Poster L-19

Computational detection of tissue- and process-specific cis-regulatory modules by a maximum specificity method

Authors:
Peter Van Loo (*Department of Human Genetics, Flanders Interuniversity Institute for Biotechnology and University of*)
Stein Aerts (*Department of Human Genetics, Flanders Interuniversity Institute for Biotechnology and University of*)
Bart De Moor (*Bioinformatics group, Department of Electrical Engineering, University of Leuven, Belgium*)
Yves Moreau (*Bioinformatics group, Department of Electrical Engineering, University of Leuven, Belgium*)
Peter Marynen (*Department of Human Genetics, Flanders Interuniversity Institute for Biotechnology and University of*)

Short Abstract: We present the first method for CRM discovery based on whole genome optimization. We applied this method on 892 genes and predicted CRMs in 218 of those. These predictions are supported both by leave-one-out cross-validations and by validations in independent test sets.

Long Abstract:
Metazoan transcription regulation occurs through the concerted action of multiple transcription factors that bind cooperatively to cis-regulatory modules (CRMs). Existing methods for the computational detection of cis-regulatory modules (CRMs) require prior knowledge of the regulatory system under study. The user is expected to provide the size of the CRMs and the number of transcription factors involved. The methods that have delivered biologically validated results (only in Drosophila) all demand a focused collection of position weight matrices (PWMs) on top of this, and hence can only be applied to well-studied processes.
Here we present ModuleMiner, a novel, generally applicable CRM detection algorithm, focused on human, that does not require any prior knowledge. The algorithm models similar CRMs, both proximal and distal, in a set of co-regulated genes as a combination of motifs (PWMs), supplemented by a number of “shape and size” parameters. ModuleMiner selects the model that is the most specific for the set of genes under study, relative to all other genes in the genome. The resulting CRM model can be used to scan the complete genome for additional similar CRMs (this step takes advantage of the well established procedure to find target genes of known PWM combinations). This genome-wide scanning, combined with leave-one-out cross-validation and binomial statistics, permits the representation of the probability of successful CRM detection by a p-value, and hence provides a powerful tool to validate ModuleMiner’s findings on any set of co-regulated genes in silico.
We applied ModuleMiner to 10 microarray clusters and successfully ($p < 0.05$) identified CRMs in 8 of them. These CRMs direct the expression of tissue-specific genes (heart, lymphocytes, liver) or of genes related to specific processes (extracellular matrix/inflammation, mRNA processing/protein synthesis, energy metabolism, mitochondia

and mitosis). Our predictions are further supported by the identification of additional similar CRMs in independent test sets, by their correspondence with literature, by their conservation in four mammalian species (human, mouse, rat and dog), and by the identification of relevant new target genes.

The liver-specific cluster contains 63 genes, which were divided into a training set of 50 genes and a test set of 13 genes. We identified similar CRMs, encompassing binding sites for HNF1, HNF4, LYF1 and FREAC4, in 17 genes of the training set. In the independent test set, we found an additional 6 similar CRMs ($p < 0.001$). Genes encoding for HNF1-alpha and HNF4-gamma were among the top scoring target genes, in support of a regulatory feedback loop.

In the mRNA processing and protein synthesis cluster, we found 2 types of CRMs. The first was found in 17 (out of 50) training set genes and in an additional 22 (out of 123) test set genes ($p < 0.05$), while the second was found in 15 training set genes and in 34 test set genes ($p < 10\text{-}6$).

Conclusion:

We present the first method for CRM discovery based on whole genome optimization. We applied this method on 892 genes and predicted CRMs in 218 of those. These predictions are supported both by leave-one-out cross-validations and by validations in independent test sets.