Utilizing Machine Learning to Accelerate Automated Assignment of Backbone NMR Data

Joel Venzke, David Mascharka, Paxten Johnson, Rachel Davis, Katie Roth, Leah Robison, Adina Kilpatrick and Timothy Urness*

Department of Mathematics and Computer Science and the Department of Physics and Astromony,
Drake University, Des Moines, Iowa

**Joel Venzke: joel.venzke@drake.edu*
Timothy Urness: timothy.urness@drake.edu

February 25, 2015

Abstract

Nuclear magnetic resonance (NMR) spectroscopy is a powerful method for studying the three-dimensional structure of molecules, including proteins. These methods provide useful data for the sequencing of amino acids into primary protein structures. Unfortunately, current manual techniques for analyzing these NMR datasets can take months to assign and are highly susceptible to error. Many programs have been developed to sequence protein backbone structures, with various strengths and weaknesses. The algorithm detailed in this paper works to overcome the obstacles others have met by utilizing machine learning to predict amino acid type, thereby increasing assignment speed. The program also generates place-holders to accommodate missing data and recurring amino acids with problematic chemical characteristics, namely proline. This approach is able to speed up the process of assignment by almost three orders of magnitude while maintaining high accuracy when compared to a previously assigned set.

KEYWORDS
Protein Sequencing, Machine Learning, NMR, Artificial Intelligence, Proteins, Bioinformatics

INTRODUCTION

The analysis of Nuclear Magnetic Resonance (NMR) data, in particular the sequence-specific assignment of backbone and side-chain protein resonances, is an error-prone and time-consuming step during protein structure determination by NMR spectroscopy.[1] NMR spectroscopy is a powerful method for obtaining atomic-resolution three-dimensional protein structures, as well as assessing changes in protein conformations or motions due to mutations or interactions with ligands or other biomolecules. Determining a proteins structure is essential for understanding its function and alterations in function, which often lead to disease. This paper describes a computational algorithm that utilizes machine learning in the process of automating the assignment of backbone protein NMR resonances.

BACKGROUND

NMR spectroscopy experiments generate information on several variables that can be used in the determination of protein structures. In particular, essential information is provided by the chemical shifts determined by 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. Measuring the

chemical shifts of all or most of the nuclei in a biomolecule is the first step in determining its structure by NMR spectroscopy. 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 amide nitrogen (N), attached hydrogen (H), and the alpha and beta carbons (C_α and C_β) of each residue. The chemical shift values are measured using various three-dimensional NMR experiments, and then matched to the individual residues in the protein in a process called sequential assignment.

Triple resonance experiments are the method of choice for proteins and other large biomolecules, because they can greatly decrease the amount of spectral overlap in data. Typical experiments for the assignment of backbone resonances include HNCA, HN(CO)CA, HNCACB, and CBCA(CO)NH or HN(CO)CACB.[2] These experiments transfer magnetization over the protein polypeptide chain, and thus connect different spin systems through covalent bonds. A spin system essentially contains all resonances belonging to a particular amino acid. The order of elements in experiment names indicates the order in which magnetization is passed down the line of nuclei, resonances in parentheses being used only for transfer to the next nucleus. All these experiments produce spectra with common H-N resonance correlations, and provide connectivities between neighboring residues. For example, the HNCACB experiment identifies the chemical shifts corresponding to the C_α and C_β nuclei of each residue in the protein chain (residue i) , as well as the immediately preceding residue (residue $i-1$). In this experiment, resonances corresponding to residue i can usually be distinguished from the $i-1$ values due to their higher intensity. However, ambiguities can arise if the intensities are comparable or if chemical shift values overlap. Such uncertainty happens frequently and can be resolved by using additional experiments. For example, an experiment CBCA(CO)NH yields the chemical shifts of the preceding residue only, unambiguously identifying i and $i-1$ values. Using all inter-residue connectivities, a chain of correlations through the protein backbone chemical shifts can be established. Sequentially-linked chemical shift values reflect the ordered linear arrangement of individual residues in the protein sequence. The pattern is then matched to the protein sequence through certain residues with characteristic C_α and C_β chemical shift ranges that uniquely identify them. Thus, each measured chemical shift is assigned to a protein residue and can be used to infer structural information about the biomolecule.

SEQUENTIAL ASSIGNMENT STRATEGIES

The sequential assignment of backbone chemical shifts can be done manually or in an automated fashion. Both approaches follow similar strategies. The process starts by identifying the chemical shifts of i and $i-1$ residues from a set of three-dimensional NMR experiments. In the early days of NMR spectroscopy, these chemical shifts were measured using a ruler.[3] Nowadays, many computer programs, such as NMRPipe[4] and SPARKY,[5] can be used to process, analyze and visualize NMR data. After grouping chemical shifts into spin systems, sequential spin systems are linked into increasingly larger segments by matching $i-1$ and i values. In parallel, the spin systems are classified into possible residue types, using a set of established chemical shift values for each of the 20 common amino acids found in proteins. For example, alanines have C_α and C_β chemical shifts ranging from approximately 50 to 5 ppm (parts per million), and 17 to 25 ppm, respectively; glycines have unique C_α values in the 45 ppm region; and threonines have distinctive C_β values in the 68-73 ppm range.[6] Using the residue type information, the linked segments from the resonance sequential walk are iteratively mapped onto the protein primary sequence.

When performed manually, the assignment process is time-consuming and prone to errors. If chemical shifts values overlap, multiple matches between i and $i-1$ values are possible. The classification of spin systems into amino acid types also has a high potential for error, as the spin systems can look identical or very similar even for very different amino acids. In recent years, advances in computer technology, accompanied by the increased use of NMR spectroscopy in drug design and structural genomics initiatives, have created a push for the automation of various steps in the NMR structure determination process.[7] These automated methods decrease the time needed to complete the assignment, and attempt to minimize the risk of human error and subjectivity. A wide variety of algorithms and software packages for the assignment of backbone chemical shifts exist. Amongst a few of the most well known are GARANT,[3] AutoAssign[5] and MARS.[2]

The GARANT program has three key elements.[8] It starts with matching measured chemical shift values to values expected from the protein sequence. It then evaluates the correctness of these matches using a scoring function, and optimizes the score using an evolutionary algorithm in conjunction with a local opti-

mization routine.[3] This optimization is similar to simulated annealing which makes repeated selections until the optimal solution is reached (8). AutoAssign uses a combination of artificial intelligence techniques and statistical methods to obtain the assignment[9] . This program utilizes four basic steps, generally following the overall assignment strategy described above, but providing customized and novel solutions for each step. It first ranks and sorts spin systems. Secondly, it creates lists of possible amino acid types using Bayesian statistical methods. In step 3, AutoAssign links sequential spin systems into fragments, starting from spin systems that have a unique one-to-one match. Lastly, the fragments are mapped onto the protein sequence. Steps three and four are interleaved and run iteratively, such that sequential connectivities are established in parallel with the identification of amino acid types. The software MARS optimizes the search by simultaneously assessing the local and global quality of the assignment.[2] This program breaks the assignment into segments, linking and mapping fragments consisting of up to five residues, and using them as units in the search. The most accurate model is selected in combination with protein secondary structure prediction. The reliability of fragment mapping is tested by performing multiple runs with chemical shifts modified by the addition of artificial noise.

These and other programs have paved the way for the automation of protein backbone assignments and made significant advances toward complete automation of protein structure determination by NMR spectroscopy. However, several issues remain. Some strategies may abandon a promising assignment too early if it has a higher error than other assignments, and thus the optimal solution will not be selected in the end.[2] Occasionally, a local minimum can throw off the optimization process.[2] Further issues arise when assigning larger protein chains. Some programs need days to finish the assignment in these cases; even then, the calculations can still have very high errors.

Our algorithm utilizes elements from these existing programs, but employs novel concepts from machine learning to accelerate the assignment process.

Machine Learning

Machine learning provides algorithms that learn from attributes in the input data to increase performance. Supervised learning, a field of machine learning, builds a model based on numerous data elements and their respective labels. The result is a mathematical model that predicts a label given a set of input attributes. By discovering patterns and trends in the data, machine learning algorithms provide excellent solutions for building models to generalize from large amounts of collected data with potentially many attributes, a task that is often difficult or impossible by other means.

Machine learning offers a natural solution to the problem of determining amino acid type from NMR chemical shifts values. The overlap between the normal range of C_α and C_β values between many amino acids makes it difficult to infer the type of residue based solely on chemical shift information. Machine learning algorithms offer a unique approach to this problem and achieve excellent accuracy.

There are several supervised machine learning algorithms that can be applied to improve automated assignment strategies. The J4.8 algorithm[10] builds a decision tree model. This means that the data is split based on a comparison to an attribute of the data, with a branch for each possible outcome of the test. At the end of the tree, the leaf, is the predicted label. In our case this is the amino acid type. To classify a new value, a datapoint begins at the root of the tree and moves through until a leaf is encountered. The encountered leaf is the predicted value for that datapoint. To construct the tree, the attribute test used at each branch is the one that partitions the set in the most useful manner.

The Logistic Model Tree, or LMT,[11] is another tree-based algorithm. In contrast to the J4.8 algorithm, LMT constructs a tree with logistic regression functions at each branch rather than an attribute test. Logistic regression attempts to model class probabilities with linear functions (Hall). The weights used for each function can be learned to split each class in a way somewhat analogous to the split in J4.8 described above.

The Decision Table algorithm[12] is comprised of a set of features and a set of labeled data. To classify a new datapoint, the set of labeled data is searched for a match with the new point, considering only the features in the feature set. If no matches are found, the majority class of the labeled data is used. Otherwise, the majority class of the matches is used.

Methods

Our goal is to create an automated program for the assignment of protein backbone chemical shifts that could deliver quality results in a small amount of time without the use of a supercomputer. Our program implements group sorting, machine learning, filtered amino acid selection, and careful cost calculation. The algorithm completes resonance assignment in six steps: (1) the NMR chemical shifts and the protein sequence are used as input data that fills data structures; (2) the protein sequence is processed and empty data structures are initialized for missing data; (3) the machine learning model assigns possible amino acid types to each spin system; (4) filtering is applied to locate chemical shift data that corresponds to residue 1 in the protein sequence; (5) the search continues with the best assignments until all chemical shift data is assigned; (6) the best solution is recorded and the process is terminated.

Before the algorithm begins the assignment process, machine learning is used to build a model for predicting amino acid type. After the model is trained, the assignment process consisting of the pre-processing steps (steps 1 to 3) and the search (steps 4 to 6) begins. Pre-processing is where our research has made significant advances in order to decrease assignment time by predicting amino acid types. This allows the data to be filtered, significantly reducing the branching factor. The algorithm is then able to intelligently search the remaining possibilities for the best assignment.

Machine Learning Data Collection

The training dataset for the machine learning algorithm was obtained from BMRB. We initially identified 9,736 datasets containing chemical shifts for the C_α and C_β resonances of 689,977 residues. However, a preprocessing step was necessary in order to remove extreme values for each amino acid and improve both accuracy and generalization. Removing outliers prevents the algorithms from fitting extraneous data, improving future performance. By looking at statistics available on the BMRB site, we excluded chemical shift values outside three standard deviations of the mean for each amino acid type. This gave us 681,363 pairs of C_α and C_β values to use for training.

Pre-processing

Step 1 consists of reading the chemical shift values and the protein amino acid sequence from an input file. We use the C_α and C_β values for residues i and $i-1$ to create an object we will refer to as a tile. A tile holds information corresponding to a single residue in the protein sequence to be assigned. In step 3, amino acid types will be assigned to each tile.

Step 2 of the algorithm converts the protein sequence from the abbreviations into C_α and C_β chemical shift values, using statistics provided by the Biological Magnetic Resonance Bank,[6] a database of NMR chemical shifts hosted by the University of Wisconsin at Madison. These statistics provide average chemical shift values for each of the twenty common amino acids found in proteins, using 4944897 chemical shifts published in BMRB as of February 2, 2015. For example, we assign alanine a C_α chemical shift of 53.19 ppm (parts-per-million) and a C_β chemical shift of 18.96 ppm. Then the protein sequence is analyzed. The algorithm looks for prolines in the protein sequence. Since prolines do not generate chemical shift data from these experiments, special tiles are created to handle this case. Identifiers in the proline tiles ensure that the tile is placed only when the corresponding residue in the protein sequence is a proline. This identifier limits the number of possibilities where the tile can be assigned. The length of the protein sequence is then compared to the total number of tiles created thus far. If fewer tiles exist than the overall number of residues, blank tiles are created to fill the difference. Blank tiles can fit in any location in the assignment. Large amounts of missing data will deteriorate the algorithms performance.

Step 3 assigns possible amino acid types to each tile.The C_α and C_β values for residue i in each tile are processed by our machine learning model, producing a list of probabilities that a tile represents each amino acid. The probabilities correspond to confidence levels used for filtering during the assignment process.

Preprocessing the dataset takes a minimal amount of time (less than a second on a standard laptop) and drastically reduces the time required to assign dataset without affecting accuracy. The search for the optimal assignment then begins.

The Search

The algorithm begins an intelligent search through filtered combinations of all possible chemical shifts. The search begins by placing the first tile, which is selected based on a filtering process. Only titles that correspond to the first amino acid based on a confidence level threshold (0.4% match or better) are placed at residue one.

At this point, the cost of assignment is generated. The cost of placing a tile consists of two parts: (1) the difference between the tiles residue $i-1$ values and the previous tiles residue i values and (2) the difference between the residue i values and the values predicted from the protein sequence. In the case of blank and proline tiles, a fixed cost is added in place of the above calculation. Since initially only one tile is placed, the cost is based solely on a comparison to the protein sequence, and the search moves on to step 5.

In step 5, the algorithm selects the assignment with the lowest cost to continue the search process. If the assignment is a solution, the algorithm moves on to step 6. A solution is reached when all tiles have been placed in an assignment. If the assignment is not a solution, the amino acid type of the next residue in the protein sequence is retrieved. Any unplaced tile that corresponds to that amino acid type with a confidence level above the threshold is placed at the next location in the sequence. In the special case that the next amino acid type is proline, only the special proline tiles are considered for placement. The cost is then adjusted to include the newly placed tile. Step 5 is repeated until a solution has been reached.

In the final step the search records the solution. The solution assignment is formatted for easy reading and output along with the algorithms performance. Then the algorithm terminates.

Results

The chemical shift dataset used in this study consisted of C_α and C_β values for the 59-residue (our data set is 62 long) long C-terminal domain of the Tfg1 subunit of the yeast transcription factor TFIIF. The chemical shifts were previously obtained by Kilpatrick et al. from HNCACB and CBCA(CO)NH experiments, and manually assigned to 99% completeness. The previously assigned data was cut into sub-sections, randomized, and used for algorithmic analysis.

The results of assigning this dataset with different filtering methods are shown in Figure 1. For sequences of up to 62 residues a smooth trend for all filtering methods tested was seen when graphing the generated nodes (or different assignments) against the sequence lengths. Interestingly, a comparison of the methods indicates that our filtering process results in a significant decrease in number of generated nodes compared to an unfiltered generic search algorithm, as seen in Figure 1. Since the most time-consuming part of the search is node generation, there is a direct correlation between assignment time and number of nodes generated. The graph indicates that the LMT algorithm has the best performance, outperforming the method with no filter by almost two orders of magnitude, without loss of assignment accuracy. This not only accelerates assignment, but also allows for larger datasets to be assigned.

Figure 2 shows the impact of proline identification on our algorithms performance. The same unfiltered and LMT data from Figure 1 is plotted for comparison. The large jump in assignment time from 32 to 33 residues shows the performance impact of missing data. In the sequence of the studied protein, residue 33 is a proline, which lacks NH (check denotation previously in paper) spin systems. As a consequence, HNCACB and CBCA(CO)NH experiments do not provide C_α and C_β chemical shifts for this residue. If the algorithm identifies and deals with prolines as a special case, only one more node is generated for the 33-residue case. However, if the proline is not dealt with separately, the algorithms performance is significantly impacted. Without proline checking, the proline tile is placed in every position in the assignment. The result is a major increase in the branching factor that leads to the jump observed in Figure 2. With proline checking, prolines are no longer problematic to the assignment process. This indicates that our algorithm can produce accurate assignments with reliable and reasonably fast assignment times even when data is missing.

Our algorithm using LMT filtering with proline checking can complete a 62-residue assignment in approximately 40 minutes on a 2.3 GHz Intel Core i7 processor with 8Gb of RAM. In less than 1 second the LMT model can complete an assignment of 43 residues, compared to 27, 29 and 30 for the no filter, decision table and J4.8 models, respectively.

The use of proline checking and utilizing machine learning to filter data has shown to be extremely effective in accelerating the assignment. The reduction of nearly 3 orders of magnitude in assignment time

allows for approximately twice as many residues to be assigned in the same time period. Our algorithm has completed assignments of up to 62 amino acids in less than an hour. The solution matched the manually assigned sequence, ensuring the accuracy of our algorithm. Given more time, our algorithm can complete longer assignments. However, with the overall exponential trend, the length of time required to complete assignments overcomes the usefulness of the automated assignment process for long protein sequences.

Conclusions and Future Directions

Our algorithm has made significant advances in the field of automated assignment for protein backbone chemical shifts. We implemented machine learning to filter NMR data in order to reduce the branching factor in a search-based algorithm. This increased our assignment rate by approximately three orders of magnitude while still maintaining the accuracy of a comparative manually assigned solution.

In the future, we will improve the overall performance of our algorithm by optimizing cost calculations, handling missing data better, and sequencing chunks of data at a time (puzzle building). (Tim, should we even talk about this?)(Tim help us)

One of our main focuses for the future is handling missing data in a more efficient manner. By examining characteristics of the amino acids in the sequence, we hope to predict where missing data will end up in the final assignment. This would reduce the number of assignments attempted.

We are currently acquiring backbone chemical shifts for a new protein.This new data will include additional chemical shifts that can be used in the assignment process, such as the backbone carbonyl and H-alpha values. This data will be used to further validate and improve our algorithm. Since the cost calculation is crucial to the effectiveness of our algorithm, additional chemical shifts may prove invaluable to the success of the algorithm for longer protein sequences or incomplete datasets. We are also investigating methods of predicting the final cost of an assignment in order to remove unrealistic assignments early on. Having this information available will help optimize our cost calculation, resulting in a further decreased assignment times.

In order to accommodate longer protein sequences while still producing accurate assignments, we will also look at assigning separate sections of the protein sequence at one time. By chunking the amino acids into smaller sections and incorpo, we hope to decrease current assignment times.

ACKNOWLEDGEMENTS

The authors of this paper would like to thank Drake University for being a conductive space to research, Iowa State University for their use of their NMR spectroscopy machine, John Emmons for kicking off this research, and our ever supportive mentors Tim Urness and Adina Kilpatrick.

References

[1] Linge, J. P.; Habeck, M.; Rieping, W.; Nilges, M. *Bioinformatics* 2003.

[2] Jung, Y.; Zweckstetter, M. *Journal of Biomolecular NMR* **2004**, *30*, 11–23.

[3] Güntert, P. **2009**, *38*, 129–143+.

[4] Delaglio, F.; Grzesiek, S.; Vuister, G.; Zhu, G.; Pfeifer, J.; Bax, A. *Journal of Biomolecular NMR* **1995**, *6*, 277–293.

[5] Zimmerman, D. E.; Kulikowski, C. A.; Huang, Y.; Feng, W.; Tashiro, M.; Shimotakahara, S.; ya Chien, C.; Powers, R.; Montelione, G. T. Automated Analysis of Protein NMR Assignments Using Methods from Artificial Intelligence. 1997.

[6] Ulrich, E. L. et al. *Nucleic Acids Research* **2008**, *36*, 402–408.

[7] Moseley, H. N. B.; Monleon, D.; Montelione, G. T. *[6] Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data*; Elsevier, 2001; Vol. 339; pp 91–108.

[8] Bartels, C.; Gntert, P.; Billeter, M.; Wthrich, K. *Journal of Computational Chemistry* **1997**, *18*, 139–149.

[9] James, T. L. *Nuclear Magnetic Resonance of Biological Macromolecules, Pt.B*; Elsevier: San Diego, CA, 2001.

[10] Quinlan, R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers: San Mateo, CA, 1993.

[11] Landwehr, N.; Hall, M.; Frank, E. *Machine Learning* **2005**, *95*, 161–205.

[12] Kohavi, R. The Power of Decision Tables. 8th European Conference on Machine Learning. 1995; pp 174–189.

ABOUT THE STUDENT AUTHORS

Joel Venzke (say some things) is not at this meeting, so his biography will be vandalized by other people while he cannot defend himself. His biography is going to be written in comic sans, since it fits him so well and he cant choose his own font. Joel was born on January 6th 1994. It was a cold winter day in Minnesota when he was brought into this world. His love for science started at a young age as he liked to take apart Easy Bake Ovens in order to put them back together. Ironically that sparked his love for petit fors, especially with blackberry jam. Yum-o! After his toaster phase, Joel started experimenting with cooking. He honed his skills and currently is making jerky, of the beef variety. Literally, he is right now. We hope he can put his love for jerky aside in the near future so he can continue his research. He better damn bring us some jerky. Also possibly take some photos along the way.

David Mascharka is a sophomore pursuing a B.S. in Computer Science and Mathematics and B.A. in Philosophy. David also pursues his love for mathematics as the president of the math club in which is he able to hone his leadership skills.Plus he is a pretty cool guy. David is a really special guy and is going to do great things in the future. Just watch. He actually will. He really has no idea what to put here, and so will change this later. Sure, sure it will.

Paxten Johnson is on track to graduate in the Spring of 2016 with a B.S. in Physics, Computer Science, and Mathematics. Aside from conducting this research for the past two years, she has been involved with Air Force ROTC, Delta Gamma Fraternity, and the Drake Honors Program. She hopes to use the knowledge and experience gained from this research to help her advance into the military intelligence branch of the Air Force.

Rachel Davis is a mere semester away from her B.S. in both Computer Science and Mathematics. This is the second year conducting this sequencing research. Her research includes not only this primary structuring of proteins, but also a collaborative tertiary structuring algorithm. Currently she is applying to summer internships closely relating to her interests and expertise.

Katie Roth is currently pursuing a B.A. in Mathematics and Computer Science. She has been working on this research for two years, and enjoys it immensely. Katie is the treasurer of the Women in Mathematics and Computer Science and an active member of Math Club. She is currently applying for summer research experiences, as well as internships.

Leah Robison is on her way to graduate with a B.S. in Environment Science and a minor in Computer Science. Apart from studying abroad in Denmark for a semester, she has been involved with this research group for the past two years. She hopes to apply the knowledge gained from this research to future studies in the field of Environmental Science.

PRESS SUMMARY

Manually assigning nuclear magnetic resonance data to a backbone protein sequence is time consuming and error prone. Although current algorithms have made advances in this area, a student research group at Drake University is improving the process by utilizing machine learning to identify amino acids before assignment. Using both old and new methods has resulted in an algorithm that is fast and accurate.