Journal of
Applied & Computational Mathematics

Gao, J Applied Computat Mathemat 2012, 1:1
DOI: 10.4172/2168-9679.1000e103

Editorial Open Access

Mathematical Approaches to the NMR Peak-Picking Problem

Xin Gao*

Mathematical and Computer Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia

Nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography are two experimental techniques used to determine the three-dimensional structures of proteins. NMR spectroscopy has the unique ability to capture proteins in vivo. Currently, protein structure determination by NMR follows the procedure proposed by Kurt Wüthrich in 1986 [1]. This procedure consists of peak picking, resonance assignment, nuclear overhauser effect (NOE) assignment and structure calculation steps. Among the four steps, peak picking is time consuming and requires extensive expert knowledge. Computational methods designed to automate and improve this step are still needed. The inputs to the peak picking problem are an NMR spectrum or a set of spectra, whereas the outputs are the lists of peaks (signals) identified from these spectra.

Mathematically speaking, each NMR spectrum is stored in the form of a multi-dimensional matrix, where the indices for each dimension are the discrete chemical shift value of a certain atom and the entries of the matrix are the intensity values of the signals. A d-dimensional NMR spectrum captures the coupling of d atoms and stores them in a d-dimensional matrix. Each peak in the spectrum represents the coupling of d atoms that have chemical shift values equal to the corresponding coordinates. Because there are a finite number of atoms in a protein, the number of peaks is also finite. Ideally, the non-signal regions of the spectrum should have intensity values equal to zero. However, due to various sources of errors in NMR experiments, such as chemical shift degeneracy, sample impurity, water bands and artifacts, all the entries in the matrix have non-zero real values. The peak shapes, which should be Gaussian like in the ideal case, are very bumpy as well.

The peak-picking problem is the main bottleneck in automated NMR protein structure determination. The accuracy requirement for automatic peak-picking methods is high because the peak-picking results are the inputs for the following steps in the structure determination procedure. The peak-picking step is always a lengthy process in NMR labs. It is usually done manually or semi-automatically. Among all the signals, the strong and obvious peaks are easy to identify by computational methods. The main difficulty, however, is to identify weak or strongly overlapping peaks, as well as eliminating strong but fake peaks.

The computational peak-picking problem has been studied for more than a decade. There are three main computational challenges: (1) how to de-noise the given NMR spectrum; (2) how to identify peaks in the de-noised spectrum; and (3) how to select the true peaks from a set of predicted peaks. Below, we discuss each of these challenges and propose possible solutions.

- How to de-noise a given NMR spectrum. There are two ways to de-noise a given NMR spectrum, hard de-noising or soft de-noising. Hard de-noising assumes that the intensity values of signals are higher than those of random noise. Therefore, such methods try to estimate the noise level by calculating the standard deviation of the noisy regions [2,3]. All data points with intensity values that are lower than the noise level are eliminated and the spectrum becomes a set of disconnected components. Apparently, such methods will eliminate weak peaks that are embedded in the noise level. On the other hand, soft de-noising tries to smooth the entire spectrum without eliminating any data points. The smoothing is tricky in the sense that it would be best to smooth out the highly frequent regions, i.e. the noise, but leave the infrequent regions, i.e. the signals. Special wavelets have been successfully applied to identify this trade off [4].

- How to identify peaks in the de-noised NMR spectrum. If soft de-noising is applied, the spectrum is smoothed, which suggests the use of the brute force method to identify all the local maxima [4]. If hard de-noising is applied, the spectrum is still noisy but disconnected, which leads to decomposition methods that process each disconnected component separately. Singular value decomposition performs well in this setting [2]. However, it is possible that matrix factorization-based methods could move the peak locations or even merge strongly overlapping peaks together.

- How to select true peaks from a set of predicted peaks. Any peak-picking method will end up with a large set of predicted peaks. Some of these peaks are true peaks, but others are fake. There can be various sources for the fake peaks. For instance, side chain groups can cause fake peaks that also have good peak shapes. Noisy signals can also be predicted as candidate peaks. The problem is to select the true peaks from a mixed set of true and fake peaks. This is a common scenario in many bioinformatics problems. In order to solve such problems, a confidence score function is needed to rank the true predictions on the top and a method to automatically select the correct number of predictions is also required. It has been shown that volume-based confidence score functions significantly outperform intensity-based score functions in the peak picking setting [4], although the volume-based method is not able to rank all the true peaks on the top. The problem of selecting the correct number of peaks to include all true peaks is a multi-testing problem in statistics. Multi-testing methods should therefore be applicable to this problem.

Although there has been progress in using computational methods to resolve the peak-picking problem, time-consuming and expensive manual work is still required. With continuous advances in high-throughput techniques, rapid protein structure determination processes are urgently needed to advance structural genomics research. Automatic peak picking is an indispensable step toward such a goal. More effective methods or the proper combination of existing methods

***Corresponding author:** Xin Gao, Mathematical and Computer Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, 23955-6900, Saudi Arabia, E-mail: Xin.Gao@kaust.edu.sa

Received March 17, 2012; **Accepted** March 20, 2012; **Published** March 24, 2012

Citation: Gao X (2012) Mathematical Approaches to the NMR Peak-Picking Problem. J Applied Computat Mathemat 1:e103. doi:10.4172/2168-9679.1000e103

should be developed to solve this problem. Moreover, there should be open access to any practically useful method so that the broad scientific community will be served.

References

1. Wuthrich K (1986) NMR of proteins and nucleic acids. John Wiley and Sons, New York.
2. Alipanahi B, Gao X, Karakoc E, Donaldson L, Li M (2009) PICKY: a novel SVD-based NMR spectra peak picking method. Bioinformatics 25: i268-i275.
3. Koradi R, Billeter M, Engeli M, Güntert P, Wüthrich K (1998) Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY. J Magn Reson 135:288-297.
4. Liu Z, Abbas A, Jing BY, Gao X (2012) WaVPeak: picking NMR peaks through wavelet-based smoothing and volume-based filtering. Bioinformatics.