



## Finding Regulatory Elements from Microarray Data

**Ben-Zaken Zilberstein, C. and Geiger, D.**  
Department of Computer Science, Technion

Microarrays yield enormous amounts of data about gene expression patterns in different cells of different species. The mechanism underlying a particular expression pattern is usually not yet known. Therefore, the development of tools that could reveal regulatory elements involved in an anomalous gene expression pattern is important in understanding the mechanisms underlying pathologies such as human cancer. Since our knowledge about the human genome is not yet complete, it is more convenient to develop these tools by investigating a well known system like the yeast genome, as we did herein. Regulatory elements involved in a certain expression pattern could be revealed by comparing the promoters of a co-regulated group of genes in order to find common sequence patterns differentiating this group from the others. This can be done by local multiple alignment of these promoters in order to detect common consensus sequences (Church et al, 2000) or by detection of over-represented oligonucleotide sequences (van Helden et al, 1998). Both analysis methods detect new candidate regulatory sites as well as sites that have already been characterized. We developed a computational tool to reliably predict short oligomers as regulatory elements by examining the promoters of a co-regulated group of genes.

**Results** We used microarray data containing the expression levels of all yeast genes 15 minutes after the introduction of rapamycin into the cell media, which closely simulate starvation of the cell (Hardwick et al, 1999).

We clustered together a group of genes known to be involved in the amino-acid metabolism that were also at least two fold over expressed 15 minutes after the rapamycin was introduced. We hypothesized that the genes in the over-expressed group share common regulatory elements and we searched this group for any such over-represented oligomers. This was accomplished by the comparison of two frequencies for each candidate oligomer. The first was the frequency of this oligomer in the promoters of the over-expressed genes and the second was its frequency in a control group of 1500 randomly chosen promoters. The higher the ratio between these two frequencies, the higher the chance that a certain oligomer has a biological role which may have caused the over-expression. In order to derive a cutoff for the ratio values to help assess their significance, we performed a simulation calculating this ratio for a randomly chosen group of the same size as that of the over-expressed genes. We repeated this simulation 20 times and used the highest ratio value obtained as the cutoff. When analyzing putative





over-represented oligomers, only those exhibiting ratios much higher than the cutoff were selected as regulatory elements. We identified the GCN4 binding site with an extremely high frequency ratio of 10.00 whereas the cutoff was 5.00. GCN4 is a master gene involved in amino-acid metabolism and has already been reported as central in the over-expression of amino-acid metabolism genes during cell starvation (Marton MJ, 2001). All yeast genes that are well known to be regulated by GCN4 (e.g. HIS4, TRP4, ILV1, ILV2, ADE4, ARG1, ARG8, HIS3). were also over-expressed as expected. Another regulatory element we isolated was the CPF1 binding site. CPF1 is involved in the regulation of methionine synthesis. All genes that are regulated by this element (e.g. PGK1, CYT1, and MET16) were also over expressed as expected. We performed the same analysis on a group of genes that participate in the carbohydrate metabolism, and identified oligomers that exhibited ratios far above the cutoff (e.g. GGGGACT, GGGGG, GGGGC and CGCGC). No literature was found regarding those oligomers. We therefore propose them as regulatory elements of the carbohydrate metabolism during cell starvation.

**Discussion** Our method has succeeded in identifying some important regulatory elements involved in the process of gene over-expression during cell starvation by analysis of data derived from a microarray. This enables us to throw light on the regulatory mechanism underlying the over-expression phenomenon by suggesting some of the elements involved. Our encouraging results are due to two aspects: First, we focus on genes that react quickly to a certain

change (e.g. elevation of rapamycin in the cell media) because this suggests they are responding to a relatively specific common cause. This common cause is likely to co-occur with a small over-represented group of regulatory elements suggesting that our method is effective in these cases. Second, we use statistical simulation to derive a cutoff value, reducing the number of false-positives to below 5%, making our method very reliable.

## References

- [1] Hardwick J.S., Kuruvilla F.G., Tong J.K., Shamji A.F. and Schreiber S.L.  
Rapamycin-modulated transcription defines the subset of nutrient-sensitive signaling pathways directly controlled by the Tor proteins. *Proc. Natl. Acad. Sci USA*, 1999 Dec 21;96.
- [2] Hughes J.D., Estep P.W., Tavazoie S., Church G.M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J.Mol.Biol.* 2000, Mar 10;296(5):1205-14.
- [3] Natarajan K., Meyer M.R., Jackson B.M., Slade D., Roberts C. Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. *Hinnebusch A.G., Marton M.J. J. Mol. Cell Biology* 2001, 21(13):4347-68.
- [4] van Helden J., Andre B., Collado-Vides J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J.Mol. Biology* 1998, 281(5):827-42.

e-mail: [chaya@cs.technion.ac.il](mailto:chaya@cs.technion.ac.il),

e-mail: [dang@cs.technion.ac.il](mailto:dang@cs.technion.ac.il)