Inferring transcription factor targets from multiple heterogeneous data sources

Xiaofeng Dai[1] and Harri Lähdesmäki[1]

[1]*Department of Signal Processing, Tampere University of Technology, Tampere, Finland*

Transcriptional regulation is a central control mechanism for many biological processes, such as development and cell cycle. Transcriptional regulation is largely controlled by transcription factors (TF) that bind gene promoters in a sequence specific manner. Thus, revealing genome-wide protein-DNA interactions is one of the key problems in understanding transcriptional regulation at mechanistic level.

Computational transcription factor binding site (TFBS) predictions rely on sequence specificities that are taken from a database (JASPAR, TRANSFAC), obtained as an output from a motif discovery method or, more recently, measured using high-throughput techniques. Sequence specificities alone, however, are not sufficiently informative to accurately predict TF targets. A natural way to improve TF target predictions is to incorporate additional data into statistical inference of TFBSs. We have recently developed a probabilistic TFBS prediction method that is able to make use of practically any additional genome-level information source [1]. Statistical data fusion becomes more challenging when several information sources need to be combined in a meaningful way.

Here we extend our previously published method [1] by incorporating novel data sources into TFBS prediction and by developing a new method for multiple data fusion. In particular, we use evolutionary conservation, nucleosome positioning data from a recently published method, regulatory potential and DNA duplex stability to improve TFBS predictions. These data sources are informative of binding sites because functional binding sites are typically conserved and free of stable nucleosomes, regulatory DNA sites have different characteristics than neutral sites, and different TFs can bind DNA in a single or double strand manner, Some of these individual data sources have already been shown to improve *de novo* motif discovery, but we demonstrate how these multiple data sources can be combined to make joint statistical inference of TF targets. Integration of those data sources that have a probabilistic interpretation is relatively straightforward [1]. For other cases, we convert the raw data into probabilities, or priors, by applying a previously proposed Bayesian transformation method. In addition, for efficient use of DNA duplex stability data, we develop a simple heuristic that can assess the binding preference (single or double stranded DNA) for a TF from a set of known binding sites.

Results on a carefully constructed test set of verified binding sites in the mouse genome (ABS, ORegAnno) demonstrate that principled data fusion can significantly improve the performance of TF target prediction methods. Our statistical data fusion method can gain valuable new insights into genome-wide models of transcriptional regulatory networks.

[1] Lähdesmäki H, Rust AG and Shmulevich I. (2008) Probabilistic inference of transcription factor binding from multiple data sources. *PLoS ONE*, Vol. 3, No. 3, e1820.