filtered based on the structure obtained by combining the high-resolution backbone (Section 4.2) and the ensemble of chosen rotamers. The final NOE assignments are fed into standard structure determination programs, such as XPLOR/CNS[2] or CYANA[4] for the structure calculation.

The following notations will be used in the description of our NOE assignment algorithm HANA (Algorithm 4.1). Let $Y = \{p_1, \ldots, p_w\}$ denote the set of experimental NOESY peaks, where w is the total number of NOESY peaks. Let A_i denote the set of atom triples that are assigned to peak p_i. Let $\mathcal{A} = \{a_1, \ldots, a_q\}$ denote the set of all atoms (including all protons) in the protein, where q is the total number of atoms. Let $L = \{\omega(a_1), \ldots, \omega(a_q)\}$ denote the set of chemical shifts for all atoms, where $\omega(a_i)$ is the chemical shift of atom a_i. Let δ_j denote the error tolerance in the jth dimension for the initial NOE ambiguous assignment, where $j = 1, 2, 3$. Let n be the number of residues in the protein, and let t be the maximum number of rotamer in a residue. Let $\mathbf{r_{ij}}$ denote the rotamer j at residue i, where $i = 1, \ldots, n$, $j = 1, \ldots, t$. Let u denote the NOE upper-limit distance bound. Let $\mathbf{P}$ denote the structure after combining the ensemble of chosen rotamers with the backbone computed from RDC-EXACT, and let $d(\| a_1 - a_2 \|, \mathbf{P})$ denote the minimum Euclidean distance between atoms a_1 and a_2 over all pairs of chosen rotamers in the three-dimensional structure $\mathbf{P}$. Let $B_{ij} = \{b_1, \ldots, b_m\}$ denote the set of back-computed NOE peaks for rotamer $\mathbf{r_{ij}}$, where m is the total number of back-computed NOE peaks, and $b_i = (\omega(a_1), \omega(a_2), \omega(a_3))$ denote the back-computed NOE peak for an expected NOE (a_1, a_2, a_3) from rotamer $\mathbf{r_{ij}}$. Let s_{ij} denote the similarity score of rotamer $\mathbf{r_{ij}}$ based on the extended Hausdorff measure. Let R_i denote the ensemble of top k rotamers chosen at residue i.

The details of HANA are as follows (Algorithm 4.1). In Phase 1 (namely initial NOE assignment), for each cross peak (p_1, p_2, p_3) in the NOESY spectrum, we search the resonance list and assign triple(s) of atoms (a_1, a_2, a_3) to (p_1, p_2, p_3) such that $p_1 - \delta_1 \leq \omega(a_1) \leq p_1 + \delta_1$, $p_2 - \delta_2 \leq \omega(a_2) \leq p_2 + \delta_2$, and $p_3 - \delta_3 \leq \omega(a_3) \leq p_3 + \delta_3$. In the rotamer selection phase, we first place all rotamers $\mathbf{r_{ij}}$ into backbone by rotation and translation computed based on the coordinates of H^N, C^α and N atoms. Then for each proton a_3 in rotamer $\mathbf{r_{ij}}$, we search the backbone structure and find all backbone protons a_1 that are within the NOE upper-bound limit from proton a_3 (an extra 2.5 Å is added as the correction of the upper-bound for every methyl group). Next for each expected NOE (a_1, a_2, a_3), we back compute its expected NOE peak $(\omega(a_1), \omega(a_2), \omega(a_3))$ based on the mapping between each atom name a and corresponding chemical shift $\omega(a)$ in the resonance list. Let $B_{ij} = \{(\omega(a_1), \omega(a_2), \omega(a_3))\}$ denote the set of all back-computed NOE peaks for rotamer $\mathbf{r_{ij}}$. We next call the function **Hausdorff_Score** to compute the match score between the NOE pattern B_{ij} of rotamer $\mathbf{r_{ij}}$ and the experimental NOESY spectrum Y. Finally we pick the top k rotamers with highest similarity scores at each residue i. In Phase 3 (namely filtration of ambiguous NOE assignment), we first place the top k rotamers (selected in the second phase) at each residue into backbone, and then obtain a protein structure $\mathbf{P}$. Note

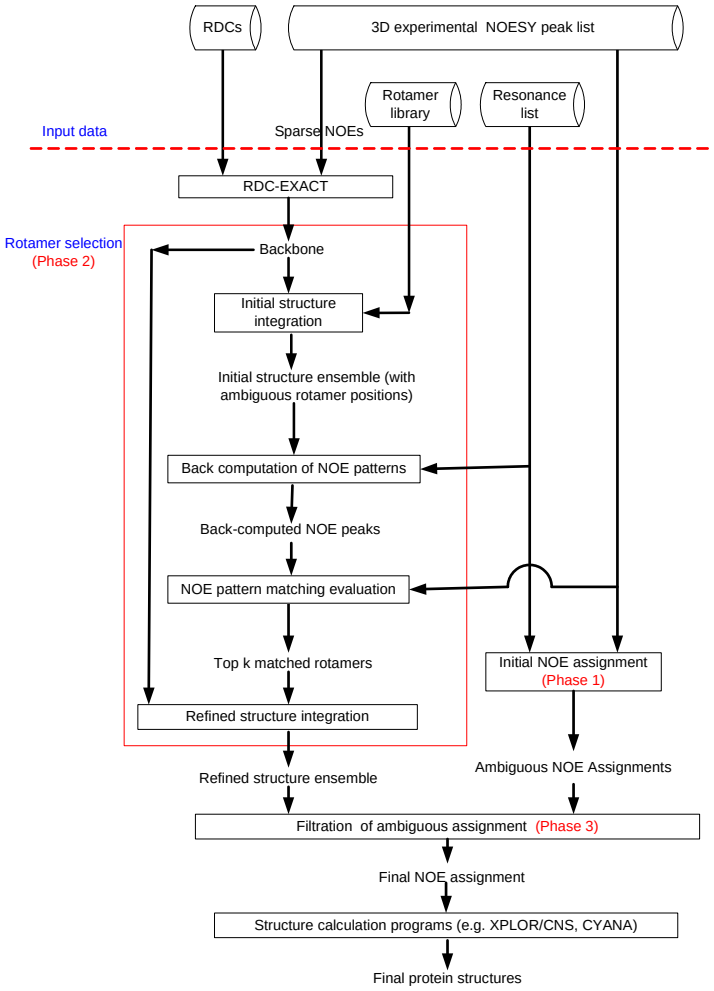

Fig. 1. Flow chart of our NOESY interpretation approach for the structure determination.

that each sidechain atom in structure $\mathbf{P}$ has k possible positions from the top k chosen rotamers. Next, for each initial NOE assignment (a_1, a_2, a_3) obtained in the first phase, we measure the Euclidean distance between protons a_1 and a_3 in structure $\mathbf{P}$. Recall that $d(\| a_1 - a_2 \|, \mathbf{P})$ stands for the minimum Euclidean distance between atoms a_1 and a_2 over all pairs of chosen rotamers in structure $\mathbf{P}$. In HANA, an NOE assignment (a_1, a_2, a_3) (from the initial NOE assignment in Phase 1) is pruned, if $d(\| a_1 - a_2 \|, \mathbf{P})$ is larger than the NOE upper-bound limit.

5. Proof of Theorem 5.1

In this section, we give the details of the proof for Theorem 5.1. We first restate the theorem and then provide the proof.

Theorem 5.1. *Suppose that* $m_f\mu_t - m_t\mu_f \geq \max(m_f, \sqrt{m_f m_t}) \cdot 4\sqrt{\mu_t \ln m_t}$. *Then with probability at least* $1 - m_t^{-1}$, *our algorithm chooses the true rotamer* $\mathbf{r_t}$ *rather than the false rotamer* $\mathbf{r_f}$.

Proof: Let X, Y be random variables as defined in Lemma 5.1 in Ref. 15. Based on our algorithm, the true rotamer $\mathbf{r_t}$ is chosen if and only if the similarity score of the true rotamer $\mathbf{r_t}$ is larger than that of the false rotamer $\mathbf{r_f}$, that is, $\frac{X}{m_t} > \frac{Y}{m_f}$. Thus, our goal is to prove $\Pr(\frac{X}{m_t} > \frac{Y}{m_f}) \geq 1 - m_t^{-1}$. We first calculate the upper bound of $\Pr(\frac{X}{m_t} \leq \frac{Y}{m_f})$, the probability that the false rotamer $\mathbf{r_f}$ is chosen. Since event $\{\frac{X}{m_t} \leq \frac{Y}{m_f}\}$ is equivalent to the union of $\{X \leq$

Algorithm 4.1 Hausdorff-based NOE Assignment (HANA)

Given L, Y, backbone, rotamer library. /* L is the assigned resonance list, and Y is the experimental NOESY spectrum. */

Phase 1 (Initial NOE Assignment):

for $i \leftarrow 1$ **to** w **do** /* w is the number of experimental peaks in the NOESY spectrum. */
  $A_i \leftarrow \emptyset$; /* Initialization of NOE assignment for each NOESY peak. */
end for
for $i \leftarrow 1$ **to** w **do**
  for $j \leftarrow 1$ **to** q **do** /* q is the number of protons in the protein. */
    $a'_j \leftarrow$ heavy atom bond-connected to a_j;
    for $k \leftarrow 1$ **to** q **do**
      if $|p_{i1} - \omega(a_j)| < \delta_1$ **and** $|p_{i2} - \omega(a'_j)| < \delta_2$ **and** $|p_{i3} - \omega(a_k)| < \delta_3$ **then**
        $A_i \leftarrow A_i \cup \{(a_j, a'_j, a_k)\}$;
      end if
    end for
  end for
end for

Phase 2 (Rotamer Selection):

for $i \leftarrow 1$ **to** n **do** /* n is the number of residues in the protein. */
  $R_i \leftarrow \emptyset$; /* Initialization for the set of chosen rotamers at residue i. */
  for $j \leftarrow 1$ **to** t **do** /* t is the maximum number of rotamers per residue. */
    $B_{ij} \leftarrow \emptyset$; /* Initialization for the back-computed NOE pattern for rotamer j at residue i. */
    $s_{ij} \leftarrow 0$; /* Initialization for the similarity score of the back-computed NOE pattern B_{ij}. */
  end for
end for
for $i \leftarrow 1$ **to** n **do**
  for $j \leftarrow 1$ **to** t **do**
    structure $\mathbf{P} \leftarrow$ rotate and translate rotamer $\mathbf{r_{ij}}$ into backbone;
    for each proton $a_3 \in \mathbf{r_{ij}}$ **do** /* $\mathbf{r_{ij}}$ is the rotamer j at residue i. */
      for each proton $a_1 \in$ backbone **do**
        $a_2 \leftarrow$ heavy atom bond-connected to a_1;
        if $d(\| a_1 - a_3 \|, \mathbf{P}) < u$ **then**
          /*$d(\| a_1 - a_3 \|, \mathbf{P})$ is the Euclidean dist. betw. protons a_1 and a_3 in $\mathbf{P}$, and u is the NOE upper-bound. */
          $B_{ij} \leftarrow B_{ij} \cup \{(\omega(a_1), (\omega(a_2), (\omega(a_3))\}$
        end if
      end for
    end for
    $s_{ij} \leftarrow$ **Hausdorff_Score**(B_{ij}, Y); /* Compute the similarity score between B_{ij} and Y (see Algorithm 3.1). */
  end for
  sort all rotamers $\{\mathbf{r_{ij}} | j = 1, \ldots, t\}$ in descending order of scores s_{ij};
  $R_i \leftarrow$ top k rotamers in $\{\mathbf{r_{ij}} | j = 1, \ldots, t\}$;
end for

Phase 3 (Filtration of Ambiguous NOE Assignment):

for $i \leftarrow 1$ **to** n **do**
  for each rotamer $\mathbf{r} \in R_i$ **do** /* R_i is the set of chosen rotamers from Phase 2. */
    structure $\mathbf{P} \leftarrow$ rotate and translate $\mathbf{r}$ into backbone
  end for
end for
for $i \leftarrow 1$ **to** w **do**
  for each $(a_1, a_2, a_3) \in A_i$ **do** /* A_i is the set of initial NOE assignments from Phase 1. */
    if $d(\| a_1 - a_3 \|, \mathbf{P}) > u$ **then**
      $A_i = A_i \setminus \{(a_1, a_2, a_3)\}$
    end if
  end for
end for
return $A_1 \cup \ldots \cup A_w$

$i\} \bigwedge \{Y \geq \frac{m_f}{m_t} i\}$ for all $1 \leq i \leq m_t$, that is, $\{\frac{X}{m_t} \leq \frac{Y}{m_f}\} = \bigcup_{i=1}^{m_t} \{X \leq i\} \bigwedge \{Y \geq \frac{m_f}{m_t} i\}$. Thus, we have

$$\Pr(\frac{X}{m_t} \leq \frac{Y}{m_f}) = \sum_{i=1}^{m_t} \big(\Pr(X \leq i) \cdot \Pr(Y \geq \frac{m_f}{m_t} \cdot i) \big). \tag{7}$$

Let $\gamma = \frac{m_t \mu_f + m_f \mu_t}{2 m_f}$. For any $1 \leq i \leq \gamma$, we have

$$\begin{aligned} & \Pr(X \leq i) \cdot \Pr(Y \geq \frac{m_f}{m_t} \cdot i) \\ & \leq \Pr(X \leq i) \qquad (\text{since } \Pr(Y \geq \tfrac{m_f}{m_t} \cdot i) \leq 1) \\ & \leq \Pr(X \leq \gamma). \qquad (\text{since } i \leq \gamma) \end{aligned}$$

Since $m_f \mu_t - m_t \mu_f \geq \max(m_f, \sqrt{m_f m_t}) \cdot 4\sqrt{u_t \ln m_t} > 0$, we have $m_f \mu_t > m_t \mu_f$. Thus, we obtain $0 < \gamma =$

$\frac{m_f\mu_t+m_t\mu_f}{2m_f} < \mu_t$. By Equation (5) from the Chernoff tail bounds (Details of the tail bounds are provided in the optional **Appendix 2**), we have

$$\begin{aligned}\Pr(X \le \gamma) &= \Pr(X \le \frac{m_f\mu_t + m_t\mu_f}{2m_f}) \\ &\le \exp\big(-\frac{(\mu_t - \frac{m_f\mu_t+m_t\mu_f}{2m_f})^2}{2\mu_t}\big) \\ &= \exp\big(-\frac{(m_f\mu_t - m_t\mu_f)^2}{8m_f^2\mu_t}\big).\end{aligned}$$

Since $m_f\mu_t - m_t\mu_f \ge 4m_f\sqrt{\mu_t \ln m_t}$, we have

$$\begin{aligned}\Pr(X \le \gamma) &\le \exp(-\frac{16m_f^2\mu_t \ln m_t}{8m_f^2\mu_t}) \\ &= \exp(-2\ln m_t)\end{aligned}$$

Thus, we have

$$\Pr(X \le \gamma) \le \frac{1}{m_t^2} \tag{8}$$

For any $\gamma < i \le m_t$, we have

$$\begin{aligned}&\Pr(X \le i) \cdot \Pr(Y \ge \frac{m_f}{m_t} \cdot i) \\ &\le \Pr(Y \ge \frac{m_f}{m_t} \cdot i) \qquad \text{(since } \Pr(X \le i) \le 1\text{)} \\ &\le \Pr(Y \ge \frac{m_f}{m_t} \cdot \gamma). \qquad \text{(since } \gamma < i\text{)}\end{aligned}$$

Since $m_f\mu_t > m_t\mu_f$, we have

$$\begin{aligned}\frac{m_f}{m_t} \cdot \gamma &= \frac{m_f}{m_t} \cdot \frac{m_t\mu_f + m_f\mu_t}{2m_f} \\ &= \frac{m_t\mu_f + m_f\mu_t}{2m_t} \\ &> \frac{m_t\mu_f + m_t\mu_f}{2m_t} = \mu_f\end{aligned}$$

Then by Equation (6) from the Chernoff tail bounds (**Appendix 2**), we have

$$\begin{aligned}\Pr(Y \ge \frac{m_f}{m_t} \cdot \gamma) &= \Pr(Y \ge \frac{m_f\mu_t + m_t\mu_f}{2m_t}) \\ &\le \exp\big(-\frac{(\mu_f - \frac{m_f\mu_t+m_t\mu_f}{2m_t})^2}{\mu_f + \frac{m_f\mu_t+m_t\mu_f}{2m_t}}\big) \\ &= \exp\big(-\frac{(m_f\mu_t - m_t\mu_f)^2}{2m_t(m_f\mu_t + 3m_t\mu_f)}\big).\end{aligned}$$

Since $m_f\mu_t - m_t\mu_f > 0$, we have $2m_t(m_f\mu_t + 3m_t\mu_f) < 2m_t(m_f\mu_t + 3m_f\mu_t) = 8m_fm_t\mu_t$. From the condition $m_f\mu_t - m_t\mu_f > 4\sqrt{m_tm_f} \cdot \sqrt{\mu_t \ln m_t}$, we have

$$\begin{aligned}\Pr(Y \ge \frac{m_f}{m_t} \cdot \gamma) &\le \exp(-\frac{16m_tm_f\mu_t \ln m_t}{8m_fm_t\mu_t}) \\ &= \exp(-2\ln m_t)\end{aligned}$$

Thus, we have

$$\Pr(Y \geq \frac{m_f}{m_t} \cdot \gamma) \leq \frac{1}{m_t^2}. \tag{9}$$

By equations (8) and (9), we obtain $\Pr(X \leq i) \cdot \Pr(Y \geq \frac{m_f}{m_t} \cdot i) \leq \frac{1}{m_t^2}$ for any $1 \leq i \leq m_t$. Thus, we have

$$\begin{aligned} \Pr(\frac{X}{m_t} \leq \frac{Y}{m_f}) &= \sum_{i=1}^{m} \Pr(X \leq i) \cdot \Pr(Y \geq \frac{m_f}{m_t} \cdot i) \\ &\leq \sum_{i=1}^{m_t} \frac{1}{m_t^2} = \frac{1}{m_t}. \end{aligned}$$

Therefore, $\Pr(\frac{X}{m_t} > \frac{Y}{m_f}) \geq 1 - m_t^{-1}$. ■

6. Time Complexity Analysis

We will analyze the time complexity of our NOE assignment algorithm HANA (Algorithm 4.1). We first restate Theorem 5.3 and then provide the proof.

Theorem 5.3. HANA *runs in $O(tn^3 + tn \log t)$ time, where t is the maximum number of rotamers at a residue and n is the total number of residues in the protein sequence.*

Proof: To analyze the algorithmic complexity of our NOE assignment algorithm, we first recall some notations defined previously. Let n be the number of residues in the protein sequence, and let w denote the total number of cross peaks in the experimental NOESY data. Let t denote the maximum number of rotamers for every amino acid in the rotamer library. Let ξ denote the maximum number of atoms per residue. Let q be the total number of atoms in the protein, then $q = O(\xi n)$.

The running time of the initial NOE assignment phase is bounded by $O(wq^2)$ steps. In Phase 2, the initialization in lines $1-7$ takes $O(tn)$ time. Since the number of protons in the backbone is bounded by $O(n)$, the total number of protons in a rotamer is less than ξ, the loop in lines $11-19$ needs $O(\xi n)$ steps. The function **Hausdorff_Score** takes $O(mw)$ time to compute the similarity score between the back-computed NOE pattern B_{ij} and the experimental NOESY spectrum Y, where m is the number of back-computed NOE peaks in B_{ij}. Hence, the loop in lines $9-21$ runs in $O\big(t(n\xi + mw)\big)$ time. Sorting all rotamers and selecting top k rotamers in lines $22-23$ only requires $O(t \log t)$ time. Thus, the overall running time for Phase 2 is $O(tn) + n \cdot O\big(t(mw + \xi n)\big) + n \cdot O(t \log t) = O\big(tn(mw + \xi n) + tn \log t\big)$. In Phase 3 (namely the filtration of ambiguous NOE assignment), placing all rotamers into the backbone (in lines $1-5$) takes $O(kn)$ time. In worst case, $|A_i|$ is bounded by $O(q^2)$, where q is the total number of atoms in the proteins. Hence the total running time for lines $6-12$ is $O(wq^2)$. Thus, Phase 3 runs in $O(kn + wq^2)$ time. Therefore, the overall running time for HANA is $O(wq^2) + O\big(tn(mw + \xi n) + tn \log t\big) + O(kn + wq^2) = O\big(wq^2 + tn(mw + \xi n) + tn \log t\big)$.

In general, it is safe to assume the number of atoms in a residue is a constant, that is, $\xi = O(1)$. Thus, $q = O(\xi n) = O(n)$. Also, since each proton can only have NOE interactions with a constant number of other protons within 6.0 Å distance, we have $w = O(n)$ and $m = O(n)$. Therefore, the running time of HANA is $O(tn^3 + tn \log t)$ in the worst case. ■

References

1. D. Angluin and L. G. Valiant. Fast probabilistic algorithms for hamiltonian circuits and matchings. In *Proceedings of the ninth annual ACM Symposium on Theory of Computing*, 1979.
2. A. T. Brünger. X-PLOR, Version 3.1: a system for X-ray crystallography and NMR. *Journal of the American Chemical Society*, 1992.
3. A. A. Canutescu and R. L. Dunbrack Jr. Cyclic coordinate descent: A robotics algorithm for protein loop closure. *Protein Science*, 12:963–972, 2003.

4. P. Güntert. Automated NMR protein structure calculation with CYANA. *Meth. Mol. Biol.*, 278:353–378, 2004.
5. R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.
6. A. Saupe. Recent results in the field of liquid crystals. *Angew. Chem.*, 7:97–112, 1968.
7. A. Shehu, C. Clementi, and L. E. Kavraki. Modeling protein conformational ensembles: from missing loops to equilibrium fluctuations. *Proteins: Structure, Function, and Bioinformatics*, 65(1):164–79, 2006.
8. N. R. Skrynnikov and L. E. Kay. Assessment of molecular structure using frame-independent orientational restraints derived from residual dipolar couplings. *J. Biomol. NMR*, 18(3):239–252, 2000.
9. 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.
10. 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.
11. 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.
12. L. Wang and B. R. Donald. An Efficient and Accurate Algorithm for Assigning Nuclear Overhauser Effect Restraints Using a Rotamer Library Ensemble and Residual Dipolar Couplings. *The IEEE Computational Systems Bioinformatics Conference (CSB), Stanford CA. (August, 2005)*, pages 189–202, 2005.
13. 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.
14. W. J. Wedemeyer, C. A. Rohl, and H. A. Scheraga. Exact solutions for chemical bond orientations from residual dipolar couplings. *J. Biomol. NMR*, 22:137–151, 2002.
15. J. Zeng, C. Tripathy, P. Zhou, and B. R. Donald. A Hausdorff-Based NOE Assignment Algorithm Using Protein Backbone Determined from Residual Dipolar Couplings and Rotamer Patterns. In *Proceedings of the 7th Annual International Conference on Computational Systems Bioinformatics (CSB), Stanford CA*, In Press. (2008).