

Automatic detection of putative *cis*-regulatory motifs

Mark Robinson
Biocomputation Group, STRC
University of Hertfordshire
AL10 9AB, UK

m.1.robinson@herts.ac.uk

Hamid Bolouri
Biocomputation Group, STRC
University of Hertfordshire
AL10 9AB, UK

H.Bolouri@Herts.ac.uk

Eric H. Davidson
Biology 156-29
California Institute of Technology
CA91125, USA

Davidson@caltech.edu

ABSTRACT

It is generally acknowledged that the evolution of metazoan morphologies was achieved primarily by the modification of the genetic regulatory networks that control development [1]. The spatial and temporal expression levels of each gene are primarily controlled by transcription factors that bind combinatorially to highly specific *cis*-regulatory binding sites [2]. Locating transcription factor binding sites within the *cis*-regulatory regions of genes is a crucial step towards determining the structure of genetic regulatory networks and understanding gene expression profiles. Current sequence searching algorithms designed to locate *cis*-binding elements utilize three primary methods:

- a) For those transcription factors with known consensus binding sites (see for example the Transfac database <http://transfac.gbf.de/TRANSFAC/>), one may search for consensus binding sites and rank them by goodness of match.
- b) Where sequenced genomes of evolutionarily close species are available, phylogenetic footprinting can identify highly conserved sequences in regulatory regions of homologous genes [3].
- c) Correlation of nucleotide word frequencies within non-coding regions of either homologous genes or genes sharing an expression profile [4][5].

Here, we propose and demonstrate a methodology complementary to the above approaches. Our approach, uses known characteristics of regulatory domains of eukaryotic genes to look for statistically significant nucleotide words in regulatory regions of genes.

To analyze regulatory regions of single genes, we exploit the tendency of transcription factors to bind to multiple *cis*-regulatory sites (e.g. by dimerization) and search for statistically significant multiple occurrences (doubles, triples, etc) of putative binding sites within user-specified distances of one another.

We also compare the regulatory domains of co-expressed genes to test the hypothesis that they are co-regulated (e.g. genes in a gene battery [2]). Our analysis is predicated on the assumption that co-regulated genes will share multiple, spatially clustered, binding sites for one or more

shared transcription factors; and that the co-occurrence of these binding sites will be at a frequency higher than random expectation.

We are also exploring cases where the bounds of the regulatory region of the gene or genes of interest cannot be estimated accurately. In such cases, we allow for searches of long non-coding DNA sequences (10 to 50K bps) and search for patterns that occur only in a spatially confined region of the total sequence and with particular spatial relationships.

We are currently applying our algorithms to the analysis of a number of genes involved in Endomesoderm specification in sea urchin Embryos [6] and intend to make our software freely available to researchers as open source software as soon as we have completed the current “field testing”.

REFERENCES

- [1] Davidson, E.H., Genomic Regulatory Systems: development and evolution, ISBN 0-12-205351-6, Academic Press, 2001.
- [2] Arnone, M.I, and Davidson, E.H. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124 (1997), 1852-1864.
- [3] Fickett, J.W, and Wasserman, W.W. Discovery and modeling of transcriptional regulatory regions. *Current Opinion in Biotechnology*. 11 (2000), 19-24.
- [4] Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J.W. and Lawrence, C.E. Human-mouse genome comparisons to locate regulatory sites. *Nature genetics*, 26 (2000), 225-228.
- [5] Bussemaker, H.J., Li, H. and Siggia, E.D. Regulatory element detection using correlation with expression. *Nature genetics*, 27 (2001), 167-171.
- [6] See “Coming to grips with gene regulation”, *Science* v293, Aug 3 2001, p789.