Accelerating Biomolecular Nuclear Magnetic Resonance Assignment with A*

Joel Venzke, Paxten Johnson, Rachel Davis, John Emmons, Katherine Roth,
David Mascharka, Leah Robison, Timothy Urness and Adina Kilpatrick
Department of Mathematics and Computer Science
Drake University
Des Moines, IA
joel.venzke@drake.edu

Abstract

Nuclear magnetic resonance (NMR) spectroscopy is a powerful method for studying the three-dimensional structure of molecules such as proteins. The post-genomic era has created the need to gather functional and structural information of unknown proteins encoded by newly discovered genes. Unfortunately, current manual techniques for analyzing NMR datasets can take days to months to assign, and are prone to error. The current goal of this 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. NMR images yield structural information that needs to be analyzed rapidly and accurately. Accurate assignment of nontrivial NMR datasets will provide the necessary framework for continued advancement in the fields of structural biology and proteomics. The algorithm utilizes A* search to sequence amino acids produced from NMR datasets of proteins. The program is given a set of nodes, each corresponding to a single amino acid in the protein chain, from an input file. Each node is placed in the initial node position of the protein chain. The node that most correctly fits the reference protein chain is used to generate child nodes. Unplaced nodes are sequentially added to the end of a list to create child node lists. The node list with the smallest error is then expanded further. The error is a running total of how well the nodes fit the reference protein chain, coupled with the accuracy of the match of the placed child. This process of assigning the child nodes and calculating costs repeats until every node is placed and the best node list contains the lowest cost of all generated lists. Assignments of test sets have shown this method can be used to assign NMR datasets, but there remains need for improvement. Research is ongoing as we continue to refine our algorithm. Future goals include adjusting for potentially missing data, sorting nodes into amino acid groups, and testing with larger datasets. Upon completion, this algorithm will allow for great advancements in the areas of structural biology and proteomics.

1 Introduction

Knowledge of a protein's structure can explain the function of a protein and how alterations to its function can lead to disease. With the use of nuclear magnetic resonance (NMR) spectroscopy, the structure of molecules, such as proteins, can be studied. However, NMR spectroscopy generates large amounts of data that need to be analyzed. This process can take months to complete and is prone to error. To allow for major advances in structural biology, a faster and more accurate method of analysis needs to be developed. Our research focuses on developing an algorithm to automate the assignment process. We hope our attempts can minimize the human error and decrease the time consumption associated with manual assignment, increasing the overall effectiveness of the operation.

1.1 Nuclear Magnetic Resonance Spectroscopy

Several types of NMR variables can be used in the analysis of proteins' 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. An important set of chemical shifts in a protein are those corresponding to the nuclei in the backbone of the protein polypeptide chain, including the nitrogen, attached hydrogen, and the alpha and beta carbon atoms (C_α and C_β) of each residue shown in Figure 1. These chemical shifts are measured using various three-dimensional NMR experiments, and then matched to the individual residues in the protein in a process called sequential assignment.

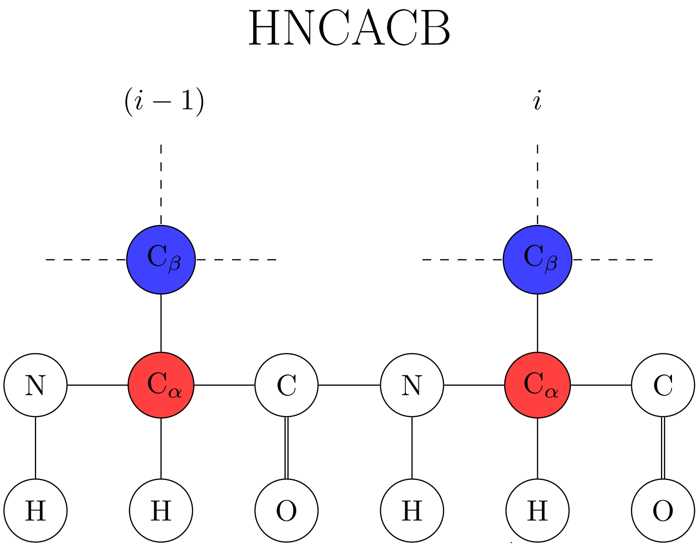

Figure 1: NMR

A prerequisite of the assignment process is data collection using experiments that can provide connectivities between neighboring residues. For example, the HNCACB experiment identifies the chemical shifts corresponding to the C_α and C_β nuclei of one residue (residue i), as well as the C_α and C_β chemical shifts of the immediately preceding residue (residue $i-1$). In this experiment, the values corresponding to residue i can usually be distinguished from the $(i-1)$ values by their higher intensity. However, ambiguities can arise if the intensities are comparable or if the chemical shift values are overlapping. These ambiguities can be resolved by using additional experiments, such as CBCA(CO) NH, which yields the chemical shifts of the preceding residue only. Using all inter-residue connectivities, a chain of correlations through the protein backbone chemical shifts can be established. This pattern of sequentially linked chemical shift values reflects the sequential linear arrangement of the individual residues in the protein sequence. The 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 to uniquely identify them. Thus, each measured chemical shift is assigned to a protein residue, and this information can then be used to infer structural information about the biomolecule.

2 Assignment

2.1 Data Collection

Determining the structure of a protein requires many highly involved steps. The first, which can take at least five days, is protein production. NMR experiments are then performed using the produced proteins in order to gather the data necessary for the assignment process. This process usually takes another one to two days for every spectrum involved. After all the data has been gathered, the assigning of the protein's structure, the most lengthy of the steps, will begin [1].

2.2 Manual Assignment

In most cases, backbone NMR data is assigned manually. However, with hundreds of chemical shift values, this process can take anywhere from 20 days to 9 months with perfect data [3]. To complicate matters even more, experience has shown most data sets to have missing or false peaks, requiring whomever is doing the actual assignment to either skip large portions of the protein sequence or to guess where a certain chain is located in the sequence. As a result, this method is extremely error prone [1].

2.3 Algorithm Design

Our algorithm, which is designed to decrease the error and time consumption associated with manual assignment, begins by reading in an expected amino acid sequence and the

NMR chemical shift data. It then changes the amino acids to their expected C_α and C_β values and stores them in a list that represents the protein chain. A tile containing the C_α and C_β values for the amino acid residue i and residue $(i-1)$ is then created. Next, missing data is accounted for. If there are fewer tiles than the number of amino acids in the protein chain, placeholder tiles are generated to make up the difference. Before the search for the best solution begins, the tiles are grouped by the types of amino acids they most likely represent in an attempt to accelerate the assignment process. These groups are created using the expected C_α and C_β values shown in Figure 2. Tiles that do not match are grouped with the placeholder tiles.

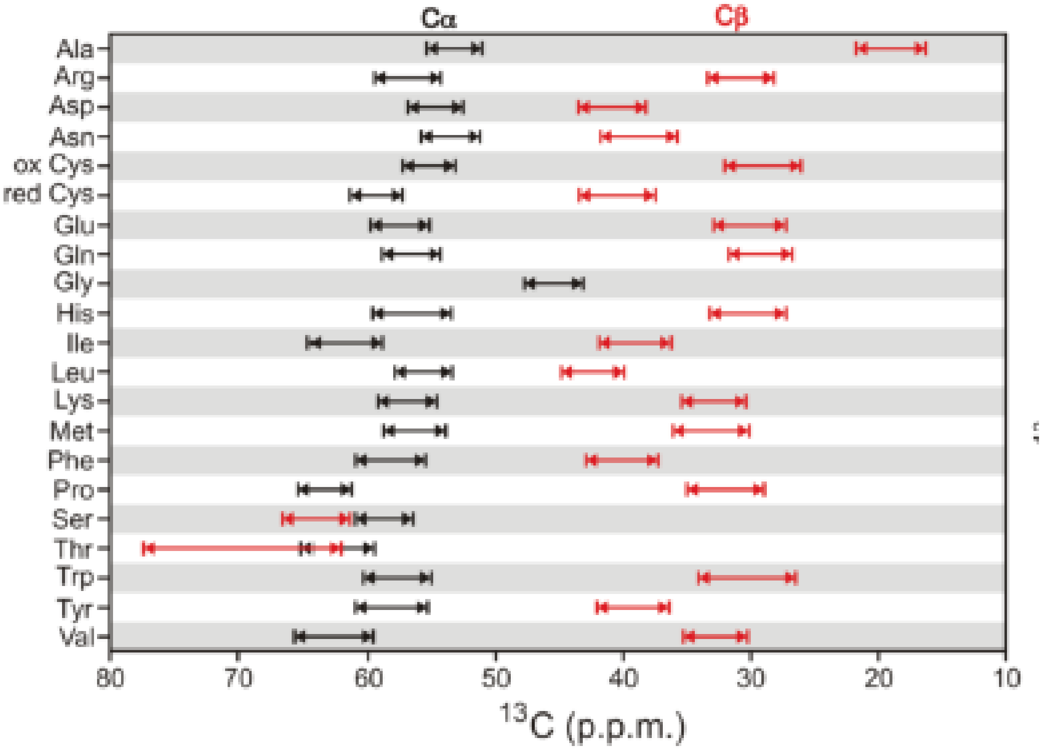

Figure 2: Expected Carbon Shift Values [2]

To start the search, all the tiles matching the first amino acid in the reference protein chain are placed in the first position of a node. All placeholder tiles are also added to this position. The tile is then compared to the first amino acid in the protein chain to generate a cost. The cost is the sum of the differences generated by placing any previous tiles in the node, how well the current tile matches the corresponding amino acid in the protein chain, and the disparity between the tile's residue $(i-1)$ values and the residue i values of the tile placed before it. All of the generated nodes are placed in an array of possible solutions, also called the frontier.

The node in the frontier with the lowest cost is removed from the array and used to generate child nodes, which is accomplished simply by adding a tile to the end of the node. A child node is generated for all unplaced tiles that are either in the placeholder group or in the group matching the corresponding amino acid in the protein chain. The cost of placing the tile is then calculated. Next, the child nodes are added to the frontier. This process of generating child nodes is repeated until a goal state is reached.

Once a node has every tile placed, it is said to be in the goal state. The first node that reaches this state is stored as the current best solution. Every node that achieves the goal state is compared to the current best solution. The node with the lowest cost then replaces the previously stored best solution.

The final solution has been found once the current best solution's cost is lower than the cost of any other nodes in the frontier. The search is completed. The algorithm then outputs the cost of the best solution followed by the assigned chemical shift values.

3 Conclusions

3.1 Results

Our algorithm has been tested with NMR datasets of varying size from experiments conducted by Dr. Kilpatrick. When compared to manually assigned data, the algorithm produces the same, optimal ordering of amino acids. Initially, we focused our research on accurate assignment at the cost of speed. Thus, at the start of our research, our algorithm was only able to assign smaller datasets with a chain of ten to fifteen amino acids. Shifting our focus to optimizing the algorithm's speed while maintaining the accuracy we had managed to achieve, we were able to improve runtime. With the increased performance, we were able to correctly assign forty amino acids in less than 30 seconds. The results of our testing are shown in Figure 3.

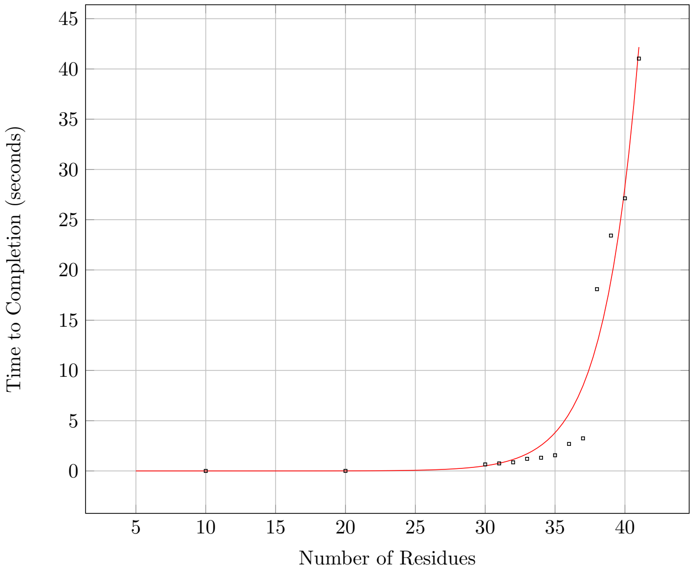

Figure 3: Time of Assignment

Since we are using real data to test the algorithm, some data points are missing. Missing data can lead to dramatic increases in assignment time. The jump in runtime between 37 and 38 residues, seen in Figure 3, is one such example. Residue 38 is missing the C_α and C_β values for residue $(i - 1)$. This missing data forces the algorithm to explore a greater number of possibilities to ensure accurate assignment, incidentally increasing the time it takes for the best solution to be found.

3.2 Future Goals

In future research, we will look into maximizing speed and efficiency with the goal of assigning full amino acid sequences. In order to achieve this, we will experiment with both parallelization and machine learning. The latter of these two will assist in optimizing the cost calculation while the former will increase the speed of assignment. We believe these implementations will help our algorithm handle datasets with numerous missing data points, validating our results even when given incomplete data.

References

[1] Babak Alipanahi, Xin Gao, Emre Karakoc, Frank Balbach, Shuai Cheng Li, Guangyu Feng, Logan Donaldson and Ming Li, *Error tolerant NMR backbone resonance assignment and automated structure generation.*, Journal of bioinformatics and computational biology, **9** (2011), 15–41.

[2] Sean Cahill and Mark Girvin. *Introduction to 3d triple resonance experiments*. 2012.

[3] Peter Gntert, *Automated structure determination from NMR spectra*, European Biophysics Journal, **38** (2009), 129–143.

[4] Flemming M. Poulsen. *A brief introduction to nmr spectroscopy of proteins*. 2002.