TieDIE Tutorial

Version 1.0.1

Evan Paull <epaull@soe.ucsc.edu>

October 9, 2013

Contents

A Signaling Pathway Example

Introduction

TieDIE is an algorithm that finds subnetworks connecting genomic perturbations to transcriptional changes in large gene interaction networks. To do this, the algorithm generates a "diffusion" kernel describing the flow of information in the master network, using code written in SciPy (`http://www.scipy.org/`) or MATLAB (`http://www.mathworks.com/products/matlab/`). The computation of the diffusion kernel is computationally intensive–particularly for a large pathway–but, once computed, the kernel can be saved to a file and re-used when running the TieDIE algorithm on the same master pathway. It is recommended that kernels for large networks be generated using MATLAB, which is considerably faster and more memory-efficient, but a free SciPy implementation is provided for those without a MATLAB license. The entire program is written in python 2.7.X and requires the numpy-1.7+ package to be installed before running. TieDIE should run on any UNIX system, and has been tested on Linux and Mac OS, at this point. Windows compatibility is not supported at this point.

TieDIE Input Format

An example network and input is shown under examples/GBM.test. Three input files are in this directory:

- pathway.sif: A signaling pathway extracted from the 2008 TCGA GBM Nature paper supplement [4]). The .sif format is compatible with Cytoscape [**?**], and uses a 3-column **source interaction target**. For example:
  - MDM2 inhibits> TP53
  - TP53 activates> MDM2
  - ...
- upstream.input: Input heats for the "upstream" set of genes, weighted by frequency of mutation or amplification. The 3-column format should contain a column for the gene name, the input heat, and the expected functional effect of the perturbation (+/-). For example:
  - PIK3CA 0.5 +
  - RB1 0.2 -
  - TP53 0.3 -

- downstream.input: Input heats for the "downstream" set of transcriptional responses. The format is the same as the upstream.input file, but the third column should be interpreted as the inferred activity.

Generating TieDIE Input

The TieDIE algorithm can be used to generate sub networks for any two (or more) input sets, but the standard analysis is to use high-throughput genomic data to identify a set of genomic events and connected them to a set of changes in transcription factor activity, inferred with gene-expression data. In this example, the "source" set represents a set of mutated, or otherwise functionally altered genes weighted by the relative confidence that each is effected in the given dataset. The "target" set of genes represent the set of gene-expression changes thought to be (at least partially) caused by the source set of perturbations. Generating the source input set and weights is relatively straitforward, given the output of a variant-calling algorithm, typically in the form of a MAF file (`https://wiki.nci.nih.gov/display/TCGA/Accessing+MAF+files`). For differential analysis in particular, a proportions test can be used to compare the frequency of events between two sample-groups, and produce inputs that define the "source" set of a TieDIE network analysis (see Figure 1 for an illustration).

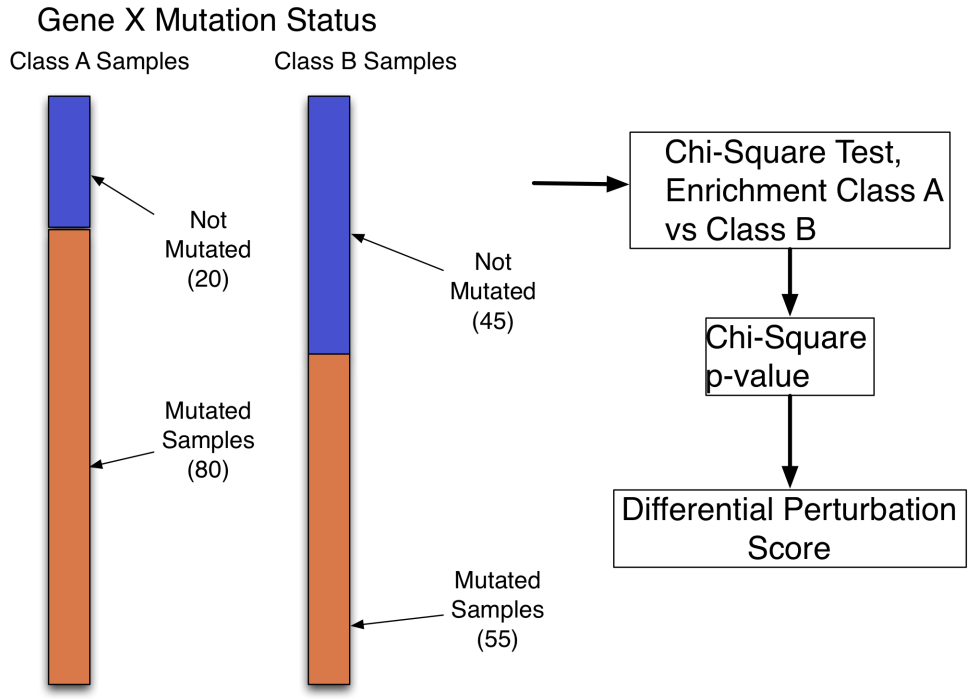

Figure 1: Pearson's Chi-squared test can be used to guage the significance of differing event frequency between groups. For example, the test reports a p-value based on the frequency of mutation events and the sample sizes between Class A and Class B, which can be converted to a "differential perturbation" score by a log transformation. This score or "input-heat" should represent the relative confidence that each gene is mutated or otherwise functionally effected in the dataset of interest.

The "target" set for TieDIE can be generated from gene expression data: for example, with microarray expression, Significnce of Microarrays (SAM) [2] can be used to rank the genome from most to least differentially expressed. SAM uses a statistical model that is suitable for microarray data and code is available via a freely available R package, while other algorithms can be used for RNA-Seq data such as edgeR [5] and DESeq [1]. The ranked genome can then be used to find transcription factors with a regulated set of genes that cluster to one end of this ranking: for example, Gene Set Enrichment Analysis [7] is a statistical technique to find the significance of such sets, and can be used to find "master regulator" genes that are likely to be controlling expression. Figure 2 (below) illustrates this method, and for further explanation of the "master regulator" algorithm, see the paper by Lim, WK and Califano, A [3].

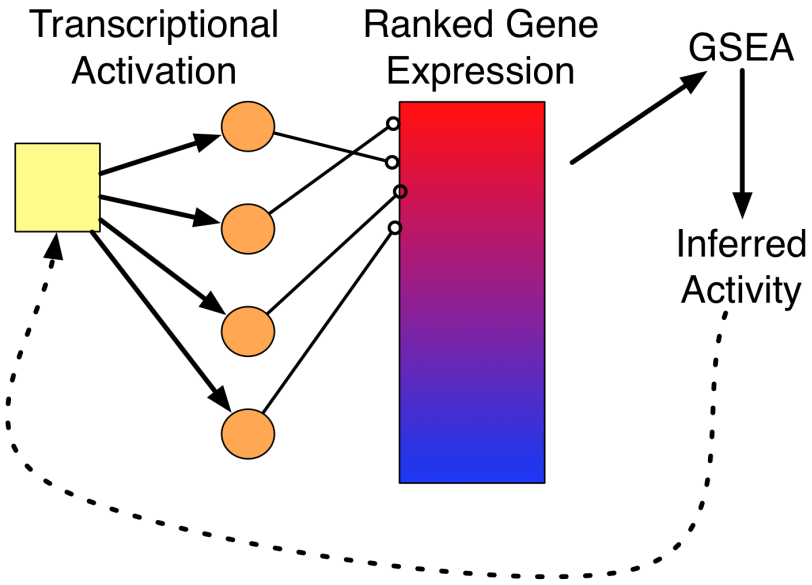

Figure 2: The "master regulator" algorithm considers the set of genes activated by a given transcription factor as input to Gene Set Enrichment Analysis (GSEA). If rankings of that set are closer to one end of that list than expected by chance, the algorithm would infer that the activity of the transcription factor is significantly effected.

Running TieDIE

To run the test, change directory to this folder and type "make". The program should run in a few seconds and report to stderr, a sub directory with the output will be created. Important output files are:

- report.txt: Network statistics and a summary report for the TieDIE sub network. Also includes the results of the permutation test.
- tiedie.sif: Connecting TieDIE sub network.
- tiedie.cn.sif: Connecting sub network after filtering for logically consistent paths.
- heats.NA: Node-attribute file formatted for Cytoscape input (visualization).

Visualization

Cytoscape

There are many software packages available for network visualization, but we've found the Cytoscape package [6] to be well suited for network visualization. The TieDIE output network (tiedie.sif) can be directly imported to Cytoscape 2.8 or later, and the node attribute file (heats.NA) along with the supplied properties file (vizmap.props) can be used to color nodes by linker heat, as shown in the figure below.

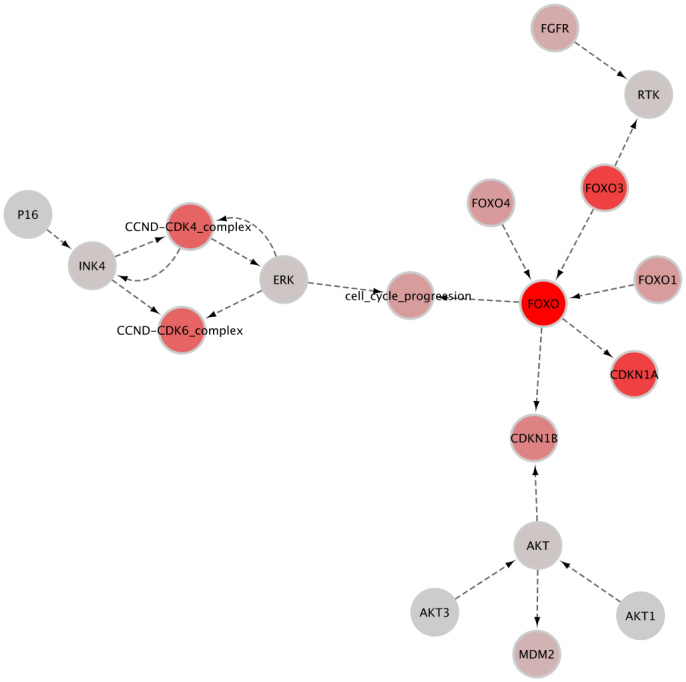

Figure 3: An small network example: sources are colored in blue, targets are red. Nodes and edges captured in the TieDIE Solution are outlined in green.

More detailed visualizations can also be performed in Cytoscape: for example, nodes can be colored or shaped based on source or target node status, and nodes/edges can be highlighted, as shown in the visualization below.

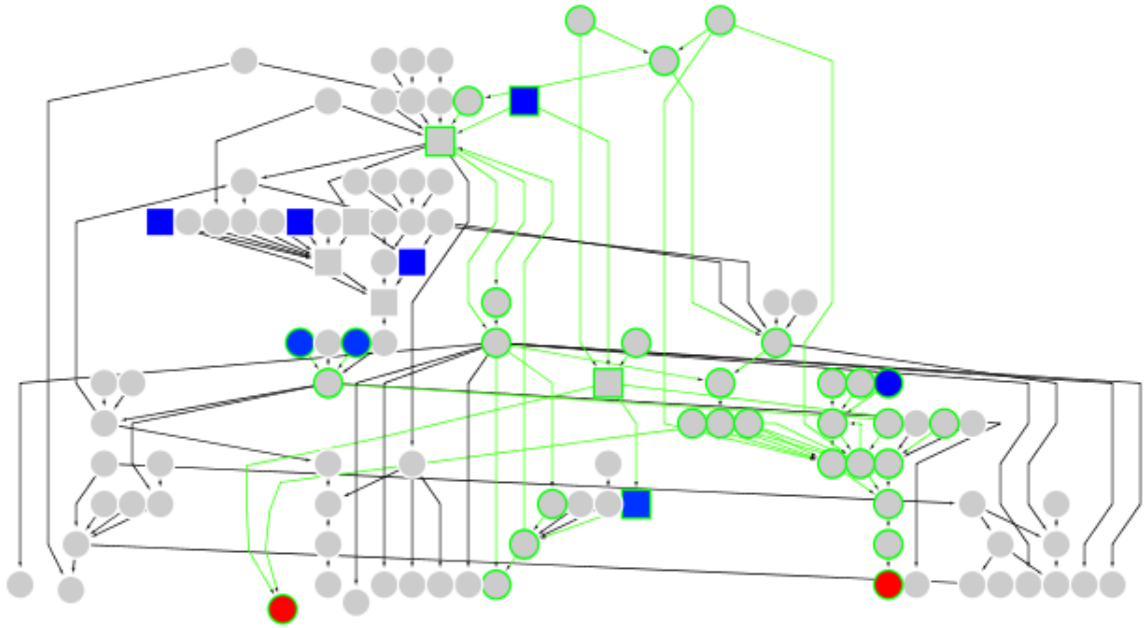

Figure 4: An small network example: sources are colored in blue, targets are red. Nodes and edges captured in the TieDIE Solution are outlined in green.

Additional Visualization Options

The more complex images shown in the TieDIE paper (Figures 4,5) allow visualization of node-specific data within an embedded network view, and are referred to as "circleplots". This code is available at `https://github.com/ucscCancer/paradigm-scripts` and requires the original perturbation-matrix and gene-expression data to run. The circlePlot.py program produces a set of PNG format images that can be imported to Cytoscape. Unfortunately, there's no simple way to bypass the GUI loader of Cytoscape and programatically load images, although this is rarely an issue with small networks. The code is documented by the authors, and comments or feature requests should be sent to them, or to me (<`epaull@soe.ucsc.edu`>).

Contact:

Questions, feature requests or requests for clarification should be sent to the author of this document, at <epaull@soe.ucsc.edu>. I will try to respond promptly and completely to all comments and requests, so feel free to contact me!

Bibliography

[1] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[2] Gilbert Chu, Balasubramanian Narasimhan, Robert Tibshirani, and Virginia Tusher. Significance analysis of microarrays (sam) software. *Nature*, 5:436–442, 2002.

[3] WK Lim, E Lyashenko, and Andrea Califano. Master regulators used as breast cancer metastasis classifier. *Pac Symp Biocomputing*, pages 504–15, 2009.

[4] The Cancer Genome Atlas Research Network*. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455, October 2008.

[5] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[6] Paul Shannon and et. al. Ideker, Trey. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 2003.

[7] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.