**Accelerating Automated Assignment of Backbone NMR Data with Machine Learning Filters**

Structural biology has become a popular topic of research over the last few decades. With the completion of the human genome project, there has become an increase in the demand for protein structures. The structure provides functional information for a protein. Pharmacists can use this information to develop vaccines and antibiotics to prevent and cure the onset of disease. A powerful way to obtain structural information about a protein is through nuclear magnetic resonance (NMR) spectroscopy. Unfortunately, current techniques for analyzing NMR datasets can take anywhere from a few weeks to many months to process. They are also prone to human error [1]. To alleviate this problem, we have set out to automate the assignment process.

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 used 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 sequence (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.

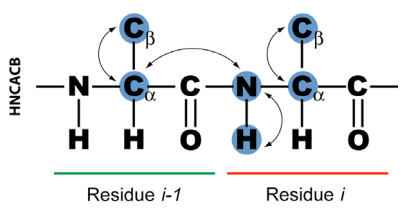


Figure 1. Backbone structure of a linear protein sequence

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 previous residue (residue *i-1*) (Figure 1) [2]. Another experiment call CBCA(CO)NH yields chemical shift values for the preceding residue only. Each experiment is done independently, and the CBCA(CO)NH experiment is used to ensure the residue *i*  and *i-1*  are properly recorded. 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).

|  |  |  |  |
| --- | --- | --- | --- |
| 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

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.

The sequential assignment of backbone NMR data is usually done manually. It is a very time consuming task, taking anywhere from a few weeks to months. It also is prone to human error [1]. Automation of this process will allow structural biologists to continue making advances in their research while simultaneously reducing human error in the assignment process.

Attempts at creating an algorithm to automate assignment of backbone NMR datasets have been made, but there is room for major improvements 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, machine learning and statistics.

The current algorithm is a search-based algorithm. It consists of two major parts: pre-processing and assignment. The pre-processing step receives the data and preforms filtering in order to prepare it to be assigned. The assignment then uses a uniform cost-based search to find the optimal order of the NMR dataset.

Pre-processing receives an NMR dataset and a protein sequence to assign the data set to. The protein sequence is converted to expected Cα and Cβ values, which will be used for reference later. Next, the NMR dataset is processed. Each amino acid it the protein sequence will have a corresponding set of four values (Cα and Cβ for residue *i* and *i-1*). These values will be stored in a single object called a tile. Once the NMR dataset has been organized into tiles, filtering can begin.

Filtering is preformed on each tile to eliminate the need to try unrealistic assignments. Our filtering system uses a logistic model tree (LMT) classification model to classify eachtile. Our model has been trained on over 68,000 amino acids with their published Cα and Cβ values. The classification process provides a confidence rating from 0 to 1 that the provided Cα and Cβ values correspond to on each of the 20 amino acids. Any tile that matches an amino acid with a 0.40 percent confidence level or better is used in the assignment process.

The NMR dataset and protein sequence completes pre-processing in less than one second. Then assignment process begins.

The assignment step uses a uniform cost search as the base. The search begins by comparing all of the tiles to the amino acid type of the first amino acid in the protein sequence. Any *tiles* that match the amino acid with a confidence rating of 0.40 percent is placed in the first location in the assignment. Next, a cost of assignment is calculated. This is done by taking a difference between the tile’s residue *i* values (Cα and Cβ) and the corresponding values in the protein sequence. These errors are then added to the total cost of assignment.

Next, the algorithm looks through all of the possible assignments, finds the one with the lowest cost and removes it from the list. The algorithm then looks through all of the tiles that have not been assigned and compares it to the next amino acid in the protein sequence. Any tiles that match the amino acid with a confidence rating of 0.40 percent is added to the end of the assignment. A cost for adding the tileis calculated. It compares the tile to the protein sequence as before, and it compares the tile’s residue *i-1* values to the residue *i* values of the tileplace before it in the assignment. The cost of placing the last tileis added to the total cost. All of the new assignments are added to the list of possible assignments, and the process of addingtiles repeats.

Once all of the tileshave been placed in a particular assignment, it is considered a possible solution. The process of adding new tiles continues until a possible solution has the lowest cost. At this point, the search is terminated and the optimal solution is recorded.

Our algorithm has been able to assign protein sequences to more than than 50 amino acids. This was made possible by the 100-fold increase in speed that filtering provides. Research is ongoing as we work toward assigning larger protein sequences.

[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.