Molecular Systems Biology (2006) doi:10.1038/msb4100072

www.molecularsystemsbiology.com
2006.0030

molecular systems biology

NEWS AND VIEWS

Promoting human promoters

Itay Furman and Yitzhak Pilpel

Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel

Molecular Systems Biology 6 June 2006; doi:10.1038/msb4100072

The effect of transcription factors on human gene expression can now be quantified owing to a new computational approach. The method was successfully applied to the cases of the cell cycle program, and for liver-specific gene expression.

Transcriptional regulation of gene expression plays a major role in the acquisition of cell identity during embryogenesis and shapes cellular response to various stimuli. Understanding how transcription regulatory networks are encoded in the genome represents one of the major challenges in contemporary genomics. The genome era opened the door to the investigation of a systems-map of transcriptional regulation. A static view of the map is provided by the wiring scheme of the network, which is encoded by combinations of *cis*-regulatory sequences (or motifs) within genomic regulatory regions. On the other hand, probing the transcriptome with expression microarrays provides snapshots of the network output and reveals its dynamics. A pioneering study that systematically established a link between sequence and expression (Tavazoie *et al*, 1999) was based on the notion that coexpressed transcripts should also be coregulated. A search for common motifs in promoters of coclustered genes revealed shared sequence motifs among similarly expressed genes (Tavazoie *et al*, 1999). While very effective, the method has shortcomings: deciding on the number, size, and tightness of clusters is not straightforward since we do not know the degree of coherence of genes that belong to the same transcriptional program *a priori*. Consequently, correlation between clusters and motifs is not a one-to-one relationship (Bussemaker *et al*, 2001); often many genes in a cluster do not contain any known motif, and not all genes that contain a motif belong to the cluster from which it was derived. Furthermore, motif combinatorics could not be easily deduced. For instance, two motifs may be derived from a cluster either because they truly synergize in regulating the cluster's genes, or simply because they form alternative regulatory programs that converged onto a similar pattern (Pilpel *et al*, 2001). A way to overcome these obstacles was to reverse the flow of information, by starting with candidate motifs and testing their regulatory effect on expression (Figure 1). While providing good results in the simple case of yeast (Bussemaker *et al*, 2001; Pilpel *et al*, 2001), a major challenge was the more complicated case of mammalian promoters. In a recent study, currently published in *Molecular Systems Biology*, Michael Zhang and co-workers made a considerable advance in the analysis of transcriptional regulation in human cells (Das *et al*, 2006).

Key to the analysis made by the authors is the controlled use of the multivariate adaptive regression splines (MARS) methodology (Friedman, 1991). MARS is a sophisticated algorithm that adaptively fits data to statistical models, which account for response thresholds and response strengths. Moreover, it can inherently deal with more complex interaction terms, corresponding here to the effect of multiple regulatory motifs. Clearly, these are desirable properties when analyzing expression data in response to binding to sequence motifs. However, sophistication comes at a price: if the input data (motif scores) and the response data (gene expression) are sufficiently large and noisy, MARS will often produce biologically nonsignificant results despite its internal controls. The work discussed here presents a computational protocol that is a step forward in overcoming this formidable challenge when facing mammalian transcription data. The authors start with a set of 521 known motifs, and generate a score for each one describing its match to each gene based on the similarity of the promoter sequence to that motif. Next, each motif is assigned a score based on its ability to explain (or predict) the expression data all by itself (a 'reduction in variance' score). The motifs are then sorted by decreasing order based on that score, and most interestingly, they seem to represent a bimodal population, with mediocre motifs separated from high scoring ones by a discernable and, thus, useful gap. The list of sorted motifs is used to prepare sublists of prioritized motifs, and MARS is run on such sublists. The final output from MARS is a minimal set of motifs that, when considered individually, in pairs, or in triplets, produce the best prediction of the condition-specific expression levels. The identified motifs are thus good candidates for biologically significant control elements regulating the genes involved in those physiological processes that were active when the microarray snapshot was taken.

The authors demonstrated the utility of the approach for cases in which only a few expression profiles are available and, therefore, clustering is relatively ineffective. For example, beginning by modeling expression levels as measured in liver cells, they identified three single motifs and five motif pairs that are good predictors of expression in this tissue, suggesting a role in liver-specific expression. Reassuringly, most of the blindly discovered motifs and motif pairs were already implicated in determining liver-specific expression, with two pairs being novel in this context. Choosing one of the major transcription factors that is associated with one of the motifs, HNF-1, they went on to use the optimal model to identify HNF-1 targets in an extended set of genes. They found 38 such targets, of which 29 had experimental support, whereas the other nine displayed strong HNF-1 binding characteristics.

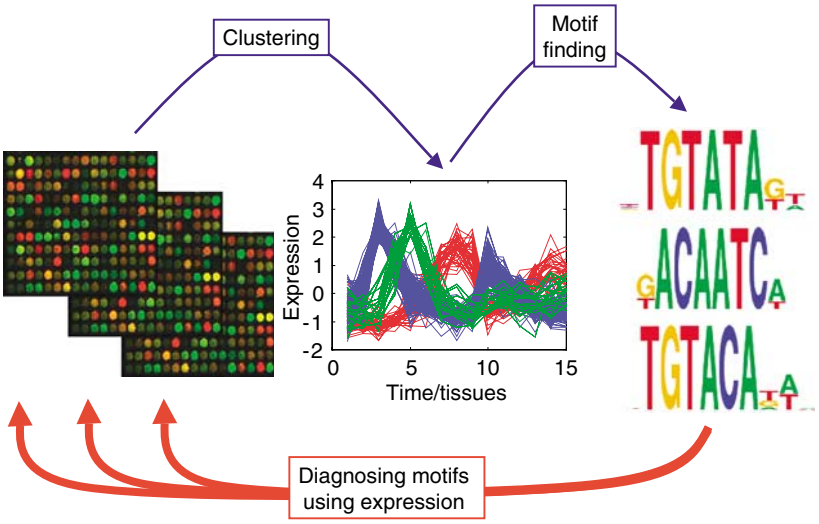

Figure 1 The classical approach (blue arrows, top) for deciphering transcription network is to start with expression data typically gathered at multiple time points during a process or in different tissues (left), to cluster coexpressed genes along the time profile (middle), and then to look for shared motifs within each cluster (right). In the approach (red arrows bottom) developed by Das *et al* in human cells, and described in previous works in yeast (Bussemaker *et al*, 2001; Pilpel *et al*, 2001), the process is reversed—starting from known motifs, and by-passing the clustering stage, motifs or combinations thereof are analyzed for their impact on gene expression under a specific condition (time point, tissue, treatment, etc.).

Even for time-course expression profiles, the Das *et al* approach may be advantageous over the clustering approach when searching for stage-specific regulators, for example. Taking human cell cycle data with 19 time-point profiles, but analyzing each time point individually, the authors identified some 20 individual motifs and 10 motif pairs that are involved in specific stages of the process. Many of these were known; for some of the rest, the authors presented experimental evidence that related them to cell cycle. Specifically, they provided additional supporting evidence that the well-known cell cycle regulator, E2F, regulates nonoverlapping sets of genes in the G1/S and G2/M phases of the cell cycle.

Regulatory motif identification is one side of the coin, and identification of functional targets of such regulators is the other. Discriminating true targets from false positives remains a major challenge when the binding sites are degenerate, as is often the case for mammals. In the present study, Das *et al* demonstrated the potential of their method to do such discrimination by identifying and validating new potential direct E2F targets, two of which are known to play a role in hepatocellular carcinoma progression.

What are the challenges that are still ahead of us on our way to fully deciphering expression regulation? To begin with, the interaction terms as implemented in MARS may correspond to an 'AND' gate—that is, two transcription factors are required simultaneously to induce transcription—but are less appropriate for other types of interaction (such as 'OR' gates), which are clearly frequently operating in transcription networks. The authors make a significant step forward by proposing that the linear function they utilize is a proxy for the gene's transcription induction function. However, when more detailed molecular processes will be considered in the future, the systematic mapping between statistical models and kinetic models is likely to be more difficult. And beyond transcription, other stages in the gene expression process, that are heavily regulated too, are much less understood. For example, steady-state mRNA levels reflect a balance between transcript production and degradation. While promoters contain information needed to tune mRNA synthesis, other genic regions, such as the 3′-untranslated regions, are involved in determining the transcript stability. Future models should combine information from both regions to predict accurately steady-state mRNA levels. Furthermore, models of gene expression should incorporate the various stages in translation control as well. Experimental technologies to probe these processes are beginning to emerge (cf. Wang *et al*, 2002; Arava *et al*, 2003) and new bioinformatic tools have to be developed to extract efficiently regulatory principles from the new data. For that purpose, more insight will be required regarding the relationship between the statistical models and the underlying kinetics and thermodynamics. With regard to analyzing transcription, we are in relatively good shape, as demonstrated by the present paper, when it comes to the other challenges we are only at the beginning of the road.

References

Arava Y, Wang Y, Storey JD, Liu CL, Brown PO, Herschlag D (2003) Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* **100:** 3889–3894

Bussemaker HJ, Li H, Siggia ED (2001) Regulatory element detection using correlation with expression. *Nat Genet* **27:** 167–171

Das D, Nahlé Z, Zhang MQ (2006) Adaptively inferring human transcriptional subnetworks. *Mol Syst Biol* doi:10.1038/msb4100067

Friedman J (1991) Multivariate adaptive regression splines. *Ann Statist* **19:** 1–67

Pilpel Y, Sudarsanam P, Church GM (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet* **29:** 153–159

Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999) Systematic determination of genetic network architecture. *Nat Genet* **22:** 281–285

Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO (2002) Precision and functional specificity in mRNA decay. *Proc Natl Acad Sci USA* **99:** 5860–5865