Finding transcription factor binding sequences

| ArticleContext | : | 130592244 |

Rachel Brem

Abstract

DNA sequences bound by transcription factors can be identified by fitting simple models to expression array data.

Significance and context

It is now routine to use expression arrays to measure the mRNA levels of thousands of genes in parallel. But analysis of these data has proved a computational challenge. Bussemaker *et al.* have used array data from the yeast *Saccharomyces cerevisiae* to identify transcription factor (TF) binding sites upstream of coding regions. The authors incorporated putative motifs into a mathematical model and asked "Does this model explain measured mRNA levels across the yeast genome?" If the answer was yes, they concluded that the motifs they studied were biologically relevant TF-binding sites. Because it models multiple motifs at once, their technique can tease out the contributions from many TF-binding motifs that operate on a single gene. This may prove an advantage over other array analysis methods, such as clustering.

Key results

Bussemaker *et al.* applied their technique in several contexts. First, they identified 11 TF-binding motifs *de novo* by fitting to an mRNA data set measured across the yeast cell cycle. Of these 11 motifs, all but two have been previously identified. Next, the authors studied the time dependence of expression control at about 70 known motifs. Their calculation found the F parameter, a measure of the strength of up- or down-regulation mediated by the TF (see methodological innovations for an explanation of the model). F parameters of 13 motifs were seen to vary during the yeast cell cycle or sporulation time courses, of which nine have not been previously associated with oscillations in yeast. Lastly, Bussemaker *et al.* focused on a single motif - the middle sporulation element (MSE), for which a consensus motif is known. A search through sequence space around the MSE consensus, fitting to sporulation mRNA data, yielded several distinct motifs never before classified.

Methodological innovations

The technique used in this work modeled the expression of all yeast genes in terms of a linear combination of effects from TFs binding to upstream motifs. In the model, the effect of motif M on the expression of gene G was calculated as the number of times M appears in the 600 base-pair region upstream of G, multiplied by a parameter F; the latter represents the strength of up- or down-regulation conferred by TF binding. Rather than fitting parameters for thousands of motifs at once, Bussemaker *et al*. fit one term at a time iteratively yeast expression data. Each motifs fit was assessed by computing the variance among mRNA levels predicted by the model and comparing it to the true variance in experimental data. If the difference was statistically significant, Bussemaker *et al*. concluded that the motif incorporated in the model binds strong TFs, whose action is largely responsible for mRNA levels across the genome.

Links

Results from this paper are available at the REDUCE (regulatory element detection using correlation with expression) website maintained by the authors.

Reporter's comments

The strength of this work is as an alternative to clustering for analysis of array data. It is not clear, however, whether the iterative single-motif fit used here is statistically sound. Also, the paper did not provide much interpretation for the study of variation across time. F parameters for some motifs were observed to change across a time course (at significance levels that Bussemaker *et al*. did not report). Can one confirm that these TFs are able both to repress and to activate transcription?

Table of links

Nature Genetics

REDUCE

References

1. Bussemaker HJ, Li H, Siggia ED: Regulatory detection using correlation with expression. Nat Genet. 2001, 27: 167-174. 1061-4036