Breast and Prostate Cancer Expression Similarity Analysis by Iterative SVM Based Ensemble Gene Selection

Darius Coelho
Dept. of Computer Science
Stony Brook University & SUNY Korea
Stony Brook, NY 11790
darius.coelho@stonybrook.edu

Lee Sael*
Dept. of Computer Science
SUNY Korea & Stony Brook University
Incheon, Korea
sael@sunykorea.ac.kr

ABSTRACT

Epidemiologic and phenotypic evidences indicate that breast and prostate cancers have high pathological similarities. Analysis of pathological similarities between cancers can be beneficial in several aspects such as enabling the knowledge transfer between the cancer studies. To gain knowledge of the similarity between breast and prostate cancer pathology, common genes that are affected by the two carcinomas are investigated. Gene expression data extracted from RNA-seq experiments, provided through TCGA consortium, are used for gene selection. Gene selection was performed using an iterative SVM based ensemble feature selection approach. Iterative SVM-based gene selection methods enable correlated gene expressions to be considered simultaneously and ensemble approach stabilizes the selection. As results of the analysis, two genes, Transglutaminase 4 (TGM4) and complement component 4A (C4A), were selected as commonly altered genes. Direct relationships of the two genes to the two cancers are not confirmed. However, TGM4 is known to be associated with adenocarcinomas and C4A with ovarian cancer. Thus provides evidence that they are pathologically important genes for the two cancers.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and genetics; H.3 [**Information Systems**]: Miscellaneous

Keywords

Breast cancer, Prostate cancer, RNA-seq, Gene Selection

1. INTRODUCTION

Many studies provide evidences that there are epidemiologic and phenotypic similarities between the breast invasive carcinoma (breast cancer) and prostate adenocarcinoma (prostate cancer) as extensively review by C. López-Otín and E. P. Diamandis [6].

Rates of incidences for breast and prostate cancer are parallel in various regions. That is, studies as early as 1990's show that breast and prostate cancer have high co-occurrence rate in various countries [2]. Also, breast and prostate cancer are both strongly influence by steroids. Moreover, removal of the gonadal reduces the risk and anti-estrogen are shown to be beneficial and possible preventive of the cancer [8]. Also, similar dietary pattern, such as fat consumption, also influence both cancer.

To investigate the similarities of breast and prostate cancer in a genomics perspective, we examine the genes that distinguish cancer samples from normal samples based on gene expression patterns. One of the methods for selection of these genes is application of feature selection [3]. The aim of feature selections in classification problems is to find a small set of features prior to classification with the end goals to best discriminate the difference between the diseased and the control samples. In an iterative feature selection approach, feature selection and classification is done alternatingly with a defined feature selection cost function. The cost function can be defined independently or dependently to the classification results. With an iterative feature selection approach, influenced genes can be selected for each cancer types that also provide low classification errors. Also, the resulting analysis of the common genes between the two cancer types will provide information on the similarity of their pathology.

There are two major technologies for measuring gene expression levels: DNA microarray and RNA-seq. The major difference between the two experiments are whether the gene expression levels are measured after hybridization to microarray chips or genes are sequenced after pulling out expressed genes from a sample. DNA microarray has been extensively used in the past years and is still being used due to low cost. However, RNA-seq is considered to be inherently better since genes with mutations or variations can be detected without any prior knowledge of the sequence pattern [7].

We use RNA-seq results as gene expression data and SVM-based iterative feature selection algorithm. The algorithm is combined with ensemble approach to identify genes that are common to breast and prostate cancer. Ensemble approaches were shown to increase the stability and robustness of gene selection in the microarray data analysis [4, 1]. Details of the gene selection pipeline are provided in the Methods section.

*To whom correspondence should be addressed.

DTMBIO'13, November 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2512089.2512099 ...$15.00.

2. METHODS

2.1 Data Set

Gene expression information for breast invasive carcinoma (BRCA) and prostate adenocarcinoma (PRAD) was acquired from The Cancer Genome Atlas (TCGA) data repository (https://tcga-data.nci.nih.gov/tcga/). For each of the BRCA and PRAD dataset, level 3 gene expression data extracted from RNA-seq experiments of solid tumor (TCGA sample type code 01) and solid tissue normal (TCGA sample type code 11) are used. BRCA dataset consists of 370 samples out of which 316 are cancer samples and 54 are normal samples. PRAD dataset consists of 218 samples out of which 175 are cancer samples and 43 are normal samples.

Both BRCA dataset and PRAD dataset have imbalanced number of cancer and normal samples. SVM is affected less by the "imbalanced data problem" compared to classification algorithms such as K-nearest neighbour. However, in the current run of linear SVM on the imbalanced datasets, biased selection towards the cancer samples occurred for the BRCA dataset. To reduce the bias and stabilize the gene selection, ensemble of classification on subsets of cancer samples is used. That is, first, the cancer and the normal samples are divided into testing sets and training sets. The training sets of the cancer samples are further divided to five equal numbers of subsets. Each one of the subsets is used as a sub-training set or as a validation set. Normal samples are divided to five subsets, four sub-training sets and a validation set. Five separate runs of SVM were performed with a subset of cancer samples and four subsets of normal samples. The final genes selected are the genes that are consistently present in the five selections. The numbers of samples are shown in the table 1.

Table 1: Data Samples

		BRCA		PRAD	
		Cancer	Normal	Cancer	Normal
Training	sub-train	52×5	8×5	26×5	6×5
	validation				
Testing		56	14	45	13
Total		316	54	175	43

2.2 Gene-Set Selection Pipeline

2.2.1 Classification with Linear SVM

Linear support vector machine (SVM) is used to learn the weights of genes. One of the strength of SVM is its capability of handling non-linear decision boundaries. However, current datasets have small number of samples compared to the number of features, which makes it difficult to exploit their non-linear characteristics. Thus, we choose to exploit only the maximum margin separation characteristics of SVM.

Linear discriminant function, $f(\boldsymbol{x})$, is calculated as the weighted sum of feature values of $\boldsymbol{x}$ with additive bias, b.

$$f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b$$

where $\boldsymbol{w} = \sum_n \alpha_n y_n \boldsymbol{x}_n$ and $b = \frac{1}{N_s} \sum_{n \in s} (y_n - \boldsymbol{w} \cdot \boldsymbol{x}_n)$ with s being the index of samples with non-zero αs and N_s being the sum of those samples.

Linear SVM trains the weight vector, $\boldsymbol{w}$, of the linear discriminant function, $f(\boldsymbol{x})$, by maximizing the margin between the boundary and support vectors on the training set $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$. For $l2$-norm soft margin classification, the optimization problem for learning the weights are as follows:

$$\begin{aligned} \text{minimize} \quad & \tfrac{1}{2}\|w\|^2 + C \textstyle\sum_{n=1}^{N} \xi_n \\ \text{subjected to} \quad & t_n f(\boldsymbol{x}_n) \geq 1 - \xi_n \\ & \xi_n \geq 0 \end{aligned}$$

The variable ξ is the slack variable for transforming the inequality constraint to an equality form. The constant C is the value associated with the strength of relaxation. The value of C is often selected by cross validation. However, in this experiment, C had little or no effect on the performance of classification. Thus, default value of $C = 100$ is used. Several implementations of SVM are publicly available. Among them, Kernlab [5], which is an R extension package, is used for SVM training and testing.

2.2.2 Feature Ranking

The rank of a gene can be evaluated by the change in the cost function caused by setting the weight of the given feature to zero. The cost function of a linear SVM can be simplified to $J = \frac{1}{2}\|\boldsymbol{w}\|^2$, which is the soft margin optimization function with the slack variable term dropped. Slack variable term depends on the sample and does not reflect the effect of the features in the classification. The minimization of $\|\boldsymbol{w}\|^2$ is equivalent to finding set of features with w_i^2, which justifies the uses of w_i^2 as the ranking criterion [3].

2.2.3 Iterative Gene Set Selection with Recursive Feature Ranking

Gene selection procedure is adopted from Recursive Feature Elimination (RFE) method introduced by Guyon et al. [3]. The original RFE algorithm is modified by integrating a section-wise removal procedure. In each iteration of the linear SVM-RFE, one or more features, i.e. genes, are removed and weights are retrained. The output of the linear SVM-RFE is gene set ranked list r that contains the list of genes or gene sets that have been removed in each iteration. With this information, gene subset ranking $G_0 \subset G_1 \subset \ldots \subset G_k$, where G_0 contains all genes and G_k contains one gene, can be generated. Then remaining genes are retrained on sub-training set and are used to measure the accuracy of classification on a validation set.

2.2.4 Ensemble Gene Set Selection

After the generation of gene subset ranking for each of the five sub-training samples, final gene set is selected with ensemble feature selection approach. There are various types of ensemble based feature selection methods [9]. However, due to the limitation in the number of available samples, we take a simple approach of voting. That is, first, find the gene-set-size threshold by selecting a gene subset that has the lowest stable type I error (false positive rate: FPR) and type II error (false negative rate: FNR) evaluated with the gene subset ranking of all five trained sets on the validation sets. After the gene-set-size threshold has been selected, gene subset of selected size is extracted for each of the five trained sets: G_{T1}, G_{T2} ... G_{T5}. Genes in the five sets, $g \in \{G_{E1} \cup G_{E2} \cup \ldots \cup G_{E5}\}$, are voted by the number of

occurrence in the selected sets. New ensemble gene subsets are generated by grouping the genes by the number of votes v: $E_{v=5}$, $E_{v\geq 4}$, $E_{v\geq 3}$, $E_{v\geq 2}$, and $E_{v\geq 1}$. Theses ensemble gene subsets are retrained with randomly selected cancer training samples and normal training sample of equal sizes (BRCA: 40/40 and PRAD: 30/30).

3. RESULT

Selection of gene sets that best discriminates the cancer sample from normal samples by the gene expression patterns is performed. Three groups of samples are used for the selection: BRCA and PRAD. We will first look at performance measures of intermediate and final results. Then analysis result of the common genes in BRCA, and PRAD will be discussed.

3.1 Performance Measures

To validate whether the selected genes play a part in explaining the breast and/or prostate cancer pathology, false positive rate (FPR), false negative rate (FNR), and accuracy of classification results are evaluated. For convenience, cancer tissues are considered to be positives data and normal tissues negative data.

3.1.1 Linear SVM-RFE

False positive rate (FPR), false negative rate (FNR), and accuracy of five sub-training results evaluated on the sub-training validation sets (Table 1) are plotted for each of the iteration steps. Figure 1 shows FPRs and FNRs for BRCA and PRAD. The x-axis is the number of remaining gene sets after removing selected number of genes in the iterative gene selection procedure. The gene set starts off with all the genes and decrease till only one gene remain in a set. The FPR/FNR patterns of BRCA and PRAD are drastically different. In PRAD dataset (Fig. 1B), FPR and FNR start around 0.1 and reduces gradually till gene set size of 1800. Then they become unstable as the gene set size reduces. In BRCA dataset (Fig. 1A), FPR and FNR start around 0.2 and 0.5, increases up to a certain point, and then decreases till a single gene is left.

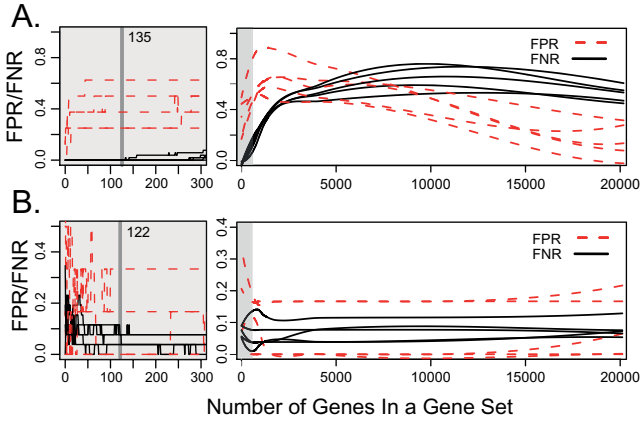

Figure 1: False positive rate and false negative rate of five trained weights on five validation sets. A. BRCA. B. PRAD.

Accuracy plots in Figure 2 also shows similar trends. To find the threshold for determining the gene-set-size for BRCA and PRAD datasets, stabilizing accuracy region with the highest accuracy just before or near peak in the accuracy values were determined and used. We selected gene set size of 135 for the breast cancer and 122 for the prostate cancer. The selected region is shown as gray vertical line in Figures 1 and 2.

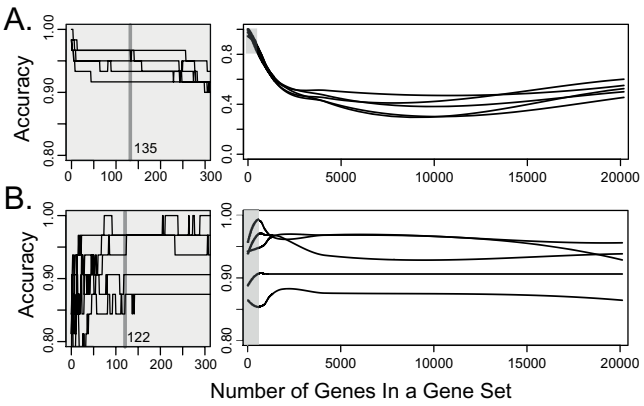

Figure 2: Accuracy measure of five trained weights on five validation sets. A. BRCA. B. PRAD.

3.1.2 Selected Gene Subsets

Using the gene-set-size of 135 for the breast cancer and 122 for the prostate cancer, five gene sets, G_{T1}, G_{T2}, ..., G_{T5}, are selected for each of BRCA and PRAD data sets. Weights of the selected gene sets are retrained on a random sample of balanced training set. The retrained weights are used to evaluate the performance on the reserved test sets for BRCA and PRAD separately. The Table 2 summarized the performance measures for the five groups.

Table 2: Performance measures of gene subset of selected size

	BRCA			PRAD		
	FPR	FNR	Acc	FPR	FNR	Acc
G_{T1}	0.57	0.00	0.89	0.08	0.11	0.90
G_{T2}	0.57	0.00	0.89	0.08	0.09	0.91
G_{T3}	0.57	0.02	0.87	0.08	0.07	0.93
G_{T4}	0.50	0.00	0.90	0.08	0.16	0.86
G_{T5}	0.64	0.00	0.87	0.15	0.11	0.88
Min	0.50	0.00	0.87	0.08	0.07	0.86
Max	0.64	0.02	0.90	0.15	0.16	0.93

3.1.3 Ensemble Gene Sets

Ensemble gene sets, $E_{v=5}$, $E_{v\geq 4}$, $E_{v\geq 3}$, $E_{v\geq 2}$, are generated by voting. The vote value of each gene, v, are the number of occurrence of the gene in the selected gene sets, G_{T1}, G_{T2}, ..., G_{T5}. Table 3 shows average performance of weights trained on three separate training sets and evaluated on the reserved testing sets. Goodness of the ensemble methods can be evaluated by comparing the performance of the original selected gene sets, G_{T1}, G_{T2}, ..., G_{T5}. For BRCA dataset, classification using 10 genes that occurred in all G_{T1}, G_{T2}, ..., G_{T5}, i.e., $E_{v=5}$, has the best accuracy. BRCA ensemble set $E_{v\geq 3}$ has equal to the maximum accuracy of the original selected gene sets. PRAC ensemble gene

sets, except for $E_{v=5}$, have the accuracy that is equal to or better than any of the ordinal selected gene sets. In the following sections, we will discuss the results using ensemble set $E_{v\geq 3}$, which BRCA and PRAC showed equal to maximum accuracy of original set, and $E_{v\geq 2}$, which have similar size of gene set as the original sets.

Table 3: Performance measures of ensemble gene subsets

	BRCA				PRAD			
	size	FPR	FNR	Acc	size	FPR	FNR	Acc
$E_{v=5}$	10	0.43	0.00	0.91	26	0.08	0.10	0.91
$E_{v\geq 4}$	10	0.43	0.00	0.91	62	0.08	0.07	0.93
$E_{v\geq 3}$	52	0.43	0.01	0.90	104	0.08	0.07	0.93
$E_{v\geq 2}$	129	0.55	0.02	0.87	162	0.05	0.06	0.94
$E_{v\geq 1}$	487	0.55	0.04	0.86	260	0.03	0.07	0.94
Min	-	0.43	0.00	0.86	-	0.0	0.06	0.91
Max	-	0.55	0.04	0.91	-	0.08	0.08	0.94

3.2 Common Genes Analysis

Major objective of current work is in the analysis of common genes that are active discriminators of both breast cancer and prostate cancer. Unlike initial expectation of finding significant number of common genes, only two common genes are present in the $E_{v\geq 2}$ set and nine in the $E_{v\geq 1}$ set.

Although only two genes are found that show significantly common expression between breast cancer and prostate cancer, pathway similarity can provide addition information in many cases. For this, we search the Ingenuity Pathway Analysis (IPA) database (Ingenuity Systems, www.ingenuity.com) to find functional information of the two genes: C4A and TGM4. C4A is known to be a part of acute phase response signaling, which is associated with rapid inflammatory response for protection against microorganisms. It is also known to be part of LXR/RXR activation, which is associated with regulation of cholesterol, fatty acid, and glucose homeostasis. Associations of TGM4 on any of the canonical pathways are yet not known. This could be the characteristics of the underlying biology or due to lack of information. However, testosterone and dihydrotestosterone, which are both sex steroids, have known association with breast cancer as well as prostate cancer. This is a strong indication the TGM4 is also associated with breast and prostate cancer.

4. CONCLUSIONS

Current study was initiated with a hypothesis that similarity of breast cancer and prostate cancer, which was strongly suggested by the epidemiologic and phenotypic evidences, can also be present in the gene expression pattern. Gene expression extracted from RNA-seq experiment for breast invasive carcinoma (BRCA) and prostate adenocarcinoma (PRAD) retrieved from TCGA database was used to provide evidence of the hypothesis. Iterative SVM based ensemble gene selection method was used to select genes that discriminate cancer samples from normal samples. The ensemble gene sets $E_{v\geq 3}$ were able to achieve accuracy of 90% for BRCA and 93% for PRAD. However, only two genes where common in the $E_{v\geq 3}$ of BRCA and $E_{v\geq 3}$ of PRAD gene sets.

Of the two common genes, Transglutaminase 4, did not have any directly known associations to the breast cancer nor the prostate cancer. Complement component 4A also did not have directly known association with breast cancer nor the prostate cancer. However, possibilities of association could be found through guilt by association in pathway analysis. This indicates that they could be common genes associated with various types of cancer. Although this information can be important in itself, further study needs to be done to conclude that there are significant pathological similarity in the genomic level exclusively for breast cancer and prostate cancer. As future work, similarity analysis between various types of cancer using various genomic information will be executed to extend the knowledge of cancer pathology.

5. ACKNOWLEDGMENTS

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the "IT Consilience Creative Program" (NIPA-2013-H0203-13-100). L.S. and D.C. thank the TCGA, the investigators, and the institutions who constitute the TCGA research network for making the BRCA and PRAD data available.

6. REFERENCES

[1] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3):392–8, Feb. 2010.

[2] K. K. Carroll. Dietary fats and cancer. *The American Journal of Clinical Nutrition*, 53(4 Suppl):1064S–1067S, 1991.

[3] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1-3):389–422, Jan. 2002.

[4] Z. He and W. Yu. Stable Feature Selection for Biomarker Discovery. pages 1–23, 2010.

[5] A. Karatzoglou, T. U. Wien, A. Smola, K. Hornik, and W. Wien. kernlab-an S4 package for kernel methods in R. *Journal of Statistical Software*, pages 1–20, 2004.

[6] C. López-Otín and E. P. Diamandis. Breast and prostate cancer: an analysis of common epidemiological, genetic, and biochemical features. *Endocrine Reviews*, 19(4):365–96, Aug. 1998.

[7] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–17, Sept. 2008.

[8] M. Osborne, P. Boyle, and M. Lipkin. Cancer prevention. *Lancet*, 349 Suppl:SII27–30, 1997.

[9] E. Tuv, A. Borisov, G. Runger, and K. Torkkola. Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination. *Journal of Machine Learning Research*, 10:1341–1366, 2009.