MARSMotif: Deciphering Transcriptional Subnetworks from Microarray Expression Data using Regression Splines

Debopriya Das[1*] and Michael Q. Zhang[2]
[1]Life Sciences Division, E.O. Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.
[2]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724. USA
[*]Corresponding author: ddas@potternexus.lbl.gov

High-throughput expression profiling technologies have raised the possibility of a global understanding of gene regulation in a wide variety of biological systems, including disease cells. By computationally integrating the expression profiles with genomic sequence data, it is possible to obtain a snapshot of the functional *cis*-regulatory motifs and their associated target genes. Classical approaches for deciphering *cis*-elements proceed by clustering microarray data across a large number of conditions. Such approaches, while useful, have severe limitations however. Over the past few years, a powerful set of methods based on regression approaches have been developed, where motif occurrence frequencies and binding strengths are correlated with expression profiles, and clustering is unnecessary. Functional motifs exhibit a statistically significant fit; inactive motifs do not.

In this demo, we will present a suite of tools based on regression splines that permit discovery of active combinations of *cis*-regulatory motifs and the segments of global regulatory networks, i.e. the subnetworks, which they belong to. Specifically, the algorithms are based on MARS (Multivariate Adaptive Regression Splines), which uses linear splines as basis functions. Linear splines comprise a complete set of basis functions, and hence, allow comprehensive modeling of gene expression signals. More importantly, they mimic switch-like behavior intrinsic to transcriptional response, and thus, provide a natural framework to model such response data. The tools that we will demonstrate are:

- *MARSMotif*: It is intended for discovery of non-degenerate or weakly degenerate motifs, as are typically observed in simpler eukaryotes[1].
- *MARSMotif-M*: This generalizes MARSMotif and is best used when the motifs can have arbitrary degeneracy, as is often the case in mammals[2].
- *GapFinder*: This allows rapid detection of the most dominant transcription factors (TFs) under the tested biological condition.

In recent studies, we have demonstrated that these tools outperform the competing methods, and work with comparable accuracies in both low eukaryotes and mammals. In addition, interacting combinations of motifs can be inferred with very high confidence, in contrast to other approaches. Furthermore, we have shown that MARS-based tools are suitable for identifying direct targets of TFs, even when their cognate *cis*-motifs are very degenerate. Target identification in such instances has been quite challenging, hindering prediction of testable hypotheses. Finally, these methods are effective in modeling condition-specific regulation of genes by a specific TF. They are applicable to both microarray expression data and ChIP-on-chip binding data[3].

The programs will be available for download from http://rulai.cshl.edu/licensing/index1.htm or http://vision.lbl.gov/Resources/.

References

1. Das D, Banerjee N & Zhang MQ (2004) Interacting models of cooperative gene regulation. *Proc Natl Acad Sci U S A.* **101:** 16234-16239
2. Das D, Nahlé Z & Zhang MQ (2006) Adaptively inferring human transcriptional subnetworks. *Molecular Systems Biology (Nature)* **2:** 2006.0029
3. Smith AD, Sumazin P, Das D & Zhang MQ (2005) Mining ChIP-chip data for transcription factor and cofactor binding sites. *Bioinformatics* **21**(Suppl 1)**:** i403-i412 (*ISMB 2005*)