Correlation motif model for integrative analyses of genomic data

Ying Ying Wei

Department of Statistics, The Chinese University of Hong Kong

Abstract

In the era of high-throughput technologies, genomic data in public repositories are rapidly growing. To illustrate, to date, more than 1,000,000 samples have been stored in Gene Expression Omnibus and ArrayExpress; meanwhile, over 2,500 ChIP-seq samples are deposited in the ENCODE project and the Sequence Read Archive. This large volume of data provides unprecedented opportunities to improve detection for weak signals by integrating multiple genomic datasets. Here, we propose a scalable correlation motif approach for integrative analysis of multiple high-dimensional genomic datasets. The approach adopts a flexible Bayesian hierarchical mixture model to capture the latent correlation structures embedded in the data and substantially improves signal detection for low-signal-to noise ratio data. The applications are illustrated by differential gene expression detection when the expression datasets have only a small number of replicate samples as well as allele-specific protein-DNA binding detection from ChIP-seq data. Moreover, the proposed model is applicable to heterogeneous data type integration and allows parallel computing.