Automated Assignment of Backbone NMR Data using Artificial Intelligence

John Emmons[σ τ], Steven Johnson[τ], Timothy Urness*, and Adina Kilpatrick*
Department of Computer Science and Mathematics
Department of Physics
Drake University
Des Moines, Iowa 50311
john.emmons@drake.edu

Abstract

Nuclear magnetic resonance (NMR) spectroscopy is a powerful method for the investigation of three-dimensional structures of biological molecules such as proteins. Determining a protein structure is essential for understanding its function and alterations in function which lead to disease. One of the major challenges of the post-genomic era is to obtain structural and functional information on the many unknown proteins encoded by thousands of newly identified genes. The goal of this research is to design an algorithm capable of automating the analysis of backbone protein NMR data by implementing AI strategies such as greedy and A* search.

[σ] Primary author
[τ] Undergraduate researcher
* Faculty adviser

1 Nuclear Magnetic Resonance (NMR)

Nuclear magnetic resonance is a phenomenon in which atomic nuclei absorb electromagnetic radiation at frequencies related to their chemical properties and the local molecular environment. Biophysicists use this property to gain structural knowledge of biomolecules, including proteins, DNA and RNA. NMR spectroscopy is currently the only method that allows the determination of atomic-level structures of large biomolecules in aqueous solutions similar to their *in vivo* physiological environments.

Several types of NMR variables can be used in the analysis of protein structures. In particular, essential information is provided by the chemical shifts of NMR-active nuclei present in proteins, including hydrogen and isotopes of carbon and nitrogen. The chemical shift is a quantifier for the deviation in the resonant frequency of a nucleus from its value in a structure-free environment, and therefore provides information on the local conformation. Determining the chemical shifts of all or most of the nuclei in a biomolecule is the first step in determining its structure.

1.1 NMR Assignment Methodology

An important set of chemical shifts in a protein are those corresponding to the backbone nuclei, including the nitrogen, attached hydrogen, and the alpha and beta carbon atoms (C_α and C_β) of each of the residues that constitute the building blocks of the linear protein chain (Figure 1). These signals are measured using various NMR experiments, and then matched to the individual residues in the protein in a process called sequential assignment.

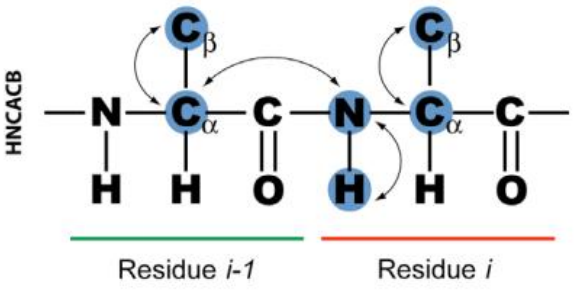

Figure 1: HNCACB NMR experiment

A prerequisite of the assignment process is data collection using experiments that can provide connectivities between neighboring residues [1]. One such experiment, called HNCACB, yields signals corresponding to the C_α and C_β nuclei of one residue in the protein (residue i) , plus the C_α and C_β signals of the immediately preceding residue (residue i - 1) (Figure 1). A second experiment, CBCA(CO) NH, can be used to yield the chemical shifts of the preceding residue only. This information is not independent, but helps distinguish unambiguously between signals from residue i versus residue i - 1.

Chemical Shift (ppm)	Residue i-1	Residue i	Residue i+1
C_α (self)	66.770	55.393	59.224
C_β (self)	38.056	17.975	29.006
C_α (preceding)	58.701	66.743	55.335
C_β (preceding)	29.070	38.067	17.927

Figure 2: Sequentially matched backbone carbon signals from HNCACB chemical shifts

Analysis of all inter-residue connectivities allows the linking of signals from each backbone atom with signals from their preceding neighbor, creating a pattern of sequentially linked chemical shift values which reflects the sequential linear arrangement of the individual residues in the protein sequence (Figure 2). This pattern is then matched with the protein sequence, by using the fact that certain residues have characteristic C_α and C_β chemical shift ranges which uniquely identify them. Thus, each measured chemical shift is assigned to a location in the protein, and this information can then be used to infer structural information about the biomolecule.

2 Manual Procedure

The sequential assignment of backbone NMR data is typically done manually. However, the process is very time-consuming (manual assignment of NMR datasets can take days to months) and is error-prone [2]. Common difficulties in manual data assignment arise from missing or ambiguous data, as well as spectroscopy artifacts. Therefore, in many cases the data analysis procedure is slow and nontrivial.

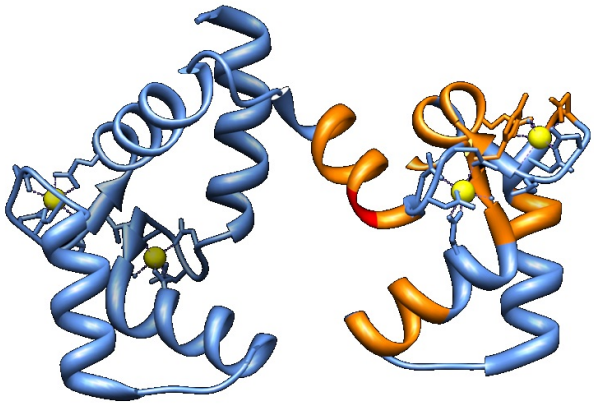

Figure 3: V91G calmodulin mutant investiated during the summer of 2012

One example is a chemical shift dataset acquired as part of a related research project during the summer of 2012 (Figure 3). A sample of the calcium-binding protein calmodulin, with a point mutation at position 91 in the sequence, produced spectra with many missing and

ambiguous data that required several weeks to assign. This experience inspired our team to develop an automatic assignment procedure.

3 Automated Algorithm

Several different approaches to creating algorithms capable of automated assignment of NMR datasets have been tried; nonetheless, there is still room for improvement in the field. Since last fall, I have been developing an algorithm capable of rapid, accurate assignment of small, yet non-trivial backbone NMR datasets, using techniques from artificial intelligence and statistics.

The current iteration of the algorithm assigns data in a three-step process: (1) the protein sequence is searched for short stretches (subsets) containing easily identifiable residues (those with high or low C and C chemical shifts); (2) these subsets are matched with sequentially linked chemical shifts; (3) an iterative greedy search algorithm completes the assignment, putting all residues in a sequential order consistent with the protein sequence. Steps (1) and (2) reduce data complexity before computationally expensive methods are used in step (3).

In step (1), the protein sequence is examined for subsets of the full-length sequence containing residue(s) with highly unique chemical shifts. The algorithm will record a subset if it contains one or several adjacent highly unique residue(s).

In step (2), the recorded subsets are matched with chemical shifts corresponding to their residue type(s). Statistical methods are used to ensure these chemical shifts are correctly linked in an i to i - 1 pattern. This process is done iteratively, starting with a very low error tolerance which is gradually increased until all recorded subsets are matched with sets of chemical shifts. This produces sequentially linked residues that are subsets of the full-length protein sequence. Each subset can be abstracted and treated as a single pseudoresidue with C_α and C_β of i and i – 1 corresponding to the chemical shifts of the residues making up the front and back of the subset, respectively.

In step (3), all chemical shifts corresponding to pseudoresidues and the remaining individual residues are placed in sequential order using an iterative greedy search algorithm. In this method, an arbitrary starting residue or pseudoresidue is first chosen. All remaining residues and psuedoresidues are added behind this starting point, such that the residue or psuedoresidue producing the lowest error in its C_α and C_β i - 1 chemical shifts is always placed next. This process is repeated until all residues and psuedoresidues have acted as the starting point. The generated sequence that produces the lowest total error (the sum of the errors between adjacent residues) is chosen as the correct assignment.

4 Preliminary Results

Assignment of small test datasets has proven that this process can correctly analyze NMR data; however, much room for improvement still exists. Research is expected to continue in the spring of 2013 to verify that the algorithm can correctly assign larger datasets, by incorporating aspects of machine learning.

5 Future Research

The accurate assignment of nontrivial NMR datasets is necessary for sustained advancement in the fields of structural biology and proteomics. As NMR technology advances, structural studies of very large molecules will become possible, with the price of increased complexity in data assignment. This research directly confronts the issue of nontrivial NMR datasets and is also consistent with the interests of our group as students and faculty of computer science, physics, and mathematics.

References

[1] Y.S. Jung and M. Zweckstetter, *Mars - robust automatic backbone assignment of proteins*. J Biomol NMR **30**, 11-23, 2004.

[2] B. Alipanahi, X. Gao, E. Karakoc, S. Li, F. Balbach, G. Feng, L. Donaldson, and M. Li *Error tolerant NMR backbone resonance assignment and automated structure generation*. J Bioinform Comput Biol **9**, 15-41, 2011.