**Automated Assignment of Backbone NMR Data using Artificial Intelligence and Statistical Methods**

Nuclear magnetic resonance (NMR) spectroscopy is a powerful method for studying the three-dimensional structure of molecules such as proteins. The data produced is used to study the function of a protein and alterations to its function that can lead to disease. Knowledge of a protein’s structure allows for further understanding of the function of a protein and alterations to its function that can lead to disease. The post-genomic era has brought the need to gather functional and structural information of unknown proteins that are encoded by newly discovered genes. NMR produces the structural information, but it needs to be analyzed and assigned rapidly and accurately. Unfortunately, current techniques for analyzing NMR datasets can take a few days to months to assign and are prone to error [1]. The current goal of my research is to develop an algorithm to automate the process of assigning nontrivial NMR datasets, in an attempt to minimize human error and accelerate a time consuming task.

Nuclear magnetic resonance is the absorbance of electromagnetic radiation at frequencies by atomic nuclei based on chemical properties and local molecular environment. Biophysicists use NMR properties to study the structure of biomolecules, such as proteins, DNA and RNA. NMR spectroscopy is the only method use today that is able to determine the atomic-level structures of large biomolecules in aqueous solutions similar to their *in vivo* physiological environments.

NMR spectroscopy produces many variables that can be used to analyze a protein’s structure. Our research focuses on the chemical shift values of NMR-active nuclei present in proteins, including hydrogen and isotopes of carbon and nitrogen. The chemical shift value measures the change in the resonate frequency of a nucleus from its structure-free environment. From these values, information about the surrounding structure can be deduced. Determining the chemical shift values of the nuclei in a biomolecule is the first step to determining its structure.

The chemical shift values pertaining to ‘backbone’ nuclei, including the nitrogen, attached hydrogen, and the alpha beta carbon atoms (Cα and Cβ) make up a residue used as a building block of a linear protein chain (Figure 1). NMR experiments are preformed to obtain the signals for each residue. The process of sequential assignment is used to match the individual residues to the protein chain.

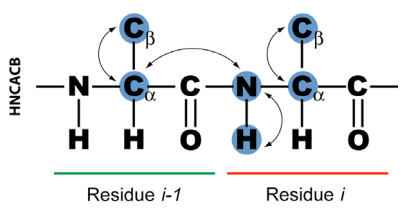


Figure 1. An HNCACB NMR experiment

Before the assignment process can take place, and experiment must be preformed to provide connections between neighboring residues. An experiment called HNCACB yields the signals that correspond to the Cα and Cβ nuclei of on a residue in a protein (residue *i*) and the Cα and Cβ of the next residue (residue *i*-1) (Figure 1) [2]. Another experiment call CBCA(CO)NH yields chemical shift values for the preceding residue only. The experiment is done independently, and it can be used to distinguish the residue *i* from residue *i*-1. Analyzing the inter-linking data produced by these experiments provides insight into the sequential linear arrangement of the residues in a linear protein chain (Figure 2). Using the fact that certain residues have characteristic Cα and Cβ chemical shift ranges, protein chains can be matched to individual residues. This allows each measured chemical shift value to be assigned to location in a protein. The resulting information provides insight into the structural information of a biomolecule.

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

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

The sequential assignment of backbone NMR data is usually done manually. It is a very time consuming task, taking anywhere from a couple days to months. It also is prone to human error [1]. The common difficulties of this task come from missing or ambiguous data. Much of the time the assignment process is slow and nontrivial. The goal of our research is to automate the assignment process.

Attempts at creating an algorithm to automate assignment of backbone NMR datasets have been made, but there is room for improvement in this field. I am involved in ongoing research with a team of computer scientists that are working on developing an effective, efficient, and scalable algorithm capable of rapidly assigning non-trivial backbone NMR data sets with high accuracy using techniques developed in artificial intelligence and statistics.

The current algorithm is a multi-step process. It uses methods of statistical analysis and the computer’s artificial intelligent to assign NMR datasets quickly and accurately.

Our algorithm takes in as input the expected protein backbone sequence of a protein chain. This will act as the characteristic pattern to which NMR data will be assigned. In our algorithm a *tile* is a container for the signals that correspond to the Cα and Cβ nucleifor residue *i-1 and i*. The goal of the algorithm is to automatically fill tiles that match the characteristic backbone sequence with tiles that have been filled with corresponding NMR data.

After reading in the NMR data and placing it in tiles, the program checks for missing data value. If missing data is detected, placeholders are used to fill the gap. Once the data is checked, the assignment process begins.

The algorithm generates a set of nodes to begin the assignment process. Each node has a different starting tile place in the first spot on the protein chain. The computer’s artificial intelligence is used to select the node with the best starting tile. The node with the best starting tile is used to generate child nodes. To generate child nodes, an unplaced tile is added to the end of the current list of tiles in the parent node. After a child node has been created for every remaining unplaced tile, the parent node is deleted.

Every tile placed after the first tile in a node generates a “cost.” This cost essentially measures the confidence that a tile belongs in that particular location along the protein chain. Statistical methods are used to compare the Cα and Cβ nuclei signals for the *i* residue to the characteristic pattern and the Cα and Cβ nuclei signals for the *i*-1residue from the tile placed before it. The cost that is generated aids the computer in selecting the next node to generate child nodes from.

The process of generating child nodes and calculating costs repeats until at least one node has every tile placed. Artificial intelligence searching algorithms are employed and are used to determine if it is the optimal solution. If it finds it is not, the algorithm continues with other nodes. However, if the node is the optimal solution, it is saved and the algorithm terminates.

Assignments of test datasets have shown this method can be used to assign NMR datasets, but there is still need for improvement. Research is ongoing as we continue to refine the way cost is assigned. We plan on testing this method with larger datasets to ensure their correct assignment. We will also look into grouping tiles by their residue characteristics to limit the number of nodes generated.

Accurately assignment of nontrivial NMR datasets will provide the necessary framework for sustained and advances in the fields of structural biology and proteomics. Advances in NMR technology will look to increase the size of the molecules they analyze. The result will be increased complexity in data assignment. My research works on tackling the challenges present with assigning nontrivial NMR data sets while sparking my interest in making scientific contributions as a student of physics, computer science and mathematics.

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

[2] Y.S. Jung and M. Zweckstetter (2004). Mars – ‘’robust automatic backbone assignment of proteins. *J Biomol NMR* **30**, 11-23.