Algorithms for the Multiple Sequence Alignment Problem

Multiple Sequence Alignment (MSA) is the problem of finding as many common features as possible among a sequence of DNA or protein sequences taken from a family of species. MSA is one of the most fundamental computation problems in molecular biology/bioinformatics that many biological modeling methods depend on, including phylogenetic trees, profiles and structure prediction. A fast and accurate computer algorithm for MSA will greatly benefit and boost the advancement of study in these methods. However, the MSA problem in its general form has been shown to be NP-hard [1], indicating that general efficient computer algorithms for this problem are very unlikely to exist. Thus researchers have been trying to develop either approximation algorithms or ad hoc heuristic algorithms for MSA that work well for a particular settingin practice. For example, Gusfield[2] devised an algorithm with approximation ratio 2 using the sum-of-pairs criteria to evaluate the goodness of an alignment. Both the Long Run algorithms by Jiang and Li [2] and the Expansion algorithm by Bonizzoni et al. [3] can be also applied to the MSA problem with guaranteed approximation ratios. On the other hand, many heuristic algorithms have been developed to work for a special set of data. Notredame[4] provides a recent survey.

Unfortunately, with all these efforts none of the algorithmsso far derives software that provides a decent general solution to the MSA problem in the biological sense, esp. for DNA/protein sequences with large non-coding gaps [5]. Biologists are still largely depend on human experts to provide solutions of concrete instances ofthe MSA problem, which in many cases are very time consuming. In light of this, much further study is still needed for the MSA problem, both theoretically and experimentally, to produce more acceptable solutions for biological studies. This has become more and more urgent as we are facing unprecedented volumes of data in the field of bioinformatics.

We propose a systematic study of the performance of current software for MSA to find out their strengths and weaknesses.We also propose to work closely with human experts to identify heuristic rules that mimics human intuition of finding the best alignment among a set of DNA sequences, which current software cannot do. Base on these, we will develop good heuristic algorithms that can be implemented intogenerally more usable software for the MSA problem. Our effort will include close collaboration with Dr. Andrea Schwarzbach in the Department of Biology, who is an expert in evolutionary biology and phylogenetics. We will also conduct further theoretical algorithmic study on the MSA problem to facilitate the design of faster and better approximation and/or heuristic algorithms.

References

1. L. Wang and T. Jiang. On the complexity of multiple sequence alignment. *Journal of Computational Biology*1(4): 337-348, 1994.
2. T. Jiang and M. Li. On the Approximation of Shortest Common Supersequences and Longest Common Subsequences. *SIAM Journal on Computing* 24(5):1122-1139, 1995.
3. P. Bonizzoni, G.D. Vedova and G. Mauri. Experimenting an Approximation Algorithm for the LCS. Discrete Applied Mathematics, 110: 13-24, 2001.
4. C. Notredame. Recent Evolutions of Multiple Sequence Alignment Algorithms. Computational Biology 3(8): 1405-1408, 2007.
5. A. Schwarzbach, L. Lei, L. Tang and L. Zhang. Private communications.