Mining for gene expression biclusters in large-scale RNA-seq data compilations

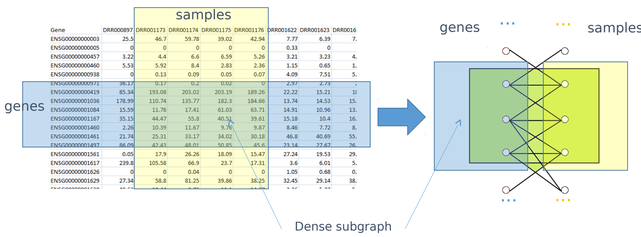
Arinjoy Basak, Clark Cuccinel, Alexandra Cummings, Jose Cadena, Andrew Warren, Anil Vullikanti, Allan Dickerman
Biocomplexity Institute, Virginia Tech

Introduction

Routine application of high-throughput sequencing have led to an accumulation of large volumes of public RNA-seq data. Finding novel biological patterns in these data remains a significant challenge, and prior methods, e.g., those based on finding bicliques, do not scale to large datasets. We develop a new approach based on quasicliques in signed networks.

Methods

We represent the gene expression data as a bipartite network network of genes and samples with weighted edges. The edge weight represents expression level, which can be both positive and negative.



For a cluster C consisting of a set G of genes and a set S of samples, we define its quasiclique density $f_\alpha(C)$ as:

$f_\alpha(C) = sum\ of\ all\ edge\ weights\ in\ C - \alpha|G||S|$

Objectives:

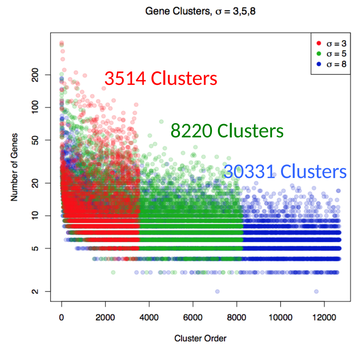
(1) Find cluster *C* that maximizes $f_\alpha(C)$

(2) Find sequence of clusters C_1, C_2, … in decreasing score

Data acquisition: Processed human RNA-seq data was downloaded from ArrayExpress (http://www.ebi.ac.uk/arrayexpress) scaled as transcripts per million (TPM). A matrix of 14,656 runs or samples (columns) and 56,729 genes (rows) was normalized row-wise to a mean of zero and a standard deviation of 1. The normalized matrix was then in units of standard deviations on a per-gene basis. The normalized matrix was transformed into a sparse bi-partite graph by thresholding at a given standard deviation (σ_t = 3,4,5 and 8). Values above the selected threshold were represented as edges in a graph connecting genes on one side and samples on the other. These edges were weighted by the log of the normalized expression value ($\log_{10}(\sigma)$).

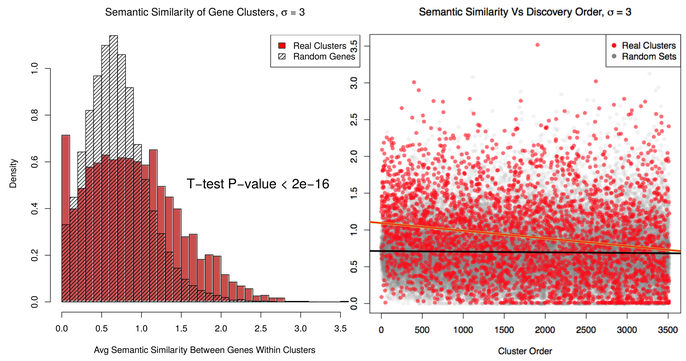
Results

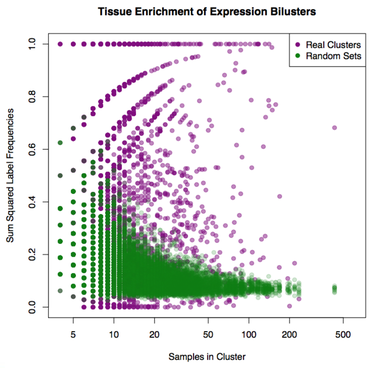
Network analyses were performed at σ = 3,4,5 and 8 and allowed to run for 48 hours each. These runs produced 3514, 5908, 8220, and 30331 clusters, respectively. Clusters selected early in each run averaged larger than clusters selected later in the run, with respect to both genes and samples.



Distributions of cluster sizes found when sparsifying the expression matrix to various thresholds. Higher thresholds result in sparser graphs which can be processed further in the 48-hour run time used here. Larger gene clusters are found in early rounds, trending toward 10 genes or less in very late rounds.

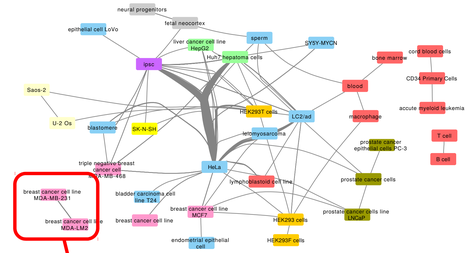
Gene Ontology (GO) semantic similarity was significantly higher for actual gene clusters than for random sets.



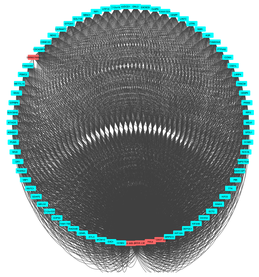


Tissue and cell-line identifiers for RNA-seq samples were used to test sample clusters for significant structure relative to randomized gene sets. The test statistic was sum of squared frequencies of the labels per cluster. Again, clusters were significantly more supported by external knowledge than random sets.

Finally, tissues and cell types which were found to be linked by co-occurrence in biclusters were connected in a network to apply this work of interpreting biological relationships.



Relationship between tissue and cell types based on sharing expression biclusters. Sample types are shown linked by and edge when they each occur in at least 1/3 of the samples united in a bicluster. Cluster_0009 is example of such a link between cell types is highlighted by the red box and discussed below.



_Term	NS	name	depth	study_count	p_bonferroni
:0022402	BP	cell cycle process	2	22	8.02E-07
:1903047	BP	mitotic cell cycle process	3	16	1.58E-05
:0051276	BP	chromosome organization	4	13	3.77E-05
:0051301	BP	cell division	3	12	0.0002
:0006996	BP	organelle organization	3	24	0.0038
:0006259	BP	DNA metabolic process	6	14	0.0127
:0006974	BP	cellular response to DNA damage stimulus	4	13	0.0280
:0051726	BP	regulation of cell cycle	4	16	0.0299
:0007049	BP	cell cycle	3	10	0.0470

Bicluster_0009 from the 3σ run links two breast caner cell lines: MDA-LM2 and MDA-MB-231. This bicluster includes 70 genes and 46 samples. Gene Ontology functional enrichment on those 70 genes reported as significantly enriched, at adjusted P-value < 0.05, the GO terms listed above.

Conclusion

Our method for finding quasicliques on signed graphs scales to the challenge of large-scale data mining of gene expression data sets extending above ten thousand samples for the ~56 thousand genes in the human genome.

The biclusters returned are measurably more consistent with previous biological knowledge, as represented by the Gene Ontology and protein-protein interactions.

These nuclei of coordinated gene action can be directly used to interpret relationships between biological samples.

Fully exploiting the information inherent in these biclusters will require extensive bioinformatic integration.