An Exact Algorithm For Determining Protein Backbone Structure From NH Residual Dipolar Couplings

Lincong Wang[*] **Ramgopal R. Mettu**[*] **Ryan Lilien**[*,†] **Bruce Randall Donald**[*,‡,§,¶,‖]

Abstract

We have developed a novel algorithm for protein backbone structure determination using global orientational restraints on internuclear bond vectors derived from residual dipolar couplings (RDCs) measured in solution NMR. The algorithm is a depth-first search (DFS) strategy that is built upon two low-degree polynomial equations for computing the backbone (ϕ, ψ) angles, exactly and in constant time, from two bond vectors in consecutive peptide planes.

1. Introduction

With the availability of gene sequence data for hundreds of organisms it becomes critical to develop efficient algorithms to compute protein structures as accurately as possible using only very sparse constraints. One way to achieve this is to develop algorithms whose key components are analytic expressions computable *in constant time*. Here we present an algorithm for determining a protein backbone structure from global angular (orientational) restraints for internuclear vectors derived from backbone RDCs measured in two aligning media by solution nuclear magnetic resonance (NMR) spectroscopy. NMR spectroscopy, capable of deriving various geometric restraints, has become a major experimental tool for structural genomics. However, the traditional nuclear Overhauser effect (NOE) based NMR methods for determining structures require months of time to record spectra and to assign a large number of NOE distance restraints, in particular those restraints involving sidechain protons. In contrast, global angular restraints, derived from the recently developed NMR experiments for measuring RDCs in weakly aligned proteins, can be assigned much faster.

Recently algorithms have been developed for computing a protein fold using RDCs alone or with RDCs aided by other restraints such as chemical shifts or sparse NOEs [1, 2, 3]. However, these algorithms either require several sets of RDCs and an ^{15}N,^{13}C doubly-labeled sample, or rely heavily on molecular dynamics or statistics from the Protein Databank (PDB). By comparison, our algorithm requires less data (only backbone NH RDCs in two media and very sparse distance restraints), depends much less on statistics from the PDB (only the averages for the backbone (ϕ, ψ) angles) and does not rely on molecular dynamics.

2. Algorithm

Previously, there were no known analytic expressions to compute either internuclear vectors or backbone (ϕ, ψ) angles from two sets of RDCs. Hence, previous protocols were forced to rely on grid searches such as one dimensional grid search for NH directions [5] or two-dimensional grid search for (ϕ, ψ) angles [2, 3]. Analytic expressions become crucial for developing fast algorithms to determine structures precisely with RDCs. We have shown that the direction of an internuclear vector can be computed by solving a quartic equation. Therefore, analytic expressions for exact solutions always exist, and in contrast to the grid search method [5] NH directions can be computed *exactly* and *in constant time*, i.e., with no dependence upon grid resolution. Furthermore, we have derived two quadratic equations to compute backbone (ϕ, ψ) angles from two consecutive NH vectors, *exactly* and *in constant time*, again in contrast to the previous grid search methods [2, 3].

[*]Dartmouth Computer Science Department, Hanover, NH 03755, USA.

[†]Dartmouth Medical School, Hanover, NH 03755

[‡]Dartmouth Chemistry Department, Hanover, NH 03755, USA.

[§]Dartmouth Biological Sciences Department, Hanover, NH 03755, USA.

[¶]Dartmouth Center for Structural Biology and Computational Chemistry, Hanover, NH 03755, USA.

[‖]Corresponding author: 6211 Sudikoff Laboratory, Dartmouth Computer Science Department, Hanover, NH 03755, USA. Phone: 603-646-3173. Fax: 603-646-1672. Email: `brd@cs.dartmouth.edu`

This work is supported by the following grants to B.R.D.: National Institutes of Health (R01 GM 65982), National Science Foundation (IIS-9906790, EIA-0102710, EIA-0102712, EIA-9818299, and EIA-9802068), and the John Simon Guggenheim Foundation.

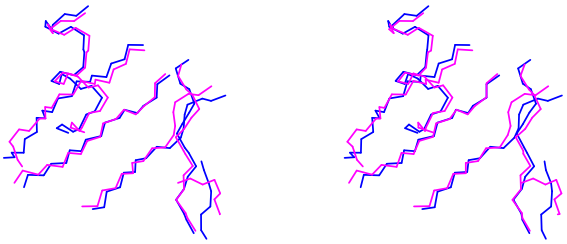

Figure 1. The RDC-derived structure (black) and X-ray structure [4] (gray) of ubiquitin have an RMSD of 1.21 Å.

Our algorithm for computing backbone structures combines these analytic solutions with a depth-first search (DFS) strategy. Briefly, the algorithm first solves the quartic equation to obtain every possible NH vector solution of an m residue structural fragment after the alignment tensors are computed using an identified α-helix. Gaussian errors are added to the experimental RDC values to ensure that solutions for every NH vector exist. Next, it searches over the cross-product of all NH vectors to find a set of feasible conformation vectors. A *feasible* conformation vector is defined as $(\phi_1, \psi_1, \phi_2, \psi_2, \cdots, \phi_{m-1}, \psi_{m-1})$ with every (ϕ_i, ψ_i) $(1 \leq i < m)$ in the favorable Ramachandran region. We note that each element in the cross-product can simply be viewed as a path in a tree. We perform DFS on this tree employing a pruning strategy based on limiting (ϕ, ψ) angles to the Ramachandran region appropriate for the given secondary structure type. From the set of feasible conformation vectors, our algorithm outputs the vector that simultaneously maximizes structural agreement with experimental RDCs and minimizes deviation from standard secondary structure geometry. Finally, the oriented secondary structure elements are positioned with a limited number of hydrogen bonds and NOE distance restraints to compute the backbone fold.

3. Results and Analysis

In practice, our implementation runs quickly and produces accurate structure; we are able to generate a backbone structure of human ubiquitin with an RMSD of 1.21 Å vs. the X-ray structure (see Figure 1). Compared with other RDC-based fold determination algorithms, our exact method uses fewer restraints but achieves similar or better accuracy. We give empirical evidence that under a Gaussian error model for RDCs our algorithm has an expected running time that is considerably lower than the worst-case exponential running time.

The running time of our algorithm was about 45 minutes to compute the entire backbone structure of the 39-residue portion of ubiquitin (one α-helix and five β-strands). The average number of samples from the Gaussian error distribution needed such that the quartic equation has real solutions for every residue of a fragment is rather small: ranging from only a few hundreds to a few thousands. Similarly, sampling roughly 2000 points in each error distribution is sufficient to generate conformations with pairwise backbone RMSD less than 0.50 Å for both the helix and each of the five strands. Despite the worst-case exponential running time, the search for an optimal conformational vector takes, in practice, only several minutes for either the helix or any of the five strands. We observed that a vast majority of the paths in the search tree were pruned after examining just a few residues. This is likely due to the fact that solutions (ϕ_i, ψ_i) for the residue i of a fragment depend not only on the two RDC values for i but also on $(\phi_1, \psi_1, \cdots, \phi_{i-1}, \psi_{i-1})$ of all the previous residues. Thus, the probability that a conformation vector is pruned increases rapidly as its search path increases in length.

In summary, we believe our algorithm is general enough to be applied to determine the folds of many proteins since a large proportion of protein backbones contain regular α-helices and β-sheets. Our method can also be extended to compute other dihedral bond angles based on RDCs from different internuclear vectors and hence is quite general. Our algorithm can be adapted to compute the orientations of sidechains from their RDC data alone, which will be helpful for computing, on a genomic scale, the complete structures of large proteins by NMR.

References

[1] A. C. Fowler, F. Tian, H. M. Al-Hashimi, and J. H. Prestegard. Rapid determination of protein folds using residual dipolar couplings. *J. Mol. Biol.*, 304(3):447–460, 2000.

[2] A. W. Giesen, S. W. Homans, and J. M. Brown. Determination of protein global folds using backbone residual dipolar coupling and long-range NOE restraints. *J. Biomol. NMR*, 25:63–71, 2003.

[3] 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(47):11791–11796, 2001.

[4] S. Vijay-Kumar, C. E. Bugg, and W. J. Cook. Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.*, 194:531–544, 1987.

[5] 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.