Computational prediction and ranking of mammalian transcriptional regulatory modules using dense comparative genomics

Bilenky M.[1*], Robertson G.[1], Dagpinar M.[1], He A.[1], Bainbridge M.[1], Varhol R.[1], Thiessen N.[1], Teague K.[1], Griffith O.L.[1], Sleumer M.C.[1], Li Y.Y.[1], Fjell C.[1], Warren[1] R.L., Zhou J.[2], Sander J.[2], Marra M. and Jones S.J.M.[1]

1 Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC, Canada

2 Department of Computing Science, University of Alberta, Edmonton, AB, Canada

Transcriptional regulation of a eukaryotic gene typically involves the coordinated binding of multiple transcription factors to closely located DNA sequences that form *cis* regulatory module(s). We describe a computational system that predicts such modules genome wide and ranks them.

Transcription factor binding sites (TFBS) were predicted as overrepresented DNA motifs found by multiple algorithms in sets of (up to 39) related promoter sequences from currently available vertebrate genomes. Motif scoring was performed by a multivariate function that included phylogenetic weighting and was optimized by simulated annealing using large set of experimental data from TRANSFAC and ORegAnno databases. Motif significance was assigned by modeling neutral promoter evolution. Genome-wide set of ~200K significant motifs were grouped by 1) clustering based on pair-wise motif sequence similarity using OPTICS algorithm, and 2) annotating motifs using their similarity to binding sites for known transcription factors. We found that 16% of significant conserved motifs were similar with a p-value <1e-4 to at least one of 177 known TFBS models, while less than 2% of random promoter motifs could be annotated with the same p-value threshold. Using motif grouping, we predicted *cis* regulatory modules, as statistically significant patterns of motifs occurring in multiple promoters. For annotation based groups of motifs with stringent selection criteria (p-value<5e-4) we found ~300 different patterns containing at least three conserved motifs and occurring in at least ten genes. Predicted modules were ranked by the genome-wide properties of their associated genes; e.g. co-expression, Gene Ontology, and protein-protein interaction data.
We predicted modules in an independent set of promoters used in high throughput luciferase expression assays, and demonstrated that genes associated with modules that have same signature that highly ranked modules predicted genome-wide had highly correlated expression profiles.

Predicted motifs and modules for promoter regions of ~15K human, ~16K mouse and ~7K rat protein-coding genes are available at www.cisred.org.
The database is useful for various work related to regulation of gene expression, high throughput genome experiments, such as wide ChIP-PET, ChIP-CHIP, as well as for studying the evolution of conserved genomic elements.