Algorithms and Analytic Solutions using Sparse Residual Dipolar Couplings for High-Resolution Automated Protein Backbone Structure Determination by NMR

Anna Yershova[1], Chittaranjan Tripathy[1], Pei Zhou[2], Bruce R. Donald[1,2]

[1]Department of Computer Science, Duke University

[2] Department of Biochemistry, Duke University Medical Center

Abstract

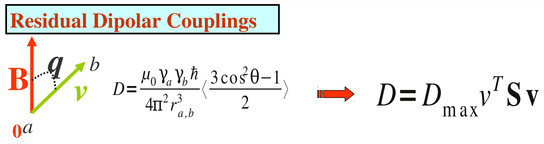
In traditional protein structure determination from solution Nuclear Magnetic Resonance (NMR) data, the nuclear Overhauser effect (NOE) distance restraints are primarily used in a simulated annealing/molecular dynamics (SA/MD) protocol to compute the protein structure. It is only in the final stages of structure computation, the orientational restraints such as residual dipolar couplings (RDCs) and residual chemical shift anisotropies (RCSAs) are incorporated to refine the structure. While SA/MD protocols may perform adequately in a data-rich setting, it is difficult to determine protein structures accurately using only sparse data, since SA/MD protocols provide no guarantee on the uniqueness or global optimality of the solutions, and can get trapped in local minima. Sparse data arises not only in high-throughput settings, but also for larger proteins, membrane proteins and symmetric protein complexes. Sparse-data algorithms require guarantees of completeness and correctness to ensure that solutions are not missed and local minima are evaded.

Here we present algorithms that use global orientational restraints from a minimal amount of RDC (and RCSA) data and a sparse set of NOEs to compute high-resolution protein backbone fold in linear time. Our methods perform simultaneous structure determination and packing of beta-sheets using RDCs and NOEs that involve only amide and methyl protons from ^{1}H-^{13}C-ILV methyl-labeled proteins. Results from beta-sheet computation for human ubiquitin show that our algorithm can compute beta sheets with backbone RMSD < 0.86 Å from X-ray reference structure (PDB Id: 1UBQ). Our tests on different combinations of RDCs for human ubiquitin and Z-Domain of Staphylococcal Protein A (SpA) demonstrate the ability of our algorithms to compute the global backbone fold accurately from sparse data. These results show that our structure determination algorithms and software can be successfully applied to compute the high-resolution protein backbone from sparse NMR data, which can be used to do automated NOE assignments [2].

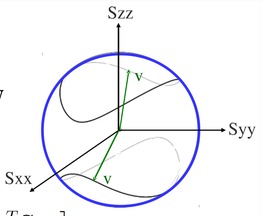
Our methods determine the solution structures by explicitly computing the intersections of algebraic curves representing discrete RDC values. However, if additional RDC data can be measured, the algebraic curves no longer generically intersect. Our new algorithms address this situation and show that globally optimal structures can still be computed analytically as points closest to all of the algebraic curves representing the RDCs. We present new algorithms that expand the types and number of RDCs from which analytic solutions are computed. We evaluate the performance of our algorithms on real experimental NMR data for four proteins: human ubiquitin, DNA-damageinducible protein I (DinI), the Z domain of staphylococcal protein A (SpA) and GB1. The results show that our algorithms are able to successfully determine high-resolution backbone structures from a limited amount of NMR data.

RDCs and CSAs

Residual Dipolar Couplings

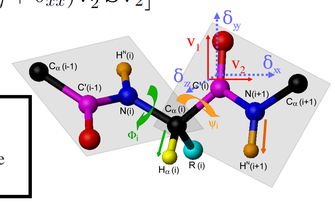


$$D = \frac{\mu_0}{4\pi^2}\frac{\gamma_a \gamma_b \hbar}{r_{a,b}^3}\left\langle \frac{3\cos^2\theta - 1}{2} \right\rangle \Rightarrow D = D_{\max} v^T \mathbf{S} v$$



Carbonyl Residual Chemical Shift Anisotropies

$$\Delta c = \frac{1}{3} D_{\max}\left[(2\delta_{xx} + \delta_{yy})\mathbf{v}_1^T \mathbf{S}\mathbf{v}_1 + (2\delta_{yy} + \delta_{xx})\mathbf{v}_2^T \mathbf{S}\mathbf{v}_2\right]$$
$$= \lambda_1 \mathbf{v}_1^T \mathbf{S}\mathbf{v}_1 + \lambda_2 \mathbf{v}_2^T \mathbf{S}\mathbf{v}_2$$



RDCs and CSAs provide global orientational restraints on inter-nuclear bond vectors and peptide planes. An RDC, for example, constrains the bond vector to lie on the two discrete algebraic curves carved on a unit sphere.

The NH (or CH) bond vector orientations from NH (or CH) RDCs in one aligning medium can be solved from a system of quartic equations obtained for the backbone dihedral φ-ψ angles from the RDC equations [1, 2]:

$$F(u) = L_0 + L_1 u + L_2 u^2 + L_3 u^3 + L_4 u^4 = 0$$

The NH (or CH) bond vector orientations from NH (or CH) RDCs in two different aligning media can be solved by univariate function minimization. The backbone φ-ψ angles can be computed by minimizing univariate RDC-RMSD fit function [3]:

$$\sigma(u) = \sqrt{\frac{1}{2}\sum_{k=1}^{2}(F_k(u))^2}$$

High-Resolution Backbone from Sparse Data

Schematic of RDCAnalytic Engine

Φ-defining	Ψ-defining
C^{α}-C'	N-H^N
C^{α}-C^{β}	C'-N
C^{α}-H^{α}	C-H^N
	C'-RCSA
	^{15}N-RCSA
	^{1}H-RCSA

RDCs in One Medium
One Φ-defining RDC/RCSA and one Ψ-defining RDC/RCSA per residue

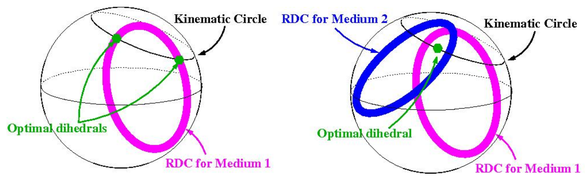
RDC-RCSA Data | Sparse NOE Table | SSE Boundary | Rotamer Library → RDCAnalytic

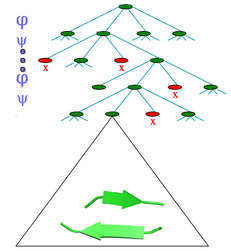
RDCs in Two Media
Two or more Φ-defining RDC/RCSA and two or more Ψ-defining RDC/RCSA per residue

RDC-RCSA Data | SSE Boundary → RDCAnalytic

RDCAnalytic Algorithm

1. BRANCH:
   solutions to the RDC equations and minimization problem for each φ-ψ
2. BOUND:
   Ramachandran filter, RDC-RMSD cutoff, steric clashes





Simultaneous Structure Determination and Packing of Beta Strands

1. Let E = {E_1, …, E_n} be a beta sheet specification, where E_i are strand specs
2. Initialize the partial sheet S ← Φ
3. Compute E_1 using RDCAnalytic
4. S ← S U { E_1 }
5. ∀ E_i ∈ E, 2 ≤ i ≤ n
6. ∀ E_{leaf} , a solution at the leaf node of RDCAnalytic φ-ψ solution tree
7. Pack(S, E_{leaf} , T_{nr}) and test if it scores better than the previously packed sheet
8. Let $E_{starBest}$ be the strand with best packing and RDC satisfaction score
9. S ← S U { $E_{starBest}$ }
10. Return S

A linear combination of following scores is used to measure packing quality:
- Over all RDC rmsd
- Rms deviation from ideal strand geometry
- Score from hydrogen bond formation
- Score from steric checker
- NOE rmsd

Conclusions and Future Work

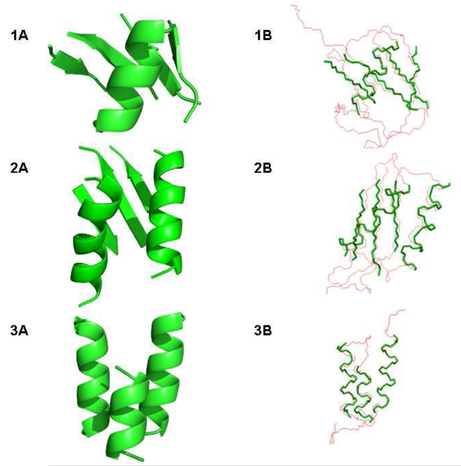
Our algorithms for structure determination from sparse NMR data have provable guarantees on correctness, completeness, accuracy and time complexity of the obtained structures with respect to the RDC RMSD function, and require less data compared to other algorithms in the literature. We have been working on the following extensions:

1. Computing structures of large proteins using RDCs and RCSAs and sparse (ambiguous) NOEs.
2. Extension of RDCAnalytic to incorporate the analytic solutions that has been derived for three planar RDCs per residue and compute the backbone apiece peptide plane wise.
3. Extension of RDCAnalytic to incorporate the analytic solutions from other combinations of RDC and RCSA data.
4. Algorithms to compute side-chain conformations from RDCs.

Funding

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

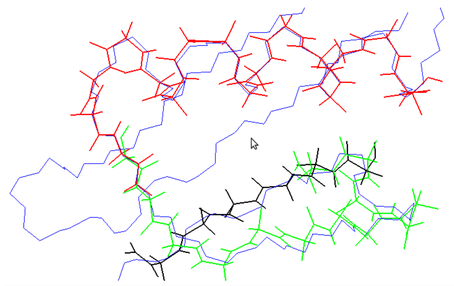
Results: RDCs in One Medium



Left: The global folds of ubiquitin (1A), DinI (2A) and SpA (3A), computed by RDCANALYTIC, using Ca -C′ and N-HN RDCs and a sparse set of NOEs. For ubiquitin and DinI experimental Ca -C′ and N-HN RDCs are used. For SpA simulated Ca -C′ and N-HN RDCs are used.
Right: (1B) Overlay of the ubiquitin global fold (green) computed by RDC-ANALYTIC with the X-ray structure (red). The backbone RMSD is 1.17 Å. (2B) Overlay of the global fold of DinI (green) computed by RDC-ANALYTIC with the Model 1 (red) of the reference structure (PDB id: 1ghh). The backbone RMSD is 1.11 Å. (3B) The global fold of SpA computed by RDC-ANALYTIC is overlaid on the Model 1(red) of the reference structure (PDB id: 1q2n). The backbone RMSD is 0.83 Å.

Protein	RDCs[a] used & RMSD (Hz)	Alignment Tensor (S_{yy}, S_{zz})	Backbone RMSD (Å) vs. X-ray/NMR structure
Ubiquitin α:23-33	C^{α}-H^{α}: 0.817, N-H^N: 0.227 C^{α}-C′: 0.036, N-H^N: 0.243 C^{α}-C′: 0.033, C′-N: 0.026	14.195, 25.213 15.002, 24.771 15.485, 25.566	0.429 0.430 0.239
Ubiquitin[c,d] α:23-33 β:2-7, 12-17, 41-45, 65-70	C^{α}-H^{α}: 1.11, N-H^N: 0.740 C^{α}-C′: 0.129, N-H^N: 0.603	15.230, 24.657 14.219, 25.490	1.276 1.172
DinI[c,d] α:18-32,58-72 β:2-8, 39-44, 49-53	C^{α}-C′: 0.483, N-H^N: 1.203	10.347, 33.459	1.111
SpA[d] α:8-17, 24-36, 41-54	(run1) C^{α}-H^{α}: 0.458,N-H^N: 2.11 (run2) C^{α}-H^{α}: 0.678,N-H^N: 0.543 [b](run3)C^{α}-C′: 1.237,N-H^N: 1.049	8.008, 23.063 8.146, 24.261 7.676, 22.961	1.063 1.577 0.834

Results: RDCs in Two Media



Overlay of the gb1 global fold (red, green, and black) computed by RDC-ANALYTIC with the NMR structure PDB id:1P7E (blue). Several portions of the protein were computed in a divide-and-conquer fashion. The backbone RMSD of the portions are the following:
p1:23-29(red) - 0.35 Å,
p2:39-51(green) - 0.49 Å,
p2:51-55(black) - 0.54 Å.

Protein	RDCs[a] used & RMSD (Hz)	Alignment Tensor (S_{yy}, S_{zz})	Backbone RMSD (Å) vs. X-ray/NMR structure
Ubiquitin α:25-31	2xC^{α}-H^{α}: 0.93, 2xN-H^N: 0.32	16.9, 23.2; 7.0, 52.4	0.403
loop:54-58	2xC^{α}-H^{α}: 2.2, 2xN-H^N: 0.7	16.9, 23.2; 7.0, 52.4	0.409
loop:59-64	2xC^{α}-H^{α}: 1.9, 2xN-H^N: 1.2	16.9, 23.2; 7.0, 52.4	0.652
loop/β:64-70	2xC^{α}-H^{α}: 3.1, 2xN-H^N: 1.2	16.9, 23.2; 7.0, 52.4	0.49
β:2-7	2xC^{α}-H^{α}: 2.6, 2xN-H^N: 1.4	16.9, 23.2; 7.0, 52.4	0.64
β:11-17	2xC^{α}-H^{α}: 2.6, 2xN-H^N: 1.5	16.9, 23.2; 7.0, 52.4	0.50
β:41-45	2xC^{α}-H^{α}: 2.2, 2xN-H^N: 0.8	16.9, 23.2; 7.0, 52.4	0.44

References

[1] L. Wang and B. R. Donald. *J. Biomol. NMR*, 29(3):223–242, 2004.
[2] J. Zeng, J. Boyles, C. Tripathy, L. Wang, A. Yan, P. Zhou, and B. R. Donald *J. Biomol. NMR*, [Epub ahead of print] PMID:19711185, 2009.
[3] A. Yershova, C. Tripathy, Zhou, and B. R. Donald, Algorithms and Analytic Solutions using Sparse Residual Dipolar Couplings for High-Resolution Automated Protein Backbone Structure Determination by NMR, accepted to WAFR 2010.