

From Genome to Candidate Cis-regulatory Networks: A Bioinformatics Approach

*Rajanikanth Vadigepalli, Praveen Chakravarthula, Daniel E. Zak,
James S. Schwaber, Gregory E. Gonye*

Motivation:

Present technological enhancements have resulted in public databases containing data sets of various types: gene expression, protein-DNA interaction and transcription factor (TF) activity data, protein-protein interactions, and genomic sequence and ontology information. The analysis of these large volumes of information holds the promise of identification of the nonlinear dynamic function of the biochemical regulatory networks. Various attempts at reverse engineering the gene regulatory networks from microarray data alone have met with limited success. Given N time points of expression of X genes, typical algorithms allow any gene to regulate any other gene ($X \times X$ parameters) and try to estimate the actual interaction network from gene expression data ($N \times X$ values). However, this renders the identification problem either intractable or sub-optimal, as the data available is much less than that required by the number of interactions being estimated (N much smaller than X ; $N \sim 10-20$ and $X \sim 100-1000$). Objective of this study is to generate constraints on interactions based on Transcription Regulatory Elements (TRE) present on promoters of genes in order to improve the network identification.

Methods:

The analysis of promoters for the genes of interest for known TRE's will directly provide a good candidate set of network interactions. To this end, we have developed PAINT: Promoter Analysis and Interaction Network Tool. PAINT is available at www.dbi.tju.edu/dbi/tools/paint. The path from a list of genes to candidate interactions involves: identifying and retrieving promoters for the genes from genome sequence assembly, inspecting these promoters for known TRE's using motif finding/matching algorithms, processing this information to create a candidate interaction matrix (CIM) whose ij -th element represents whether i -th gene (row) can be regulated by j -th TF (column).

PAINT currently consists of (1) a database of predicted promoter sequences of known or predicted genes in the Ensembl annotated mouse genome database (www.ensembl.org), and (2) various modules that can retrieve and process the upstream sequences for known TRE's. A good estimate of promoters, or equivalently, a good estimate of the location of Transcription Start Site (TSS) for each gene significantly improves CIM prediction. The 5' Untranslated Region (UTR) sequence of each gene has to be considered to properly estimate TSS location. To this end, PAINT utilizes alignments with full-length clone sequences from RIKEN that contain 5' UTR data (fantom2.gsc.riken.go.jp). The start of the Open Reading Frame (ORF) is considered as an initial estimate of TSS (TSS-temp) which is then refined based on the alignment of 5000 base pairs (bp) upstream from TSS-temp with the RIKEN clone sequences. The alignments are filtered for those that have less than 3 mismatches. The alignment that is closest to the start of clone sequence and is 5' most on the upstream sequence is selected. TSS for 25% of the genes are

identified through this procedure. For the remaining genes, Eponine, a TSS prediction tool, is employed to identify TSS within the 5000 bp of TSS-temp (servlet.sanger.ac.uk:8080/eponine). TSS for approximately 5% of the genes are estimated based on predictions from Eponine. For the remaining genes, TSS-temp is considered as the final TSS estimate. For each gene, a 2000 bp upstream sequence of the TSS estimate is retrieved and placed in the promoter database.

PAINT utilizes MatInspector (www.genomatix.de) for identification of TRE's on promoters. Clustering analysis is available in PAINT utilizing the R software for statistical analysis. The CIM can be visualized after clustering genes by shared similar TRE's and/or by TRE's found on similar genes. The dissimilarity metric used is the binary distance - fraction of elements that are dissimilar between two rows/columns when considering only the elements for which at least one element is non-zero in either row/column (Jaccard coefficient). A network layout diagram is also produced using GraphViz to aid in visual analysis (www.research.att.com/sw/tools/graphviz).

The frequency of appearance of TRE's in genes of interest needs to be assigned statistical significance. This paves way for further experimental study on limited set of TF's based on highly significant occurrences of corresponding TRE's. PAINT constructs empirical reference distributions (ERD) of TRE's on promoters through random re-sampling from the promoter database. These ERD's are used to test the null hypothesis that the TRE's found in the experimentally-selected group of promoters can be explained by occurrence on randomly selected promoters from the database. The TRE's may then be prioritized for further study by their hypothesis test p-values.

The CIM generated by PAINT can be incorporated into methods such as Bayesian networks or linear dynamic model identification as constraints to render the algorithms tractable for large-scale systems. The PAINT output can be loaded into clustering software such as Cluster and TreeView (rana.lbl.gov), as well as into network visualization tools such as Cytoscape (www.cytoscape.org) and Pajek (vlado.fmf.uni-lj.si/pub/networks/pajek) for further analysis. PAINT consists of a MySQL database and Perl modules and can be easily wrapped for use in UNIX shell scripts, perl programs and PHP scripts for web interfaces. In addition, PAINT also has a Open Agent Architecture module allowing it to function as an 'agent' for BioSPICE, the simulation and analysis software platform for systems biology being developed under DARPA's BioComp program (www.biospice.org).

Case studies:

Application of PAINT is demonstrated in two separate case studies; the first one involving 50 up regulated and 50 down regulated genes in neuroblastoma cell differentiation, and the second study involving 575 differentially expressed genes in neuroblastoma cells upon activation of the AT1 receptor by angiotensin II. In both examples, the candidate interactions generated by PAINT reduce the number of computed parameters by six to seven fold, thus improving the accuracy and computational tractability of the network identification methods. Preliminary analysis of CIM's indicates that the similarity of genes based on the TRE's is a reasonable indicator of similarity of expression. However, significant number of genes that share similar TRE's do differ in expression. These results indicate combinatorial aspect of transcriptional

regulation (common TRE's indicates only potential, not absolute, coregulation) and the need for combining TF activity data with gene expression information for improved accuracy of regulatory network identification.