Inferring Biological Meaning from Cap Analysis Gene Expression Data

HRYSOULA PAPADAKIS

1. Introduction

This project is inspired by the recent development of the Cap analysis gene expression (CAGE) method, which introduces new benefits to standard gene expression techniques such as being higher-throughput and more accurately mapping the transcription start sites.[7] Building off of previous serial analysis of gene expression (SAGE) methods, I aim to extract biological meaning from the gene expression patterns, use module networks to find modules of co-regulated genes, and lay the foundations for future work in predicting gene expression using regression trees corresponding to regulation programs.[1]

2. Methods

2.1. *Dataset*

The data comes from CAGE experiments in 246 human cancer cell lines measuring 30,970 transcripts. It is arranged in a matrix where the rows correspond to the genes; the columns to the cell lines; and the values are tags-per-million, a count proportional to the fraction of reads in the cell line experiment that overlapped the transcript of the gene. For the purposes of this analysis, I consider both the matrix and its transpose; in the former the genes are the samples and the experiments are the variables, in the latter the samples are the experiments and the genes the variables.

2.2. *Filtering*

Since the data was accumulated from a number of sources and contained high variability, I applied a series of filters to remove noise and produce a dataset from which strong biological signals could be recovered. No normalization was required since the tags-per-million count is already normalized by the total number of reads in a given experiment. I experimented with various thresholds for dropping genes by computing basic statistics at each stage, and settled on the following procedure.

Gene entries Many of the genes appear multiple times, which I collapse by summing the values per experiment over all copies of a gene. There are a resulting 18,698 unique genes.

Missing values Many values in the matrix were zeros, indicating that for a given experiment, the gene transcript was not observed at all. This is unexpected and most likely due to anomalies in the experiments where either the tags are inaccurate or the transcript was not tested. In order to avoid imputing values or determining a safe cut-off for number of zeros tolerated per gene, I dropped all genes with any zero values.

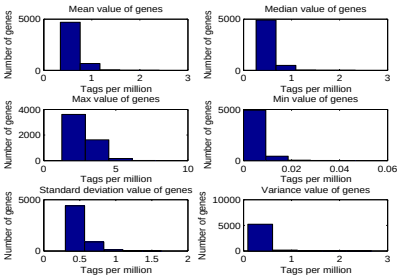
Total reads The number of total reads in the experiments varied significantly. I removed all experiments with fewer than 500,000 reads, as low read count introduces uncertainty.

Gene ID The validation of my results depends on gene ontology (GO) term enrichment analysis, where the annotations map from EntrezGene IDs to functional GO-terms. Since the dataset genes are labeled with HGNC symbol IDs, I dropped genes for which there was not a mapping to a corresponding EntrezGene ID.

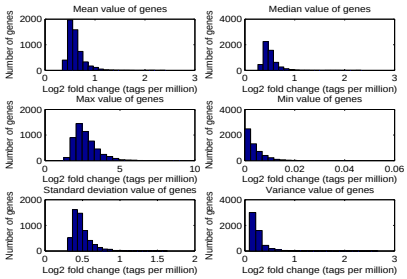
Transformation I computed the $\log_2$ transform of the data, first coercing values less than two to two. I then computed the log fold change $\log(x_{gene,exp}) - \log(mean_{gene}(experiments))$. I removed genes with low variance (standard deviation < 0.3).

2.3. *Final data*

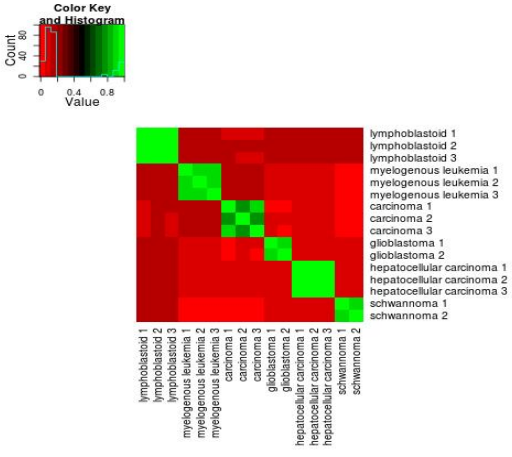
After all stages of filtering, the remaining data consisted of 5426 genes and 243 experiments. To determine the quality of this filtered dataset, I examined the basic statistics and performed correlation analysis. The the filtered data (Fig. 1(b)) has a cleaner distribution than the original data (Fig. 1(a)). The experiments include a handful of replicates for the same cell line, which are indeed highly correlated as expected (Fig. 1(c)). Similarly, genes from the same gene family should be highly correlated. In order to test this, I overlapped my gene set with a set of 220 known transcription factors and used the resulting 55 identified transcription factors as a proxy for genes. As desired, the genes within each family are highly correlated (Fig. 1(d), where the genes are ordered by families).



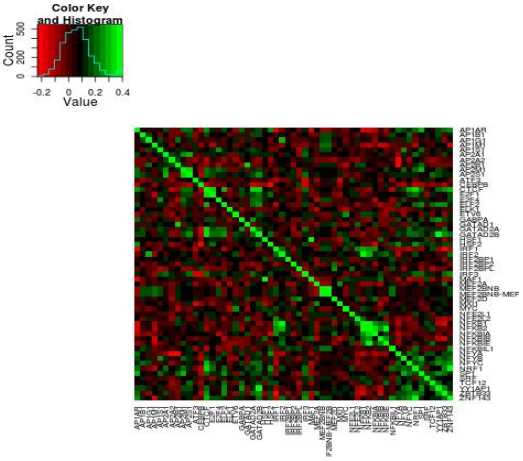
(a) Unfiltered: Mean, median, min, and max of genes



(b) Filtered: Mean, median, min, and max of genes



(c) Correlation of cell line replicates



(d) Correlation of transcription factors

Fig. 1: Basic Statistics and Correlation Analysis of Data

2.4. *Clustering of genes*

The first question I was interested in was whether or not the genes would cluster into biologically meaningful clusters based on the expression levels across cell lines. To cluster the genes, I used MATLAB's version of k-Means with values of $k = 20, 30, 50, 75, 100$ and correlation as the distance metric. To validate the results, I used GO term enrichment analysis, which computes the enrichment of each annotation term using the Pearson correlation coefficient. Terms with a p-value $< 1 \times 10^{-3}$ are considered significant. In increasing the value of k, I was hoping to see an increase in specificity of function in the clusters. I additionally compute motif enrichment of known transcription factors for the clusters. This is computed by intersecting the coordinates of the known TFs with a 200kb window around the transcription start site of each gene. I then calculate the p-values using a hypergeometric test and apply the Bonferroni correction. Significance is considered as p-value < 0.01.

2.5. *Clustering of experiments*

I was also interested if by clustering the cell lines, the types of cancers would form distinct clusters. To cluster the experiments, I used k-Means with $k = 15, 25, 50, 100$ and correlation as the distance metric. To validate the results, I consider the number of times the replicates cluster together. For smaller values of k, I expect to see multiple types of cancer clustering together (which I can validate from the literature), and in increasing k, I expect the clusters to become specific to one cancer type. As an annotation of the cell lines with cancer terms did not exist, I annotated them by hand from the literature and looked for clusters with enrichments. As the annotations were not well-defined, I did not perform a stringent enrichment analysis as in the gene clustering case.

2.6. *Module network analysis*

In addition to inferring biological relationships and evolution, I was interested in finding co-regulated genes (genes that are enhanced or repressed by a common set of regulators). To find modules of co-regulated genes, I used Genomica (citation), which uses an Expectation Maximization (EM) algorithm for finite Bayesian networks:

Repeat until convergence {
(E-Step) For each gene, determine which module's regulatory program best predicts the gene's expression
(M-Step) Learn the regression trees for a fixed partitioning of genes into modules
}

It has been previously demonstrated that this method is successful in producing modules of co-regulated genes for S. cerevisiae.[5] To validate the resulting modules, I examine the regression trees for the regulation programs of a subset of clusters for known regulators from the literature. Future work will include GO term enrichment analysis as well as motif enrichment analysis on a subset of clusters.

3. Results

3.1. *Clustering of genes*

As anticipated, when clustering the genes using small values of k, the clusters were enriched for general functional terms, such as 'regulation of metabolic process' and 'nucleotide receptor activity'. This indicates that the gene clusters correspond to broad gene categories that each include a significant number of genes (Table 1). For large values of k, however, the clusters were enriched for more specific functions or pathways, such as 'sensory organ development' and 'viral reproduction' (Table 2). Using Genomica, I was able to look at the expression pattern of particular gene clusters for interesting trends. Figure 2 shows the expression pattern for Set 10 for k=75. Down-regulated genes are in green whereas up-regulated are in red. SFT2D1, one of the genes in this set, is correlated with Lymphoma according to disease association studies.[4] In particular, it is shown to be down-regulated in lymphoblastic leukemia. This is directly mirrored in the expression data, where the cell lines for cancers of lymphocytes (AML, lymphoma), show significant down-regulation for this gene set compared to other cell lines.

The motif enrichment analysis was inconclusive due to too stringent criteria for significance as well as the large size of the motif dataset. Of the 14.8 million transcription factor (TF) motifs, I selected 210,000, randomizing over the motifs and limiting the number of motifs selected per TF. In the future, I will try using a larger subset of the motifs and use the False Discovery Rate rather than the Bonferroni correction to report significance.

Table 1: Selected GO term enrichment for k=20

Cluster	Selected Gene	Enriched GO term	PValue	Set Size	Percent Set Hits
Set 12	ATF6	negative regulation of metabolic process	7.94E-05	47	12.77
Set 12	RPL5	regulation of transcription from RNA polII promoter	1.57E-04	47	12.77
Set 6	DAG1	potassium ion transport	2.06E-04	137	4.38
Set 10	ATPAF2	lymphocyte proliferation	2.48E-04	54	5.56
Set 18	UNG	nucleotide receptor activity	2.52E-04	72	4.17

Table 2: Selected GO term enrichment for k=75

Cluster	Selected Gene	Enriched GO term	PValue	Set Size	Percent Set Hits
Set 51	PLOD3	gamma-aminobutyric acid signaling pathway	8.68E-06	53	5.67
Set 10	SFTD21	lymphocyte proliferation	2.48E-04	54	5.56
Set 64	ZFYVE16	embryonic development	3.29E-04	36	8.34
Set 18	PSMD9	viral reproduction	3.43E-04	72	4.17
Set 74	MRS2	sensory organ development	4.84E-04	73	5.48

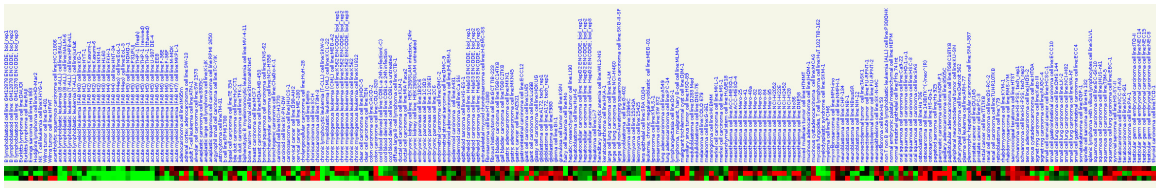


Fig. 2: Expression pattern for SFT2D1. Cancers of lymphocytes are clustered to the left and shown down-regulation.

3.2. *Clustering of experiments*

As hypothesized, the clusters of cell-lines did contain biological coherency for certain values of k. For k=15, the cell lines clustered according to broad categories of cancer, with carcinomas and sarcomas generally being split into distinct clusters. Notable results include a cluster that was enriched for different types of tumors (testicular, schwannoma); a cluster of acute myeloid leukemia (AML); and a cluster of cancers of the nervous system (merkel cell, medulloblastoma, neuroblastoma). For k=25, the clusters became more specific to a particular type of cancer. Notable clusters included mesothelioma; cancer of the uterus and embryo; cancer of lymphocytes; germ cell tumors; epithelial cancers; colon and liver cancers; and squamous cancers. For k=50 and k=100, there were too few cell lines per cluster and the results became random and uninformative. For all values of k, however, the cell line replicates were clustered together in every case.

3.3. *Module network analysis*

Some of the modules of co-regulated genes created by Genomica were indeed functionally coherent, representing regulation programs known from the literature. One example of this is shown in Figure 3, which depicts a regression tree rooted at HSF2, with a down-regulated subtree rooted at SP1 and an up-regulated subtree regulated by E2F4. The SP1 subtree also contains NFKB1. In the image, the leftmost columns are the myeloid leukemia cell lines, which are known to be regulated by a network including both SP1 and NFKB1. In fact, the promyelocytic leukemia protein inhibits SP1, which supports the down-regulation indicated in the regression tree.[2]

4. Conclusion and Future Work

I demonstrated that it is possible to extract biological meaning for the purposes of understanding, as well as potentially classifying, both genes and human cancer cell lines. I additionally showed that module network

Table 3: Clustering cell lines k=15

Cluster label type	Selected cell lines	Total in cluster	Fraction labeled
Tumor	*testicular germ cell, schwannoma*	12	0.5
Sarcoma	*bone marrow, pagetoid sarcoma*	9	0.67
Nervous system	*merkel cell, medulloblastoma, neuroectodermal*	20	0.7
Mesothelioma	*mesothelioma*	15	0.8
Myeloid leukemia	*acute myeloid leukemia, chronic myeloblastic leukemia*	17	1.0

Table 4: Clusters for k=25

Cluster label type	Selected cell lines	Total in cluster	Fraction labeled
Uterus/embryo	*leiomyoma, mixed mullerian*	10	0.5
Squamous	*oral squamous, bronchial squamous*	14	0.86
Kidney/liver	*clear cell, embryonic kidney, renal cell*	8	0.88
Colon/Liver	*colon carcinoma, hepatocellular carcinoma*	4	1.0
Germ cell tumor	*teratocarcinoma, testicular germ cell*	7	1.0
Lymphocyte	*Burkitt's lymphoma, chronic lymphocytic leukemia*	11	1.0

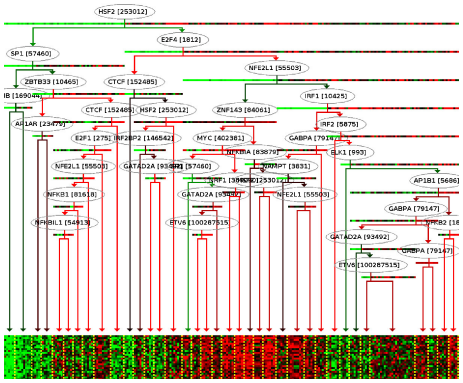


Fig. 3: Regulation tree for cluster 3098

analysis succeeds in finding accurate regulation programs. Since the module network analysis produced meaningful clusters of co-regulated genes, future work will predict gene expression using this method.

Acknowledgments

I would like to thank Sofia Kyriazopoulou-Panagiotopoulou for the project idea, mentoring, and access to the data.

References

1. A. Kundaje et al., *A predictive model of the oxygen and heme regulatory network in yeast*, (PLoS Computational Biology, Vol 4, Issue 11, Nov 2008).
2. S. Liu et al., *Sp1/NFKappaB/HDAC/miR-29b regulatory network in KIT-driven myeloid leukemia*, (Cancer Cell, Vol 17, Issue 4, April 2010).
3. C. Plessy et al., *Linking promoters to functional transcripts in small samples with nanaCAGE and CAGEscan*, (Nature Methods, Vol 7, No. 7, July 2010).
4. Nextbio disease atlas. (2011). Retrieved from www.nextbio.com.
5. E. Segal et al., *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data*, (Nature Genetics, Vol 34, No. 2, June 2003).
6. E. Segal et al., *Learning module networks*, Stanford Robotics Department. Retrieved from robotics.stanford.edu.
7. T. Shiraki et al., in *Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage*, (PNAS, Vol 100, No. 26, Dec 2003).