An Iterative Relaxation Technique for the NMR Backbone Assignment Problem

W. L. Hsu, J. M. Chang, W. C. Chou, J. B. Chen, K. P. Wu, T. Y. Sung,
Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C.
and C. F. Chang, W. J. Wu and T. H. Huang
Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan, R.O.C.

Abstract

NMR spectroscopy is one of the popular experiments to determine protein structures. An important stage of protein structure determination by using NMR is protein backbone resonance assignment (or backbone assignment for short). Due to the messiness and disorder of NMR spectral data, backbone assignment is usually a tedious and time-consuming manual work. This raises a great interest in developing an efficient and automatic method to perform backbone assignment.

NMR spectral data is usually transferred into spin systems. A spin system contains the chemical shifts of atoms within a residue. To perform backbone assignment, we usually need to know the (partial) sequential order of spin systems; the procedure of figuring out such an order is called connectivity determination. Different NMR experiments provide different chemical shift information that can be used in connectivity determination. Consider the following two experiments.

Given the i-th \( N_i \) and \( H^N_i \) pair in a peptide sequence, the experiment HNCACB provides four peaks of the form \((C^?, N_i, H^N_i, +/\)) in a spin system, where the question mark indicates that we do not know whether the carbon is located in the \((i-1)\)-th residue or the \(i\)-th residue. The fourth element is peak intensity in which \( C^\alpha \) has a positive value and \( C^\beta \) has a negative one. So we only know that two of the four peaks are associated with \( C^\alpha \) and the other two peaks are associated with \( C^\beta \). Another experiment CACB(CO)NH, also for the i-th \( N_i \) and \( H^N_i \), provides two peaks of the form \((C^?,_{i-1}, N, H^N_i, +)\) in a spin system, where the question mark indicates that we do not know whether the carbon is \( C^\alpha \) or \( C^\beta \), since the intensity values of this experiment are always positive. So we only know that the two peaks are associated with the \((i-1)\)-th residue.

Though each of these two experiments provides only partial information about atoms, combining them can provide more complete information for determining the order of two consecutive spin systems, which is the building block of connectivity determination. Some other experiments such as HSQC can also help us to produce better connectivity determination results. Biologists can decide which experiment should be performed to best suit their needs.

Connectivity determination is usually followed by backbone assignment, which assigns chemical shifts to corresponding atoms in a protein sequence. Given a perfect connectivity relation among spin systems, much work has been done to obtain a good backbone assignment result. However, due to the nature of NMR experiments, the spectral data usually contains two kinds of noises: false positive, namely, a visible peak may be a noise; or false negative, namely, a peak maybe missing. We need a way to select useful peaks to perform connectivity determination and backbone assignment; we call this task peak picking. False positives and false negatives make peak picking difficult. For example, for a missing \( C^\beta \) chemical shift, we do not know whether the value is really lost, or the corresponding peak is associated with a Glycine that contains no \( C^\beta \) atom.

We propose an iterative algorithm that is equipped with two operations: grouping and linking. Grouping is responsible for peak picking and part of connectivity determination. Ideally, those peaks with the same \( H^N \) and \( N \) chemical shifts can be grouped together; peaks belonging to the same group can be used to determine the order of two consecutive spin systems. However, in real situation, grouping is a difficult task due to false positives and false negatives. We sometimes add hypothetic peaks to tackle false negatives (see next paragraph) and use linking operation to remove false positives. Linking is responsible for part of connectivity determination and backbone assignment. Given a protein sequence and partial connectivity information, we try to link connected spin systems as much as possible. False positive spin systems may create conflicts in the linking stage. To find a good assignment on a noisy dataset, we model the backbone assignment problem as a maximum independent set problem. Although the problem is NP-complete, there are heuristic methods that obtain pretty good results.

There are three main stages to perform grouping and linking iteratively: perfect, weak false negative and strong false negative. Our algorithm performs a iterative relaxation as follows. In the perfect stage, we use actual peaks to generate
spin systems; no hypothetical peak is considered. The generated spin systems are then used to perform linking, which generates a partial backbone assignment. In the weak false negative stage, we generate more spin systems by adding a few hypothetical peaks. These new spin systems along with the partial backbone assignment generated in the perfect stage are used to generate new backbone assignment. To generate a better result, the assignment of the perfect stage is sometimes modified in the weak false negative stage. Similarly, in the strong false negative stage, we add a lot of hypothetical peaks to generate even more spin systems, and these new spin systems along with the partial backbone assignment result generated in the weak false negative stage are used to generate more complete backbone assignment.

We use actual experiment data and synthetic data to evaluate our algorithm. There are two experimental datasets from Dr. T. H. Huang’s labs in IBMS, Academia Sinica obtained by performing HSQC, HNCACB and CACB(CO)NH experiments. These datasets contain more than 50% false positives and false negatives. Compare to the best manual solution, the precision and recall of our algorithm on the first dataset “sbd” are 91.43% and 74.42%, respectively; and those on the second dataset “lbd” are 80.65% and 62.50, respectively. There are also eight synthetic datasets that are generated according to eight datasets chosen from the BioMagResBank (BMRB). The average precision and recall of our algorithm on these eight datasets are 98.70% and 90.86%, respectively.