Genome-wide Analysis of Functions Regulated by Sets of Transcription Factors

Szymon M. Kiełbasa*[1], Nils Blüthgen*, Hanspeter Herzel

Humboldt University, Institute for Theoretical Biology,

Invalidenstraße 43, 10115 Berlin, Germany

[1] s.kielbasa@itb.biologie.hu-berlin.de

* both authors contributed equally

Abstract: We present a pipeline for inferring biological functions regulated by a combinatorial interaction of transcription factors. Using a robust statistical method the pipeline intersects the presence of transcription factor binding sites in gene upstream sequences with Gene Ontology terms associated with these genes. Positional frequency matrices for the transcription factors constitute the input of the pipeline and significantly enriched biological processes are reported as the output.

We demonstrate the usage of the pipeline using two groups of transcription factors: a cell-cycle related family of E2F factors and a NFAT/AP-1 pair involved in immune response. In both cases the reported results match well the experimental knowledge. Furthermore, for the NFAT/AP-1 composite element novel functions are predicted.

INTRODUCTION

Transcriptional regulation of gene expression is a main mechanism governing the temporal and spatial organization of biological processes in eukaryotic organisms. External and internal stimuli are transduced into activities of transcription factors controlling major biological processes within the cell. Single factors controlling expression in prokaryotic cells evolved into a sophisticated regulatory machinery in eukaryotes, involving the combinatorial action of transcription factors. This phenomenon gives rise to enormously complex responses equipping cells with means to respond adequately to surrounding information [FMZ+02]. Considering the importance of transcriptional regulation for most biological processes, understanding of the regulatory network is a major challenge in the postgenomic era. Automated inference of the regulatory network could result in predictive models of interactions among genes, that allow a better understanding of gene expression patterns and could hint to targets for drug design.

The fundamental building block of the regulatory network is the binding of transcription factors to regulatory cis-sequences located in the promoter regions of genes. Experimentally identified cis-sequences for a single transcription factor can be aligned to find the sequence recognized by the factor. Alignments of such cis-sequences yield positional frequency matrices, which are further used to model the binding affinity of the factor towards

DNA sequences [St98]. An in-vitro method, SELEX-SAGE, allows to discover potential binding sites for transcription factors in vitro and to build positional frequency matrices without knowing their target genes. For numerous transcription factors corresponding frequency matrices have been constructed (e.g. [HWR⁺98]) for the prediction of transcriptional regulation. Despite advances in this field genome-wide scans of binding sites are difficult to interpret since false, nonfunctional predictions dominate. [WK03] estimate that a simple search using only a single frequency matrix results in only one functional site per 1000 predictions. More recent solutions of this problem take into account further biological properties, like clustering of the functional binding sites [FLW03] and possible conservation of cis-sequences in evolution [DWR⁺03]. These approaches can reduce the number of non-functional predictions by about two orders of magnitude [WK03].

There are attempts [WF98, KW01, KMIWK02] to use binding sites predicted in a gene upstream sequence to predict a function of the gene under consideration. This approach originates from the concept that genes sharing the same regulatory elements carry out also the same biological function. Thus it is limited to extend functional families with already well known regulatory properties. New functions of transcription factors (possibly in another cellular context, stage of development, tissue, etc.) are unlikely to be inferred.

Utilizing the growing systematic functional annotation provided by the Gene Ontology [ABB⁺00] we propose a novel, more function oriented, genome-wide analysis to predict the biological function of transcription factors. Our approach starts from transcription factor frequency matrices and predicts corresponding target genes using established methods. As stated above, many of these predictions are likely to be false. We assume that these false targets are scattered randomly over the biological processes. Contrarily, the true functional predictions are likely to be involved in a few, rather specific biological processes. Consequently we associate significantly enriched biological processes in the predicted target genes as biological functions of the transcription factors under consideration. This approach allows to discover novel regulatory functions without bias towards known functions of a transcription factor. In the following we discuss the details of our prediction pipeline and apply our pipeline to a single transcription factor family (E2F) and to a set containing the transcription factors NFAT and AP-1.

MATERIALS AND METHODS

Prediction Pipeline

We have constructed a pipeline to predict biological functions controlled by a set of transcription factors, as shown in Fig. 1. A set of positional frequency matrices is used in the Cluster-Buster program [FLW03] to predict clusters of binding sites in upstream regions of human genes. We assume, that genes having such clusters in their upstream sequences are potentially regulated by the factors. These genes are annotated with terms describing biological processes from the Gene Ontology [ABB⁺00]. Subsequently we identify biological processes which are significantly associated with the potential target genes with

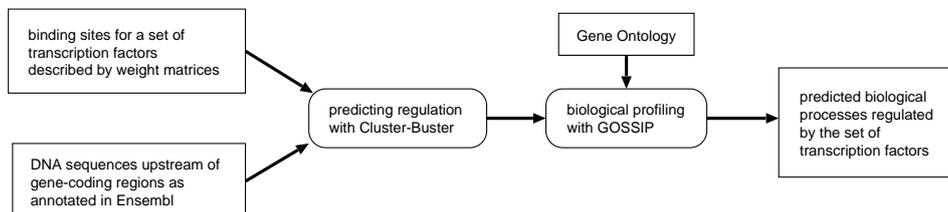GOSSIP [BBC$^+$04]. Details of these steps of the pipeline are discussed below.

Figure 1: Data processing in our pipeline: 1,000 bp upstream regions of 16,032 human genes predicted by Ensembl were extracted (utilizing HomGL [BKCH04]) giving a collection of 15,362 potential promoter regions. These sequences together with positional frequency matrices of transcription factors binding sites were analyzed by Cluster-Buster to find a subset of genes which are potentially regulated by these factors. Subsequently, GOSSIP [BBC$^+$04] utilizes the Gene Ontology to find biological processes associated with this subset. Finally, significant processes are mapped to the hierarchy of the Gene Ontology graph.

Positional frequency matrices

As an example of our method, we have studied two sets of transcription factors. The first set contains only a single positional frequency matrix corresponding to the E2F family of transcription factors. It is taken from the Transfac database [HWR$^+$98] with the accession number M00427 (V$E2F_Q6). As a second set we use the matrices describing sites recognized by AP-1 and NF-AT, as published by Kel et al. [KKMBW99].

Upstream Regions

The human part of the UniGene database partitions human transcript sequences into 118,517 UniGene clusters derived from both known genes and ESTs. We could map 16,032 of the clusters to genes reported by the Ensembl database [CEA$^+$04] by using HomGL [BKCH04]. For these genes we extracted 1,000 bp long DNA sequence upstream of the predicted transcription start sites of the genes. We found, that only 15,362 of these sequences were unique, and we chose them as predicted promoter regions, since the multiple occurrence of the same promoter sequence could bias the further analysis of enriched processes.

Predicting genes regulated by transcription factors

Computational prediction of genes regulated by a single transcription factor results in many false positives. Eukaryotic regulation is typically achieved by several regulators, and predictions of clusters of several binding sites for one or more transcription factors may reduce the rate of false predictions by more than an order of magnitude. Therefore we assume here, that a gene is regulated by factors under study if a cluster of several binding sites corresponding to some of the factors is predicted in the gene upstream region. We selected the publicly available Cluster-Buster program [FLW03] to perform the task of cluster searching. The program takes all gene upstream sequences and a set of matrices as input and returns a list of probable locations of functional clusters in these sequences. To keep our pipeline simple we used the Cluster-Buster program with all parameters set to their default values.

Functional profiling

The Gene Ontology [ABB$^+$00] specifies a controlled, hierarchical vocabulary for annotating genes. It can be represented by a directed acyclic graph. We utilize the Gene Ontology (GO) to test whether a biological process is significantly associated with a group of genes predicted to be regulated by the set of transcription factors under consideration. This analysis requires four data sources: The group of potentially regulated genes as the test group, all genes with upstream regions as a reference group, GO annotations for these both gene sets, and the Gene Ontology.

We annotate the groups of genes with HomGL [BKCH04]. Annotations are usually given as terms close to the leafs of the graph implying a series of more general annotations upward in the GO graph. For each term in the ontology we test whether this particular term is enriched in the test group as compared to the reference group. To do this we categorize each gene in two distinct ways: first, whether it is annotated with the term under consideration or not, and second, whether it belongs to the test group or not. We apply Fisher's exact test that allows to detect and quantify associations between the two distinct categorizations. But since the number of terms in the Gene Ontology is very high, problems arising from multiple testing cannot be left aside. We adjust the p-values to control the number of false positives by calculating the false discovery rate (FDR), which quantifies the ratio of the expected number of false positives among the positives. For this specific problem it can be calculated exactly [BBC$^+$04]. Assuming, that incorrectly predicted target genes by the prior step of our pipeline do not favor any processes but scatter randomly, the FDR of our pipeline predictions is controlled reliably. In the further analysis, we apply a threshold of $FDR \leq 0.05$ that keeps the expected ratio of false predictions under five percent. The analysis is been performed by the software package GOSSIP [BBC$^+$04].
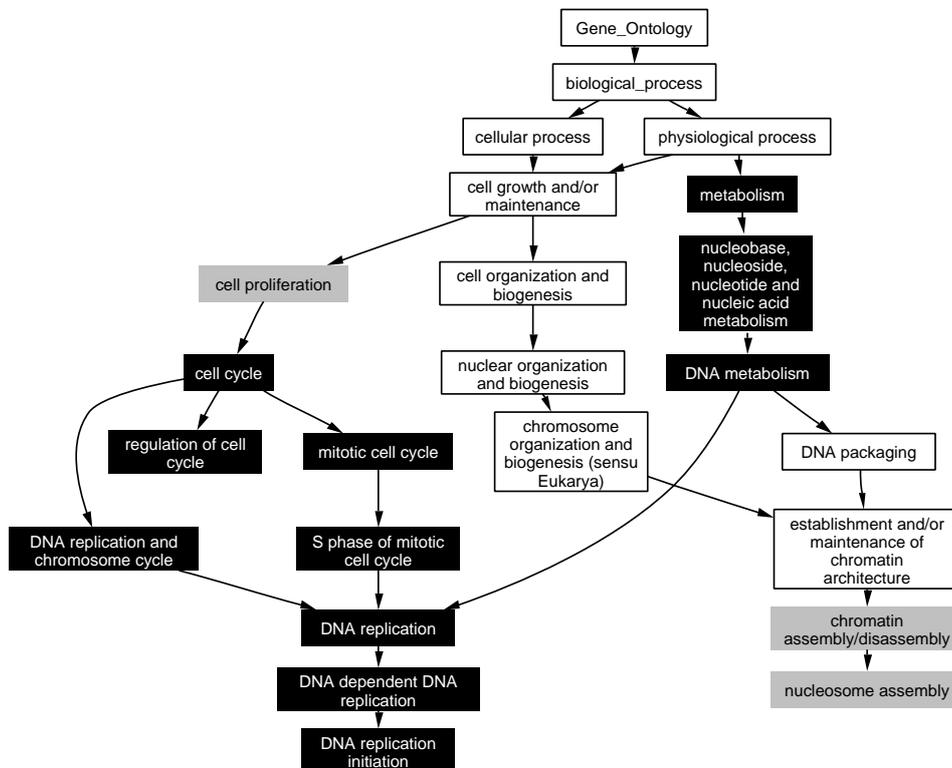
Figure 2: A part of the Gene Ontology highlighting the biological processes that are significantly enriched in 273 genes with a cluster of E2F-binding sites predicted in their upstream regions . Black boxes show those processes with $FDR \leq 0.01$, gray boxes with $FDR \leq 0.05$ (see Methods for details).

RESULTS

Processes regulated by E2F

E2F's are transcription factors known to be involved in the regulation of the S-phase of the mitotic cell cycle [HS97]. We choose this transcription factor to test whether our pipeline can predict the function regulated by a single family of transcription factor. The Cluster-Buster program predicted 273 genes to contain a cluster of binding sites for E2F. In these 273 target genes, many biological processes related to the cell cycle are significantly enriched (see Figure 2). The more specific processes are related to DNA replication and S-phase of the mitotic cell cycle. Therefore the prediction of our pipeline yields exactly what we know from experiments: E2F is regulating the synthesis phase of the cell cycle.
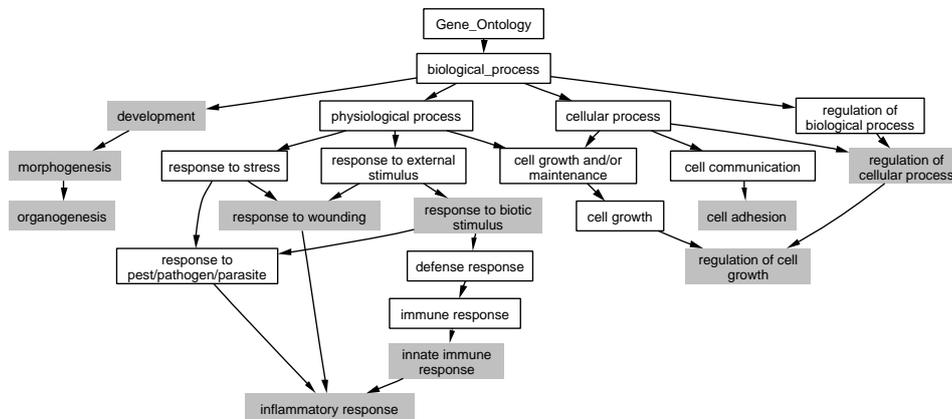
Figure 3: A part of the Gene Ontology highlighting the biological processes that are significantly enriched in 273 genes with a cluster of NFAT and AP-1 binding sites predicted in their upstream regions. Grey boxes display those processes with $FDR \leq 0.05$ (see Methods for details).

Combinatorial regulation by AP-1 and NFAT

Kel et al. [KKMBW99] have studied the composite binding of NFAT and AP-1 transcription factors. NFAT plays a role in the regulation of cytokines and other genes during immune response. NFAT co-operates with AP-1 to integrate calcium and PKC-signal transduction. They concluded that the combination of NFAT and AP-1 exhibits a significantly higher specificity than individual factors towards genes induced upon T-cell activation.

We applied our pipeline to each of these matrices separately and found no significantly enriched process within the predicted targets of individual factors (NFAT: 2716 predicted target genes – no process significant, AP-1: 921 predicted target genes – no process significant). However, studying both factors together we found significantly enriched biological processes within the predicted 1913 target genes (see Fig. 3). Many of these overrepresented biological processes are related to immune response: inflammatory response, response to wounding, innate immune response, response to biotic stimulus which is in good agreement with the results of Ref. [KKMBW99]. Moreover we found the processes organogenesis, morphogenesis, development, regulation of cell growth, regulation of cellular process, and cell adhesion as novel processes which might be regulated by the combinatorial action of NFAT and AP-1.

DISCUSSION

The availability of whole-genome sequences and the growing systematic annotations like the Gene Ontology provide the means for more function oriented data mining that goes beyond the single gene level. In this article we propose an approach, which opens a door

to infer a biological function regulated by the combinatorial interaction of transcription factors. Contrary to other widespread techniques our method does not intend to predict which factors control genes of similar expression profile. Instead our search only requires a set of positional frequency matrices representing a set of transcription factors to predict in silico their biological function. First, using Cluster-Buster [FLW03] we predict a list of potential target genes for a set of transcription factors. Afterwards, a rigorous statistical test for over-representation implemented in GOSSIP [BBC$^+$04] is applied to all biological processes provided by the Gene Ontology. Therefore the search is not biased by any prior knowledge related to the factors and gives a chance to detect novel regulatory associations.

Additionally the pipeline provides a list of genes supporting the evidence of the reported enriched processes. This gene list can be understood as the cross section of the list of genes regulated by the studied factors and a list of genes annotated with at least one of the overrepresented terms. Due to the filtering nature of the cross section, the final gene list is expected to have less false predictions than the primary list of potentially regulated genes.

Two well studied examples presented in this paper illustrate the pipeline's utility. As expected, the clusters of sites related to the E2F transcription factors family are significantly enriched with terms related to the synthesis phase of the mitotic cell cycle. These terms have been detected despite the fact, that only clusters of binding sites corresponding to a single positional frequency matrix have been searched. Although separate studies of AP-1 and NFAT show no significant biological processes, the combinatorial interaction of NFAT and AP-1 yields several significant terms. Besides terms describing aspects of immune response where NFAT/AP-1 is known to be involved in, several significant terms hint to novel hypothesis that require experimental validation. Notably, no tuning of parameters was necessary within these studies.

Our method could be further improved by a better selection of promoters, for example by using databases like dbTSS [SYSN04] and EPD [SPD$^+$04]. Our analysis does not include distal enhancer regions, but might be included when sufficient knowledge about enhancer regions has been accumulated.

We show in this paper, that the tools are now in hand to infer the regulatory complexity using genome-wide studies of transcription factor binding sites. In forthcoming studies we plan to investigate systematically combinations of transcription factors to get a better understanding of the combinatorial regulation in eukaryotic organisms.

ACKNOWLEDGMENTS

References

[ABB+00]   Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., and Sherlock, G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 25(1):25–29. 2000.

[BBC+04]   Blüthgen, N., Brand, K., Cajavec, B., Swat, M., Herzel, H., and Beule, D.: Biological profiling utilizing gene ontology. *submitted*. 2004.

[BKCH04]   Blüthgen, N., Kielbasa, S., Cajavec, B., and Herzel, H.:   HOMGL-comparing genelists across species and with different accession numbers. *Bioinformatics*. 20(1):125–6. 2004.

[CEA+04]   Curwen, V., Eyras, E., Andrews, T., Clarke, L., Mongin, E., Searle, S., and Clamp, M.: The Ensembl automatic gene annotation system. *Genome Res*. 14(5):942–50. 2004.

[DWR+03]   Dieterich, C., Wang, H., Rateitschak, K., Luz, H., and Vingron, M.: CORG: a database for COmparative Regulatory Genomics. *Nucleic Acids Res*. 31(1):55–7. 2003.

[FLW03]   Frith, M., Li, M., and Weng, Z.: Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res*. 31(13):3666–8. 2003.

[FMZ+02]   Fessele, S., Maier, H., Zischek, C., Nelson, P., and Werner, T.: Regulatory context is a crucial part of gene function. *Trends Genet*. 18(2):60–3. 2002.

[HS97]   Herwig, S. and Strauss, M.: The retinoblastoma protein: a master regulator of cell cycle, differentiation and apoptosis. *Eur J Biochem*. 246(3):581–601. 1997.

[HWR+98]   Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A., Kel, O., Ignatieva, E., Ananko, E., Podkolodnaya, O., Kolpakov, F., Podkolodny, N., and Kolchanov, N.: Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res*. 26(1):362–7. 1998.

[KKMBW99] Kel, A., Kel-Margoulis, O., Babenko, V., and Wingender, E.:   Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. *J Mol Biol*. 288(3):353–76. 1999.

[KMIWK02] Kel-Margoulis, O., Ivanova, T., Wingender, E., and Kel, A.: Automatic annotation of genomic regulatory sequences by searching for composite clusters. *Pac Symp Biocomput*. S. 187–98. 2002.

[KW01]   Krivan, W. and Wasserman, W.: A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res*. 11(9):1559–66. 2001.

[SPD+04]   Schmid, C., Praz, V., Delorenzi, M., Perier, R., and Bucher, P.: The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. *Nucleic Acids Res*. 32 Database issue:D82–5. 2004.

[St98]   Stormo, G.: Information content and free energy in DNA–protein interactions. *J Theor Biol*. 195(1):135–7. 1998.

[SYSN04]   Suzuki, Y., Yamashita, R., Sugano, S., and Nakai, K.: DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Res*. 32 Database issue:D78–81. 2004.

[WF98]   Wasserman, W. and Fickett, J.: Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol*. 278(1):167–81. 1998.

[WK03]   Wasserman, W. and Krivan, W.: In silico identification of metazoan transcriptional regulatory regions. *Naturwissenschaften*. 90(4):156–66. 2003.