Journal of Statistics: Advances in Theory and Applications
Volume 8, Number 1, 2012, Pages 1-13

USING DATA MINING TOOLS TO FIND SIMILARITIES IN GENETIC PREDICTORS FOR COLON CANCER RECURRENCE

JOHN IHRIE and MINDY HONG

Department of Statistics
North Carolina State University
Raleigh
North Carolina
USA
e-mail: jdihrie@ncsu.edu

Department of Math and Computer Science
Emory University
Atlanta
Georgia
USA
e-mail: mrhong@emory.edu

Abstract

Prognosis predictors based on gene lists have been proposed to supplement existing methods for predicting risk of recurrence in colon cancer patients. Currently, staging systems are used to assess risk in individual patients, but these systems often lack accuracy. Genetic predictors might improve risk assessment; however, different research teams often obtain dissimilar gene lists. In this study, web-based data mining tools are used to explore similarities of seven gene lists that are difficult to discern at the gene level. These lists are

2010 Mathematics Subject Classification: 62P10, 62-07, 62F07.

Keywords and phrases: biological networks, biological pathways, colon cancer, colorectal cancer, data mining, Genes2Networks, genetic predictor, graph theory, positive matching index, prognosis predictor, WebGestalt.

Received August 21, 2012

examined at three levels: gene, pathway, and network. WebGestalt is applied to identify statistically significant pathways in each list; Genes2Networks is then employed to search for relevant networks for each possible pair of lists and to create a network for all seven lists combined. Finally, the positive matching index is used to compare each list with each other list at all three levels. Even though gene sets showed little or no similarities at the gene level, similarities were generally greater at the pathway and network levels. Four non-list genes (AR, EGFR, GSN, and CEBPB) are identified in the combined-list network that might play a role in colon cancer recurrence. The results help support the widely held belief that biological networks play an important role in disease behaviour and suggest that these seven prognosis predictors might be more similar than they appear. Comparing genetic prognosis predictors might help scientists better understand the underlying biology of colon cancer and gene-based prediction.

1. Introduction

Colorectal cancer is the third highest cause of cancer-related death, as well as the fourth most common form of cancer in the United States. Although methods such as chemotherapy and surgery can be used to extend survival, recurrence has been shown to occur in approximately 20% of stage II colon cancer patients. As a result, researchers have proposed genetic prognosis predictors (PPs) to supplement existing methods for predicting risk of recurrence in colon cancer patients [2, 3, 9, 10, 12, 17]. The ability to accurately predict risk is important for physicians to determine the best treatment for an individual [2, 3, 9, 10, 12, 17]. Physicians use staging systems based on readily observable tumor characteristics to assess risk; however, these systems often lack accuracy [9, 10, 12, 17]. Genetic PPs can improve risk assessment, but different research teams often obtain dissimilar gene lists. Similarities in PPs that are difficult to discern at a superficial level are then explored. In this study, PPs are analyzed at three levels: gene, pathway, and network.

Past studies have indicated that different gene signatures are in actuality quite similar when examined past the individual gene level. As a result, web-based tools that examine pathway and network levels are used to analyze gene signatures in an effort to reveal undisclosed correlations. The web-based gene set analysis toolkit (WebGestalt) [20]

performs functional enrichment analyses of gene lists and identifies statistically significant biological pathways. Genes2Networks [4] searches protein-protein interaction databases and creates networks of interconnected genes (the connections are interactions). All possible pairs of seven PPs are then examined by using the positive matching index (PMI) [8], a similarity coefficient such that $0 \leq \text{PMI} \leq 1$; it is purported to have better characteristics than traditional similarity indices such as the Jaccard coefficient. All seven PPs are then combined into one list and the resulting network from Genes2Networks is treated as a graph with genes as nodes (vertices) and interactions as edges. The graph is analyzed to compare topology parameters (node degree and betweenness centrality) from graph theory. Node degree and betweenness centrality are two measures of vertex centrality in a graph [19]. These measures are related to how significant a node is in determining a graph's layout [19]. Vertex betweenness centrality is also a measure of how much control a node has over interactions between other nodes in the network [19]. These parameters could help to determine which genes are more important in the network.

The results from the pair-wise comparisons show that similarities are generally greatest at the network level, followed by the pathway level. Similarities, if any, are small at the gene level. The network created from all PPs suggests that some of the seed (original list) genes might be more important than others. Furthermore, the network suggests the possible existence of important contributing genes (intermediate nodes) that are not found in the seed lists.

2. Methods

2.1. Data sets

Seven gene lists (PP1-PP7) from six different published articles were used. Since gene nomenclature and identification change rapidly, some ambiguity exists in the identification of gene symbols. The following counts (lists available upon request) of gene symbols were identified from

the given tables in the referenced papers for PP1 to PP7, respectively: 25 from Table 3 [3], 66 from Table 1 [2], 29 from Table 2 [2], 34 from Table 5 [9], 45 from Table S2 [10], 119 from Table S2 [12], and 34 from Table 1 [17].

2.2. Data mining and statistical methodology

The lists were individually uploaded to WebGestalt and tested for statistically significant pathways by using Wikipathways analysis. The parameters follow: Organism = hsapiens, Gene ID type = gene symbol, Reference set = entrezgene, Significance level = .05, Statistical method = hypergeometric, Multiple test adjustment = BH, and Minimum number of genes for a category = 2.

WebGestalt determines significant pathways by comparing how many genes from the uploaded list are in a given pathway with how many genes would be expected, based on the reference gene set [20]. Suppose a list has n total genes with k genes being in a particular pathway. Suppose the reference set (all human genes in this analysis) has m total genes with j genes in that pathway. If the gene lists were a random sample from the reference set, we would expect the number of genes in that pathway to be $k_e = (n/m)^* j$ for that list [20]. If $k > k_e$, the pathway is considered enriched [20]. To test the significance of the enrichment, WebGestalt uses a hypergeometric test [20]. To calculate the probability of observing k or more pathway-genes in the list, assuming the lists were a random sample (i.e., to find the raw p-value), WebGestalt uses the following calculation [20]:

$$P = \sum_{i=k}^{n} \frac{\binom{m-j}{n-i}\binom{j}{i}}{\binom{m}{n}}.$$

WebGestalt also allows for multiple comparison adjustment (Benjamini & Hochberg was used for this analysis) [20]. Only pathways with adjusted p-values less than .05 were used in PP comparisons.

The lists were then combined into all possible pairs, uploaded to Genes2Networks, and analyzed by using default parameters. All intermediate nodes (genes or proteins) identified were used in subsequent analyses regardless of relative rank (*Z*-score). All seven lists were also combined and analyzed. Genes2Networks uses a binomial proportions test to calculate *Z*-scores for testing the significance of intermediates [4]. Suppose a is the number of links from an intermediate to nodes in the seed list, b is the number of links for the intermediate in the reference (background) network, c is the total number of links in the determined network, and d is the total number of links in the reference network. Genes2Networks uses the following calculation [4] for the *Z*-score:

$$z = \frac{\frac{a}{c} - \frac{b}{d}}{\sqrt{\frac{\frac{b}{d} \cdot \left(1 - \frac{b}{d}\right)}{d}}}.$$

2.3. Pair-wise comparisons

Pair-wise comparisons of gene lists, pathway lists, and interconnected genes in networks were made by using the PMI. Dissimilarities between PPs were measured as distances between gene lists, calculated as 1-PMI (similar to the Jaccard distance). A PMI distance of 0 means the sets are identical, whereas a distance of 1 means the sets are completely different. The PMI was based on the contingency table [8] in Figure 1.

		PPY	
		Present	Absent
PPX	Present	e	f
	Absent	g	h

Figure 1. Contingency table comparing gene symbol lists X and Y.

Heat maps were used to visualize the results (Figure 4). When calculating PMI distances for the list pairs, the values of e, f, and g depended on the level of comparison. At the gene (pathway) level, e was defined as the number of genes (pathways) common to both seed lists; f and g were defined as the numbers of genes (pathways) present in one seed list but not the other. Since the PMI only considers positive attributes, h is never included in the calculation (h was also irrelevant in this application).

At the network level, only seed genes and intermediates in the resulting pair-wise networks were considered. Intermediate nodes that provided a path to connect seed genes from PPX without "passing through" seed genes from PPY were included in the list for PPX, and vice versa. The rationale for this method was that these intermediates provided paths for seed genes of the same list to interact; therefore, they were considered part of that list. Intermediate nodes that connected seed genes from PPX to seed genes from PPY were included in both lists. The count of genes (both seed and intermediate) common to both lists was defined as e, while f and g were the counts of genes in one list but not the other.

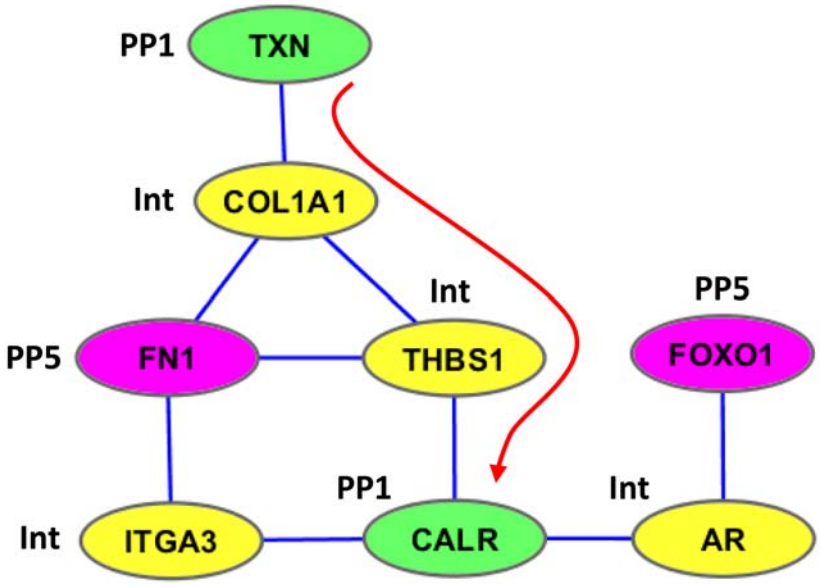

Figure 2. Network created from PP1 and PP5, including intermediates (Int).

Cytoscape [14] 2.7.0 was used to better visualize and analyze the results from Genes2Networks. For the PP1/PP5 network (Figure 2), *e* consisted of all the intermediates (COL1A1, THBS1, ITGA3, and AR); *f* consisted of TXN, COL1A1, THBS1, and CALR (arrow path), while *g* consisted of only FN1 and FOXO1 (CALR prevented a path of intermediates to exist between the seed genes).

2.4. Combined-list network

After the seven lists were combined and analyzed with Genes2Networks, the Cytoscape plugin NetworkAnalyzer [1] was used to calculate the degree and betweenness centrality (BC) of each node (gene or protein) in the resulting network (Figure 3). Node degree is the number of edges connected to a node. BC is calculated as follows [19]:

$$BC(\nu) = \sum_{s \neq \nu \neq t} \frac{\sigma(s, t, v)}{\sigma(s, t)},$$

where $\sigma(s, t, v)$ is the number of shortest paths between nodes s and t that contain node ν, and $\sigma(s, t)$ is the total number of shortest paths between s and t [19]. NetworkAnalyzer normalizes this to a value between zero and one. The network was treated as an undirected graph (i.e., the direction of interaction between genes was ignored), and the six small disjoint groups (Figure 3, upper right) were not used in the calculations because the nodes must be connected to analyze paths. Intermediates with larger node degree or BC were further explored with a literature search for possible findings by other researchers.

3. Results

3.1. Data mining and statistical methodology

WebGestalt identified the following pathway counts (lists available upon request) for PP1-PP7: 1, 12, 2, 4, 11, 13, and 3, respectively. The most common pathways were focal adhesion (4 PPs), apoptosis (3 PPs), IL-6 signaling pathway (3 PPs), and myometrial relaxation and contraction pathways (3 PPs).

Genes2Networks found networks (only PP1/PP5 shown, Figure 2) for all gene-list pairs except PP1/PP3 and PP3/PP4. The combined-list network (Figure 3) contained 193 nodes (88 seed, 105 intermediate) and 382 edges (interactions).

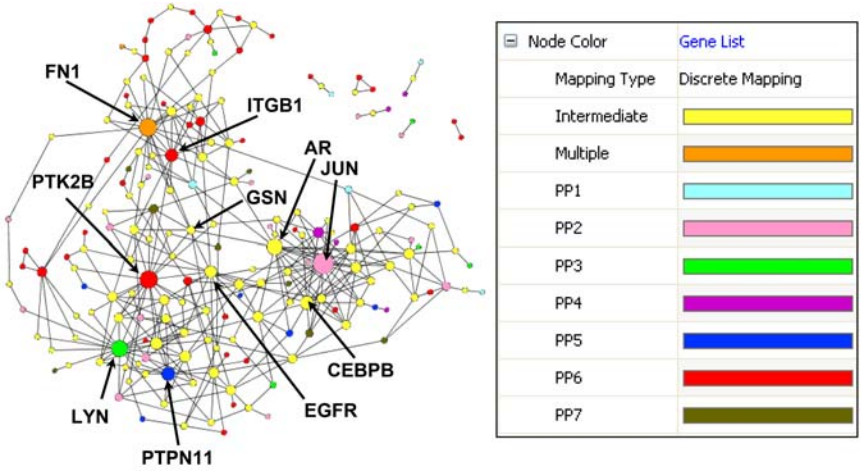

Figure 3. Combined-list network resulting from all seven lists.

The connected (non-disjoint) part of the network contained 177 nodes (76 seed, 101 intermediate) and 371 edges.

3.2. Pair-wise comparisons

Heat maps (Figure 4) show the pair-wise PMI distances. For example, the PMI distance between PP2 and PP4 at the pathway level is 0.83.

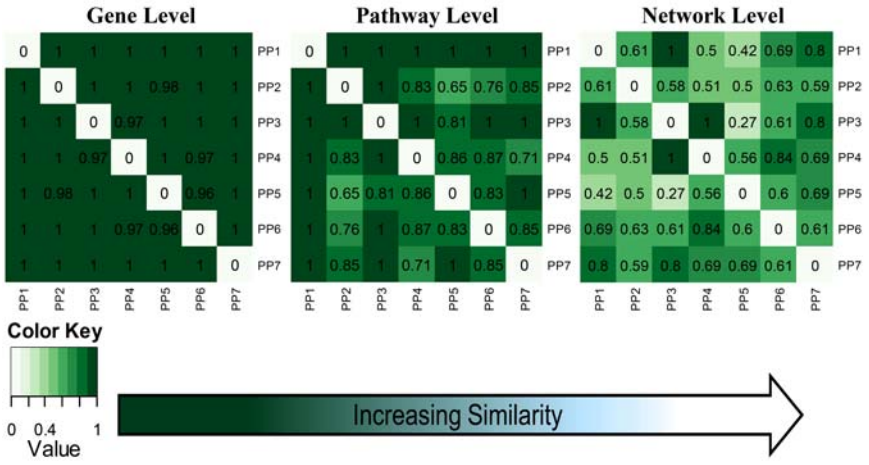

Figure 4. Heat maps of PMI distances at each level.

3.3. Combined-list network

In the combined-list network (Figure 3), four intermediates (AR, EGFR, GSN, and CEBPB) and five seed genes (FN1, PTK2B, JUN, LYN, and ITGB1) had relatively high degree and/or BC (Figure 5).

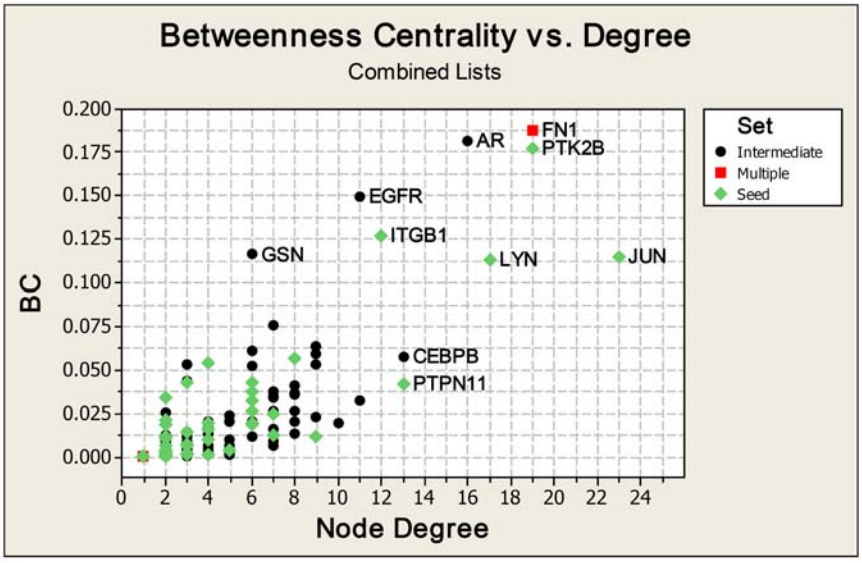

Figure 5. Measures of node (vertex) centrality for combined-list network.

FN1 was in two different PPs. The Genes2Networks *Z*-score ranks of AR, EGFR, GSN, and CEBPB were 96, 80, 47, and 62 out of 101, respectively.

4. Discussion

Genetic predictors might help to determine risk for colon cancer recurrence, but differences in gene lists should be examined. The PPs showed little or no similarities by PMI distance (Figure 4) at the gene level. The pathway level showed more similarities, while the network level showed the greatest similarities; however, some PMI distances did not change. In fact, PP3/PP4 had a slightly larger PMI distance at the pathway level than at the gene level (no resulting network), which could be due in part to the small numbers of pathways compared to seed genes. As a comparison, the above procedure was performed with Jaccard distances. The results (not shown) were similar but less pronounced.

Since node betweenness centrality measures how much control a node (gene or protein) exerts over interactions between other nodes [19] and node degree measures the number of connections to other nodes, the genes identified in Figures 3 and 5 might play an important role in colorectal cancer. The moderate to poor *Z*-score rankings of the intermediates (AR, EGFR, GSN, and CEBPB) suggest they might have only been in the network because of the structure of the reference network [4]; however, the AR (androgen receptor) gene was previously shown to be associated with colorectal cancer risk [16]. Also, the EGFR (epidermal growth factor receptor) gene is often targeted in the treatment of metastatic colorectal cancer [11]. The GSN (gelsolin) gene was found to be jointly differentially expressed with certain genes in colon cancer tumor samples [7]. The CEBPB (CCAAT/enhancer binding protein, beta) gene was shown to be correlated with Keratin23 expression in colon tumors; Keratin23 was shown to be related to the microsatellite instability (MSI) of the tumors [5].

Further support for these results might be obtained by using WebGestalt analysis methods other than Wikipathways. Also, more genetic predictors for colorectal cancer could be considered. Further analyses could be performed on the combined-list network using a mixed (directed and undirected edges) and weighted graph, although Cytoscape does not presently have that option. Such an analysis might result in more meaningful centrality measures. Furthermore, studies involving a PP with some or all of the seven combined lists could be performed to assess possible improvements in prognosis accuracy.

These results help support the widely held belief that biological networks play an important role in disease behaviour. Comparing multiple PPs might help scientists better understand the biology behind genetic prediction. Furthermore, web-based data mining and network analysis tools provide a convenient method for finding possible hidden similarities among PPs.

Acknowledgement

We would like to thank Dr. Don Hong for supervising this project at Middle Tennessee State University, as well as Dr. Bing Zhang for his guidance of the project at Vanderbilt University. We would also like to thank the National Science Foundation (NSF) for scholarship and grant support.

References

[1] Y. Assenov, F. Ramirez and S. E. Schelhorn et al., Computing topological parameters of biological networks, Bioinformatics 24 (2008), 282-284.

[2] A. Barrier, A. Lemoine and P. Y. Boelle et al., Colon cancer prognosis prediction by gene expression profiling, Oncogene 24 (2005), 6155-6164.

[3] A. Barrier, P. Y. Boelle and F. Roser et al., Stage II colon cancer prognosis prediction by tumor gene expression profiling, J. Clin. Oncol. 24 (2006), 4685-4691.

[4] S. I. Berger, J. M. Posner and A. Ma'ayan, Genes2Networks: Connecting lists of gene symbols using mammalian protein interactions databases, BMC Bioinformatics 8 (2007), 372.

[5] K. Birkenkamp-Demtroder, F. Mansilla and F. B. Sorensen et al., Phosphoprotein Keratin23 accumulates in MSS but not MSI colon cancers in vivo and impacts viability and proliferation in vitro, Molecular Oncology 1 (2007), 181-195.

[6] J. L. Chen, J. Li, W. M. Stadler and Y. A. Lussier, Protein-network modelling of prostate cancer gene signatures reveals essential pathways in disease recurrence, Journal of the American Medical Informatics Associations: JAMIA 18 (2011), 392-402.

[7] M. Dettling, E. Gabrielson and G. Parmigiani, Searching for differentially expressed gene combinations, Genome Biology 6 (2005), R88.

[8] D. A. Dos Santos and R. Deutsch, The positive matching index: A new similarity measure with optimal characteristics, Pattern Recognition Letters 31 (2010), 1570-1576.

[9] S. Eschrich, I. Yang and G. Bloom et al., Molecular staging for survival prediction of colorectal cancer patients, J. Clin. Oncol. 23 (2005), 3526-3535.

[10] K. Garman, C. Acharya and E. Edelman et al., A genomic approach to colon cancer risk stratification yields biologic insights into therapeutic opportunities, PNAS 105 (2008), 19431-19436.

[11] V. Heinemann, S. Stintzing and T. Kirchner et al., Clinical relevance of EGFR- and KRAS-status in colorectal cancer patients treated with monoclonal antibodies directed against the EGFR, Cancer Treatment Reviews 35 (2009), 262-271.

[12] R. Jorissen, P. Gibbs and M. Christie et al., Metastasis-associated gene expression changes predict poor outcomes in patients with dukes stage B and C colorectal cancer, Clin. Cancer. Res. 15 (2009), 7642-7651.

[13] Y. H. Lin, J. Friederichs, M. A. Black, J. Mages and R. Rosenberg et al., Multiple gene expression classifiers from different array platforms predict poor prognosis of colorectal cancer, Clinical Cancer Research 13 (2007), 498-507.

[14] P. Shannon, A. Markiel and O. Ozier et al., Cytoscape: A software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003), 2498-2504.

[15] M. Shi, D. R. Beauchamp and B. Zhang, A network-based gene expression signature informs prognosis and treatment for colorectal cancer patients, (2011).

[16] M. L. Slattery, C. Sweeney and M. Murtaugh et al., Associations between vitamin D, vitamin D receptor gene and the androgen receptor gene with colon and rectal cancer, International J. of Cancer 118 (2006), 3140-3146.

[17] J. J. Smith, N. G. Deane and F. Wu et al., Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer, Gastroenterology 138 (2009), 958-968.

[18] Y. X. Wang, T. Jatkoe, Y. Zhang, M. G. Mutch and D. Talantov et al., Gene expression profiles and molecular markers to predict recurrence of dukes' B colon cancer, Journal of Clinical Oncology 22 (2004), 1564-1571.

[19] J. Yoon, A. Blumer and K. Lee, An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality, Bioinformatics 22 (2006), 3106-3108.

[20] B. Zhang, S. Kirov and J. Snoddy et al., WebGestalt: An integrated system for exploring gene sets in various biological contexts, Nucleic Acids Res. 33 (2005), W741-W748.

■