Research Focus

Yeast expression-array analysis goes molecular

Thomas Werner

GSF-Research Center for Environment and Health, Institute for Experimental Genetics, Ingolstädter Landstrasse 1, D-85764 Neuherberg & Genomatix Software GmbH, Landsbergerstr. 6, D-80339 München, Germany

Combining expression-array analysis with molecular mechanisms of transcription control is an approach that is still in its infancy. Currently, the best understood eukaryotic organism is *Saccharomyces cerevisiae*, and it is with the genome of this organism that most progress in molecular array analysis has been achieved. Molecular analysis, in the form of *de novo* motif detection, has recently been combined directly with expression-level analysis, bypassing clustering solely by statistics of expression levels or by restricting analysis to known transcription factor binding sites. This has identified several sets of transcription factors and their target genes, complementing similar approaches. These studies might significantly enhance similar analyses for human and mouse expression arrays, demonstrating the huge potential of integrated sequence and expression-level analysis.

The use of expression-array analyses gained momentum following the arrival of full genomic sequences of eukaryotic organisms, starting with *Saccharomyces cerevisiae* in 1996 and reaching an initial peak with the release of the human genome sequence draft in 2000 (published in 2001 [1]). This provided whole genomic sequences and a genome-wide high-throughput method to study gene expression [2]. However, it quickly became clear that there was a wide gap between the expression data and their interpretation in functional terms, with no direct link to the genomic sequence. There is, of course, a link to cDNA, but not in the genomic (e.g. regulatory) context. Cluster analysis of corresponding cDNAs from the microarrays, with respect to expression level or even change in expression level in response to treatment of cells, often left the majority of the data without interpretation. Most clusters contained genes that did not make immediate sense as coexpressed groups. There are several reasons for this frustrating experience. First, it is impossible to look at cellular processes such as the transcription of specific functionally related genes as isolated events. The picture we get from microarray experiments is like an aerial view of the morning traffic in a megacity, with hundreds or thousands of different traffic streams intermingled, rather than an orderly entourage of vehicles moving towards one destination, each clearly separable from the other traffic. Second, gene expression and its changes are governed to a large extent by gene promoters and other regulatory regions but not by the cDNAs used as tags for identification of the genes, which requires analysis of expression levels in relation to promoters, not cDNAs [3]. Consequently, it is imperative that we focus on regulatory sequence analysis (e.g. promoters) to determine why genes are co-regulated. Analysis of promoters rather than cDNAs allows the study of the molecular basis of transcription. However, promoters must be grouped in a meaningful manner before analysis, to distinguish the many simultaneous but distinct events occurring in every microarray experiment [4]. Promoter analysis itself then usually focuses on transcription factor binding sites, which are intricately connected with transcription control in virtually all organisms [5].

Such studies are most advanced in *S. cerevisiae* owing to the availability and almost complete annotation of its genome. Finding the physical location of yeast promoters is simple because most are located upstream of the open reading frame (ORF), with only a short untranslated 5′ region (5′-UTR) between the coding region and the promoter. Consequently, yeast promoters have been used for genome-wide analyses in several studies [6].

However, there is an inherent problem in analyses based solely on statistics of expression levels (or differences), and in focusing on predefined binding sites. Transcription factors can, in principle, bind many more sequences than they act on functionally. In a pure *in silico* approach starting with transcription factor binding sites this introduces a significant background, which cannot be avoided by computational means alone. By contrast, statistical clustering of genes by expression data cannot distinguish between genes co-regulated by a common factor or coexpressed by different mechanisms. The unrelated genes detected by both methods jeopardize biologically meaningful pattern detection and/or correlation in such sets of genes because there is no consistent pattern present.

New approaches for integrated yeast expression-array analysis

Wang *et al.* recently demonstrated a powerful and elegant approach to tackling the problem of gene grouping and promoter analysis, using microarray expression data in several ways [7]. They linked the expression data directly to the pattern definition process, first by selecting genes with a similar transcription factor core motif based on the REDUCER program (published previously by Bussemaker *et al.*, who developed this method for *in silico* pattern detection based on expression data correlation [8]). After the initial motif definition by REDUCER, Wang *et al.* identified genes on the microarray that were likely to be regulated by the transcription factor corresponding to the

Corresponding author: Thomas Werner (Werner@genomatix.de).

initial motif, based on expression profiles. They identified the genes by the presence of the motif in the promoter of candidate genes. New matches were weighted by the fit of the expression level of that gene and were used to enhance the motif (a gene containing a motif from a high-expression group that is expressed only at a low level is probably not regulated by the transcription factor specific for that motif). The subset of activated transcriptional modules (the combination of a transcription factor and its target genes) was then estimated by statistical analysis of expression ratio and frequency of the particular motif in the promoters. Finally, these results were checked against data from expression experiments in which a specific transcription factor was either absent or overexpressed. Wang *et al.* could therefore prevent genes that contained potential, but functionally irrelevant, binding sites from affecting the results, using the pattern definition to exclude genes that were coexpressed by other means. They identified several known transcription factors and their binding sites, as well as biologically verified target genes, with this approach, opening the way to a general understanding of transcriptional networks in yeast on a molecular level.

The work by Wang *et al.* was not the first attempt to integrate transcription factor binding-site analysis with interpretation of yeast expression data. Pilpel *et al.* had already attempted to elucidate the molecular connection of transcription control in yeast by starting from known binding-site motifs and finding all yeast genes that had a particular binding site and were expressed in a similar manner [9]. They then identified the promoters that contained synergistic pairs of any two binding sites. Synergism was defined as a better coherence between genes containing such pairs in their promoters and their expression patterns compared with that of genes containing only one of the motifs in the promoter.

Zhu *et al.* based their approach on the idea that transcription factors should be expressed themselves in a similar way to their target genes and thus could be used as tags to find those target genes [10]. They took transcription factors and clustered genes around them by expression level. They then applied motif-finding programs (Gibbs Sampler) to the promoters of these potential target genes. They were often able to detect the binding sites of the respective transcription factor in the promoters of genes clustered around it in expression space.

Lee *et al.* have taken expression analysis in *S. cerevisiae* one step further, to the automatic reconstruction of regulatory networks [11]. They combined genome-wide chromatin immune precipitation (ChIP), which directly identifies proteins bound to promoter regions *in vivo*, with expression data to identify what they called network motifs such as autoregulatory loops (in which a transcription factor binds its own promoter) and closed loops containing several factors (e.g. A binds B, B binds C, C binds A). They then used such network motifs to reconstruct automatically a network regulating the cell cycle based solely on cell-cycle-associated expression data.

However, as powerful as these approaches are, they all focus primarily on connections with and between known transcription factors or transcription factor binding sites, whereas the approach by Wang *et al.* does not use known binding sites; it begins with *de novo* motif detection, linking known transcription factors only during a later stage. In principle, they could also carry out the whole analysis for an unknown factor. The work of Wang *et al.* goes beyond the earlier study of Bussemaker *et al.*, as they went on to identify target genes and active transcriptional modules. The approach of Wang *et al.* is not bound by current knowledge about transcription factors and can be applied to unrestricted whole-genome analysis.

From array to network – future perspectives

This work on yeast transcriptional regulation shows elegantly that there is a way to sort out and understand gene expression patterns observed with high-throughput methods on a molecular level, at least in yeast. The crucial part in all of these studies is the integration of expression analysis with genomic information and transcription factor binding-site analysis (Fig. 1). This is all-important because it is the only way to separate and understand the many interconnected regulatory events and cascades that appear as a confusing tangle in the initial array results. Such approaches are divide-and-conquer strategies that have already been extremely successful in many applications. By finding the means to subdivide problems, even the most complex systems can be tackled and finally understood. It should be noted that each of the discussed approaches has unique merits and all are compatible in principle, so one all-inclusive strategy would be possible (if one could weed out the redundancies that such a system would inevitably generate). For example, Wang *et al.* took advantage of predefined *ab initio* motifs (from the REDUCER approach of Bussemaker *et al.*), extending the motifs by adding new sequences that were ranked by similarity of their expression profiles with the profiles of already included sequences. This produced motifs that were maximally associated with similar expression levels.

The approach of Zhu *et al.* focused more on individual transcription factors because close matching between the expression profiles of particular transcription factors and other genes was interpreted as an association measure that subsequently defined the motif. This ensures the finding of transcription factor motifs in promoters that are tightly co-regulated with the transcription factor, which could be interpreted as more likely immediate targets. Lee *et al.* have already extended their approach beyond expression profiles because they include ChIP, which allows them to go from putative binding sites to proven binding sites in their bioinformatics analysis.

The most complex biological system is probably the human body. With a genome of 3×10^9 base pairs and a far more complex organization of transcriptional regulation than yeast, understanding this system is not an easy task. However, some of the major obstacles, such as obtaining human promoters, have recently been solved [12]. Although this is a crucial development, it is only a first step not the solution to the whole problem. There has also been significant progress in understanding the more complex modular design of mammalian transcription control [13–15]. For example, major histocompatibility complex class I genes are induced by interferon or tumor necrosis factor α via a promoter module that requires two binding sites

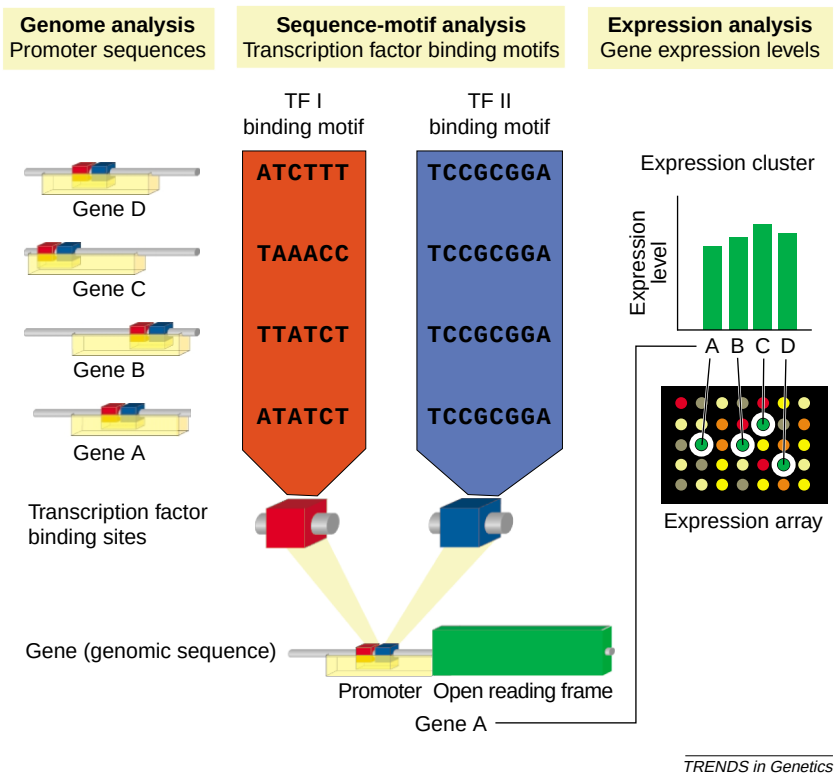

Fig. 1. Overview of integrated expression data and promoter analysis. Clusters of genes with similar expression levels or profiles from expression arrays (right side) can be used to select genes (generic gene with promoter shown as gene A, bottom center). Promoter sequences from four genes of a cluster are shown on the left side of the figure, each containing two transcription factor binding sites (TF I and TF II; red and blue cubes). Bioinformatics methods can be employed to extract the actual binding-site sequences from the promoters, as shown in the center of the figure (the red site is variable, whereas the blue site has a fixed sequence). Such binding site collections can then be represented as transcription factor motifs (in the form of weight matrices or consensus sequences) to be used in the location of additional potential target genes.

(IRF1 and NfkB). This module is found at different locations in responsive promoters, and the whole module can be present in both strand orientations in different promoters [14]. This complicates pattern recognition significantly, despite being one of the simpler examples in mammals. Although problems in humans are different from those in the yeast system in this and other respects (e.g. modules act in a cell-specific or tissue-specific manner), the approaches of Wang *et al.* and others might be adapted, at least in part, to mammalian expression analysis. Thus, reminiscent of the genome sequencing history, research in *S. cerevisiae* could again pave the way to overcoming similar problems in the human system. With this in mind, the recent developments in the understanding of yeast transcription control on both cellular and molecular levels might be pivotal.

Acknowledgements

This work was in part supported by DFG grant 'Informatic methods for the analysis and interpretation of large amounts of genomic data' (Grant#2370/1–1).

References

1 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
2 Pollack, J.R. and Iyer, V.R. (2002) Characterizing the physical genome. *Nat. Genet.* 32 (Suppl), 515–521
3 Zhou, A. *et al.* (2003) Identification of NF-κ B-regulated genes induced by TNFα utilizing expression profiling and RNA interference. *Oncogene* 22, 2054–2064
4 Werner, T. (2001) Cluster analysis and promoter modelling as bioinformatics tools for the identification of target genes from expression array data. *Pharmacogenomics* 2, 25–36
5 Hampsey, M. (1998) Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiol. Mol. Biol. Rev.* 62, 465–503
6 Wyrick, J.J. and Young, R.A. (2002) Deciphering gene expression regulatory networks. *Curr. Opin. Genet. Dev.* 12, 130–136
7 Wang, W. *et al.* (2002) A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* 99, 16893–16898
8 Bussemaker, H.J. *et al.* (2001) Regulatory element detection using correlation with expression. *Nat. Genet.* 27, 167–171
9 Pilpel, Y. *et al.* (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* 29, 153–159
10 Zhu, Z. *et al.* (2002) Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm. *J. Mol. Biol.* 318, 71–81
11 Lee, T.I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298, 799–804
12 Werner, T. (2001) The promoter connection. *Nat. Genet.* 29, 105–106
13 Firulli, A.B. and Olson, E.N. (1997) Modular regulation of muscle gene transcription: a mechanism for muscle cell diversity. *Trends Genet.* 13, 364–369
14 Klingenhoff, A. *et al.* (1999) Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* 15, 180–186
15 Konig, S. *et al.* (2002) Modular organization of phylogenetically conserved domains controlling developmental regulation of the human skeletal myosin heavy chain gene family. *J. Biol. Chem.* 277, 27593–27605

doi:10.1016/S0168-9525(03)00197-5