17. Discovering regulatory modules from heterogeneous information sources.

Karen Lemmens [1] T. De Bie [1], P. Monsieurs [1], K. Engelen [1], B. De Moor [1], N. Cristianini [2], and K. Marchal [3]

[1] K.U.Leuven, ESAT-SCD, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium
[2] U.C.Davis Dept. of Statistics, 360 Kerr Hall, One Shields Ave, CA 95616, US
[3] K.U. Leuven, Department of Microbial and Molecular Systems and ESAT-SCD, Kasteelpark Arenberg 20, 3001 Leuven-Heverlee, Belgium

We present ReMoDiscovery, a combinatorial method for the inference of regulatory transcriptional modules from 3 independently obtained heterogeneous data sources (i.e. microarray data, ChIP-chip data and motif data). By applying our method to publicly available yeast data, we demonstrated its biological relevance.

Nowadays, data representative for different cellular processes are being generated at large scale. Based on these "omics" data, the action of the regulatory network that underlies the organism's behavior can be observed. Whereas until recently bioinformatics research was driven by the development of methods that deal with the analysis of each of these data sources separately, the focus is now shifting towards integrative approaches treating several data sources simultaneously.

Indeed, by analyzing each of these "omics" data separately, different aspects of the cellular adaptation are studied independently from each other. However, combining data allows gaining a more holistic insight into the network studied.
Besides an added biological value, the simultaneous analysis of coupled datasets also has a technical advantage. High-throughput experiments are characterized by a low signal/noise ratio. Moreover, because of technical and biological limitations, only a restricted number of independent experiments are available (technical replica's). The amount of information of a high-throughput experiment thus is limited. Coupled datasets provide information on the same biological system and combining them will thus increase the confidence in the final analysis result (higher specificity).

We present "ReMoDiscovery" [De Bie et al., 2005], an integrative method for inference of transcriptional modules from 3 independently acquired heterogeneous data sources: ChIP-chip data (chromatin immunoprecipitation on arrays) provide information on the direct physical interaction between a regulator and the upstream regions of its target genes; motif information as obtained by phylogenetic shadowing describes the DNA recognition sites of these regulators in the promoter regions of the target genes; and microarray experiments identify the expression behavior of the target genes in the conditions tested. Combining these 3 types of "omics" data allows reconstructing the structural composition of the basic building blocks of transcriptional networks i.e. transcriptional modules.

Our approach distinguishes itself from previous work in that most existing approaches exploit the availability of heterogeneous data sources in a sequential or an iterative way. The method takes the different data sources into account in a highly concurrent way. By doing so it allows correlating a set of regulators with their corresponding regulatory motifs and elicited profiles in a very natural and direct way.

Our methodology comprised 2 steps. The first step, the seed construction step, aims at finding maximal

gene sets that have a minimal number of regulators and motifs in common and that share a similar expression profile. A regulatory module comprises these sets of genes that meet the 3 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, identifying them with a naïve approach is prohibitive, even for the smallest genomes. To solve this combinatorial problem, we relied on ideas similar to those that laid the foundations for the Apriori algorithm, [Agrawal et al., 1993].

Because the seed construction method can be rather conservative in recruiting genes (each of the genes in the module has to satisfy all 3 of the constraints) we extend in a second step the seed with additional genes: genes of which the profile is correlated with the average profile of the seed are recruited. The number of additional genes recruited depends on the threshold for the required minimal correlation with the average seed profile. The correlation coefficient for which the statistical overrepresentation of the seed motifs and seed regulators in the additionally recruited genes is most pronounced, will be selected.

Using our method on publicly available yeast data allowed demonstrating the biological relevance of the inference, showing a very high specificity (low false positive detection rate). We applied the algorithm described above on the yeast cell cycle and stress datasets [Spellman et al., 1998 ; Gasch et al., 2000], the genome-wide location data performed by Harbison et al. (2004) and the motif data of Kellis et al. (2003). The identified regulatory modules are involved in a wide range of biological processes including cell cycle related processes, various stress responses, ribosome biogenesis.

References

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.

De Bie T, Monsieurs P, Engelen K, De Moor B, Cristianini N, Marchal K. (2005). Discovering transcriptional modules from motif, chip-chip and microarray data. Pac. Symp. Biocomput. 2005:483-494.

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell, 11:4241-4257.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA. (2004). Transcriptional regulatory code of a eukaryotic genome. Nature, 431:99-104.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature, 423: 241-254.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell, 9:3273-97.