Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Array Platform Modeling and Analysis (A)

Integrative Pathway Analysis Using Graph-Based Learning with Applications to TCGA Colon and Ovarian Data

Andrew E. Dellinger,[1,2] Andrew B. Nixon,[3] and Herbert Pang[2,4]

[1]Department of Mathematics and Statistics, Elon University, Elon, NC, USA. [2]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA. [3]Department of Medicine, Division of Medical Oncology, Duke University School of Medicine, Durham, NC, USA. [4]School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.

ABSTRACT: Recent method development has included multi-dimensional genomic data algorithms because such methods have more accurately predicted clinical phenotypes related to disease. This study is the first to conduct an integrative genomic pathway-based analysis with a graph-based learning algorithm. The methodology of this analysis, graph-based semi-supervised learning, detects pathways that improve prediction of a dichotomous variable, which in this study is cancer stage. This analysis integrates genome-level gene expression, methylation, and single nucleotide polymorphism (SNP) data in serous cystadenocarcinoma (OV) and colon adenocarcinoma (COAD). The top 10 ranked predictive pathways in COAD and OV were biologically relevant to their respective cancer stages and significantly enhanced prediction accuracy and area under the ROC curve (AUC) when compared to single data-type analyses. This method is an effective way to simultaneously predict binary clinical phenotypes and discover their biological mechanisms.

KEYWORDS: multi-dimensional genomic data, integrative analysis, clinical outcome prediction, serous cystadenocarcinoma, colon adenocarcinoma

SUPPLEMENT: Array Platform Modeling and Analysis (A)

CITATION: Dellinger et al. Integrative Pathway Analysis Using Graph-Based Learning with Applications to TCGA Colon and Ovarian Data. *Cancer Informatics* 2014:13(S4) 1–9 doi: 10.4137/CIN.S13634.

RECEIVED: November 12, 2013. **RESUBMITTED:** March 17, 2014. **ACCEPTED FOR PUBLICATION:** March 18, 2014.

ACADEMIC EDITOR: J.T. Efird, Editor in Chief

TYPE: Technical Advance

FUNDING: Authors disclose no funding sources.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: herbert.pang@duke.edu

This paper was subject to independent, expert peer review by a minimum of two blind peer reviewers. All editorial decisions were made by the independent academic editor. All authors have provided signed confirmation of their compliance with ethical and legal obligations including (but not limited to) use of any copyrighted material, compliance with ICMJE authorship and competing interests disclosure guidelines and, where applicable, compliance with legal and ethical guidelines on human and animal research participants. Provenance: the authors were invited to submit this paper.

Introduction

The wealth of publicly available genomic data can be more greatly leveraged if more than one genomic data type is available per patient. For example, genomic data types like copy number variations (CNVs), single nucleotide polymorphisms (SNPs), DNA methylation levels, and microRNA (miRNA) expression can all inform gene expression or function. Integrating these genomic data types into a single model can better inform researchers about the nature of the gene networks, functional pathways, and biological interactions involved in disease. Each genomic data type used in an integrative method gives information on a different aspect of biology, such as mutation, regulation, and expression. Integration of these data types, which are functionally connected, can form a more biologically realistic model and enhance the accuracy of integrative models' predictions. The Cancer Genome Atlas (TCGA)[1] contains such a set of multiple genomic data types for several cancers like serous cystadenocarcinoma (OV)[2] and colon adenocarcinoma (COAD)[3] used in this study. TCGA datasets have been the data source of many recent investigations into how to best leverage the integration of multiple genomic data types to discover new biological connections.[4–6]

Integration of multiple genomic data types has been used to build interaction[7,8] and coexpression networks[9] of genes that have biological significance to cancer. Genes in these networks were tested for overrepresentation within a biological pathway, revealing important cancer mechanisms. Pathway analysis integrating genomic data types has previously been performed

by testing individual genes to infer pathway activities and rank pathways.[10] It has also been performed by pathway-based analyses on single data types followed by Fisher's method to combine the results into a single *P*-value.[11]

The method presented herein is the first to integrate multiple genomic data types with graph-based learning in a pathway-based manner, which utilizes the entire pathway instead of single genes or the entire set of genes in the dataset. Unlike other integrative methods, pathway-based analyses ask not only how well these data types perform as a binary classifier but also what biological mechanisms underlie the differences between the two classes. Pathway-based analyses also have the advantage of incorporating prior biological knowledge,[12] unlike methods whose results are used to find relevant pathways by overrepresentation of significant genes. This prior knowledge has been advocated for genomic analyses by Chasman,[13] Peng et al,[14] and Ritchie.[15] Pathway-based analyses also reduce computational burden and increase the interpretability of the results.[16] To identify significant pathways, previous studies evaluated all individual genes either to build a custom pathway via their relationships or to evaluate differential expression or overrepresentation within known pathways. This is the only known study of multi-dimensional genomic data to perform graph-based classification at a pathway level.

Unlike Tsuda et al,[17] the integrative pathway analysis algorithm uses data from the set of genes in a pathway, not from the entire genome. Pathway analysis produces a computational benefit in reducing complexity and makes the results more interpretable by incorporating prior biological knowledge.[16] It uses a Monte Carlo hold-out internal validation method. It accepts Spearman correlation (not presented) and Gaussian kernel distance weight matrices. It also accepts patient-level data useful in clinical studies both to classify patients and to discover biological mechanisms.

Integrative pathway analysis was conducted on OV and COAD datasets to predict stage and determine the biological mechanisms underlying advancing stage. The algorithm allows these pathways to be ranked based on the predictive power of multiple –omics data types. As the first integrative genomic study on COAD, pathways not previously described as advancing stage are presented.

Materials and Methods

Datasets. The pathway dataset includes 171 KEGG[18] pathways and 347 BioCarta[19] pathways. Data from TCGA[1] were at level 3. This means methylation data are a probe's beta value [0,1]; gene expression data are an average of all probes covering a gene; and SNP is a mean value of all probes in a segmented chromosomal region. Genes within a segmented chromosomal region were given that region's value.

Patients included in the analysis of the OV dataset were limited to those with all of the following: Illumina Infinium HumanMethylation27 data, Agilent Human Genome CGH 1×1 M CN data, Agilent 8×15 K human miRNA-specific data, Agilent 244 K Custom Gene Expression Array G4502A-07 data, and Affymetrix Genome-Wide Human SNP Array 6.0 data. Patients also had to have known cancer stage. Patients who met all these criteria were grouped into advanced (stage IV, n = 80) and non-advanced (stages IA–IIIB, n = 70) stages.

Patients included in the analysis of the COAD dataset were limited to those with known cancer stage and all of the following: Illumina Infinium HumanMethylation27 data, Agilent 244 K Custom Gene Expression Array G4502A-07 data, and Affymetrix Genome-Wide Human SNP Array 6.0 data. To create two nearly equal size groups, patients who met all these criteria were grouped into early (stages IA–IIB, n = 83) versus late (stages III–IVA, n = 61). Cancer stages are often classified or dichotomized as either "advanced" vs "non-advanced" stage of disease or "early" versus "late" stage of disease. Both allow researchers to better understand the disease mechanisms. Here we demonstrate two different ways to dichotomize the two diseases of interest.

Algorithm. The algorithm was run with one and three data types. The algorithm for single data-type analysis differs from the algorithm for integrative analyses in that no weighting of data types is used, as described below.

In graph-based semi-supervised learning, the set of patients is denoted as a vector y of size n, where n is the total number of patients in the dataset. A patient y in {–1,1} is either –1, denoting early stage, or 1, denoting advanced stage. The rest of this Materials and Methods section is repeated over 50 seeds, sampling without replacement at each iteration, where the training set T is 70% of the dataset and the test set V is 30% of the dataset.

For the patients in the test set, $y = 0$. Let there be P pathways, denoted by $p = 1,\ldots,P$. Let W_E, W_M, and W_S denote the distance weight matrices of E expression, M methylation, and S SNP, respectively. The expression matrix, E, is based solely on the Agilent 244 K microarray data and does not include expression of miRNA. The matrix is composed of distance measures w_{ijp}, which describe the strength of the relationship between patients i and j in pathway p with weight $w_{ijp} \geq 0$ when an edge is present. When $w_{ijp} = 0$, there is no edge between patients i and j for pathway p. w_{ijp} equals the r value of the Pearson correlation between i and j if $r \geq 0.5$; otherwise, the value is zero. This means the closer the two patient samples are in terms of the feature measures in a pathway, the larger the weight edge. The data used to calculate w_{ijp} are columns i and j in T.

Let the following be $n \times n$ matrices: I the identity matrix; D_E, D_M, and D_S the diagonal matrices, where $D = \text{diag}(d_i)$, $d_i = \Sigma_j w_{ijp}$; and L_E, L_M, and L_S the Laplacian matrices, $L = D - W$. For multiple data-type analysis, use a nonlinear optimization program to calculate a vector of optimal α values, where α is the set of weights for analyses using multiple data types. Optimization is by gradient descent, which requires the dual objective function and its derivative. For our question of interest, the primal objective function is $\min_\alpha(y^T(I + (\alpha_{Ep}{}^* L_{Ep}$

$+ \alpha_{Mp}{}^{*}L_{Mp} + \alpha_{Sp}{}^{*}L_{Sp}))^{-1}y)$ s.t. $\Sigma\alpha_k \leq c$, $k = 1, \ldots, m$, where m is the number of networks in the algorithm. The dual objective function is $\max_{\alpha}((y^Ty) - y^{T*}(I + (\alpha_{Ep}{}^{*}L_{Ep} + \alpha_{Mp}{}^{*}L_{Mp} + \alpha_{Sp}{}^{*}L_{Sp}))^{-1}y)$ s.t. $\Sigma\alpha_k \leq c$, $k = 1,\ldots,m$, where m is the number of networks in the algorithm. The solution of the primal objective function is equivalent to the solution of the dual objective function when the solution is fully optimized. The solution of the primal objective function is also superior to the solution of the dual objective function when finding an approximate solution.[20] Hence, the solution of the primal objective function is used herein. In this study, the number of networks is $m = 3$. There is an expression network, a methylation network, and a SNP network. c is a constant, determined as below in the analytical strategy section. k is 1, 2, or 3, the index of the relevant network Also, the derivative of the minimization function, $\partial d/\partial\alpha_j = -y^T(I + \Sigma_{k=1..m}(\alpha_k L_k))^{-1}L_j(I + \alpha^{*}L_{Ep} + \alpha_{Mp}{}^{*}L_{Mp} + \alpha_{Sp}{}^{*}L_{Sp})^{-1}y$, is used by the optimization program. The single data-type analysis will only use one diagonal matrix and one Laplacian matrix, for example D_E and L_E, without calculation of optimal α values.

The algorithm's output is a vector f_p of size n for pathway p. f_p is determined by all the available information. f_{ip} must not be too different from $f_{i'p}$ of adjacent nodes (smoothness), and f_{ip} must be close to a given label y_i in training nodes (loss). For three data types, $f_p = (I + \alpha_{Ep}{}^{*}L_{Ep} + \alpha_{Mp}{}^{*}L_{Mp} + \alpha_{Sp}{}^{*}L_{Sp})^{-1}y$. For a single data type, like SNP data, $f_p = (I + L_{Sp})^{-1}y$.

To classify each patient in V as $y = 1$ or $y = -1$, first compute the median f for the patients in T where $y = -1$ and the median f for the patients in T where $y = 1$. If the patient's f-score is closer to the median f of $y = -1$ than to the median f of $y = 1$, the patient will be classified as $y = -1$. Otherwise, the patient will be classified as $y = 1$.

The implementation of this algorithm was written in R version 2.14.0 using nonlinear optimization to find the optimal network weights.

Analytical strategy. The value of parameter c, a constant, was determined by testing multiple values over multiple pathways for the value that maximized area under the ROC curve (AUC) and accuracy. Gaussian kernel distance and Pearson correlation weight measures were determined for each pair of patients, where Gaussian kernel distance width $\sigma^2 = 1$ and Pearson weight equals the correlation coefficient r if $r \geq 0.5$ or zero otherwise. The cutoff is modeled after the network model cutoffs of Tsuda et al[17] and Deng et al,[21] which were not shown to be sensitive to the value of the coefficient cutoff.[21] The cutoff distinguishes related patients from unrelated patients. The optimization algorithm for α is a sequential quadratic programming (SQP) algorithm for nonlinearly constrained gradient-based optimization.

Accuracy of p is the proportion of correct classifications, comparing the predictions and the known status from the phenotype data. The AUC of p is a cutoff-independent performance measure and is equal to the value of the Wilcoxon—Mann–Whitney test statistic.[22]

To select the best distance measure between Pearson correlation and Gaussian kernel distance, two comparisons were made between the distances: OV accuracy and OV AUC. Five runs of the data were performed using different seeds to determine the training and validation sets. For the OV dataset, the runs consisted of the integrated analysis and individual analyses of gene expression, methylation level, and SNP data. The best distance measure for a comparison had both the integrative analysis with the most pathways above the threshold, and the integrative analysis with the most pathways both exceeding the threshold and exceeding the accuracy or AUC of all the one data-type analyses in at least three of five runs.

Significant pathways were determined over 50 runs. For OV and COAD, the integrated analysis and individual analyses of gene expression, methylation, and SNP data were compared. Significant pathways have two criteria. First, the integrative analysis exceeds a mean accuracy of 55% or a mean AUC of 0.55. Second, the mean of the integrative analysis measure (accuracy or AUC) minus the 97.5% lower confidence interval (LCI) using the standard error of the integrated measure (accuracy or AUC) over 50 runs must exceed the mean of each of the one data-type analysis measures. Pathways are ranked by the minimum difference between the LCI of the integrative measure and the mean of any single data-type measure.

Results and Discussion

Initial testing. The value of parameter c was determined by testing a glioblastoma multiforme dataset from TCGA. c = 25 was chosen as the value for all analyses, because it most frequently had the highest accuracy and AUC over five seeds in 20 pathways with 15–70 genes (data not shown). The main integrated analysis for OV and COAD included SNP, methylation, and gene expression data. Gaussian and Pearson distance measures were compared over five OV runs, as described in the Materials and Methods section. The results are as follows: 31 versus 9 pathways, respectively, with accuracy greater than 60%. Overall, 81 versus 55% of these pathways, respectively, had at least three of five runs where integrative accuracy exceeded all single data-type accuracies. The Pearson versus Gaussian AUC comparison results are as follows: 184 versus 95 pathways, respectively, with AUC >0.6. In all, 86 versus 58% of these pathways had at least three of five runs where integrative AUC exceeded all single data-type AUCs. Hence, Pearson correlation was a better measure for the OV dataset. Additionally, Pearson correlation has previously been successfully used in both pathway analyses[23] and other methods of data integration.[7,10] Thus, Pearson correlation was chosen as the distance measure for this algorithm.

COAD analyses. Patients were grouped into early stage (stages IA–IIB) and advanced stage (stage III to IVA) classes for pathway-based prediction. Pathways significantly improved by integrative analysis were discovered and ranked as described in the Materials and Methods section. There were

29 significant pathways found using accuracy and 22 using AUC. The biopeptide and Fc epsilon receptor I-mediated signaling pathways were significant in both accuracy and AUC measures. Fc epsilon receptor I may affect stage by inhibiting colorectal adenocarcinoma cell growth.[24] Biopeptide pathway genes may affect stage through colon cancer cell growth[25] and general tumor suppression.[26]

The remaining top 10 pathways from the accuracy (Table 1) and AUC (Table 2) analyses were shown to be biologically relevant to colon cancer stage as follows. The stage can be advanced through metastases spread by migration or motility, tumor growth by cell proliferation, failure of apoptosis, failure of cell cycle arrest, and other mechanisms.

In Table 1, cell-to-cell pathway function is related to stage and its genes individually are related to stage through their expression and their mediators.[27,28] The P38 MAPK signaling pathway regulates increase in stage through cell migration, apoptosis, and extravasation.[29–31] Growth hormone signaling pathway genes are correlated with stage through apoptosis and cell cycle arrest.[32,33] Toll-like receptor signaling pathway gene mutations are associated with stage through cancer growth and neoplastic progression.[34]

High concentrations of cholesterol are associated with more advanced stage;[35] thus, the cholesterol-lowering statin pathway could inhibit progression by inhibiting the growth of cancer cells[36] and decreasing polyp number and size.[37]

In Table 2, the differentiation pathway drives stage through its MAPK genes and cell cycle regulators.[29–31,38] The apoptosis and death pathways, which have many of the same genes, are related to stage through their roles in progression.[39] The transendothelial migration, integrin-mediated cell adhesion, and cell adhesion molecule pathways are interrelated by common genes and/or functions. They affect stage through metastasis, migration, and apoptosis.[40–42] Angiotensin II increases stage through its influence on tumor growth, invasion, and metastasis.[43,44]

Serous cystadenocarcinoma (OV) analyses. Patients were grouped into early stage (stages IA—IIIB) and advanced stage (stage IV) classes for pathway-based prediction. Pathways significantly improved by integrative analysis were discovered and ranked as described in the Materials and Methods section.

There were 63 significant pathways found using accuracy, and there were 192 significant pathways using AUC. The caspase and maturity onset diabetes of the young pathways are common to the top 10 pathways in both accuracy (Table 3) and AUC (Table 4). Caspases affect stage through apoptosis.[45] Ovarian cancer patients with diabetes are more likely to be diagnosed at a higher stage and have shorter survival time.[46] Estradiol synthesis[47] and overexpression in the glycolysis pathway[48] are underlying factors. The insulin receptor tyrosine kinase called anaplastic lymphoma kinase (Table 3) controls proliferation and apoptosis in ovarian cancer.[49,50]

In Table 3, the $G_{12\alpha}$ and $G_{s\alpha}$ pathways are connected to stage by G-protein regulation of gonadotropin-induced ovarian cancer cell proliferation and protein kinase C regulation of angiogenesis.[51] The ACE2 pathway is involved in tumor growth, angiogenesis, and metastasis.[52] Regulatory T cells in advanced stage ovarian cancer lead to significant immune suppression.[53]

Alzheimer's disease pathway genes are linked to previously described roles in ovarian cancer, such as insulin metabolism, caspases, and immune response. More than 60% of the genes in this pathway and 87% in the neurodegenerative disease pathway (Table 4) are apoptosis related.[54]

Figure 1 is an example of the network graph of a test set *V* in the OV dataset. This network graph was used to predict each patient's stage (early or advanced) in *V*.

In Table 4, the leukocyte transendothelial migration pathway is associated with the survival phenotype in multiple tests and studies.[55] Colony stimulating factors in the stem cell pathway give a poor prognostic outlook regarding stage and survival.[56] Interleukins regulate anchorage-independent

Table 1. Top COAD accuracy improvements by combining three data types.

PATHWAY	MEAN GENE	MEAN METHYLATION	MEAN SNP	MEAN 3 TYPES (97.5% LCI)
Cell-to-Cell	54.0	53.9	53.0	63.3 (61.5)
Biopeptides	52.8	53.0	51.8	59.5 (57.5)
Thrombopoietin	54.1	51.8	54.6	58.6 (57.1)
Cholesterol biosynthesis	57.7	58.0	57.1	62.3 (57.5)
Statin	59.1	56.7	60.5	64.7 (60.8)
Nucleotide metabolism	55.1	57.9	57.5	61.5 (58.5)
Fc epsilon receptor I-mediated signaling	57.0	55.2	50.2	61.1 (57.1)
Growth hormone signaling	58.3	53.0	53.1	62.0 (58.5)
Toll-like receptor signaling	58.0	57.3	49.0	61.4 (58.2)
P38 MAPK signaling	54.1	57.0	43.5	60.2 (57.2)

Notes: The top 10 ranked pathways using the accuracy measure in the COAD dataset. "Mean" denotes the mean accuracy of the pathway's classification of early versus advanced stage over 50 iterations. LCI is calculated as defined in the Materials and Methods section.

Table 2. Top COAD AUC improvements by combining three data types.

PATHWAY	MEAN GENE	MEAN METHYLATION	MEAN SNP	MEAN 3 TYPES (97.5% LCI)
Differentiation in PC12 cells	0.52	0.53	0.55	0.66 (0.64)
Leukocyte transendothelial migration	0.60	0.60	0.53	0.68 (0.65)
Cell adhesion molecules	0.60	0.61	0.59	0.68 (0.66)
BCR	0.53	0.48	0.53	0.61 (0.59)
Apoptosis	0.62	0.56	0.53	0.69 (0.67)
Biopeptides	0.57	0.51	0.57	0.64 (0.62)
Integrin mediated cell adhesion	0.55	0.55	0.50	0.62 (0.59)
Angiotensin II mediated activation of JNK	0.55	0.51	0.57	0.63 (0.61)
Death	0.68	0.60	0.63	0.73 (0.71)
Fc epsilon receptor I-mediated signaling	0.55	0.49	0.57	0.62 (0.60)

Notes: The top 10 ranked pathways using the AUC measure in the COAD dataset. "Mean" denotes the mean accuracy of the pathway's classification of early versus advanced stage over 50 iterations. LCI is calculated as defined in the Materials and Methods section.

growth, proliferation, and invasion.[57] The cytokine–cytokine receptor interaction pathway is enriched in patients with long survival time.[58] Ovarian cancer proliferation is stimulated by calcium signaling.[59] A great majority of colorectal cancer pathway genes drive stage, as they are Wnt,[54] caspase,[45] and diabetes related. In all,[46] 55% of top-ranked OV pathways advance cancer stage by apoptosis and 44% by cell proliferation. The top pathways in Tables 2 and 4 have the greatest difference between the larger AUC of the integrative analysis and the smaller AUC of the single data-type analyses.

Figure 2 demonstrates this fact for "maturity onset diabetes of the young" (Table 4). This pathway is a clear example of the relationship between exemplary prediction and underlying biology with regard to cancer stage. This top pathway in AUC and accuracy is known to advance ovarian cancer stage.[46]

We may observe that there may not be a high number of pathways that overlap between top pathways within each disease dataset. Accuracy measure is based on a single cutoff while AUC looks at the area under the ROC curve across multiple cutoffs. If the researcher is interested in how accurate a future sample can classify at a particular cutoff, the accuracy measure would be a better table to refer to. On the other hand, if researcher is interested in the overall performance of a test, eg, a diagnostic one, using a pathway, the AUC measure would be more suitable.

The aim of this study was to use a graph-based learning algorithm for multiple networks to find biological pathways that accurately classify disease stage. Each network is a genomic data type, like gene expression or miRNA expression. Integration of genomic networks increases the proportion of true classifications of stage in pathways critical to disease progression or status. Primarily, this method can be used to give insights into disease biology and progression. Pathway-based analysis can help researchers identify more biologically meaningful genomics markers than single-gene-based approaches. Finding these important pathways allows researchers to focus on smaller sets

Table 3. Top OV accuracy improvements by combining three data types.

PATHWAY	MEAN GENE	MEAN METHYLATION	MEAN SNP	MEAN 3 TYPES (97.5% LCI)
Caspase	53.2	53.5	54.7	60.4 (58.8)
Alzheimer's disease	56.3	52.9	56.9	62.0 (60.4)
Glycolysis and gluconeogenesis	61.4	54.9	55.0	65.6 (64.0)
ACE2	57.8	54.9	52.5	61.8 (60.4)
Maturity onset diabetes	56.6	54.1	58.0	62.1 (60.5)
Anaplastic lymphoma kinase	55.3	58.1	55.5	62.1 (60.5)
G alpha 12	54.6	46.8	51.5	58.2 (56.8)
T-cell receptor	52.6	54.2	54.0	58.5 (56.3)
Glycosphingolipid biosynthesis	58.5	51.5	58.4	62.2 (60.6)
G Alpha S	59.7	50.4	58.1	63.4 (61.7)

Notes: The top 10 ranked pathways using the accuracy measure in the OV dataset. "Mean" denotes the mean accuracy of the pathway's classification of early versus advanced stage over 50 iterations. LCI is calculated as defined in the Materials and Methods section.

Table 4. Top OV AUC improvements by combining three data types.

PATHWAY	MEAN GENE	MEAN METHYLATION	MEAN SNP	MEAN 3 TYPES (97.5% LCI)
Maturity onset diabetes of the young	0.57	0.57	0.59	0.69 (0.67)
Stem Cell	0.63	0.56	0.56	0.72 (0.70)
Cytokine-cytokine receptor interaction	0.65	0.60	0.57	0.73 (0.71)
Caspase	0.56	0.53	0.55	0.63 (0.62)
Alanine and aspartate metabolism	0.64	0.58	0.63	0.72 (0.70)
Neurodegenerative diseases	0.60	0.68	0.62	0.75 (0.73)
Histidine metabolism	0.61	0.60	0.50	0.69 (0.67)
Leukocyte transendothelial migration	0.64	0.63	0.63	0.71 (0.69)
Colorectal cancer	0.62	0.58	0.61	0.68 (0.67)
Calcium signaling	0.64	0.62	0.57	0.70 (0.68)

Note: The top 10 ranked pathways using the AUC measure in the OV dataset. "Mean" denotes the mean accuracy of the pathway's classification of early versus advanced stage over 50 iterations. LCI is calculated as defined in the Materials and Methods section.

of genes that explain the response of interest. We demonstrated the use of our tool with colon and ovarian TCGA datasets. It is well known that genes work together in groups. A graph-based classification algorithm takes into account the correlations among biomarkers in pathways and is an ideal algorithm to use for performing integrative pathway analysis.

The methodology of this study is similar to that of Kim et al[58] in that it is a graph-based classification algorithm using multiple networks, although there are significant differences across the two methodologies. In this study, the sets of data types were not preselected as in Kim et al but were the best three of five available data types. Pearson correlation coefficients performed better in the OV dataset as a measure of edge weight than the Gaussian function of Euclidean distance used in the OV dataset of Kim et al.[57] Instead of one testing–training set with an unknown number of samples withheld for

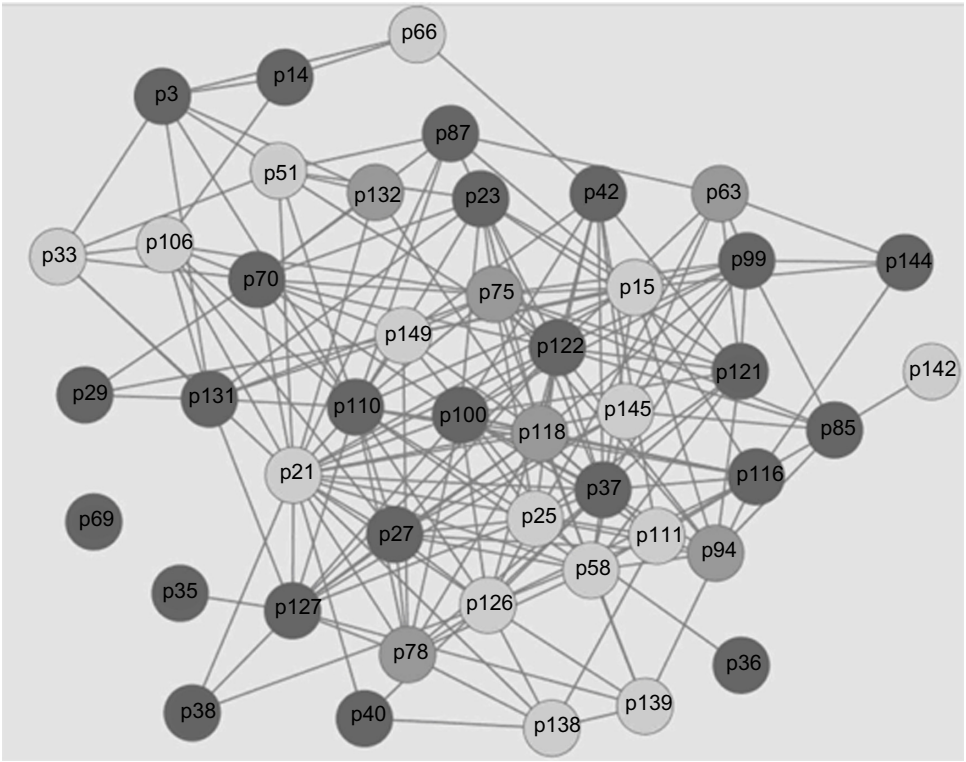

Figure 1. Caspase pathway validation network in OV. This figure represents the network of patients discovered in testing the caspase pathway in ovarian cancer. Nodes represent patients. The top 200 weighted edges are shown. Weights were determined using α and Pearson correlation coefficients of the integrated data types. Light gray nodes are incorrect integrative method predictions. Medium gray nodes are correct predictions by all data types. Dark gray nodes are correct integrative method predictions and at least one incorrect single data-type prediction.

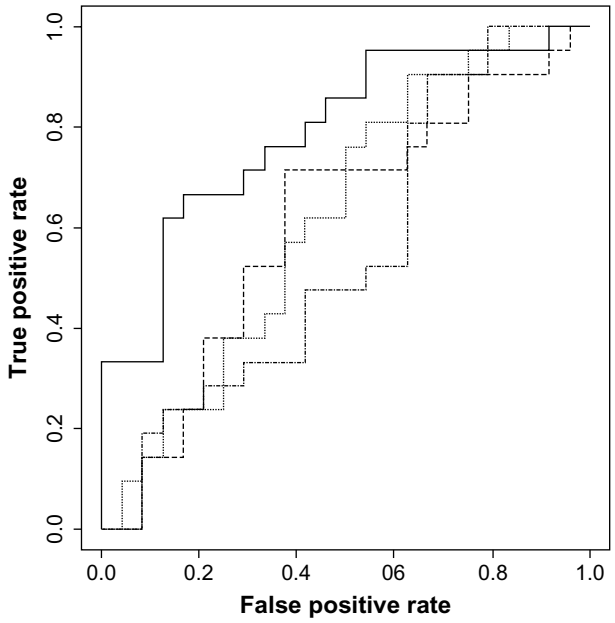

Figure 2. ROC curves for OV pathway Maturity Onset Diabetes of the Young.

Notes: ROC curves for single data type analyses SNP (dashed), gene expression (dotted), and methylation level (dot-dash), and for the three data type analysis (solid).

testing, this study used a more robust method of running 50 analyses with unique testing–training sets where 30% of samples were withheld for testing. This study had more even divisions between the two classes of OV and COAD stages (53% advanced-stage OV and 42% advanced-stage COAD) than that of Kim et al (92% advanced-stage OV). This may give a more unbiased prediction and more power to predict. Finally, this study classified patients using only the data related to the genes in a single pathway, whereas Kim et al classified patients using the entire dataset.[58]

Pathways that were significantly improved in stage prediction using integrative analysis were biologically relevant to their respective cancers. This demonstrates the success of the method in finding pathways that accurately classify stage in ovarian and colon cancers.

Difficulties presented by previous methods of pathway analysis, as described by Wang et al,[60] included genetic architecture, multiple testing, and replication of results. Other difficulties did not apply to the way the integrated analysis was conducted. A replicate dataset is not available; however, validation sets were run for each of the 50 iterations of the datasets. They determined accuracy and AUC of each pathway. Apart from a classification problem, another approach taken could be testing based, in which *P*-values will be provided instead. Moreover, a separate and independent dataset would be needed if these results were to be used in a clinical setting.

Wang et al described the difficulty presented by genetic architecture as one gene driving the entire pathway.[60] It is more likely that the opposite is true. Integrative analysis is likely to drown out the signal of one good gene with the noise of the other genes in the pathway, which are non-informative. Integrative analysis should be used in conjunction with a single-gene-based approach so that no information is overlooked. One gene that classifies patients very well is unlikely to do so significantly more in the integrative analysis than in single data-type analysis, which is the requirement for a significant pathway. This is so because, in the example of this study, a well-classified pathway in the integrated analysis driven by SNP association, gene expression, and/or methylation level would be compared to data types containing these driving factors, and the improvement should be small to none. Only if a combination of these driving factors increases the AUC or accuracy significantly more than a single data type would a gene drive the pathway, and then it would only happen if the other genes in the pathway did not drown out the signal. Integrative pathway-based analysis was an effective way to determine the mechanisms underlying advancing stage in serous cystadenocarcinoma and COAD. The graph-based semi-supervised learning algorithm, which determined these mechanisms also, significantly improved prediction of stage in these cancers compared to analysis of single–omics studies, including gene expression, methylation, and SNP studies. This algorithm can be extended to classify patient status and detect relevant biological mechanisms underlying any disease according to the chosen dichotomous clinical variable, if multiple sets of–omics data per patient are available. As more resources like TCGA[1] become available and expand their datasets, the utility of this algorithm will increase.

Conclusions

Integration of multiple genomic datasets resulted in a significant improvement over the single dataset analyses. Integration of multiple genomic datasets gave a maximum of 70% accuracy and 0.79 AUC. The pathways with these maxima are not in the tables, because the one dataset gene expression analysis for these pathways gives a similar level of accuracy or AUC. The top 10 OV and COAD pathways were all biologically relevant to their diseases, and some were known to be directly related to stage, as described above. Therefore, this algorithm is an effective method of classification and biological discovery.

To date, causes of complex genetic diseases, including cancers, have had small effect size and/or low frequency. Most studies have sought single point sources like a gene or a SNP with a single data type. They have failed to yield the expected result of an association that explains the cause of a disease in a large percentage of patients. The sources of a complex genetic disease can be found in multiple data types like SNPs and methylation level changes. These sources can spread their effects on other data types like methylation level changes to gene expression. Therefore, networking these sources using multiple types of–omics data over a pathway with a function relevant to the disease is an effective way to incorporate these

truths over multiple potential disease-associated markers into a single model with the power to determine which biological mechanisms, or pathways, have significant contribution to disease. The contribution discovered will relate to the dichotomous clinical variable used in the model. For example, if the variable is cancer stage, the mechanisms will describe the pathways differing between early and advanced stages. The network generated by the integrative pathway-based analysis for each significant pathway can become a hypothesis-generating tool in the discovery of the precise elements of the pathway contributing to the disease.

Future work with regard to this algorithm should include a decrease in computational time to make it feasible to work with more than three data types. A potential method may be to increase the sparsity of the distance matrix, if it does not impair the algorithm's ability to detect biologically relevant pathways. The algorithm could also be improved by adding a method to automatically calibrate the optimal *c* value and distance matrix. Testing of additional dichotomous variables, additional datasets, and replicate datasets will yield more information about this algorithm, when they become available. Furthermore, our methodology may be modified to evaluate the contribution from multiple pathways.[61]

Author Contributions

Conceived and designed the experiments: AD and HP. Analyzed the data: AD and HP. Wrote the first draft of the manuscript: AD and HP. Contributed to the writing of the manuscript: AD, AN, HP. Agree with manuscript results and conclusions: AD, AN, HP. Jointly developed the structure and arguments for the paper: AD, HP. Made critical revisions and approved final version: AD, AN, HP. All authors reviewed and approved of the final manuscript.

REFERENCES

1. The Cancer Genomics Atlas 2013; [cancergenome.nih.gov].
2. Bell D, Berchuck A, Birrer M, et al. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011;490:609–15.
3. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487:330–7.
4. Ivan C, Hu W, Bottsford-Miller J, et al. Epigenetic analysis of the Notch superfamily in high-grade serous ovarian cancer. *Gynecol Oncol*. 2013;128:506–11.
5. Cope L, Wu RC, Shih IM, Wang TL. High level of chromosomal aberration in ovarian cancer genome correlates with poor clinical outcome. *Gynecol Oncol*. 2013;128:500–05.
6. Wang W, Baladandayuthapani V, Morris JS, Broom BM, Manyam G, Do KA. iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*. 2013;29:149–59.
7. Zhang S, Liu CC, Li W, Shen H, Laird PW, Zhou XJ. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res*. 2012;40:9379–91.
8. Li W, Zhang S, Liu CC, Zhou XJ. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*. 2012;28:2458–66.
9. Dong H, Luo L, Hong S, et al. Integrated analysis of mutations, miRNA and mRNA expression in glioblastoma. *BMC Syst Biol*. 2010;4:163.
10. Vaske CJ, Benz SC, Sanborn JZ, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*. 2010;26:i237–45.
11. Jia P, Liu Y, Zhao Z. Integrative pathway analysis of genome-wide association studies and gene expression data in prostate cancer. *BMC Syst Biol*. 2012;6:s13.
12. Pang H, Hauser MA, Minvielle S. Pathway-based identification of SNPs predictive of survival. *Eur J Hum Genet*. 2011;19:704–09.
13. Chasman DI. On the utility of gene set methods in genomewide association studies of quantitative traits. *Genet Epidemiol*. 2008;32:658–68.
14. Peng G, Luo L, Siu H, et al. Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur J Hum Genet*. 2010;18:111–7.
15. Ritchie MD. Using prior knowledge and genome-wide association to identify pathways involved in multiple sclerosis. *Genome Med*. 2009;1:65.
16. Pang H, Datta D, Zhao H. Pathway analysis using random forests with bivariate node-split for survival outcomes. *Bioinformatics*. 2010;26:250–8.
17. Tsuda K, Shin H, Scholkopf B. Fast protein classification with multiple networks. *Bioinformatics*. 2005;21:ii59–ii65.
18. Kanehisa M, Goto S, Hattori M, et al. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*. 2006;34:D354–7.
19. Biocarta 2013; [www.biocarta.com/genes/allpathways.asp]
20. Chapelle O. Training a support vector machine in the primal. *Neural Comput*. 2007;19:1155–78.
21. Deng M, Chen T, Sun F. An integrated probabilistic model for functional prediction of proteins. *J Comput Biol*. 2004;11:463–75.
22. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. *Bioinformatics*. Oct 15, 2005;21(20):3940–1.
23. Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*. 2008;92:265–72.
24. Kershaw MH, Darcy PK, Trapani JA, MacGregor D, Smyth MJ. Tumor-specific IgE-mediated inhibition of human colorectal carcinoma xenograft growth. *Oncol Res*. 1998;10:133–42.
25. Chatzinikolaou G, Nikitovic D, Berdiaki A, et al. Heparin regulates colon cancer cell growth through p38 mitogen-activated protein kinase signalling. *Cell Prolif*. 2010;43:9–18.
26. Sun Y, Qiao L, Xia HH, et al. Regulation of XAF1 expression in human colon cancer cell by interferon beta: activation by the transcription regulator STAT1. *Cancer Lett*. 2008;260:62–71.
27. Chiu ST, Chang KJ, Ting CH, Shen HC, Li H, Hsieh FJ. Over-expression of EphB3 enhances cell–cell contacts and suppresses tumor growth in HT-29 human colon cancer cells. *Carcinogenesis*. 2009;30:1475–86.
28. Gulhati P, Bowen KA, Liu J, et al. mTORC1 and mTORC2 regulate EMT, motility, and metastasis of colorectal cancer via RhoA and Rac1 signaling pathways. *Cancer Res*. 2011;71:3246–56.
29. Hsu HH, Liu CJ, Shen CY, et al. p38α MAPK mediates 17β-estradiol inhibition of MMP-2 and -9 expression and cell migration in human lovo colon cancer cells. *J Cell Physiol*. 2012;227:3648–60.
30. Lim JH, Woo SM, Min KJ, et al. Rottlerin induces apoptosis of HT29 colon carcinoma cells through NAG-1 upregulation via an ERK and p38 MAPK-dependent and PKC δ-independent mechanism. *Chem Biol Interact*. 2012; 197:1–7.
31. Wolf MJ, Hoos A, Bauer J, et al. Endothelial CCR2 signaling induced by colon carcinoma cells enables extravasation via the JAK2-Stat5 and p38MAPK pathway. *Cancer Cell*. 2012;22:91–105.
32. Yang X, Liu F, Xu Z, et al. Growth hormone receptor expression in human colorectal cancer. *Dig Dis Sci*. 2004;49:1493–98.
33. Hohla F, Buchholz S, Schally AV, et al. GHRH antagonist causes DNA damage leading to p21 mediated cell cycle arrest and apoptosis in human colon cancer cells. *Cell Cycle*. 2009;8:3149–56.
34. Slattery ML, Herrick JS, Bondurant KL, Wolff RK. Toll-like receptor genes and their association with colon and rectal cancer development and prognosis. *Int J Cancer*. 2012;130:2974–80.
35. Miller SR, Tartter PI, Papatestas AE, Slater G, Aufses AH Jr. Serum cholesterol and human colon cancer. *J Natl Cancer Inst*. 1981;67:297–300.
36. Taylor ML, Wells BJ, Smolak MJ. Statins and cancer: a meta-analysis of case-control studies. *Eur J Cancer Prev*. 2008;17:259–68.
37. Siddiqui AA, Nazario H, Mahgoub A, Pandove S, Cipher D, Spechler SJ. The long-term use of statins is associated with a decreased incidence of adenomatous colon polyps. *Digestion*. 2009;79:17–22.
38. Senda T, Matsumine A, Yanai H, Akiyama T. Localization of MCC (mutated in colorectal cancer) in various tissues of mice and its involvement in cell differentiation. *J Histochem Cytochem*. 1999;47:1149–58.
39. Sarela AI, Scott N, Ramsdale J, Markham AF, Guillou PJ. Immunohistochemical detection of the anti-apoptosis protein, survivin, predicts survival after curative resection of stage II colorectal carcinomas. *Ann Surg Oncol*. 2001;8:305–10.
40. Enarsson K, Lundin BS, Johnsson E, Brezicka T, Quiding-Järbrink M. CD4+ CD25high regulatory T cells reduce T cell transendothelial migration in cancer patients. *Eur J Immunol*. 2007;37:282–91.
41. Yoshioka T, Nishikawa Y, Ito R, et al. Significance of integrin αvβ5 and erbB3 in enhanced cell migration and liver metastasis of colon carcinomas stimulated by hepatocyte-derived heregulin. *Cancer Sci*. 2010;101:2011–18.
42. Guebel DV, Schmitz U, Wolkenhauer O, Vera J. Analysis of cell adhesion during early stages of colon cancer based on an extended multi-valued logic approach. *Mol Biosyst*. 2012;8:1230–42.
43. Shimomoto T, Ohmori H, Luo Y, et al. Diabetes-associated angiotensin activation enhances liver metastasis of colon cancer. *Clin Exp Metastasis*. 2012;29:915–25.

44. Luo Y, Ohmori H, Shimomoto T, et al. Anti-angiotensin and hypoglycemic treatments suppress liver metastasis of colon cancer cells. *Pathobiology*. 2011;78:285–90.
45. Flick MB, O'Malley D, Rutherford T, et al. Apoptosis-based evaluation of chemosensitivity in ovarian cancer patients. *J Soc Gynecol Investig*. 2004;11:252–9.
46. van dePoll-Franse LV, Houterman S, Janssen-Heijnen ML, Dercksen MW, Coebergh JW, Haak HR. Less aggressive treatment and worse overall survival in cancer patients with diabetes: a large population based analysis. *Int J Cancer*. 2007;120:1986–92.
47. Arcidiacono B, Iiritano S, Nocera A, et al. Insulin resistance and cancer risk: an overview of the pathogenetic mechanisms. *Exp Diabetes Res*. 2012:789174.
48. Wang Y, Wu R, Cho KR, et al. Differential protein mapping of ovarian serous adenocarcinomas: identification of potential markers for distinct tumor stage. *J Proteome Res*. 2009;8:1452–63.
49. Yuan Y, Liao YM, Hsueh CT, Mirshahidi HR. Novel targeted therapeutics: inhibitors of MDM2, ALK and PARP. *J Hematol Oncol*. 2011;4:16.
50. Palmer RH, Vernersson E, Grabbe C, Hallberg B. Anaplastic lymphoma kinase: signalling in development and disease. *Biochem J*. 2009;420:345–61.
51. Usha L, Sill MW, Darcy KM, et al. A Gynecologic Oncology Group phase II trial of the protein kinase C-beta inhibitor, enzastaurin and evaluation of markers with potential predictive and prognostic value in persistent or recurrent epithelial ovarian and primary peritoneal malignancies. *Gynecol Oncol*. 2011;121:455–61.
52. Ino K, Shibata K, Yamamoto E, et al. Role of the renin-angiotensin system in gynecologic cancers. *Curr Cancer Drug Targets*. 2011;11:405–11.
53. Fialová A, Partlová S, Sojka L, et al. Dynamics of T-cell infiltration during the course of ovarian cancer: the gradual shift from a Th17 effector cell response to a predominant infiltration by regulatory T-cells. *Int J Cancer*. 2013;132:1070–9.
54. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*. 2011;39:D53–7.
55. Lee S, Kim J, Lee S. A comparative study on gene-set analysis methods for assessing differential expression associated with the survival phenotype. *BMC Bioinformatics*. 2011;12:377.
56. Toy EP, Chambers JT, Kacinski BM, Flick MB, Chambers SK. The activated macrophage colony-stimulating factor (CSF-1) receptor as a predictor of poor outcome in advanced epithelial ovarian carcinoma. *Gynecol Oncol*. 2001;80:194–200.
57. Wang Y, Li L, Guo X, et al. Interleukin-6 signaling regulates anchorage-independent growth, proliferation, adhesion and invasion in human ovarian cancer cells. *Cytokine*. 2012;59:228–36.
58. Kim S, Kon M, DeLisi C. Pathway-based classification of cancer subtypes. *Biol Direct*. 2012;7:21.
59. Bagnato A, Tecce R, Moretti C, Di CastroV, Spergel D, Catt KJ. Autocrine actions of endothelin-1 as a growth factor in human ovarian carcinoma cells. *Clin Cancer Res*. 1995;1:1059–66.
60. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet*. 2010;11:843–54.
61. Pang H, Kim I, Zhao H. Random effects model for multiple pathway analysis with applications to type II diabetes microarray data. *Stat Biosci*. 2014; [Epub ahead of print].