Progress in Nuclear Magnetic Resonance Spectroscopy 55 (2009) 101–127

Contents lists available at ScienceDirect

Progress in Nuclear Magnetic Resonance Spectroscopy

journal homepage: www.elsevier.com/locate/pnmrs

Automated NMR assignment and protein structure determination using sparse dipolar coupling constraints

Bruce R. Donald *, Jeffrey Martin

Departments of Computer Science and Biochemistry, Duke University, D101 LSRC, Research Drive, Durham, NC 27708-0129, USA

ARTICLE INFO

Article history:
Received 25 September 2008
Accepted 15 December 2008
Available online 23 January 2009

Contents

1. Introduction

The introduction of residual dipolar couplings (RDCs) for protein structure determination over 10 years ago has energized development of NMR methods. Robust automation of the complete NMR structure determination procedure has been a long-standing goal, and RDC-based algorithms may increase the consistency and reliability of NMR structural studies. It has also been recognized that structure determination based primarily on orientational restraints could be quicker and more accurate than traditional distance-restraint methods. Furthermore, NMR is increasingly important in applications where structural information is already available, so that methods which effectively automate NMR assignment of known structures would also be a substantial contribution.

Since RDCs are measured in a global coordinate frame, they enable molecular replacement-like methods that perform assignments using structural priors. Furthermore, recent methods for structure determination have exploited novel RDC equations, which combine RDC data and protein kinematics. Under fairly mild assumptions, the dihedral torsional angles of a protein can be analytically expressed as roots of these low-degree monomials. Solving these equations exactly has enabled a departure from earlier stochastic methods, and led to linear-time, combinatorially-precise

Abbreviations: NMR, Nuclear Magnetic Resonance; ppm, parts per million; RMSD, mean square deviation; HSQC, heteronuclear single quantum coherence spectroscopy; NOE, Nuclear Overhauser Effect; RDC, residual dipolar coupling; PDB, Protein Data Bank; pol η, zinc finger domain of the human DNA Y-polymerase Eta; CH, $C^{\alpha}-H^{\alpha}$; hSRI, human Set2-Rpb1 interacting domain; FF2, FF Domain 2 of human transcription elongation factor CA150 (RNA polymerase II C-terminal domain interacting protein); POF, principal order frame; SA, simulated annealing; MD, molecular dynamics; SSE, secondary structure element; C′, carbonyl carbon; WPS, well-packed satisfying; vdW, van der Waals; DOF, degrees of freedom.

* Corresponding author. Tel.: +1 919 660 6583.
E-mail address: brd+pnmrs09@cs.duke.edu (B.R. Donald).
URL: http://www.cs.duke.edu/brd (B.R. Donald).

0079-6565/$ - see front matter Published by Elsevier B.V.
doi:10.1016/j.pnmrs.2008.12.001

algorithms for NMR structure determination. These algorithms are optimal in terms of combinatorial (but not algebraic) complexity, and show how structural data can be used to produce a deterministic, optimal solution for the protein structure in polynomial time.

The coefficients of the RDC equations are determined by the data. An RDC error bound therefore defines a range of coefficients, which, in turn, yield a range of roots representing the structural dihedral angles. Hence, the RDC equations define an analytical relationship between the RDC error distribution, and the coordinate error of the ensemble of structures that satisfy the experimental restraints. Precise methods that relate the experimental error to the coordinate error of the computed structures therefore appear within reach. This article reviews these and other recent advances in NMR assignment and structure determination based on sparse dipolar couplings.

Color versions of the figures in this paper are available online at http://www.sciencedirect.com/science/journal/00796565.

1.1. Background

While automation is revolutionizing many aspects of biology, the determination of three-dimensional (3D) protein structure remains a harder, more expensive task. Novel algorithms and computational methods in biomolecular NMR are necessary to apply modern techniques such as structure-based drug design and structural proteomics on a much larger scale. Traditional (semi-) automated approaches to protein structure determination through NMR spectroscopy require a large number of experiments and substantial spectrometer time, making them difficult to fully automate. A chief bottleneck in the determination of 3D protein structures by NMR is the assignment of chemical shifts and nuclear Overhauser effect (NOE) restraints in a biopolymer.

The introduction of residual dipolar couplings (RDCs) for protein structure determination enabled novel attacks on the assignment problem, to enable high-throughput NMR structure determination. Similarly, it is difficult to determine protein structures accurately using only *sparse data*. New algorithms have been developed to handle the increased spectral complexity encountered for larger proteins, and sparser information content obtained either in a high-throughput setting, or for larger or difficult proteins. The overall goal is to minimize the number and types of NMR experiments that must be performed and the amount of human effort required to interpret the experimental results, while still producing an accurate analysis of the protein structure.

This review is tempered by our recent experiences in automated assignments [79,82,83,118,153,174], novel algorithms for protein structure determination [152,156,117,89,110,151,155,154], characterization of protein complexes [118,99] and membrane proteins [117], and fold recognition using only unassigned NMR data [82,83,78,80]. Recent algorithms for automated assignment and structure determination based on sparse dipolar couplings represent a departure from the stochastic methods frequently employed by the NMR community (e.g., simulated annealing/molecular dynamics (SA/MD), Monte Carlo (MC), etc.) A corollary is that such stochastic methods, now routinely employed in NMR structure determination pipelines [60,53,91,64], should be reconsidered in light of their inability to assure identification of the unique or globally-optimal structural models consistent with a set of NMR observations. In this vein, our review focuses on *sparse data.* While SA/MD may perform adequately in a data-rich, highly-constrained setting, it is difficult to determine protein structures accurately using only sparse data. Sparse data arises not only in high-throughput settings, but also for larger proteins, membrane proteins [117], symmetric protein complexes [118], and difficult systems including denatured or disordered proteins [154]. Sparse-data algorithms require *guarantees of completeness* to ensure that solutions are not missed and local minima are evaded.

We caution that in the context of NMR, "high-throughput" is relative, and currently not as rapid as, for example, gene sequencing or even crystallography. Hence the term "batch mode" may be more appropriate. The challenge is to develop new algorithms and computer systems to exploit sparse NMR data, demonstrating the large amount of information available in a few key spectra, and how it can be extracted using a blend of combinatorial and geometric algorithms. Moreover, because of their (relative) experimental simplicity, we hypothesize that the computational advantages offered by such approaches should ultimately obtain an integrated system in which automated assignment and calculation of the global fold could be performed at rates comparable to current-day protein screening for structural genomics using ^{15}N-edited heteronuclear single quantum coherence spectroscopy (^{15}N-HSQC).

This article reviews how sparse dipolar couplings can be exploited to address key computational bottlenecks in NMR structural biology. The past few years have yielded rapid progress in automated assignments, novel algorithms for protein structure determination, characterization of protein complexes and membrane proteins, and fold recognition using only unassigned NMR data. We review recent algorithms that assist these advances, including: (1) *Sparse-data algorithms for protein structure determination from residual dipolar couplings (RDCs)* using exact solutions and systematic search; (2) RDC-based molecular replacement-like techniques for structure-based assignment; (3) *Structure determination of membrane proteins and complexes*, especially symmetric oligomers, enabled by RDCs; and (4) *Automated assignment of NOE restraints* in both monomers and complexes, based on backbones computed primarily using sparse RDC restraints.

These define the four main themes in our review:

(1) It is difficult to determine protein structures accurately using only *sparse data.* Sparse data arises not only in high-throughput settings, but also for larger proteins, membrane proteins, and symmetric protein complexes. For *de novo* structure determination, there are now roots-of-polynomials approaches to compute exact solutions, by systematic search, for internuclear bond vectors and backbone dihedral angles using as few as 2 recorded RDCs per residue (for example NH in two media, or NH and H^{α}—C^{α} in one medium). By combining systematic search with exact solutions, it is possible to efficiently compute accurate backbone structures using less NMR data than in traditional approaches.

De novo structure determination from sparse dipolar couplings can exploit structure equations derived by Wang and Donald [152,151]. These include a quartic equation to compute the internuclear (e.g., bond) vectors from as few as 2 recorded RDCs per residue, and quadratic equations to subsequently compute protein backbone (ϕ, ψ) angles *exactly* [152,151]. The structure equations make it possible to compute, exactly and in constant time, the backbone (ϕ, ψ) angles for a residue from very sparse RDCs. Simulated annealing, molecular dynamics, energy minimization, and distance geometry are not required, since the structure is computed exactly from the data. Novel algorithms build upon these exact solutions, to perform protein structure determination, using mostly RDCs but also sparse NOEs. For example, the RDC-EXACT algorithm employs a systematic search with provable pruning, to determine the conformation of helices, strands, and loops and to compute their orientations using exclusively the angular restraints from RDCs [152,156]. Then, the algorithm uses very sparse distance restraints between these computed segments of structure, to determine the global fold.

(2) Algorithms using sparse dipolar couplings can accelerate protein NMR assignment and structure determination by exploiting *a priori* structural information. By analogy, in X-ray crystallography, the molecular replacement (MR) technique allows solution of the crystallographic phase problem when a "close" or homologous structural model is known, thereby facilitating rapid structure