

AN IMPROVED GIBBS SAMPLING METHOD FOR MOTIF DISCOVERY VIA SEQUENCE WEIGHTING

Xin Chen

School of Physical and Mathematical Sciences
Nanyang Technological University, Singapore

Tao Jiang

Department of Computer Science and Engineering
University of California at Riverside, USA
Currently visiting at Tsinghua University, Beijing, China

The discovery of motifs in DNA sequences remains a fundamental and challenging problem in computational molecular biology and regulatory genomics, although a large number of computational methods have been proposed in the past decade. Among these methods, the Gibbs sampling strategy has shown great promise and is routinely used for finding regulatory motif elements in the promoter regions of co-expressed genes. In this paper, we present an enhancement to the Gibbs sampling method when the expression data of the concerned genes is given. A sequence weighting scheme is proposed by explicitly taking gene expression variation into account in Gibbs sampling. That is, every putative motif element is assigned a weight proportional to the fold change in the expression level of its downstream gene under a single experimental condition, and a position specific scoring matrix (PSSM) is estimated from these weighted putative motif elements. Such an estimated PSSM might represent a more accurate motif model since motif elements with dramatic fold changes in gene expression are more likely to represent true motifs. This weighted Gibbs sampling method has been implemented and successfully tested on both simulated and biological sequence data. Our experimental results demonstrate that the use of sequence weighting has a profound impact on the performance of a Gibbs motif sampling algorithm.