2017

A Novel Approach for Classifying Gene Expression Data using Topic Modeling

Soon Jye Kho

Himi Yalamanchili
Wright State University - Main Campus, yalamanchili.13@wright.edu

Michael L. Raymer
Wright State University - Main Campus, michael.raymer@wright.edu

Amit Sheth
Wright State University - Main Campus, amit@sc.edu

A Novel Approach for Classifying Gene Expression Data using Topic Modeling

Soon Jye Kho[*]
Knoesis Center, Wright State University
3640 Colonel Glenn Hwy
Dayton, Ohio 45431
soonjye@knoesis.org

Hima Bindu Yalamanchili[*]
Knoesis Center, Wright State University
3640 Colonel Glenn Hwy
Dayton, Ohio 45431
himay.87@gmail.com

Michael L. Raymer
Knoesis Center, Wright State University
3640 Colonel Glenn Hwy
Dayton, Ohio 45431
michael.raymer@wright.edu

Amit P. Sheth
Knoesis Center, Wright State University
3640 Colonel Glenn Hwy
Dayton, Ohio 45431
amit@knoesis.org

ABSTRACT

Understanding the role of differential gene expression in cancer etiology and cellular process is a complex problem that continues to pose a challenge due to sheer number of genes and inter-related biological processes involved. In this paper, we employ an unsupervised topic model, Latent Dirichlet Allocation (LDA) to mitigate overfitting of high-dimensionality gene expression data and to facilitate understanding of the associated pathways. LDA has been recently applied for clustering and exploring genomic data but not for classification and prediction. Here, we proposed to use LDA in clustering as well as in classification of cancer and healthy tissues using lung cancer and breast cancer messenger RNA (mRNA) sequencing data. We describe our study in three phases: clustering, classification, and gene interpretation. First, LDA is used as a clustering algorithm to group the data in an unsupervised manner. Next we developed a novel LDA-based classification approach to classify unknown samples based on similarity of co-expression patterns. Evaluation to assess the effectiveness of this approach shows that LDA can achieve high accuracy compared to alternative approaches. Lastly, we present a functional analysis of the genes identified using a novel topic profile matrix formulation. This analysis identified several genes and pathways that could potentially be involved in differentiating tumor samples from normal. Overall, our results project LDA as a promising approach for classification of tissue types based on gene expression data in cancer studies.

KEYWORDS

Topic modeling, Latent Dirichlet Allocation, Clustering, Classification, Machine learning, Cancer, Gene expression

1 INTRODUCTION

Traditional diagnosis for cancer is based on clinical and morphological data, but these methods have been reported to have limitations in their diagnostic ability. [1, 12]. To overcome the limitations, cancer detection based on genomic data has been proposed [17, 18]. In recent years, with the wide employment of microarray and next-generation sequencing methods, increase in data volume poses both promise and challenges to researchers in identifying patterns and analyzing the data [15]. Although there has been a lot of research in the identification of differentially expressed genes associated with various types of cancer tissues, typically small sample size and high dimensionality of expression data continues to pose challenges to researchers. Understanding and interpretation of results remains a key challenge in analyzing gene expression data.

Topic modeling, a machine learning approach has shown promise in the fields of text mining and image retrieval, where it has been successfully implemented to extract information from high dimensional data [6, 14, 24]. Latent Dirichlet Allocation (LDA) is one of the most popular topic modeling approaches in text mining among others like Latent Semantic Indexing (LSI) and Probabilistic Latent Semantic Analysis (PLSA) [9, 11]. Since its emergence, researchers have implemented this approach in biomedical text mining as well [4, 25].

Given the successful implementation of topic models in discovering the useful structure of the documents, researchers are beginning to implement topic modeling approaches to analyze data other than document collections [8, 27]. Recently, there have been efforts to use topic modeling techniques in the field of bioinformatics to perform unsupervised analysis and obtain insights into high-dimensional -omics data [13, 16, 22, 23]. To gain better understanding into cancer classification and gene identification from such data, in this paper, we focus on the class prediction of breast cancer and lung cancer based on gene expression data using topic modeling.

We approached this study in three phases (Figure 1). In the first phase, we employed LDA as a clustering application. LDA has been successfully applied as a clustering algorithm for gene expression data in the past [28]. Here, we confirmed that LDA-derived document clusters based on our rank-based data transformation technique show clear separation between cancer and healthy tissue

[*] These authors contributed equally to this work

ACM-BCB '17, August 20-23, 2017, Boston, MA, USA
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-4722-8/17/08…$15.00
https://doi.org/10.1145/3107411.3107483

samples. In the second phase, we investigated the application of LDA as a classification algorithm. This involved optimizing several algorithmic parameters including the number of topics and algorithm passes with respect to classification of gene expression data. We used 10-fold cross validation to estimate classification effectiveness. We compared our results with alternative machine learning techniques including Support Vector Machine (SVM), Naive Bayes and Random Forest classification, as well as Principal Component Analysis (PCA) and Hierarchical Clustering for unsupervised learning. In the third phase, we additionally investigated the effectiveness of LDA in identifying genes differentially regulated between cancer and normal tissues. We explored the degree to which specific topics identified by the LDA algorithm correspond to gene regulatory pathways which vary between tumor and normal tissues.

The remaining of this paper is organized as follows. In Section 2, we review the related work of implementing LDA in gene expression data. In Section 3, we describe data collection and preprocessing while in Section 4 we describe our proposed method, evaluation techniques, results from two data sets and interpretation of identified differentially expressed genes.

2 RELATED WORK

One of the challenges in using LDA for genetic studies is the nature of the data. The input of topic modeling is typically a simplified bag-of-words representation of a corpus. However, gene expression data is numerical in nature, and thus needs to be transformed into text to provide appropriate input for the LDA algorithm. The most commonly used transformation method is scaling the matrix and interpreting the discrete values as gene/word occurrences. The higher the expression value, the higher the frequency of the gene in the bag of words [3, 19]. One limitation using this approach is genes that have zero expression value will not be present in the training corpus. Consequently, this transformation approach is suitable for clustering and feature reduction, but not for classification because the test set might have genes that are unseen by the trained model.

More recently, Zhao et al. [28] used a different transformation method for generating a corpus. For each gene, expression values were normalized to 0 (lower than median) or 1 (higher than median). The samples were then transformed into a bag-of-words, containing only genes with normalized value 1. This transformation method does not capture the deviation between samples. That is, two differentially expressed genes are treated the same regardless of the degree of deviation. Besides, this method is not feasible for imbalanced dataset as the cutoff point between two different classes could not be represented by median value. Rogers et al. [22] used Latent Process Decomposition (LPD), a derivative of LDA, to capture the continuous nature of gene expression data. No significant difference in terms of accuracy has been reported using LPD compared to classical approaches like pLSA [3].

To alleviate these limitations and broaden the range of applications using LDA, we propose a novel transformation strategy to generate bag of words. We adapted the gene ranking method from Wang et al. [26], which is described in detail in Section 3.2. The rationale behind this ranking is that relative ordering of gene expression is generally stable in a particular type of normal human tissue but is widely disturbed in a diseased tissue. Ranking method is able to capture the degree of deviation of gene expression across different samples and still enable gene interpretability of topics using topic profiles.

3 METHODS

3.1 Data Collection and Filtering

We obtained the mRNA-Seq data of 229 breast invasive carcinoma samples and 98 lung squamous cell carcinoma samples from The Cancer Genome Atlas (TCGA) data portal [1]. We artificially balanced the data using only the tumor samples with matched normals. Out of 229 breast cancer samples, 117 represent primary solid tumor and 112 represent solid normal tissue. Of the 98 lung cancer samples, half of the data (49) represent primary solid tumor and the other half (49) represent solid tissue normal. Both datasets have 60483 variables with an abundance of zero values – roughly 45% of the datasets. Genes that have zero expression value in more than 10% of samples were removed. This filtering step might remove some differentially expressed genes but such stringent threshold was used to prevent any bias (multiple bins having genes with exclusively zero expression value) that might arise in the preprocessing step. After filtering, a total of 23424 genes remained in the breast cancer data and 23996 genes remained in the lung cancer data.

3.2 Preprocessing

Here, we used a ranking approach to transform gene expression values into a bag-of-words. Within each sample, genes were sorted in ascending order based on their expression values. Expression values form a distribution that is divided into 20 bins. The performance of the algorithm is robust against the bin size (5, 10, 20, 30, 40 bins) and this is expected since the number of bins represents the relative ranking of expression level instead of absolute ranking (data not shown).

We chose 20 bins as a representative of the performed experiments. Actual expression values were then replaced with a quantized value (expression rank) based on its position. Value 1 indicates the group with lowest expression level and value 20 indicates the group with highest expression level. The combination of each gene with its corresponding expression rank (separated by a hyphen), generated a bag of words for each sample. For example, if gene BRCA1 was found in quantized bin 3 for a particular sample, the bag-of-words associated with that sample would include the word BRAC1-3 to represent the ranked and quantized expression of this gene. It is important to note that as a result of quantization, comparison between gene and word may not be absolute since one gene might represent many words across the samples. To extract gene-specific information from the LDA-derived model, it is thus necessary to develop a technique to reverse this one-to-many mapping and trace back genes from the corresponding words without deviating much from the expression rank (Section 4.2). After the transformation, the breast cancer data corpus had 229 documents and 23424 words within each document. The lung cancer data corpus had 98 documents and each document contained 23996 words.

[1] https://portal.gdc.cancer.gov/

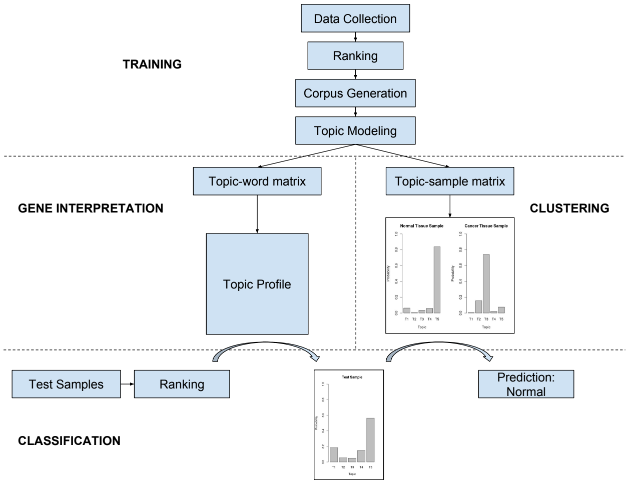


Figure 1: Flowchart of the proposed approach. Highlighted boxes represent the sequence of steps involved. Dashed lines divide the analysis into four groups: Training, Clustering, Classification and gene interpretation. Barplots represent the topic distribution of samples.

4 RESULTS AND ANALYSIS

For each dataset, LDA was implemented using an open-source Python library, Gensim [21]. The number of passes and number of topics were set to 10 and 3 respectively. These parameters were determined experimentally to provide the optimum performance (data not shown). The LDA algorithm produces two matrices: the topic-sample matrix, which expresses the computed probability of topics in a particular sample (document), and the topic-word matrix, containing the probability distribution over all available words for each topic. The LDA-derived matrices were utilized to perform analysis in 3 phases: clustering, classification and gene interpretation.

4.1 Clustering

Average topic distributions of samples from each class (normal and cancer) are depicted in Figure 2. This topic-sample matrix can be used for feature extraction/projection in a manner analogous to principal component analysis (PCA). Each column vector of the topic-sample matrix represents a single sample. There is one value in this vector for each topic, expressing the degree of association between the topic and the sample. Complete-linkage and euclidean-distance-based Hierarchical Clustering was applied on the topic-sample matrix to cluster the samples with similar probability distributions. The heatmap in Figure 2 shows that for both breast cancer and lung cancer data, samples were clearly grouped together into two classes, cancer and normal. Colors from red to green in the heatmap represent topic probabilities of the samples, ranging from 0 to 1.

The clustering result from LDA was compared with other conventional clustering and projection methods including hierarchical clustering and PCA. Hierarchical clustering was directly applied to the raw expression values for breast cancer (23424 genes) and lung cancer (23996 genes). PCA was first used to transform the original data into components that retain 90% of the variance. Hierarchical clustering was then applied on the transformed data. For both datasets, cluster purity and number of misclassified samples were calculated using Equation 1 to evaluate the clustering performance.

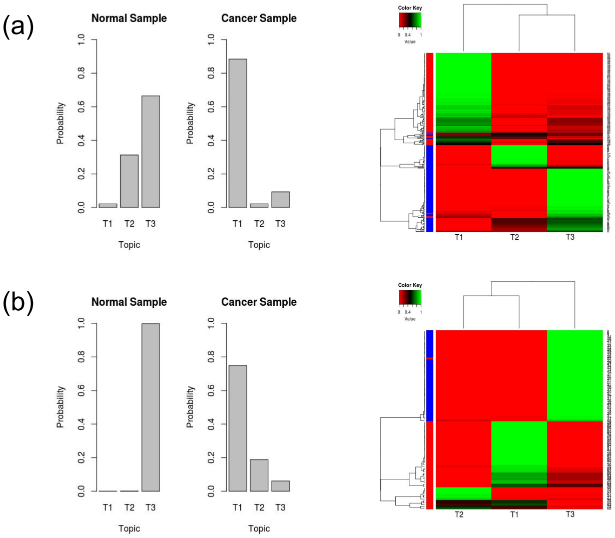


Figure 2: Topic probability distribution and Hierarchical clustering of samples using derived topic probability for (a) breast carcinoma and (b) lung squamous cell carcinoma. Bar plots represent average topic-sample probability for the 3 topics. Heat map shows the clustering of samples into two classes.

$$Purity = 1/N \sum_{i=0}^{k} max_j |c_i \cap t_j| \tag{1}$$

Where N = number of data points, k = number of clusters, c_i= a cluster in K, t_j = classification that has the maximum count for cluster c_i.

Table 1 shows the clustering results of the three methods where the number of clusters (K) is fixed at 2, 3, and 4 clusters. Overall, LDA clusters align more closely to known cancer/healthy tissue labels than those obtained by PCA and Hierarchical Clustering. One advantage of LDA-based clustering for gene expression data may be its mutually non-exclusive assumption. Most non-fuzzy clustering approaches have a mutually exclusive/independence assumption that a sample/gene is restricted to only one cluster. This assumption might not be logical for gene expression data largely when a sample/gene share characteristics with more than one cluster (that is one gene can be involved in different pathways). Thus using LDA for clustering of gene data would reflect the complex interplay between genes and pathways and improve quality of the results.

Table 1: Clustering results of LDA-derived topic probabilities, PCA reduced features, and Hierarchical Clustering.

Method	K	Breast cancer		Lung cancer	
		No. of Misclassified Samples	Cluster Purity	No. of Misclassified Samples	Cluster Purity
Latent Dirichlet Allocation (LDA)	2	11	0.952	1	0.990
	3	11	0.952	1	0.990
	4	11	0.952	1	0.990
PCA	2	112	0.512	43	0.561
	3	112	0.512	43	0.561
	4	109	0.524	43	0.561
Hierarchical Clustering	2	112	0.511	48	0.510
	3	112	0.511	19	0.806
	4	112	0.511	15	0.847

4.2 Classification

In this section, we propose a novel approach to classification using LDA-based topic modeling, and explore the effectiveness of the proposed method in distinguishing gene expression patterns in breast and lung cancer tissues from those in healthy tissue. In contrast to the work by [3] where LDA is solely used as a feature extraction method and requires running LDA on both training and testing data, the approach proposed here applies the LDA algorithm only to the training data.

First, training of the LDA algorithm proceeds in the same manner as described previously for LDA-based clustering. For testing, LDA-derived topic-sample and topic-word matrices from the trained model are employed. To revert/collapse the genes from their corresponding words, expression rank and probabilities of all words representing a gene are taken and normalized over the entire probability distribution (Equation 2):

$$exp_g = \frac{\sum_{j=1}^{j} rank_g^j \times prob_g^j}{\sum_{j=1}^{j} prob_g^j} \tag{2}$$

where exp_g represents the collapsed expression rank of a gene g, j is the total number of words that representing gene g. $rank_g \in (rank_g^1, rank_g^2, ..., rank_g^j)$ is a vector of rank extracted from words and $prob_g \in (prob_g^1, prob_g^2, ..., prob_g^j)$ is the vector of corresponding probabilities for the words.

Combining the topic-word probabilities in this manner results in a new topic-gene matrix, in which each topic is associated with a specific rank value for each gene. The resulting topic profile facilitates subsequent classification.

To classify a test sample based on the trained LDA, we need to determine topic probability distribution of the test sample. First, the same pre-processing step described for clustering is applied to the test data, resulting in a bag-of-words based on ranked gene expression for each test sample. Then similarity (MSE) between testing data and each topic profile was calculated using Equation 3. The lower the MSE, the more similar the topic.

$$MSE = \sum_{g=1}^{g} (exp_g - bin_g)^2 \tag{3}$$

Where exp_g represents the collapsed expression rank of gene g in a topic profile and bin_g represents expression rank of gene g in test sample. Gaussian normalization was performed to reduce the range of the MSE, thus avoiding overflow errors in computation, without affecting the variation between topics and the softmax function (Equation 4) is applied to ensure that the probability distribution properly sums to 1.0 [5]. The probability distribution determined from similarity is used for prediction.

$$SoftmaxOutput = \frac{e^{sMSE}}{\sum_{k=1}^{k} e^{sMSE}} \tag{4}$$

The performance of our classification approach was compared with popular supervised methods including SVM, Naive Bayes and Random Forest classifiers using 10-fold cross validation. Traditional metrics for comparison like accuracy, precision, recall and F-measure were applied. As shown in Table 2, LDA achieves a competitive performance comparable to other algorithms.

4.3 Gene Interpretation

We utilized topic profiles from the two data sets to perform pathway analysis and disease annotation of differentially expressed genes within each topic. A list of differentially expressed genes (DEGs) is extracted by computing the difference in expression rank of each topic with a baseline. For the experiments shown here, topic 3 was chosen as the baseline in both breast cancer and lung cancer datasets, since it shows the highest probability in normal samples (Figure 2). DEGs were then identified from all other topics (1 and 2 in this case). A threshold rank-difference of 5 was used to extract the significant genes and also limit the number of genes for further analysis. Thus, the extracted genes have at least 5 ranking changes between normal and tumor that could potentially represent significant dysregulation in our analysis.

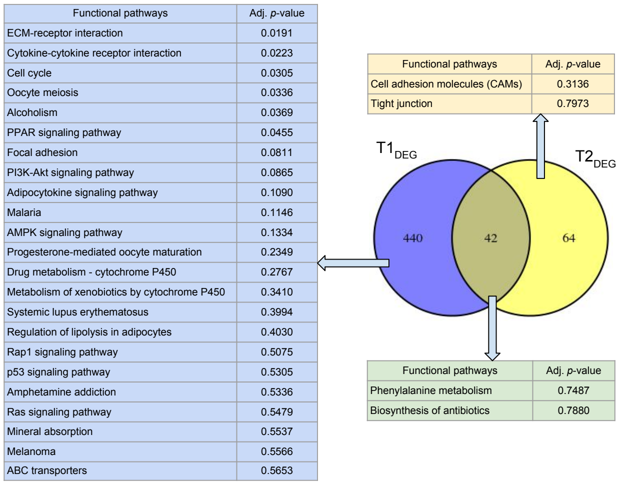
We examined the pathway enrichment of the differentially expressed genes using DAVID Bioinformatics Resources 6.8 [2] [10] to explore the KEGG pathway database. Figures 3 and 4 show the number of genes and statistically enriched pathways in the topics for breast cancer and lung cancer, respectively.

The relevance of the identified differentially expressed genes was validated using the DAVID bioinformatics suite [10] and the

[2] https://david.ncifcrf.gov/home.jsp

Table 2: Classification results of our proposed approach and three supervised algorithms: SVM, Naive Bayes, and Random Forest.

	Breast cancer					Lung cancer			
	F-measure	Accuracy	Specificity	Sensitivity		F-measure	Accuracy	Specificity	Sensitivity
LDA	0.991	0.991	1.000	0.983	LDA	0.990	0.990	1.000	0.980
SVM	0.970	0.978	0.973	0.966	SVM	0.969	0.969	0.980	0.959
Naive Bayes	0.974	0.974	1.000	0.945	Naive Bayes	0.980	0.980	0.980	0.980
Random Forest	0.987	0.987	0.991	0.983	Random Forest	0.990	0.990	1.000	0.980



Functional pathways	Adj. *p*-value
ECM-receptor interaction	0.0191
Cytokine-cytokine receptor interaction	0.0223
Cell cycle	0.0305
Oocyte meiosis	0.0336
Alcoholism	0.0369
PPAR signaling pathway	0.0455
Focal adhesion	0.0811
PI3K-Akt signaling pathway	0.0865
Adipocytokine signaling pathway	0.1090
Malaria	0.1146
AMPK signaling pathway	0.1334
Progesterone-mediated oocyte maturation	0.2349
Drug metabolism - cytochrome P450	0.2767
Metabolism of xenobiotics by cytochrome P450	0.3410
Systemic lupus erythematosus	0.3994
Regulation of lipolysis in adipocytes	0.4030
Rap1 signaling pathway	0.5075
p53 signaling pathway	0.5305
Amphetamine addiction	0.5336
Ras signaling pathway	0.5479
Mineral absorption	0.5537
Melanoma	0.5566
ABC transporters	0.5653

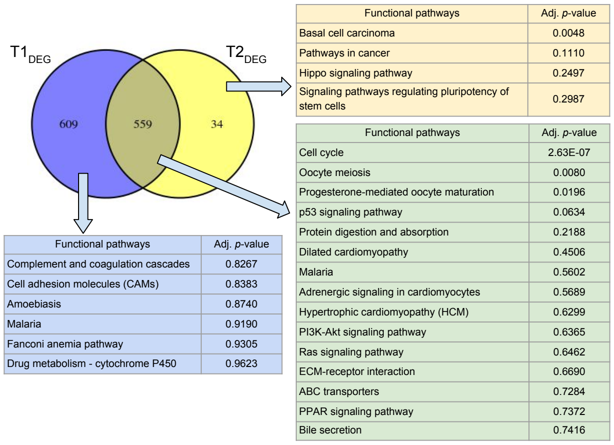
Functional pathways	Adj. *p*-value
Cell adhesion molecules (CAMs)	0.3136
Tight junction	0.7973

Functional pathways	Adj. *p*-value
Phenylalanine metabolism	0.7487
Biosynthesis of antibiotics	0.7880

Figure 3: Functional annotation of dysregulated genes in breast cancer. Venn diagram of differentially expressed genes in Topic 1, in Topic 2 for breast cancer. Tables represent the affected pathways for each subset of genes, and their adjusted *p*-value. The lower the adjusted *p*-value, the higher the significance of pathway.

NIH genetic association database (GAD) [2]. The DEGs were extracted by comparing the expression profile for normal and tumor samples. For each gene, its expression rank (exp_g) in topic profile matrix (~24000×3) was multiplied with respective topic probability distribution of normal group (3×1) and tumor group (3×1). This would generate two expression profiles: one for normal samples (~24000×1) and one for tumor samples (~24000×1). The difference in expression rank between these two profiles was computed. Top 250 DEGs was extracted both for breast cancer and lung cancer and annotated using DAVID Bioinformatics Resources 6.8 in terms of GAD disease. Table 3 shows the top 5 highly related diseases which are ranked ascendingly according to their adjusted *p*-value.

As expected, for the gene list extracted from breast cancer data corpus, 'breast cancer' is the disease with highest enrichment (lowest adjusted *p*-value) and for lung cancer data corpus, 'smoking cessation' and 'chronic obstructive pulmonary disease' are highly enriched.



Functional pathways	Adj. *p*-value
Basal cell carcinoma	0.0048
Pathways in cancer	0.1110
Hippo signaling pathway	0.2497
Signaling pathways regulating pluripotency of stem cells	0.2987

Functional pathways	Adj. *p*-value
Cell cycle	2.63E-07
Oocyte meiosis	0.0080
Progesterone-mediated oocyte maturation	0.0196
p53 signaling pathway	0.0634
Protein digestion and absorption	0.2188
Dilated cardiomyopathy	0.4506
Malaria	0.5602
Adrenergic signaling in cardiomyocytes	0.5689
Hypertrophic cardiomyopathy (HCM)	0.6299
PI3K-Akt signaling pathway	0.6365
Ras signaling pathway	0.6462
ECM-receptor interaction	0.6690
ABC transporters	0.7284
PPAR signaling pathway	0.7372
Bile secretion	0.7416

Functional pathways	Adj. *p*-value
Complement and coagulation cascades	0.8267
Cell adhesion molecules (CAMs)	0.8383
Amoebiasis	0.8740
Malaria	0.9190
Fanconi anemia pathway	0.9305
Drug metabolism - cytochrome P450	0.9623

Figure 4: Functional annotation of dysregulated genes in lung cancer. Venn diagram of differentially expressed genes in Topic 1, in Topic 2 for Lung Cancer. Tables represent the affected pathways for each subset of genes, and their adjusted *p*-value. The lower the adjusted *p*-value, the higher the significance of the pathway.

Table 3: Top 5 annotated diseases, arranged in ascending order of adjusted *p*-value, from breast and lung cancer DEGs.

Breast Cancer	Adj. *p*-value	Lung Cancer	Adj. *p*-value
Breast cancer	2.7E-2	Cleft lip, cleft palate	3.7E-1
Alzheimer's disease	1.7E-1	Smoking cessation	4.0E-1
Chorioamnionitis	2.4E-1	Chronic obstructive pulmonary disease	9.2E-1
Colorectal cancer	3.6E-1	Breast cancer	9.5E-1
Lung cancer	4.3E-1	Leukemia, lymphocytic, chronic, B-cell	9.5E-1

5 DISCUSSION

Many approaches to diagnostic classification based on mRNA expression focus primarily on differential expression. The LDA-based approach described here differs in that the focus is primarily on co-expression. Just as textual LDA attempts to group co-occurring words into topics in order to explain the topic composition of a document, gene expression LDA can be used to identify co-regulated groups of genes that together explain the overall patterns of gene expression in healthy and disease states. In an unsupervised mode, this technique has shown a somewhat surprising ability to produce gene-collections (topics) that differ significantly between cancer and healthy tissues. Augmenting this technique with class labels results in an even more capable diagnostic classifier.

Here, our approach is only tested on gene expression data from next generation sequencing. Nevertheless, we expect it to work equally with similar data from different platforms or for different tasks such as disease subtype classification [7], survival analysis and treatment prognosis prediction [20]. We have shown that by using only gene expression data, the model was able to cluster the data into topics that are biologically coherent and meaningful. An interesting next step would be to include additional lifestyle and clinical metadata in the bag-of-words representing each sample. For example, by tagging clinical study data with labels such as smoker/non-smoker, male/female, adult/child, etc., associations between subject groups corresponding to these labels, disease state, and gene expression might further be identified. Understanding associated pathways and genomic etiology in these tasks is important in stratifying patients according to their risk and provide better diagnosis.

6 CONCLUSION

Overall our approach provides a novel direction for applying the LDA algorithm to identify and group differentially expressed genes between healthy and cancer tissues of various types. A novel technique for transformation of gene expression levels to words is presented and shown to be effective. Comparative evaluation of this approach with state-of-the-art pattern classification methods confirms the effectiveness of the proposed methodology. Differential gene expression patterns associated with lung and breast cancer were identified and validated as relevant using pathway analysis and the NIH's genetic association database. Next steps include testing the effectiveness of this LDA-based classification approach on genetic and epigenetic variation patterns and also testing the effectiveness of LDA as a supervised algorithm for the more complex problem of distinguishing among cancer subtypes.

ACKNOWLEDGMENTS

This work is partially funded by the Biomedical Sciences Ph.D. program at Wright State University.

REFERENCES

[1] F Azuaje. 1999. Interpretation of genome expression patterns: computational challenges and opportunities. *IEEE engineering in medicine and biology magazine: the quarterly magazine of the Engineering in Medicine & Biology Society* 19, 6 (1999), 119–119.

[2] Kevin G Becker, Kathleen C Barnes, Tiffani J Bright, and S Alex Wang. 2004. The genetic association database. *Nature genetics* 36, 5 (2004), 431–432.

[3] Manuele Bicego, Pietro Lovato, Barbara Oliboni, and Alessandro Perina. 2010. Expression microarray classification using topic models. In *Proceedings of the 2010 ACM Symposium on Applied Computing*. ACM, 1516–1520.

[4] Halil Bisgin, Zhichao Liu, Hong Fang, Xiaowei Xu, and Weida Tong. 2011. Mining FDA drug labels using an unsupervised learning technique-topic modeling. *BMC bioinformatics* 12, 10 (2011), S11.

[5] Christopher M Bishop. 2006. Pattern recognition. *Machine Learning* 128 (2006), 1–58.

[6] David M Blei and John D Lafferty. 2009. Topic models. *Text mining: classification, clustering, and applications* 10, 71 (2009), 34.

[7] Lars Bullinger, Konstanze Döhner, Eric Bair, Stefan Fröhling, Richard F Schlenk, Robert Tibshirani, Hartmut Döhner, and Jonathan R Pollack. 2004. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *New England Journal of Medicine* 350, 16 (2004), 1605–1616.

[8] Xin Chen, Xiaohua Hu, Tze Yee Lim, Xiajiong Shen, EK Park, and Gail L Rosen. 2012. Exploiting the functional and taxonomic structure of genomic data by probabilistic topic modeling. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 9, 4 (2012), 980–991.

[9] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science* 41, 6 (1990), 391.

[10] Glynn Dennis, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Clifford Lane, and Richard A Lempicki. 2003. DAVID: database for annotation, visualization, and integrated discovery. *Genome biology* 4, 9 (2003), R60.

[11] Thomas Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning* 42, 1 (2001), 177–196.

[12] Halliday A Idikio. 2011. Human cancer classification: a systems biology-based model integrating morphology, cancer stem cells, proteomics, and genomics. *Journal of Cancer* 2, 1 (2011), 107–115.

[13] Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. 2016. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* 5, 1 (2016), 1608.

[14] Wenhan Luo, Björn Stenger, Xiaowei Zhao, and Tae-Kyun Kim. 2015. Automatic Topic Discovery for Multi-Object Tracking.. In *AAAI*. 3820–3826.

[15] Vivien Marx. 2013. Biology: The big challenges of big data. *Nature* 498, 7453 (2013), 255–260.

[16] Tomonari Masada, Tsuyoshi Hamada, Yuichiro Shibata, and Kiyoshi Oguri. 2009. Bayesian multi-topic microarray analysis with hyperparameter reestimation. In *International Conference on Advanced Data Mining and Applications*. Springer, 253–264.

[17] Matthew Meyerson, Stacey Gabriel, and Gad Getz. 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics* 11, 10 (2010), 685–696.

[18] Iver Petersen. 2011. The morphological and molecular diagnosis of lung cancer. *Dtsch Arztebl Int* 108, 31-32 (2011), 525–531.

[19] Naruemon Pratanwanich and Pietro Lio. 2014. Exploring the complexity of pathway–drug relationships using latent Dirichlet allocation. *Computational biology and chemistry* 53 (2014), 144–152.

[20] Sridhar Ramaswamy, Ken N Ross, Eric S Lander, and Todd R Golub. 2003. A molecular signature of metastasis in primary solid tumors. *Nature genetics* 33, 1 (2003), 49–54.

[21] Radim Rehurek. 2008. Gensim. (2008).

[22] Simon Rogers, Mark Girolami, Colin Campbell, and Rainer Breitling. 2005. The latent process decomposition of cDNA microarray data sets. *IEEE/ACM transactions on computational biology and bioinformatics* 2, 2 (2005), 143–156.

[23] Janne Sinkkonen, Juuso Parkkinen, Janne Aukia, and Samuel Kaski. 2008. A simple infinite topic mixture for rich graphs and relational data. (2008).

[24] Min Song and Su Yeon Kim. 2013. Detecting the knowledge structure of bioinformatics by mining full-text collections. *Scientometrics* 96, 1 (2013), 183–201.

[25] Hongning Wang, Minlie Huang, and Xiaoyan Zhu. 2009. Extract interaction detection methods from the biological literature. *BMC bioinformatics* 10, 1 (2009), S55.

[26] Hongwei Wang, Qiang Sun, Wenyuan Zhao, Lishuang Qi, Yunyan Gu, Pengfei Li, Mengmeng Zhang, Yang Li, Shu-Lin Liu, and Zheng Guo. 2014. Individual-level analysis of differential expression of genes and pathways for personalized medicine. *Bioinformatics* (2014), btu522.

[27] Weizhong Zhao, James J Chen, Roger Perkins, Yuping Wang, Zhichao Liu, Huixiao Hong, Weida Tong, and Wen Zou. 2016. A novel procedure on next generation sequencing data analysis using text mining algorithm. *BMC bioinformatics* 17, 1 (2016), 1.

[28] Weizhong Zhao, Wen Zou, and James J Chen. 2014. Topic modeling for cluster analysis of large biological and medical datasets. *BMC bioinformatics* 15, 11 (2014), S11.