BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 9 2008, pages 1208–1209
doi:10.1093/bioinformatics/btn094

Gene expression

ChIPCodis: mining complex regulatory systems in yeast by concurrent enrichment analysis of chip-on-chip data

Federico Abascal[1], Pedro Carmona-Saez[2], Jose-Maria Carazo[1] and Alberto Pascual-Montano[3,*]

[1]BioComputing Unit, National Center of Biotechnology (CSIC), [2]Integromics SL, Madrid and [3]Computer Architecture Department. Facultad de CC Físicas. Universidad Complutense de Madrid, Spain

Received on October 29, 2007; revised and accepted on March 6, 2008

Advance Access publication March 12, 2008

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Eukaryotic genes are often regulated by multiple transcription factors (TFs). Depending on the interactions among different TFs the expression of a gene can be tuned to respond to diverse environmental conditions. Chip-on-chip experiments provide a snapshot of which TF are *in vivo* bound to which genes in a particular condition, and have been applied to characterize the regulatory code of yeast under several experimental settings. ChIPCodis mines this data to provide new insights about how the expression of a particular group of genes is regulated. For a given list of yeast genes ChIPCodis determines which combinations of TFs are significantly over-represented in a series of environmental conditions.

Availabilty: http://chipcodis.dacya.ucm.es

Contact: pascual@fis.ucm.es

Supplementary information: http://chipcodis.dacya.ucm.es/sm.html

1 INTRODUCTION

The regulation of gene expression determines which of the instructions contained in the DNA are transcribed and, possibly, executed. Eukaryotic genes are often regulated by multiple TFs (Lee *et al*., 2002; Lemon and Tjian, 2000), and the repertoire of TFs is large (~200 in yeast; 1500–2000 in humans) The underlying possible combinations and interactions between TFs are immense, making organisms able to respond to diverse environmental conditions and to finely coordinate complex developmental processes.

High-throughput experimental approaches, as well as *in silico* methods, are providing lots of data related to the regulatory code of eukaryotic genomes (Hu *et al*., 2007). A good example of these techniques is the genome-wide chip-on-chip experiments that are able to determine the genes to which a TF is bound in a particular cellular state. Remarkably, this technique has been applied to the complete repertoire of yeast TFs, under diverse environmental conditions (Harbison *et al*., 2004).

Most often, regulatory systems have been studied from a global perspective, focusing on finding relevant TF combinations (Pilpel *et al*., 2001), unveiling regulatory modules (Segal *et al*., 2003), or analysing the dynamic nature of the system (Ernst *et al*., 2007). In contrast to global analyses, some approaches are focused on the analysis of particular lists of genes (e.g. co-expressed genes). Most of these tools evaluate the enrichment of individual TFs (Backes *et al*., 2007). Nevertheless, in the last few years some approaches have been developed to address the combinatorial nature of transcription regulation. The Opossum web server reduces the computational complexity of finding enriched combinations of TFs by grouping into classes those TFs that have similar binding motifs. In a next step, the statistical relevance of combinations of pairs of TFs is calculated (Huang *et al*., 2006). The Composite Module Analyst (CMA) web-server identifies combinations of TF motifs (composite modules) that are likely to act in conjunction (Waleev *et al*., 2006). The main focus of CMA is the characterization of the structure of promoters. Both approaches are based on sequence information [transcription factor binding site (TFBS)], not considering information related to the condition dependent binding of TFs.

Here we present the ChIPCodis web-server. Based on available chip-on-chip (Harbison *et al*., 2004) and TFBS data (MacIsaac *et al*., 2006) in yeast, ChIPCodis aims at discovering which combinations of TFs (not restricted to pairs) are statistically over-represented in a list of genes. In addition, since the Harbison *et al* chip-on-chip data contain *in vivo* information of which TFs are bound to which genes under a series of conditions, ChipCodis is able to search for TF co-occurrences that are condition dependent, yielding results of the type 'binding of TFs $X1, \ldots, Xn$ to the user provided genes $G1, \ldots, Gn$ is significantly over-represented under condition C'.

2 ChIPCodis

ChIPCodis is oriented to the analysis of particular lists of genes, currently restricted to yeast genes. Typically such a list will contain co-expressed genes. Alternatively, it could contain the group of genes involved in a particular biological process (e.g. the tricarboxylic acid cycle). The Association Rules Discovery (ARD) data mining technique is applied to identify combinations of TFs that frequently co-occur in the input list of genes (similarly to Carmona-Saez *et al*., 2007). The statistical significance of these associations is evaluated by comparing their frequency in the user list of genes with the background frequency. To calculate the corresponding P-values, the χ^2 test of independence and the hypergeometric distribution are available. The obtained P-values can be corrected either through simulations (similarly to Boyle *et al*., 2004) or with the False Discovery Rate method (see Supplementary Material).

*To whom correspondence should be addressed.

Enriched TF combinations can be searched in two different groups of datasets, corresponding to the analyses of (Harbison *et al*., 2004; MacIsaac *et al*., 2006). Under the 'Harbison *et al*. 203 TFs' group there are three datasets: 'Growth under rich medium', 'All available conditions' and 'Compiled'. The last two are likely to be the most interesting ones. The 'Conditions' dataset contains the dynamic information about which TFs bind which genes under each of the 14 experimental conditions tested. In the 'Compiled' dataset a TF is said to be bound to a gene if it binds that gene in at least one of the conditions. The MacIsaac *et al*. datasets are described in the Supplementary Material.

3 EXAMPLE

Firstly we conducted a general analysis of yeast expression to test the utility of ChIPCodis. In Gasch and Eisen (2002), the authors analysed the conditional regulation of gene expression in yeast. Ninety-nine clusters of genes with similar expression patterns were identified with a fuzzy clustering method. A relaxable membership threshold (k) allowed these clusters to include genes with less similar expression patterns. We analysed these 91 clusters at different k under the 'Harbison-compiled' and 'Harbison-conditions' datasets. These results are available as Supplementary Material.

The following example demonstrates the utility of ChIPCodis. Cluster #39 contains 56 genes at $k<0.06$. The analysis under the 'Harbison-conditions' dataset indicates that significantly enriched TF combinations identified by ChIPCodis are associated mainly with the nutrient deprived medium (RAPA), but also with the amino acid starvation (SM), and growth in rich media (YPD). A heatmap of enriched TF combinations versus genes shows that there are two main groups of genes as well as two main groups of TF combinations and conditions (see Supplementary Material). One group contains the genes that are active at RAPA, which are regulated by combinations of the following TFs: Dal81/2, Gat1, Gln3 and HAP2. The second group contains the genes regulated under the SM and YPD conditions, and is characterized by the Met4/31/32 and Cbf1 TFs. This second group of genes is related to the Methionine/Sulfur metabolism, also including the urea transporter *dur3*. The first group of genes, the ones regulated under RAPA, are related to the catabolism of allantoin, but also include amino acid and ammonium transporters, asparaginases and TFs, all of which can be associated with nitrogen rescue pathways. It is clear that such pathways must be important under a nutrient deprived medium such as RAPA. Interestingly, since the general function of most of the genes in this group is much conserved, it is probable that the two uncharacterized membrane proteins YDR090C and YGR125W, which are similarly regulated, are also involved in the acquirement of nitrogen. In summary, ChIPCodis allowed to determine how these 56 co-expressed genes are regulated. Two main types of regulation were found inside this cluster, linked to particular combinations of TFs and environmental conditions.

4 DISCUSSION

Understanding the programmes of gene regulation is an open and complex problem, mainly due to the combinatorial and dynamical nature of transcriptional regulatory networks. The ARD data mining technique combined with the subsequent statistical assessment of the identified TF associations aims to face up this combinatorial problem. On the other hand, the dynamic properties of the regulatory network can be modelled using chip-on-chip data obtained under different conditions, as in (Harbison *et al*., 2004). It is important to keep in mind that, although the complete repertoire of yeast TFs (203) was analysed under growth in rich medium, only a reduced set of yeast TFs was evaluated under a restricted number of environmental conditions (84 TFs were analysed in at least one of 13 environmental conditions), hence representing a partial picture of the regulatory network dynamics of yeast.

Nowadays, attempts to reconstruct the complete regulatory network of a eukaryote have been restricted to the model organism *Saccharomyces cerevisiae*. Hopefully, in the near future, regulatory networks of higher eukaryotes will also be described. In parallel, ChIPCodis will be updated to include forthcoming data.

ACKNOWLEDGEMENTS

The authors thank Rubén Nogales, César Vicente and Enrique de la Torre for their technical support. The insightful comments of three anonymous reviewers are also acknowledged. This work has been partially funded by the Spanish grants BIO 2007-67150-C03-02, S-Gen-0166/2006, TIN2005-5619, PR27/05-13964-BSCH, CSD2006-00023 and CNIT-OCNOSIS. FA is a recipient of an I3P contract of the Spanish MEC. APM acknowledges the support of the Ramón y Cajal program.

Conflict of Interest: none declared.

REFERENCES

Backes,C. *et al*. (2007) GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Res*., **35**, W186–W192.

Boyle,E.I. *et al*. (2004) GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.

Carmona-Saez,P. *et al*. (2007) GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol*., **8**, R3.

Ernst,J. *et al*. (2007) Reconstructing dynamic regulatory maps. *Mol. Syst. Biol*., **3**, 74.

Gasch,A.P. and Eisen,M.B. (2002) Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. *Genome Biol*., **3**, RESEARCH0059.

Harbison,C.T. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Hu,Z. *et al*. (2007) Genetic reconstruction of a functional transcriptional regulatory network. *Nat. Genet*., **39**, 683–687.

Huang,S. *et al*. (2006) Identification of over-represented combinations of transcription factor binding sites in sets of co-expressed genes. In Jiang,T. *et al*. (eds) *Advances in Bioinformatics and Computational Biology*, Vol. 3, Imperial College Press, London, UK.

Lee,T.I. *et al*. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.

Lemon,B. and Tjian,R. (2000) Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev*., **14**, 2551–2569.

MacIsaac,K.D. *et al*. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 113.

Pilpel,Y. *et al*. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet*., **29**, 153–159.

Segal,E. *et al*. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet*., **34**, 166–176.

Waleev,T. *et al*. (2006) Composite Module Analyst: identification of transcription factor binding site combinations using genetic algorithm. *Nucleic Acids Res*., **34**, W541–W545.