Protein Structure Determination using Sparse Orientational Restraints from NMR Data

Chittaranjan Tripathy[1], Anna Yershova[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

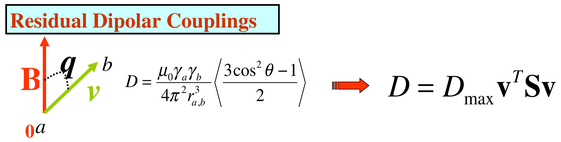
In traditional protein structure determination from solution Nuclear Magnetic Resonance (NMR) data, the nuclear Overhauser effect (NOE) distance restraints are primarily used in a simulated annealing/molecular dynamics (SA/MD) protocol to compute the protein structure. It is only in the final stages 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 minimal amount of RDC (and RCSA) data and a sparse set of NOEs to compute high-resolution protein backbone fold in linear time. We have also developed new algorithms to do 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 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].

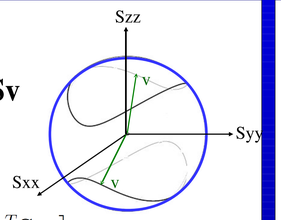
We are also developing an algorithm for structure determination of the backbone secondary structure elements, loops, and Saupe matrix elements using sparse RDC data and minimal a priori modeling. The data used in this method are NH and CH RDCs in two alignment media. The method is based on the exact computation of all of the local minima of the function determined over the structure parameters and Saupe matrix elements. The minima computed by the algorithm correspond to the structures minimizing the fit to the RDC data. We show some preliminary results obtained using the algorithm. The preliminary results demonstrate that very little NMR data contains enough information to perform truly de novo protein structure determination. Computing protein loops and the Saupe matrices is a new and attractive benefit over all of the previous approaches.

RDCs and CSAs

Residual Dipolar Couplings

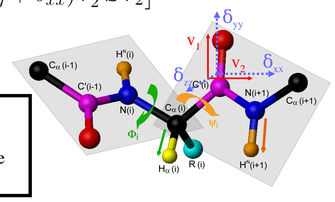


$$D = \frac{\mu_0 \gamma_a \gamma_b}{4\pi^2 r_{a,b}^3}\left\langle \frac{3\cos^2\theta - 1}{2} \right\rangle \Rightarrow D = D_{\max}\mathbf{v}^T\mathbf{S}\mathbf{v}$$



Carbonyl Residual Chemical Shift Anisotropies

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

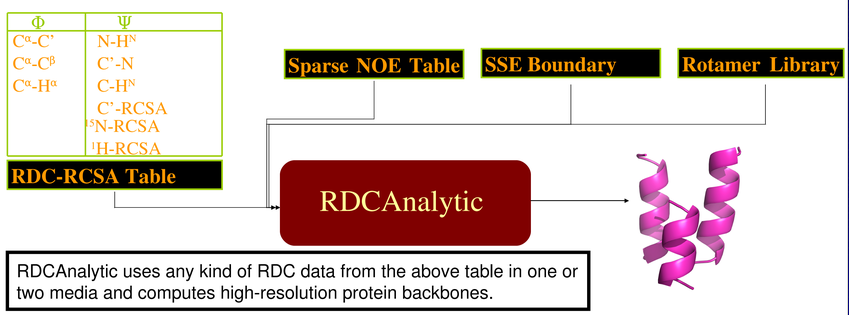


RDCs and CSAs provide global orientational restraints on inter-nuclear bond vectors and peptide planes. An RDC, for example, constrains the bond vector to lie on the two discrete algebraic curves carved on a unit sphere.

The NH (or CH) bond vector orientations from NH (or CH) RDCs in two different aligning media can be solved from a system of quartic equations. The backbone φ-ψ angles can be computed from simple quadratic equations [1, 2]. NH and CH RDCs in one medium have been used in a recent work (Zeng *et al,* 2009, in [2]) in our lab to determine high-resolution protein structure starting with a global fold calculated from exact solutions to the RDC equations. The structure of FF Domain 2 of human transcription elongation factor CA150 (RNA polymerase II C-terminal domain interacting protein) solved using the methodology in [2] has been deposited to the Protein Data Bank (PDB). The PDB Id is 2KIQ.

High-Resolution Backbone from Sparse Data

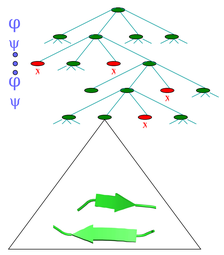
Schematic of RDCAnalytic Engine



RDCAnalytic uses any kind of RDC data from the above table in one or two media and computes high-resolution protein backbones.

Simultaneous Structure Determination and Packing of Beta Strands

1. Let $E = \{E_1, \ldots, E_n\}$ be a beta sheet specification, where E_i are strand specs
2. Initialize the partial sheet $S \leftarrow \Phi$
3. Compute E_1 using RDCAnalytic
4. $S \leftarrow S \cup \{E_1\}$
5. $\forall E_i \in E, 2 \le i \le n$
6. $\quad \forall E_{i\text{-leaf}}$, a solution at the leaf node of RDCAnalytic φ-ψ solution tree
7. $\quad$ Pack($S, E_{i\text{-leaf}}, T_{noe}$) and test if it scores better than the previously packed sheet
8. $\quad$ Let $E_{i\text{-leaf-best}}$ be the strand with best packing and RDC satisfaction score
9. $\quad S \leftarrow S \cup \{E_{i\text{-leaf-best}}\}$



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

During tree-search bad conformations are pruned by Ramachandran and real-solution filters.

Backbone Structure From Sparse RDC Restraints and Minimal a Priori Modeling

RDC Fit Function

$$(D_{max}v^{\top}Sv - D)^2$$

Single term squared

$$(D_{max}(S_{xx}v_0^2 + 2S_{xy}v_0v_1 + S_{yy}v_1^2 + 2S_{xz}v_0v_2 + 2S_{yz}v_1v_2 + S_{zz}v_2^2) - D)^2$$

Each term is a linear function of Saupe matrix elements

$$(D_{max}(-\cos\phi\sin\phi S_{xz}v_0^2 + \sin^2\phi S_{zz}v_0^2 - \cos\phi\sin\phi S_{xz}v_0^2 + \ldots) - D)^2$$

Each term is a quartic in terms of the tangent of the half angle; v' is the previous from v bond vector along the backbone with the RDC measurement

RDC Data Requirements

We require that each of the dihedral angles (φ ,ψ) is involved in at least two of the terms of the RDC fit function. For example, two RDC measurements per each NH and CH vector in the backbone would satisfy such a requirement.

Benefits and Guarantees

1. The RDC fit function satisfying the above requirements has only a finite number of minima. This means that only a finite number of different structures can satisfy the RDC data.
2. The structures can be obtained with minimal amount of modeling. Computation of beta-strands and loops is possible with no additional data.
3. Saupe matrices for each of the alignment media can be computed.

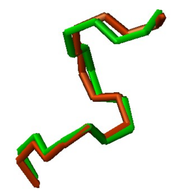
The Algorithm

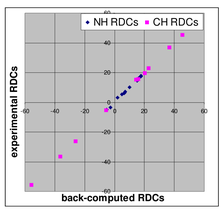
For a portion of the protein with the required amount of data do the following computations:

1. **Initialize the structure.** Take the first approximation for the structure of that portion as an ideal helix.
1. **Compute the alignment tensors.** Convert RDC fit function to a linear system of equations for the alignment tensor elements and solve it using SVD.
2. **Compute the dihedral angles.** Minimize each of the RMSD terms as a univariate function.
3. **Iterate.** Go to step (3). Repeat this process until convergence is detected.

Preliminary Results

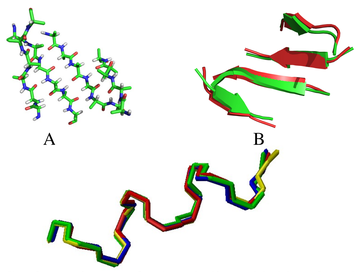
Conformation of the portion [25-31] of the helix for human ubiquitin computed using NH and CH RDCs in two media (red) has been superimposed on the same portion from high-resolution X-ray structure (PDB Id: 1UBQ) (green). The backbone RMSD is 0.58 Å.





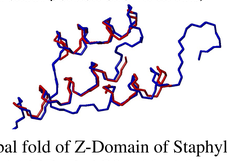
Protein	RMSD (Hz)	Alignment Tensor (S_{yy}, S_{zz})
Ubq α :25-31	C^{α} H^{α} : 0.32 NH: 0.24	(23.66, 16.48) (53.25, 7.65)

Results: Backbone Computation from Sparse Data



(A) Structure of ubiquitin beta sheet computed by our new packing algorithm using CH and NH RDCs in one medium and NOEs (assigned) of the type expected from ^{1}H-^{13}C-ILV methyl-labeled proteins. The NOEs used involve amide and methyl protons only. (B) The computed beta sheet (green) has backbone rmsd 0.864 Å wrt. the high-resolution X-ray structure (PDB id: 1UBQ) (beta sheet portion shown in red).

Conformations of the helix [23-33] for human Ubiquitin computed from different sets of RDCs. Blue backbone is from 1UBQ. Green backbone uses CH and NH RDCs, yellow backbone uses CAC and NH RDCs, and red backbone uses CAC and CN RDCs. The results are shown in the table below. Ubiquitin NMR structure has PDB id 1D3Z.



Global fold of Z-Domain of Staphylococcal Protein A (SpA) (PDB ID: 1Q2N, blue) from simulated CAC and NH RDCs (brown). The backbone rmsd with respect to the Model 1 of 1Q2N is 0.87 Å. Red is The table below shows the results of different runs with both experimental and simulated RDCs.

Protein	RDC used & RMSD (Hz)	Alignment Tensor (S_{yy}, S_{zz})	Backbone RMSD (Å) vs. X-ray structure	Backbone RMSD (Å) vs. NMR structure
(Ubq α:23-33)	C^{α}-H^{α}: 0.817, N-H^{N}: 0.227	14.195, 25.213	0.429	0.428
	C^{α}-C′: 0.036, N-H^{N}: 0.243	15.002, 24.771	0.430	0.393
	C^{α}-C′: 0.033, C′-N: 0.026	15.485, 25.566	0.239	0.317
SpA	(exp rdc run1) C^{α}-H^{α}: 0.458, N-H^{N}: 2.11	8.008, 23.063	N/A	1.117
	(exp rdc run2) C^{α}-H^{α}: 0.678, N-H^{N}: 0.543	8.146, 24.261	N/A	1.630
	(sim rdc) C^{α}-C′: 1.237, N-H^{N}: 1.049	7.676, 22.961	N/A	0.869

Conclusions and Future Work

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

- Computing structures of large protein complexes using RDCs and RCSAs and sparse (ambiguous) NOEs.
- 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.
- Extension of beta sheet computation to compute more curvy strands.
- Algorithms to compute side-chain conformations from RDCs.

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.

Funding

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