Year 2013

Thesis for Master Degree

Unavoidable ambiguity of RNA secondary structure estimation and methods to overcome it

Graduate School of Frontier Sciences, The University of Tokyo
Department of Computational Biology

47-116913 Ryota Mori

Advisor: Professor Kiyoshi Asai

Thesis for Master Degree Overview Year 2013

Unavoidable ambiguity of RNA secondary structure estimation and methods to overcome it

RNA conformation is widely regarded as an issue of importance in its biological function. For several decades, many kinds of tools for predicting RNA secondary structure are proposed by both dry and wet strategies. From the point of view of base-by-base predicting accuracy, recent methods have accomplished quite high performance for RNAs of known structure. All of these methods, however, have common latent problems such as unclear reliability or omission of information on thermal fluctuation and significant sub-optimal structures. These problems are caused by unavoidable ambiguity of current estimation style, which is called point estimation.

To solve these problems, we require an alternative method for extracting information from exact probability distribution efficiently. The basic idea is to calculate existence probability for each distance from a certain reference structure by applying a dynamic programming technique. We constructed this algorithm by adopting McCaskill model for calculation of partition function and hamming distance for difference of structures. We can compute this certain structure centered probability distribution by $O(n^3)$ time and $O(n^2 d_{max})$ memory (n: RNA sequence length, d_{max}: maximum hamming distance) with maximum parallelization. In addition, we expand the algorithm into two dimensions for comparing some structural clusters. This two dimensional distribution can be computed by $O(n^3)$ time and $O(n^2 d_{max}^2)$ memory under the abundant computational resources.

We indicate how much ambiguity exists in current estimation methods and then show several outputs of our algorithm. For example, the distribution of a certain riboswitch around its minimum free energy structure implies two distinct structural clusters. This riboswitch is known to change its conformation dynamically in the presence of SAM. Actually, SAM^+ and SAM^- structure correspond to primary and secondary peaks of our distribution respectively. Our two diminutional expansion implies that these two peaks seem to be separated by quite high potential barrier but there might exist a channel to associate these peaks.

Our proposing method yields us profound information on reliability and stability. By applying this algorithm, we can extend a range of analysis such as expression level, thermal stability, and so on.

In addition, general form of our algorithms has the possibility to be employed to various problems in bioinformatics. It is applicable to any dynamic programming problems including Hidden Markov Models, which are widely made use of in the field of bioinformatics. We discuss it by exemplifying completely different problem of RNA secondary structures.

Keywords: RNA secondary structure, reliability, sub-optimal structure