Abstract

The main topic of this PhD is the development of computational tools for the detection of *cis*-regulatory module (CRMs) using *itemset mining* techniques.

A first method ModuleDigger, is a CRM detection method to detect *cis*-regulatory modules based on set of coregulated sequences, relying on CHARM to enumerate possible motif combinations and well-designed statistical scoring scheme to prioritize biologically valid CRMs. We benchmarked ModuleDigger with existing tools and tested its validity on a real dataset. However, as ModuleDigger doesn't take into account the proximity of binding sites composed a certain CRM, it still oversimplifies the biological problem. Although it performs well in detecting the true regulatory modules it can not specify the true binding sites that compose the modules.

Therefore we developed CPModule, a CRM detection method that relies on a *constraint programming* framework for *itemset mining*. CPModule enumerates all possible CRMs that meet the following biologically motivated constraints: a certain CRM should occur in a minimal number of sequences (frequency constraint) and its composing motif sites should occur within a maximal genomic distance from each other (proximity constraint). The first constraint allows tuning the degree of overrepresentation that we expect in a set of intergenic sequences, while the second constraint reflects that sites of combinatorially acting TFs occur in each others neighbourhood. A last constraint (redundancy constraint) reduces the level of redundancy amongst the valid CRMs. Firstly, we experimentally validate our approach and compare it with state-of-art techniques using a literature existing synthetic data. Secondly, we propose CRM detection in combination with ChIP-Seq by performing a real case study on ChIP-Seq experiments of five transcription factor KLF4, NANOG, OCT4, SOX2 and STAT3 on mouse embryonic stem cell. Epigenetic information is also used to check whether surrounding chromatin stability for TFBSs is permissive for the binding of TFs.

Besides for detecting CRMs, we also developed ViTraM, a tool for visualizing expression module i.e. gene sets that are coexpressed under a specific set of conditions with or without their regulatory program (sets of transcription factors

that are responsible for the observed coregulation). It uses a input the result of biclustering or network inference algorithms. ViTraM is capable of visualizing overlapping these transcriptional/expression modules in an intuitive way by allowing for a dynamic visualization and using multiple methods for obtaining the optimal layout. In addition to visualizing multiple modules, ViTraM also allows to display additional information on the regulatory program of the modules, which consists of the transcription factors and their corresponding motifs. Information on the regulatory program is either obtained from curated databases or from the outcome of the inference tool itself. By visualizing not only the modules but also the regulatory program, ViTraM can provide more insight into the modules and facilitates the biological interpretation of the identified modules.