An Introduction to the pandaR Package

Daniel Schlauch

March, 2015

Introduction

The fundamental concepts behind the PANDA approach is to model the regulatory network as a bipartite network and estimate edge weights based on the evidence that information from a particular transcription factor i is successfully being passed to a particular gene j. This evidence comes from the agreement between two measured quantities. First, the correlation in expression between gene j and other genes. And second, the strength ofevidence of the existence of an edge between TF i and those same genes. This concordance is measured using Tanimoto similarity. A gene is said to be available if there is strong evidence of this type of agreement. Analogous to this is the concept of responsibility which similarly focuses on a TF-gene network edge but instead measures the concordance between suspected protein-complex partners of TF i and the respective strength of evidence of a regulatory pathway between those TFs and gene j.

PANDA utilizes an iterative approach to updating the bipartite edge weights incrementally as evidence for new edges emerges and evidence for existing edges diminishes. This process continues until the algorithm reaches a point of convergence settling on a final score for the strength of information supporting a regulatory mechanism for every pairwise combination of TFs and genes. This package provides a straightforward tool for applying this established method. Beginning with data.frames or matrices representing a set of gene expression samples, motifpriors and optional protein-protein interaction users can generate an m by n matrix representing the bipartite network from m TFs regulating n genes. Additionally, pandaR reports the co-regulation and cooperative networks at convergence. These are reported as complete graphs representing the evidence for gene co-regulation and transcription factor cooperation.

Example

An example dataset derived from a subset of stress-induced Yeast is available by running

```
library(pandaR)
data(pandaToyData)
```

`pandaToyData` is a list containing a regulatory structure derived from sequence motif analysis, protein-protein interaction data and a gene expression. The primary function in pandaR is called with

```
pandaResult <- panda(pandaToyData$motif, pandaToyData$expression, pandaToyData$ppi)
pandaResult
```

```
## PANDA network for 913 genes and 87 transcription factors.
##
## Slots:
## regNet   : Regulatory network of 87  transcription factors to 913 genes.
## coregNet : Co-regulation network of 913 genes.
## coopNet  : Cooperative network of 87 transcription factors.
##
## Regulatory graph contains  79431 edges.
## Regulatory graph is complete.
```

Where `pandaResult` is a 'panda' object which contains `data.frames` describing the complete bipartite gene regulatory network as well as complete networks for gene coregulation and transcription factor cooperation. Due to completeness, edgeweights for the regulatory network are reported for all mxn possible TF-gene edges. The distribution of these edge weights for these networks has approximate mean 0 and and standard deviation 1. The edges are therfore best interpreted in a relative sense. Strongly positive values indicative of relatively larger amounts of evidence in favor a regulatory mechanism and conversely, smaller or negative values can be interpreted as lacking evidence of a shared biological role. It is naturally of interest to specify a high edge weight subset of the complete network to investigate as a set of present/absent edges. This is easily performed by using the `topedges` function. A network containing the top 1000 edge scores as binary edges can be obtained with

```
topNet <- topedges(pandaResult, 1000)
```

Users may then examine the genes targeted by a transcription factor of interest.

```
targetedGenes(topNet, c("AR"))
```

```
##  [1] "AKAP10"       "CNDP2"        "CRHR1"        "HNRNPD"
##  [5] "KIAA0652"     "LOC100093631" "LOC100128811" "PRR15"
##  [9] "TCF4"         "TCP11L2"      "TMPRSS11B"    "VCX3B"
## [13] "WDR4"
```

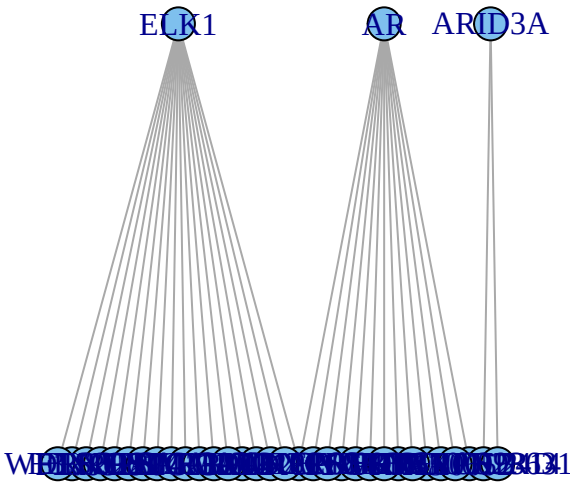
The network can be further simplified by focusing only on transcription factors on interest and the genes that they are found to regulate. The `subnetwork` method serves this function

```
topSubnet <- subnetwork(topNet, c("AR","ARID3A","ELK1"))
```

Existing R packages, such as igraph, can be used to visualize the results

```
plotGraph(topSubnet)
```

```
## Loading required package: igraph
```



References

Glass K, Huttenhower C, Quackenbush J, Yuan GC. Passing Messages Between Biological Networks to Refine Predicted Interactions, *PLoS One*, 2013 May 31;8(5):e64832

Glass K, Quackenbush J, Silverman EK, Celli B, Rennard S, Yuan GC and DeMeo DL. Sexually-dimorphic targeting of functionally-related genes in COPD, *BMC Systems Biology*, 2014 Nov 28;**8**:118