A Markov Random Field Framework for Protein Side-Chain Resonance Assignment[⋆] (Supporting Material)

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.

Below is supplementary material for the following paper:

- J. Zeng, P. Zhou, and B. R. Donald. "A Markov Random Field Framework for Protein Side-Chain Resonance Assignment." *Proceedings of the 14th Annual International Conference on Research in Computational Molecular (RECOMB)*, Lisbon, Portugal. In Press. (2010).

Appendix

The following is an appendix which provides additional information to substantiate the claims of the paper [26]. **Appendix 1** briefly reviews the high-resolution protein backbone determination from residual dipolar coupling data. **Appendix 2** provides the proof of Claim 1. **Appendix 3** describes the NMR experimental procedures and the results of backbone structure calculation from RDCs. **Appendix 4** describes how to compute NOE distance restraints using the side-chain resonance assignments computed by our algorithm.

1 Backbone Structure Determination from Residual Dipolar Couplings

Residual dipolar couplings (RDCs) provide global orientational restraints on the internuclear bond vectors with respect to an external magnetic field [20, 19], and have been used to determine protein backbone conformations [6, 20, 7, 17, 15, 18, 6, 22, 23, 16]. We applied our recently-developed algorithms [22, 23, 25, 6] to compute the backbone structures using two RDCs per residue (either NH RDCs measured in two media, or NH and CH RDCs measured in a single medium) and sparse NOE distance restraints. In our backbone determination, we first computed conformations and orientations of secondary structure element (SSE) backbones from RDC data using the RDC-EXACT algorithm [22, 23, 6]. Instead of randomly sampling the entire conformation space to find solutions consistent with the experimental data, RDC-EXACT computes the backbone dihedral angles exactly by solving a system of quartic monomial equations derived from the RDC equations [22, 23, 6]. A depth-first search strategy is applied to search systematically over all roots of a system of low-degree (quartic) equations, and find a globally optimal solution for each SSE fragment. These RDC-defined SSE backbone fragments are then assembled using a sparse set of inter-SSE NOE distance restraints [22, 23, 25]. A methyl-protonated specific isotopic labelling strategy [8, 21] is used to obtain these sparse inter-SSE NOE distance restraints, which involve only amide and methyl protons from isoleucine, leucine and valine (ILV) residues. More details on backbone structure determination using RDCs can be found in [6, 22, 23, 25]. Note that alternatively the global fold (i.e., backbone) could, in principle, be computed by other approaches, such as protein structure prediction [1], protein threading [24] or homology modeling [10, 11].

[⋆] 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.

2 Proof of Claim 1

Recall that an estimated cost function is *admissible*, if it does not overestimate the cost from every node to the goal node. An A* search algorithm is guaranteed to find the optimal solution if the heuristic cost function is admissible. In this section, we give the details of the proof for Claim 1. We first restate the claim and then provide the proof.

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

Proof. Let $(u_1^*, \cdots, u_t^*)$ be the optimal solution to our side-chain resonance assignment problem, where u_i^* is the assignment of resonance node r_i. Suppose that the A* algorithm has assigned the first $i-1$ resonances, and reached the i^{th} depth of the search tree. Let h^* be the cost from the current node to the goal node in the optimal solution $(u_1^*, \cdots, u_t^*)$. By Eq. (16) and Eq. (18) of the main article, we have

$$\begin{aligned} h &= -\ln \Pr(X_t|X_{t-1}, \cdots, X_1, H, Q) \cdots \Pr(X_{i+1}|X_i, \cdots, X_1, H, Q) && (1) \\ &= -\ln \left(\max_{\substack{u_j \in A(r_j) \\ \cdots \\ u_{i+1} \in A(r_{i+1})}} \Pr\left(\gamma(r_j, u_j)|\gamma(r_{j-1}, u_{j-1}), \cdots, \gamma(r_{i+1}, u_{i+1}), X_i, \cdots, X_1, H, Q\right) \right) && (2) \\ &\leq -\ln \Pr\left(\gamma(r_j, u_j^*)|\gamma(r_{j-1}, u_{j-1}^*), \cdots, \gamma(r_{i+1}, u_{i+1}^*), X_i, \cdots, X_1, H, Q\right). && (3) \end{aligned}$$

Since $h^* = -\ln \Pr(\gamma(r_j, u_j^*)|\gamma(r_{j-1}, u_{j-1}^*), \cdots, \gamma(r_{i+1}, u_{i+1}^*), X_i, \cdots, X_1, H, Q)$, we have

$$h \leq h^*.$$

Thus, the estimated cost function defined in Eq. (18) of the main article never overestimates the cost from the current node to the goal node, and hence is admissible. Since the conformation search space in our formulation is a tree, our A* algorithm is guaranteed to find the optimal solution given the admissible estimated cost function (i.e., Eq. (18) of the main article). □

3 NMR Experimental Procedures and Backbone Structure Determination Results

All NMR data except the RDC data of ubiquitin and GB1 were recorded and collected using Varian 600 and 800 MHz spectrometers at Duke University. The NMR spectra were processed using the program NMRPipe [5]. All NMR peaks were picked by the programs NMRView [9] or Xeasy/Cara [3], followed by manual editing. Backbone assignments, including resonance assignments of atoms N, HN, C$^\alpha$, H$^\alpha$, C$^\beta$, were obtained from the set of triple resonance NMR experiments HNCA, HN(CO)CA, HN(CA)CB, HN(COCA)CB, and HNCO, combined with the HSQC spectra using the program Paces [4], followed by manual checking. The NOE cross peaks were picked from three-dimensional ^{15}N- and ^{13}C-edited NOESY-HSQC spectra. In addition, we removed the diagonal cross peaks and water artifacts from the picked NOE peak list. The NH and CH RDC data of FF2, pol η UBZ and hSRI were measured from a 2D ^{1}H-^{15}N IPAP experiment [13] and a modified (HACACO)NH experimental [2] respectively. The C$^\alpha$C$'$ and NC$'$ RDC data of FF2 were measured from a set of HNCO-based experiments [14]. The CH and NH RDC data of ubiquitin were obtained from the Protein Data Bank (PDB ID of ubiquitin: 1D3Z). For GB1, we computed its global fold using the CH and NH RDC data from a homologous protein, namely the third IgG-binding domain of Protein G (GB3) (PDB ID: 1P7E).

The list of unassigned side-chain resonances were extracted from 3D NOESY spectra by projecting all 3D NOE cross peaks into the plane of the first and second dimensions (i.e., the dimensions of the first proton and its bond-connected heavy atom). We used the Penultimate rotamer library [12]. We first applied our recently-developed algorithm RDC-PANDA [22, 23, 25] with $3-15$ inter-SSE NOEs between amide and methyl protons of $2-6$ ILV residues as input to compute the global backbone fold from RDCs. The backbone RMSD between the computed backbone and reference structures is less than 1.3±0.6 Å, and RMSD between experimental and back-calculated RDCs for the RDC-defined backbone is 1.1±0.9 Hz for CH RDCs and 1.2±1.1 Hz for NH RDCs. These RDC-defined structures are only medium-resolution and do not contain side-chain conformations. As we demonstrated in Sec. 3 of the main article, these RDC-defined backbones provide sufficient structural information for side-chain resonance assignment. The set of distance restraints derived from side-chain resonances assigned by our algorithm enables high-resolution structure determination that both computes the accurate side-chain conformations and improve the RDC-defined backbone conformations (Table 3 of the main article).

4 Computing NOE Distance Restraints Using Assigned Side-Chain Resonances

The side-chain resonance assignments computed by our algorithm enable an NOE assignment procedure based on the NOESY graph (described in Sec. 2.2 of the main article) in our MRF framework. After applying our algorithm (described in Sec. 2.4 and Sec. 2.5 of the main article) to obtain the set of optimal side-chain resonance assignments, we used the following procedure to compute the NOE distance restraints. We first extracted a set of initial NOE assignments from the edges E in the NOESY graph, using the known backbone resonance assignments and the side-chain resonance assignments computed by our algorithm. Such a set of NOE assignments may contain noisy (i.e., spurious) NOE assignments due to experimental noise or chemical shift overlap. For each possible NOE assignment, we checked whether the distance between the coordinates of assigned side-chain proton labels in the rotamers (after being placed on the backbone) violates the NOE distance bound. An NOE assignment was pruned when the Euclidean distance between the coordinates of a pair of assigned proton labels was larger than the NOE distance calibrated from NOE peak intensity. The set of remaining NOE assignments were output for final structure determination. Note that after pruning the violated NOE assignments, two NOE restraints can still be assigned to the same NOESY peak. In this situation, these two NOEs are unified by the logical "OR" operation when being used in structure calculation.

References

1. D. Baker and A. Sali. Protein structure prediction and structural genomics. *Science*, 294:93–96, 2001.
2. 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.
3. C. Bartels, T. Xia, M. Billeter, P. Güntert, and K. Wüthrich. The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. *Journal of Biomolecular NMR*, 6:1–10, 1995.
4. B. E. Coggins and P. Zhou. PACES: Protein sequential assignment by computer-assisted exhaustive search. *Journal of Biomolecular NMR*, 26:93–111, 2003.
5. F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, and A. Bax. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *Jour. Biomolecular NMR*, 6:277–293, 1995.
6. 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.
7. 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.
8. N.K. Goto, K.H. Gardner, G.A. Mueller, R.C. Willis, and L.E. Kay. A robust and cost-effective method for the production of Val, Leu, Ile (δ_1) methyl-protonated ^{15}N-, ^{13}C-, ^{2}H-labeled proteins. *J. Biomol. NMR*, 13:369–374, 1999.
9. B. A. Johnson and R. A. Blevins. NMRView: a computer program for the visualization and analysis of NMR data. *Jour. Biomolecular NMR*, 4:603–614, 1994.
10. 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.

11. 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.
12. S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson. The Penultimate Rotamer Library. *Proteins: Structure Function and Genetics*, 40:389–408, 2000.
13. 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.
14. P. Permi, P. R. Rosevear, and A. Annila. A set of HNCO-based experiments for measurement of residual dipolar couplings in ^{15}N, ^{13}C, (^{2}H)-labeled proteins. *Journal of Biomolecular NMR*, 17:43–54, 2000.
15. J. H. Prestegard, C. M. Bougault, and A. I. Kishore. Residual Dipolar Couplings in Structure Determination of Biomolecules. *Chemical Reviews*, 104:3519–3540, 2004.
16. C. A. Rohl and D. Baker. De Novo Determination of Protein Backbone Structure from Residual Dipolar Couplings Using Rosetta. *J. Am. Chem. Soc.*, 124:2723 –2729, 2002.
17. 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.
18. F. Tian, H. Valafar, and J. H. Prestegard. A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones. *J Am Chem Soc.*, 123:11791–11796, 2001.
19. 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.
20. 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.
21. V. Tugarinov, V. Kanelis, and L. E. Kay. Isotope labeling strategies for the study of high-molecular-weight proteins by solution NMR spectroscopy. *Nat Protoc.*, 1:749–754, 2006.
22. 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.
23. 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.
24. Y. Xu, D. Xu, and E. C. Uberbacher. An efficient computational method for globally optimal threading. *J Comput Biol.*, 5(3):597–614, 1998.
25. 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.
26. J. Zeng, P. Zhou, and B. R. Donald. A Markov Random Field Framework for Protein Side-Chain Resonance Assignment. In *Proceedings of the 14th Annual International Conference on Research in Computational Molecular (RECOMB), Lisbon, Portugal*, 2010.