Graph-based Meta-Analysis of Gene Expression Biclustering

Daniel R. Chee[1] and Susan R. Atlas[2]

***Short Abstract* — Genomic datasets have the potential to shed light on the etiology of cancer through the use of biclustering to identify coherent patterns in the gene-patient expression matrix. A key step is the principled merging of overlapping biclusters to isolate unique *meta-clusters* associated with distinct dysregulated biological processes and pathways. We present a method motivated by Kruskal's algorithm wherein biclusters map to nodes and edge weights are computed from gene intersections. The resulting acyclic graph is searched to identify "hubbed" components for downstream pathway analysis. We illustrate the approach with the biclustering algorithm of Wang and Atlas applied to pediatric leukemia.**

***Keywords* — Biclustering, graph theory, greedy methods, Kruskal's algorithm, pathway analysis.**

I. Introduction

Recent years have seen a vast increase in the availability of genomic datasets with the potential to shed light on the etiology and treatment of human cancer. A key challenge is the need for new data mining approaches to analyze this data for biological significance. Biclustering is a powerful machine learning approach used to identify coherent patterns in the cancer gene expression matrix, but it often yields large numbers of biclusters, which can overlap significantly in their defining genes. Since overlapping biclusters are presumed to represent the same intrinsic biological process, it is important to further refine them into distinct groups, or *meta-clusters*, before proceeding with downstream pathway analysis. Here we present a novel graph-based approach that accomplishes this in a systematic and principled way.

II. Methodology

Our algorithm can be regarded as the final step in the genomic biclustering framework of Wang and Atlas [1,2]. This method uses a semi-supervised greedy search to identify statistically-robust patterns of up- and down-regulation within subsets of genes and samples. Overlapping biclusters are identified simultaneously and without the use of masking as in earlier 'top down' approaches [3]. Starting from exhaustively-enumerated seed patterns, genes and samples are added and deleted in order to refine biclustering patterns based on quality metrics. A supervised sign-reversal step enforces inverse correlation of expression patterns between class A and not-A patients. Only statistically robust biclusters, as determined via extensive dataset cross-validation and repartitioning, are retained.

The meta-clustering analysis is applied to this final set of biclusters. The method is an adaptation of Kruskal's algorithm [4] that constructs a weighted graph, with each bicluster corresponding to a node, and edge weights computed based on gene intersection. Only edges above a specified threshold, and that do not create a cycle, are included. The resulting graph is searched to yield a list of connected components. We select from these components only those centralized around one or two nodes with degrees higher than the rest of the nodes in that component. These "hubbed" components become the final meta-clusters that can be analyzed using curated databases to construct gene linkages and pathways [5].

III. Results and Conclusion

Meta-cluster analysis significantly reduces the space of biclusters generated by machine learning algorithms in a manner consistent with subsequent biological interpretation. As a demonstration of the method, 1175 robust biclusters derived from the UNM P9906 pre-B acute leukemia cohort (N=207 samples, 54675 genes [6]), were successfully merged into 12 unique meta-clusters. These were interpreted via pathway analysis to define possible new disease subtypes characterized by unique genomic signatures.

Acknowledgements: We gratefully acknowledge support from the UNM Initiatives to Maximize Student Diversity Program, NIH Grant No. GM-060201 (D Chee), and the UNM Cancer Center Shared Resource for Bioinformatics and Computational Biology, supported by D.H.H.S NIH/NCI P30 Grant CA 118100. We thank the UNM Center for Advanced Research Computing for the computational resources used in this work.

[1]Department of Computer Science, Department of Biology, and Center for Advanced Research Computing, University of New Mexico, Albuquerque NM, 87131. E-mail: dchee7@unm.edu

[2]Department of Physics and Astronomy and UNM Cancer Center, University of New Mexico, Albuquerque, NM 87131. E-mail: susie@sapphire.phys.unm.edu

References

[1] Wang X (2007). *Hybrid neuro-fuzzy inference models for outcome prediction in acute leukemia using gene expression and covariate data.* Ph.D. dissertation, University of New Mexico.

[2] Wang X, Chee DR, Willman CL, and Atlas SR (2012). *Robust semi-supervised biclustering for discovery of novel phenotypes in cancer expression data,* arXiv:q-bio for submission to Bioinformatics.

[3] Cheng, Y and Church GM (2000). Proc. 8th International Conference on Intelligent Systems for Molecular Biology: 93D103.

[4] IPA Version 9.0, Ingenuity Systems, Inc., Redwood City, CA.

[5] Kleinberg, J and Tardos, É. *Algorithm Design* (Addison-Wesley, San Francisco, 2006).

[6] Kang, H *et al.* (2010). Blood **115**, 1394-1405.

Nothing should be here on page 2! Please limit your abstract to a single page, and create a one-page .pdf file for submission.