FlexModule: a flexible cis-regulatory module sampler

by

HONGKAI JI Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health 615 North Wolfe Street Baltimore, MD 21205, USA hji@jhsph.edu

Abstract

Cis-regulatory modules (CRM) are genomic loci where proteins (transcription factors) bind to regulate gene expression. Multiple transcription factors can bind cooperatively to a CRM through the recognition of sequence specific binding sites (TFBS). The characterization of physical locations and binding site structures of CRMs provides useful information for studying gene regulation. When little is known about CRM structures and binding sequence specificities, de novo Gibbs motif sampler and module sampler have been proposed to search for transcription factor binding motifs and modules. When prior knowledge about the module structures is available, one may wish to incorporate the prior into the module discovery as well. Since prior knowledge varies from case to case, this integration cannot be easily done using currently available models. To simplify the integration of different types of prior information, a two layer sequence generating model is proposed. Based on t! he new model, a flexible cis-regulatory module sampler, FlexModule, is developed. Various types of prior information can be integrated into the model through a common interface. The probability landscape can be modulated by user-specified prior knowledge and then explored by a user-independent Gibbs sampler. The flexibility of the approach is illustrated through a few simulated and real data examples.