Robust, Integrated Computational Control of NMR Experiments to Achieve Optimal Assignment by ADAPT-NMR

Arash Bahrami1, Marco Tonelli1, Sarata C. Sahu2, Kiran K. Singarapu1, Hamid R. Eghbalnia3, John L. Markley1,2*

1 National Magnetic Resonance Facility at Madison, Biochemistry Department, University of Wisconsin - Madison, Madison, Wisconsin, United States of America, 2 Center for Eukaryotic Structural Genomics, University of Wisconsin-Madison, Madison, Wisconsin, United States of America, 3 Department of Molecular and Cellular Physiology, University of Cincinnati, Cincinnati, Ohio, United States of America

Abstract

ADAPT-NMR (Assignment-directed Data collection Algorithm utilizing a Probabilistic Toolkit in NMR) represents a groundbreaking prototype for automated protein structure determination by nuclear magnetic resonance (NMR) spectroscopy. With a [13C,15N]-labeled protein sample loaded into the NMR spectrometer, ADAPT-NMR delivers complete backbone resonance assignments and secondary structure in an optimal fashion without human intervention. ADAPT-NMR achieves this by implementing a strategy in which the goal of optimal assignment in each step determines the subsequent step by analyzing the current sum of available data. ADAPT-NMR is the first iterative and fully automated approach designed specifically for the optimal assignment of proteins with fast data collection as a byproduct of this goal. ADAPT-NMR evaluates the current spectral information, and uses a goal-directed objective function to select the optimal next data collection step(s) and then directs the NMR spectrometer to collect the selected data set. ADAPT-NMR extracts peak positions from the newly collected data and uses this information in updating the analysis resonance assignments and secondary structure. The goal-directed objective function then defines the next data collection step. The procedure continues until the collected data support comprehensive peak identification, resonance assignments at the desired level of completeness, and protein secondary structure. We present test cases in which ADAPT-NMR achieved results in two days or less that would have taken two months or more by manual approaches.

Citation: Bahrami A, Tonelli M, Sahu SC, Singarapu KK, Eghbalnia HR, et al. (2012) Robust, Integrated Computational Control of NMR Experiments to Achieve Optimal Assignment by ADAPT-NMR. PLoS ONE 7(3): e33173. doi:10.1371/journal.pone.0033173

Editor: Annalisa Pastore, National Institute for Medical Research, Medical Research Council, London, United Kingdom

Received October 6, 2011; Accepted February 8, 2012; Published March 12, 2012

Copyright: © 2012 Bahrami et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Funding provided by National Institutes of Health, National Center for Research Resources: P41 RR02301 and 3P41/RR02301-2551 National Institutes of Health, National Institute of General Medical Sciences: 1U54 GM74501. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: markley@nmrfam.wisc.edu

Introduction

A goal-directed experimental strategy can be defined as one that optimizes each new experimental step by analyzing the current sum of available results with the aim of achieving a particular goal. In principle, such a strategy should be superior to ones in which data are collected first, either by conventional or fast methods, and analyzed later. The idea of focusing on an end goal to guide data collection could be broadly applicable to many domains of investigation. We describe here a proof of concept implementation of this strategy to the collection and analysis of protein NMR data with the goal of achieving complete resonance assignments of the type required for automated structure determination. Our approach, ADAPT-NMR (Assignment-directed Data collection Algorithm utilizing a Probabilistic Toolkit in NMR), successfully navigates a large set of experimental options on the basis of iterative analysis of the current data and achieves efficient and complete assignments and secondary structure determination.

The initial stage in solution-state NMR spectroscopy of proteins concerns the production of labeled molecules and the identification of suitable solution conditions for data collection. These steps are analogous to the production of protein and suitably diffracting crystals for X-ray crystallography. Whereas, with crystallography, the subsequent data collection and analysis steps leading to structure determination are fairly standardized and automated, this is not yet the case with protein NMR spectroscopy. Typically, several multinuclear, multidimensional NMR data sets are collected and subsequently analyzed in separate steps leading to a structure (Figure 1). Each of these steps has been automated to some extent [1,2,3,4,5,6,7]; and in some cases, multiple steps have been pipelined to work sequentially [8,9]. However, a software pipeline is not adept at emulating the iterative nature of structure determinations performed by human experts. As a result, manual intervention in data analysis continues to be one of the main bottlenecks in structure determination by NMR. ADAPT-NMR is the first iterative and fully automated approach designed specifically for the optimal assignment of proteins. The increased efficiency results from more rapid data collection, data analysis, and data verification.
Figure 1. Conventional steps in protein structure determination by solution state NMR spectroscopy.
doi:10.1371/journal.pone.0033173.g001

Methods

Approach

ADAPT-NMR builds a fully probabilistic, yet computationally tractable, network capable of dynamically representing interrelationships among assigned attributes. Our development of this goal-directed approach required a fundamental reevaluation of our prior paradigm for NMR data collection and analysis [10,11]. ADAPT-NMR uses a probabilistic goal-seeking approach in which a set of cooperating probabilistic sub-networks, each implementing their own computational model, drive toward the optimal fashion by controlling the flow of experiments (Figure 2).

The goal-seeking tasks select two parameters at each iteration: a) the 3D NMR (1H, 13C, and 15N) experiment to be conducted in the next step, and b) the subspace (tilt angle that combines 13C and 15N frequencies to reduce data collection from 3D to 2D) used in data collection. The subspace can be considered as the projection of 3D spectra into 2D tilted planes. The cooperating sub-networks enforce three key conditions designed to enhance the stability and reliability of the global network: a) probabilistic representation, b) concise probabilistic communication, and c) simplified domain decomposition. The choice of these conditions was motivated by practical experience with Bayesian updating approaches. Condition a) requires that each subnet must represent its state by an ensemble of probabilistic variables. Condition b) requires that subnets communicate with one another by means of probability distributions over all the variables they share, and Condition c) maintains network robustness (See Text S1).

The architecture of the network integrates an extensive pseudo-energetic model (analogous to those used in biophysics and statistical mechanics) for each subnet by using a number of mathematical and machine learning [10,12] techniques. With each step in data collection and analysis, the network evolves toward an ensemble of states – a configuration – that represents the assignment of NMR resonances (chemical shifts) and secondary structural elements to groups in the covalent structure of the protein.

The pseudo-energetic system is represented by a canonical ensemble, in which the probability of each configuration $p_i$ (corresponding to microscopic state $\mathbf{s}$) is given by the Boltzmann distribution

$$p_i = \frac{1}{Z} e^{-\beta E_S}, \quad Z = \sum_i e^{-\beta E_S}$$

where $\beta$ resembles the thermodynamic variable (determined empirically), and $\mathbf{s}$ is the canonical partition function. $E_S$ the energy of microstate $s$, is the sum of individual and interaction potentials. In our model, individual potentials represent prior information, such as NMR chemical shift distributions and prior probabilities for the most recent model for peak lists, resonance assignments, and secondary structure probabilities. The interaction potentials, on the other hand, represent constraints and the consistency of variable choices. The total energy of the network in ADAPT-NMR is written as a sum of single, pairwise, and triple-wise interactions among network microstates:

$$E_S = \sum_i U_i(z_i(v_i)) + \sum_{ij} U_{ij}(z_i(v_i),z_j(v_j))$$

$$+ \sum_{ijk} U_{ijk}(z_i(v_i),z_j(v_j),z_k(v_k))$$

where $z_i(v)$ represents the state of the probabilistic variable $v$, $U_i$ represent individual potentials, and $U_{ij}$ and $U_{ijk}$, respectively, represent pair-wise and triple-wise interaction potentials. The use of triple-wise interactions is unique to our definition of the statistical model. In ADAPT-NMR, rather than seeking a single solution, which would necessitate the identification of the unique configuration that minimizes the total energy, we determine marginal probabilities for every probabilistic variable.

Approximation of the ground state, a state where probabilities are effectively stationary, relies on the implementation of algorithms of the kind used in graphical models [13,14,15]. We use the factor graph representation [16] because it is computationally efficient and it has been applied successfully in diverse areas of information technology [17]. In the factor graph representation [18], our pseudo-energetic model transforms to factorized local functions in which the probability for a configuration $s$ can be described as:

$$p_s = \prod_{i,j} \delta_{ij}(z_i(v_i),z_j(v_j)) \cdot \prod_{i,j,k} \delta_{ijk}(z_i(v_i),z_j(v_j),z_k(v_k))$$

where the factors $\delta_{ij}$, $\delta_{jk}$, and $\delta_{ijk}$ can be calculated from pseudo-energy terms as: $\delta_{ij} = e^{-\beta U_{ij}(z_i(v_i),z_j(v_j))}$, $\delta_{ij} = e^{-\beta U_{ij}(z_i(v_i),z_j(v_j))}$, and $\delta_{ijk} = e^{-\beta U_{ijk}(z_i(v_i),z_j(v_j),z_k(v_k))}$.
However, the resulting computational structure does not guarantee convergence to the ground state [13,14,15]. To overcome this, we have developed specialized algorithms in ADAPT-NMR. Detailed description of the ideas and algorithms can be found in the Text S1.

Algorithm
The algorithm is briefly summarized here (see the Text S1 for complete details). The ADAPT-NMR iteration starts with the amino acid sequence of the protein and 2D NMR data sets $^1$H-$^15$N($^15$N-HSQC) and $^1$H-$^13$C. These data sets serve as “orthogonal projections” of conventional 3D NMR spectra. Then ADAPT-NMR applies an advanced automated peak identification algorithm, and generates probabilistic “spin systems”. A spin system is defined as a group of peaks that are most likely belong to the same amino acid in the protein sequence. At this level, on-the-fly evaluations of the spin systems determine which experiment and projection (tilted plane) should be collected next. The iteration continues until the spin system quality is good enough for initial calculations of sequence-specific resonance assignments and secondary structure. Thereafter, an extended network, which takes into consideration spin systems, chemical shift assignments, and secondary structure, selects the next experiment and tilted plane. The iteration continues until the desired completeness of chemical shift assignments is achieved (Figure 2).

Spectral Acquisition. The optimum spectra (as determined in the optimization step) are collected by ADAPT-NMR, and are classified as $S_{\gamma,p}$ where $\gamma$ is the experiment identifier (for example HNCA) and $\gamma$ is the tilt angle (projection angle). Tilted angle spectra are generally collected in pairs ($S_{\gamma,j}$ and $S_{\gamma,-j}$).

Spectral Processing. The key derived measure in this step is the “conclusive probability” for each identified spectral peak, which is defined as the probability that a peak represents a real peak as opposed to an artifact or noise peak. ADAPT-NMR imports the most recently collected spectral data, co-registers all peaks by aligning all spectra, and peak picks spectra by an algorithm that assigns a probability to each peak on the basis of the noise level, peak intensity, the number of the residues in the protein, and the experiment type. Every 2D peak maintains a set of specific attributes (or properties), e.g., frequency coordinates, intensity, volume, possible back-projected 3D peak candidate, and priority weight.

ADAPT-NMR generates a candidate 3D peak with numerous attributes from every pair of peaks present in the orthogonal planes that have a common $^1$H chemical shift (within a tolerance). The 3D peak lists are updated after each step of data collection.

Spin System Generation and Update. ADAPT-NMR applies the pseudo-energetic model presented above and an iterative update algorithm to derive probabilistic spin systems from available peak lists. Spin system objects are initialized from $^15$N-HSQC peaks and have multiple attributes and properties, including eight fields that represent the chemical shifts of different classes of nuclei: $^1$H$_{\gamma,i}$, $^1$H$_{\gamma,i}$, $^1$C$_{\alpha,i}$, $^1$C$_{\beta,i}$, $^1$C$_{\gamma,i}$, $^1$H$_{\alpha,j}$, $^1$N$_{\alpha,j}$, $^1$C$_{\alpha,j}$, $^1$C$_{\beta,j}$, and $^1$C$_{\gamma,j}$, where $(i-j)$ denotes the chemical shift of the previous residue. Each field is a probabilistic variable that might have multiple chemical shift choices. The chemical shift choices and their probabilities are calculated in the probabilistic network on the basis of 3D peak attributes. A “null” state for matching is provided in order to represent the probability that no chemical shift in the data could be matched with the field. Null is a possible state for almost every probabilistic variable in ADAPT-NMR.

All attributes of spin systems are updated after each round of iteration and data collection. New spin systems are added if high probability peaks cannot be associated with any $^15$N-HSQC peaks. An important attribute of spin systems is “the probability of overlap”. In overlapped spectral regions, multiple spin systems may originate from a single $^15$N-HSQC peak. A probabilistic
The NMR data collection is continuously evaluated. The quality of spin systems is lower if the quality of the spectra is the recommended procedure. The overall quality of the spectra is the recommended procedure. The complete analysis of 2D planes, 3D peaks, and the expected number of tilted planes for the various 3D spectra are then collected, processed and analyzed on-the-fly and without any manual intervention. The output from the analysis module is automatically fed back into the spectrometer to direct the latest status of the project. These features enable “restarts”, for example in cases where the process has been interrupted for unexpected reasons, or where data collection by specific experiments or specific angles is impractical. The default initial settings of parameters for data collection and analysis have been optimized through testing on a number of proteins. The initial values are dynamically optimized by the algorithm during the data collection process. For example, the noise threshold level for the peak picking algorithm is modified on the basis of the threshold level used for the previous plane and the expected number of peaks. ADAPT-NMR gives users the ability to manually revise all parameters used in data collection and analysis.

**Results and Discussion**

We have implemented the ADAPT-NMR algorithm on Varian and Bruker spectrometers at the National Magnetic Resonance
Facility at Madison (NMRFAM). We evaluated its performance (accuracy of assignment, accuracy of the secondary structure prediction, and the total time of data collection and data analysis) on six proteins labeled uniformly with $^{13}$C and $^{15}$N that have been studied in our laboratory (Table 1). The tilted angles and experiments selected on-the-fly by ADAPT-NMR for data collection and other experimental details are provided in Table S1.

For proteins ranging from 54 residues to 109 residues, the total time for data collection (for orthogonal and tilted planes), assignment, and secondary structure designation ranged from 15 hours to 55 hours (Table 1) – a significant reduction in time and effort. No manual intervention was required, and the quality of assignment exceeded that normally required for structure calculation. To evaluate accuracy, we compared the ADAPT-NMR results with chemical shift assignments achieved by separate approaches and the secondary structure from coordinates deposited in PDB, when available. The assignments deposited in BMRB associated with a PDB structure have benefited from the later steps of structure calculation (e.g. NOE assignments). In the case of four proteins (ubiquitin (human), RI-brazzein, HSP12, and AeSCP2), we carefully peak picked and assigned the traditional 3D spectra manually; in all cases the level of completeness achieved by ADAPT-NMR equaled that of the manual backbone assignment. ADAPT-NMR yielded higher accuracy of assignment in less than 17691).

Table S1 (Figure S1) (PDB accession code 2LE4) was determined solely on the basis of the ADAPT-NMR assignments (BMRB accession code 17691).

The fact that ADAPT-NMR has access to the actual spectra and can dynamically adjust peak picking so as to optimize the assignment of spin systems gives it an advantage over automated assignment tools that deal with peak lists or spin systems. ADAPT-NMR represents a major step toward a fully automated approach for protein structure determination by NMR. Although the ADAPT-NMR algorithm has been described here as sequential, it is important to note that the implementation of the algorithm executes the data collection and data analysis steps in parallel so that subsequent steps, including assignment, do not have to wait for the data collection to be completed.

The study of aggregated, disordered, and unstable proteins has been consistently a challenge in NMR spectroscopy. Fast data collection by ADAPT-NMR might be particularly helpful in certain unstable samples (for example, samples that are stable for one or two days). Furthermore, “auto-adjustments” have been designed in the ADAPT-NMR algorithm to manage mild aggregation or proteins with small disordered regions. Examples of these adjustments include spin system splitting and iterative peak analysis as described in the algorithm section. However, in the case of severely aggregated proteins or spectra with large exchange broadening, reduced dimensionality methods like ADAPT-NMR are not generally desirable. The recommended manual screening of orthogonal planes prior to launching ADAPT-NMR serves to detect these instances. In addition, various quality measurements are executed during the data collection, and the process will stop if they do not satisfy some minimum thresholds.

A visualization tool being developed for ADAPT-NMR has a user interface that permits manual data analysis (for example, editing of the peaks picked) either on-the-fly or as a post-processing step. After each manual change, ADAPT-NMR updates the probabilistic network, and adjusts the outputs accordingly. We expect that this visualization tool will be particularly helpful with larger or disordered proteins that prove not to be amenable to the fully-automated data collection and analysis approach.

ADAPT-NMR is readily extensible, and we plan to develop versions that include other steps of structure determination. ADAPT-NMR currently accepts side chain peak lists as an optional input and provides full side chain assignment (Table S3). However the goal is to include on-the-fly data collection and analysis of side chain and NOE data. It will be relatively easy to collect less crowded side chain experiments such as HBHA-CO(NH) and C(CO)NH by the reduced dimensionality method; however, more complicated spectra (e.g., HCCCH-TOCSY) normally are not amenable to reduced dimensionality collection. The on-the-fly algorithm can be programmed to decide whether data should be collected by full 3D or reduced dimensionality. The addition of these data types, particularly 3D $^{13}$N- and $^{13}$C-NOESY, should enable ADAPT-NMR to handle larger proteins. Such extensions are achievable, because each sub-network performs the inference task separately. ADAPT-NMR can be readily integrated in an iterative fashion with structure calculation programs such as CYANA [1] or CS-Rosetta [23].

### Supporting Information

Text S1 It discusses the robustness of the ADAPT-NMR approach and contains a complete description of the algorithm. (DOC)

Figure S1 NMR solution structure of human SOX2(39–118). This structure has been deposited in the Protein Data Bank

### Table 1. Results from ADAPT-NMR data collection and backbone analysis of six proteins.

| Protein name | Amino acid residues | Time for data collection and analysis | Completeness of chemical shift assignments | Accuracy of chemical shift assignments | Accuracy of secondary structure predictions | wwPDB and/or BMRB deposition [reference] |
|--------------|---------------------|--------------------------------------|--------------------------------------------|----------------------------------------|--------------------------------------------|------------------------------------------|
| Brazzein (RI) | 54                  | 17 h                                 | 98%                                       | 100%                                   | 100%                                       | 2KGQ, 5296 [24]                          |
| Ubiquitin (human) | 76                 | 13 h                                 | 97%                                       | 100%                                   | 100%                                       | 17730 (a)                                |
| Ubiquitin (Chlorella) | 76               | 15 h                                 | 100%                                      | 100%                                   | 100%                                       | 17730 (a)                                |
| SOX2 (39–118) | 81                  | 55 h                                 | 98%                                       | 100%                                   | 100%                                       | 2LE4, 17691 (a)                          |
| AeSCP2 (complex with palmitate) | 106         | 39 h                                 | 98%                                       | 100%                                   | 100%                                       | 2KSL, 16665 [20]                         |
| HSP12 (intrinsically disordered) | 109          | 17 h                                 | 99%                                       | 100%                                   | 100%                                       | 17483[21]                                |

*This work.

doi:10.1371/journal.pone.0033173.t001
other proteins have been published. The ordered region (residues 7–67 of the domain; 45–105 in the SOX2 numbering system) has a backbone RMSD of 0.74 Å. (b) Ribbon diagram of the ordered region (residues 7–67 of the domain; 45–105 in the SOX2 numbering system). Prior X-ray (5) and NMR (6) structures of the SOX2 DNA binding domain in complexes with other proteins have been published. (TIFF)

Table S1  Experimental details: protein sample, experimental conditions, NMR experiments, and tilted planes collected by ADAPT-NMR for a) SOX2, b) AeSCP2-PA, c) HSP12, d) RI-brazzein, and e) ubiquitin. (DOC)

Table S2  Comparison of ADAPT-NMR with a pipelined approach consisting of 3D data collection, automated peak picking by SPARKY, and automated assignment by PINE-NMR. (DOC)

References
1. Gu¨ntert P (2004) Automated NMR structure calculation with CYANA. Methods in Molecular Biology 278: 353–378.
2. Volk J, Herrmann T, Wuthrich K (2008) Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH. J Biomol NMR 41: 127–138.
3. Herrmann T, Gu¨ntert P, Wuthrich K (2002) Protein NMR structure determination with automated NOE-identification in the NOE S spectra using the new software ATNOS. J Biomol NMR 24: 171–189.
4. Herrmann T, Gu¨ntert P, Wuthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J Mol Biol 319: 209–227.
5. Moseley RN, Monle´on D, Monle´on GT (2001) Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. Methods Enzymol 339: 91–108.
6. Egghabina HR, Bahrani A, Wang L, Assadi A, Markley JL (2005) Probabilistic Identification of Spin Systems and their Assignments including Coil-Helix Inference as Output (PISTACHIO). J Biomol NMR 32: 219–233.
7. Bardiaux B, Malliavin T, Ni Jeans M (2011) ARIA for Solution and Solid-State NMR. Methods Mol Biol 631: 453–483.
8. Fiorito F, Herrmann T, Wuthrich K (2008) Automated amino acid side-chain NMR assignment of proteins using 13C- and 15N-resolved 3D [1H, 1H]-NOESY. J Biomol NMR 42: 23–33.
9. Lopez-Mendez B, Gu¨ntert P (2006) Automated protein structure determination from NMR spectra. J Am Chem Soc 128: 13112–13122.
10. Bahrani A, Assadi AH, Markley JL, Egghabina HR (2009) Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy. PLoS Comput Biol 5: e1000307.
11. Egghabina HR, Bahrani A, Tonelli M, Hallenga K, Markley JL (2005) High-resolution iterative frequency identification for NMR as a general strategy for multidimensional data collection. J Am Chem Soc 127: 12328–12336.

Table S3  Sidechain assignment by ADAPT-NMR for proteins with available 3D sidechain spectra. Peak lists from HCC-TOCSY, HCCONH, HBBACONH, and GCONH spectra were provided to ADAPT-NMR. (DOC)

Acknowledgments
We thank Prof. Masatsune Kainoˇso (Tokyo Metropolitan Univ. and Nagoya Univ.) for the Chlorella ubiquitin clone.

Author Contributions
Conceived and designed the experiments: AB HRE JLM. Wrote the paper: AB JLM HRE MT. Developed the probabilistic network software: AB. Implemented the NMR experiments: MT. Independently collected and assigned manual NMR data used in verifying ADAPT-NMR automated results: SCS KKS.