Poster J-8

Inferring transcriptional modules from ChIP-chip-, motif- and microarray data

Authors:
Karen Lemmens (*Dept. Electrical Engineering, K.U.Leuven*)
Thomas Dhollander (*Dept. Electrical Engineering, K.U.Leuven*)
Tijl De Bie (*Dept. Psychology*)
Pieter Monsieurs (*Dept. Electrical Engineering, K.U.Leuven*)
Kristof Engelen (*ept. Microbial and Molecular Systems, K.U.Leuven*)
Bart De Moor (*Dept. Electrical Engineering, K.U.Leuven*)
Kathleen Marchal (*ept. Microbial and Molecular Systems, K.U.Leuven*)

Short Abstract: We present 'ReMoDiscovery', a module detection algorithm that exploits in a concurrent way three independent data sources: ChIP-chip data, motif information and gene expression profiles. When compared to published module discovery algorithms, ReMoDiscovery is fast and easily tunable. When evaluated on yeast data, biologically meaningful findings were generated.

Long Abstract:

Inferring transcriptional modules from ChIP-chip-, motif- and microarray data

K. Lemmens1, T. Dhollander1, T. De Bie3, P. Monsieurs1, K. Engelen2, B. De Moor1, K. Marchal1,2
1 Dept. Electrical Engineering, K.U.Leuven, Leuven, Belgium, 2 Dept. Microbial and Molecular Systems, K.U.Leuven, Leuven, Belgium, 3 Dept. Psychology, K.U.Leuven, Leuven, Belgium

We present "ReMoDiscovery" [1], a novel integrative method for inference of transcriptional modules from three independently acquired heterogeneous data sources: ChIP-chip data (chromatin immunoprecipitation on arrays), motif information (as obtained by phylogenetic shadowing) and microarray experiments. Combining these three types of "omics" data allows reconstructing transcriptional modules, i.e. the building blocks of the regulatory network. Our approach differs from previous work in that it takes the different data sources into account in a highly concurrent way and avoids sequential or iterative integration. By doing so, it allows correlating a set of regulators with their corresponding regulatory motifs and elicited profiles in a very intuitive, fast and direct way.

The core step of our algorithm is the seed construction which searches for all maximal gene sets that have a minimal number of regulators and motifs in common and that share a similar expression profile. A regulatory module is thus defined as such a maximal set of genes that meets the three requirements, together with the common motifs and regulators. Since the number of gene sets is exponentially large in the number of genes in the dataset, it is infeasible to detect all valid sets by an exhaustive search, even for the smallest genomes. To solve this combinatorial problem, we use hereditary constraints and rely on ideas similar to

those of the Apriori algorithm [2]. The statistical significance of the discovered seed modules is assessed.

Because it is rather conservative in recruiting genes (each of the genes in the module has to satisfy all three of the constraints), the seed construction step is followed by a seed extension step: additional genes with expression profiles that are highly correlated with the average profile of the seed are recruited. The required level of correlation with the seed profiles corresponds to the one for which the statistical overrepresentation of the seed motifs and seed regulators in the additionally recruited genes is most pronounced.
Application of our method on publicly available yeast data allowed demonstrating the biological relevance of the inference. Comparison of our results with literature revealed experimental evidence for many of our statistically significant seed modules. For modules of which no direct evidence existed so far, very often a plausible explanation for their composition could be inferred from literature and potential new links between the detected pathways and modules were derived.
In addition, we quantitatively showed significant overlap with the results of previously published modules discovery algorithms (SAMBA [3], GRAM [4]) and demonstrated improved performance of our algorithm over those algorithms.

[1] Lemmens K, Dhollander T, De Bie T, Monsieurs P, Engelen K, Smets B, Winderickx J, De Moor B, Marchal K (2006) Inferring transcriptional modules from ChIP-chip-, motif- and microarray data. Genome Biology, accepted for publication.
[2] Agrawal R, Imielinski T, Swami A (1993) Mining Association Rules between Sets of Items in Large Databases. Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp 207-216.
[3] Tanay A, Sharan R, Kupiec M, Shamir R (2004): Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proc Natl Acad Sci USA 101:2981-2986.
[4] Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, et al. (2003): Computational discovery of gene modules and regulatory networks. Nat Biotechnol 21:1337-1342.