Poster H-38

A Motif Sampler for the Discovery of Conserved Motif Pairs with Periodic Spacing

Authors:
Erik Larsson (*Department of Biomedicine, Göteborg University*)
Olle Nerman (*Department of Mathematical Statistics, Chalmers University of Technology*)
Per Lindahl (*Department of Biomedicine, Göteborg University*)

Short Abstract: Due to the turning DNA helix, distances between binding sites within cis-regulatory modules sometimes have periodic properties. We propose a motif discovery tool based on Gibbs sampling which finds motifs pairs with periodic spacing. Evaluation on simulated datasets demonstrates increased sensitivity compared to single motif and colocalization models.

Long Abstract:

Transcription factors often cooperate and form DNA-protein-protein-DNA complexes in regulatory regions. As a consequence, binding sites for different transcription factors often colocalize in promoter sequences. Such functional groups of binding sites are often referred to as cis-regulatory modules (CRMs). Significant effort has been put into the problem of de novo discovery of transcription factor binding sites from sequence data. The task, sometimes described as multiple local alignment, is difficult due to the degeneracy and low information content of many transcription factor binding motifs. In recent years, methods which take advantage of the synergistic effects of colocalization to improve discovery of regulatory motifs have been proposed. However, several studies on both genomic sequences and mutated promoters indicate that in addition to proximity, the phase between adjacent sites may often be crucial to maintain functionality of a regulatory complex. A small modification of the distance between sites may drastically disrupt function and this is explained by the turning of the DNA helix. In some cases, changing the distance of two adjacent binding sites by a multiple of the period of the DNA helix will cause functionality to be maintained, as making a complete turn on the helix will cause the same face of the binding protein to be exposed to cofactors and nearby DNA binding factors.

We propose a motif discovery tool based on Gibbs sampling which finds structured motifs consisting of periodically spaced motif pairs. In other words, the motifs are assumed to be separated by a distance which is allowed to vary in such a way that the “helical phase” is conserved. The separation is modeled as a fixed distance p (the phase) plus an integer multiple of the period T (+/- noise). The motif pair may optionally be allowed to occur on both strands and is assumed to be either present or absent in each sequence. The period T can be specified by the user or be optimized by the program, and the method is therefore not limited to finding patterns with a period equal to that of the DNA helix (~10 bp). Moreover, minimum and maximum separation (specified in periods) can be specified. The separation can be allowed to be negative, making it possible to find unordered motif pairs. The flexibility of the algorithm turns the tasks of finding colocalized motifs (e.g. positioned within 100 bp of one another with no respect to the phase) or motifs with fixed spacing (e.g. always exactly 25 bp from each other) into special cases of the full model, simply achieved by choosing

appropriate parameter settings. By favoring motif placement in highly conserved regions, the algorithm also incorporates the possibility to take advantage of interspecies conservation.

Evaluation on simulated datasets demonstrates increased sensitivity compared to a single motif model and a colocalization model. We also compare it to previously published tools such as MEME (single motifs) and CisModule (CRMs), and show that weak motifs which are part of CRMs with periodic spacing properties may be more accurately detected with our model. The method is implemented Matlab but is currently being ported to C and is intended to be made publicly available.