Gene Selection from Microarray Data for Age-related Macular Degeneration by Data Mining

Yuhan Hao and Gary M. Weiss

Abstract—**A DNA microarray can measure the expression of thousands of genes simultaneously. Previous studies have demonstrated that this technology can provide useful information in the study of molecular pathways underlying Age-related Macular Degeneration (AMD) process. Relevant genes involved in AMD are not understood completely. The microarray dataset used in this study contains 175 samples measured around 41,000 genes' expression. The samples have two classes, normal eyes and AMD eyes. Dimensionality reduction from 41,000 features is necessary before a classifier for distinguishing normal and AMD eyes can be built. In this paper three established methods are utilized to perform feature selection: Naive Bayes with feature removal, Logistic regression with L1-regularization, and Decision Tree methods (the resulting classification accuracies are 89.66%, 77.01% and 66.67%, respectively). The microarray dataset is also visualized using Principal Component Analysis (PCA). The PENK and TRAF6 genes are the only two genes which are selected by the three feature selection methods and we expect that both of them are involved in the extracellular signal transduction. Functional Annotation Clustering of genes selected by any classifier also suggests that most genes are responsible for signal transduction in individual cell and between cells. The signal transduction pathway containing the selected genes may play an important role in AMD pathogenesis .**

1. Introduction

Age-related macular degeneration is a progressive neurodegenerative disease, which primarily affects retina pigmented epithelium (RPE) cells in the retina. RPE and choroid tissues contribute to maintaining vision function. RPE layers will eliminate the shedding outer segment of photoreceptors and promote retinal adhesion stabilizing alignment. Dysfunction of RPE usually results in disruption of retinal adhesion in persistent retinal detachment or photoreceptor apoptosis [1]. Nearly 40% of people over 75 years old have some pathological signs of AMD [2]. The molecular pathogenesis of AMD is not fully understood. A microarray can measure the expression of thousands of genes simultaneously, so it allows us to analyze this disease by a top-down approach.

This study uses the microarray dataset on Gene Expression Omnibus, Series GSE29801 [3]. It contains 175 RPE tissues samples from normal people and AMD patients. Although this dataset contains information about gender and age, only gene expression data is used in this study, because the goal is to discover representative genes for AMD molecular pathogenesis. Because the microarray data consists of 175 samples with 41,000 genes as features, dimensionality reduction is required before the data can be mined to generate a classifier capable of distinguishing normal samples from AMD samples. One straightforward method for reducing the dimensionality is to set the normal samples as standard and calculate difference from AMD samples and the p-value from GEO2R by unpaired two-tailed t-test with unequal variance [4]. The p-value is adjusted by permuting class labels 1,000 times with the Fisher-Yates methods [5]. GEO2R is a web tool that is able to compare groups of samples in order to identify genes that are differentially expressed across two experimental conditions.

In this paper three established feature selection algorithms are used to select relevant genes. The first method, Naive Bayes with feature removal, tends to select the features that maximize the accuracy of the generated classification model. The second method, logistic regression with L1 regularization, forces most coefficients to zero and the features with large coefficients represent the selected features. In both cases the selected features represent a relevant-gene list. These two classification methods are generative methods that can learn non-linear boundaries. The third method that is used is a decision tree method. This method is very robust with respect to large numbers of features, and can build an effective classifier that only utilizes the most informative feature. That is, decision tree methods automatically identify the most relevant features and ignore irrelevant or redundant features. The features that appear in the decision tree form the relevant-gene list. Finally, in addition to these three methods, an ensemble feature selection method is used to aggregate the selected features from these three methods [6].

Only two genes, PENK and TRAF6, are selected by all three methods, while 36 genes are selected by one of the three methods. We believe that the selected genes, especially PENK and TRAF6, will make a contribution to molecular pathway in the process of AMD and give hints to future therapies for this disease.

The purpose of this study is not to propose a new feature selection algorithm that can help build an effective classifier, since we focus on the molecular mechanism of AMD instead of diagnosis of this disease. Rather, the goal is to utilize data mining approaches to investigate biological systems and try to identify genes hidden in the microarray data that may be important to the AMD process.

The rest of the paper is organized as follows. In Section 2 we describe the microarray dataset, the four feature selection methods, a method to visualize the dataset, and one bioinformatics tool. Section 3 describes the experiments and results. Related work is discussed in

Y. Hao is with Fordham University, Bronx, NY 10458 USA (e-mail: yhao2@fordham.edu).

G. M. Weiss is with Fordham University, Bronx, NY 10458 USA (e-mail: gaweiss@fordham.edu).

Section 4 and the conclusion is provided in Section 5.

2. Materials and Methods

This section describes the microarray dataset and the methods and tools used in this study. The dataset is described in Section 2.1, the four feature/gene selection methods are described in Sections 2.2 to 2.5, the use of PCA for visualizing the dataset using three values is described in Section 2.6, and a functional clustering method to cluster similar genes is described in Section 2.7.

2.1 Microarray Dataset

This microarray dataset of Newman et al. [3] is measured from RNA collected from eyes of the human. It consists of two types of samples, 96 from RPE/choroid of normal eyes and 79 from RPE/choroid of AMD eyes. The training dataset has 88 samples which are selected randomly (48 normal eyes and 40 AMD eyes). Each sample has 41,000 features or genes. The test dataset consists of 48 normal eyes and 39 AMD ones.

2.2 Naive Bayes Classifier and Feature Removal

The Naive Bayes classifier is used to classify eyes status as normal or AMD. It assumes that continuous features are independent given class and normality. The assumption of independence does not hold for this dataset since genes are highly correlated. The data is transformed by applying the log_2 function in order to have the data more closely resemble a normal distribution. The class is determined by joint likelihood, $P(Gene_1, Gene_2, Gene_3.../ Y)$, $Y \in \{normal, AMD\}$.

Feature removal is based on maximal test accuracy. It starts with $F=\{x_1, x_2, x_3,...\}-\{x_i^1\}$, which x_i is removed Gene i and x_i satisfies argmaxi accuracy(F). Then it enters a loop, finding higher test accuracy for removing another attribute, x_i^2: argmaxi accuracy(F), removing x_i^2 to F feature with highest accuracy increase: $F=F-\{x_i^2\}$. The loop is repeated until test accuracy ceases to increase. The attributes that exist in F are the selected genes.

2.3 Logistic Regression with L1 Regularization

The Logistic Regression classifier is appropriate when dependent variable is nominal and the independent variable is continuous. The sigmoid function is used to represent the likelihood of the class. The learning algorithm is as follow:

$$w_j \leftarrow w_j + \Delta w_j - \frac{sign(w_j)}{\lambda} \qquad (1)$$
$$\Delta w_j = \eta * x_j^i * [y^i - g(w^T x^i)]$$

The learning step, η is set to 0.0001 and the number of iterations is set to 10,000. Meanwhile, in order to find the most representative or important genes, Maximum A Posteriori L1-regularization is used to force some parameter values to zero, equivalent to the function of feature selection. In this case, λ is equal to 1/500, a bias to minimize the number of non-zero variables.

2.4 Decision Tree (J48)

A decision tree is an expressive algorithm that generates a tree-like graph with leaf nodes and edges and can handle large numbers of features. The decision tree algorithm will tend to ignore redundant and irrelevant features. The J48 decision tree algorithm, which is part of the WEKA data mining suite, is used in this study. The splitting attributes that appear in the induced decision tree classification model are the genes selected by the model.

2.5 Ensemble Feature Selection

To increase the confidence of feature selection results and decrease the bias and errors induced by unsatisfied assumptions of the model, an ensemble feature selection technique [7] is employed. Ensemble feature selection methods utilize the same basic idea of ensemble classification algorithms [8]. A number of different feature selection algorithms are used and then the selected features from each algorithm are then aggregated. This study simplifies the ensemble feature selection process, which typically ranks each selected feature and then accepts only those with a high overall ranking. However, the ensemble feature selection in this study only focuses on the intersection and union of features selected by Naive Bayes, Logistic Regression and Decision Tree methods.

2.6 Principal Component Analysis (PCA)

PCA is a widely used statistical method to decrease dimensionality. In this study it is not used as a classifier, but as a method to simplify the visualization of the data. PCA is applied to the entire dataset (training and test data) and then the first three principal component (PC) vectors are selected as the *x*, *y* and *z* values to be plotted.

2.7 The Database for Annotation, Visualization and Integrated Discovery (DAVID)

DAVID is interactive web software that is used to cluster genes by function annotations. It provides a rapid approach to reducing large lists of genes into functionally related categories, which are ordered by enrichment of the category. It is widely used among researchers from over 5,000 institutes globally and cited by more than 6,000 academic publications [9]. In this study, the input is the union of the genes selected by the three base feature selection methods. The output is Functional Annotation Clustering Report. It shows different functional categories where genes most enriched. (DAVID: david.ncifcrf.gov)

3. Results

In this section the microarray data is filtered and we obtain 86 features and transform the expression of the gene. We then select relevant genes using the three base feature selection methods and aggregate the relevant genes using ensemble feature selection. Finally, we visualize the dataset using PCA

3.1 Filtering of AMD microarray dataset

GEO2R provides a way to calculate the p-value of each gene which is used to remove genes with random changes of expression (Table 1). The genes in Table 1 are listed in the order according to their adjusted p-value. The genes with large p-value indicate that their expression does not have a stable change between two groups, and they will

lead to more variations in the model. The threshold of adj p-value 0.05 is chosen. The genes with higher 0.05 adj p-value are removed. Also, it is found that some genes are repeated several times and some lack of annotation information. Finally, 86 genes are left and take on the role of features for the classification task.

3.2 Naive Bayes classifier for Eye Status Classification

In Naive Bayes classifier, the log_2 transformation is needed for the microarray expression data, continuous variable. The distribution of the transformed dataset is closer to a normal distribution, so it is fair that we use normal distribution function to estimate $P(Gene_i / Y)$.

The accuracy of this model is 83.91%. Then feature removal method is used, which tries to achieve maximal test accuracy. There are 20 genes left and they are listed in Table 2. Genes 74 to 86 are almost selected continuously. The test set accuracy, somewhat surprisingly, increases to 89.66%.

3.3 Logistic Regression for Eye Status Classification

The training dataset for Logistic Regression has added one constant column where each cell is 1 working as a constant. In order to find the most important genes in this AMD dataset, MAP with L1 regularization (η=0.0001, λ=1/500 and the number of iterations is 10000) is used to get minimal number of non-zero variables. The accuracy of the model is 77.01% and the time to build this model is 2 minutes. A larger learning rate, η and smaller iterations are also tested, but decreases the accuracy to around 50%. The value of weight vector provides a way to determine which genes are relatively dominant in this classification. Table 2 also shows that 16 attributes are selected by logistic regression and weight of these 16 attributes are higher than 0.5 or lower than -0.5.

TABLE 1
GENES ORDERED BY ADJ. P VALUE FROM GEO2R

Gene ID	adj.P.Val	Gene.symbol
30870	0.00014	----[a]
10463	0.00197	DNM3OS
9571	0.00533	----
29804	0.00533	AKAP8L
.	.	.
44181	0.85117	----
2160	0.85117	PFN1
9502	0.85117	EPS15
15494	0.85117	YLPM1
.	.	.
9773	0.99985	LRRC72
37141	0.99985	----
38929	0.99985	DZANK1

This table only contains the small fragment of the whole dataset.
[a]Information of Gene symbol is not complete.

3.4 Decision Tree for Eye Status Classification

The J48 decision tree algorithm is used to classify normal and AMD eyes. The accuracy of the induced decision tree, shown in Fig. 1, is 66.67%. This accuracy is a little lower than for the other two classifiers, but is higher than the 52.87% accuracy of the ZeroR method, which corresponds to predicting the majority class. The splitting attributes in Fig. 1 comprise the selected genes (Table 2).

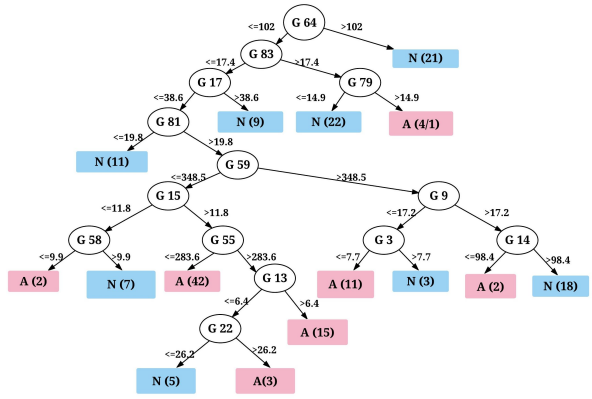

Fig. 1. J48 decision tree of classification of normal and AMD eyes. The rectangle leaf nodes represent the class of the sample and round splitting features represent selected genes. Accuracy of this tree is 66.67%. The following abbreviations are applied: G--Gene, A--AMD, N--Normal.

3.5 Ensemble Feature Selection for Three Classifiers and DAVID

Ensemble feature selection can reduce bias and variance and increase confidence. In this study, the ensemble feature selection contains three classifiers which are used to pick up candidate genes. Table 3 shows that information of genes which is selected at least twice and the first two genes, PENK and TRAF6, which are selected by all three classifiers. Moreover, there are 36 genes that at least are selected by one classifier.

Functional Annotation Clustering by DAVID is used for these 36 genes to discover what functional clusters these genes belong to. Add the gene list to DAVID and it generates numbers of functionally related categories ordered by enrichment. The first three functional categories are about extracellular region, negative regulation of cell

TABLE 2
RELEVANT-GENE LIST FROM THREE CLASSIFIERS

Naive Bayes	Logistic Regression	Decision Tree
3	7	3
29	12	9
34	14	13
46	15	14
58	27	15
62	29	17
66	31	22
71	34	55
73	35	58
74	44	59
75	58	64
77	59	79
78	62	81
79	79	83
80	85	----
82	86	----
83	----	----
84	----	----
85	----	----
86	----	----

communication, and cell surface receptor linked signal transduction. All three clustering functions indicate those genes are involved in the process of signal transduction.

3.6 Visualization of the dataset using PCA

Principal component analysis is a common technique to reduce the dimensionality of dataset by mapping variables to a new set of variables. It has been used to analyze gene expression patterns [10]. Even though PCA is used to cluster genes in the whole dataset, instead of being a classifier, it also can be used for gene selection in cancer microarray dataset [11]. In this study it is used to illustrate the distribution of samples with 86 genes and 12 genes which are selected by two classifiers (Table 3).

In Fig. 2 the *x*, *y* and *z*-axis represent PC1, PC2 and PC3 respectively. Figure 2a indicates the first, second and third principle components contain 33%, 9% and 6% of the total information. However, in the case of selected genes (Fig 2b), PC1 can already represent 97% of total information.

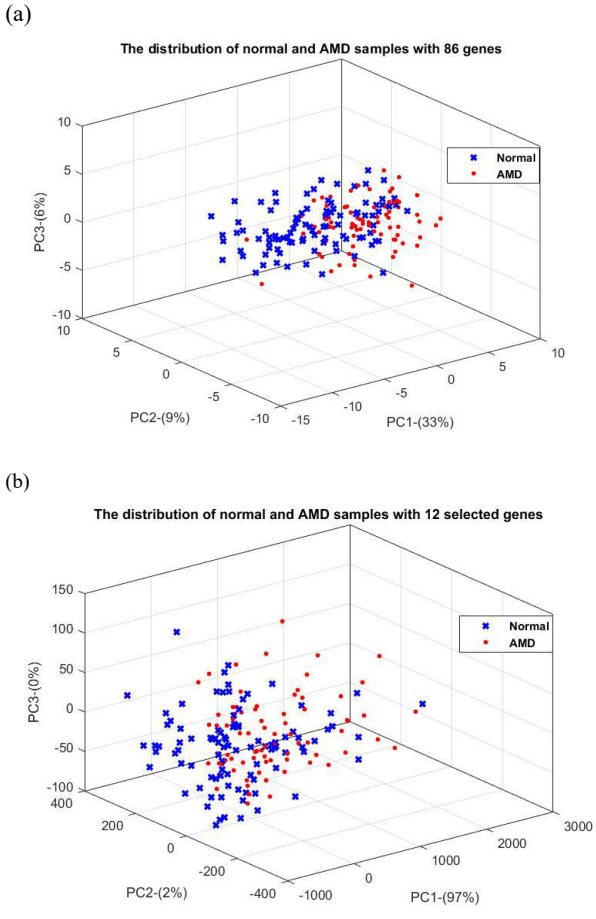

Fig. 2. Gene expression of normal and AMD eyes. It is established by PCA and x,y,z represent PC1, PC2 and PC3 respectively. Blue crosses represent normal eyes, and red dots represent AMD samples. The samples are not successfully separated under 86 genes, but PC1 under 12 selected genes can represent 97% of data. It indicates that selected genes are effective for classification of normal and AMD eyes.

These results show that the total dispersion in 86 genes PCA is much larger than in the 12 selected genes by any two classifiers. It strengthens the effectiveness of selected genes for classification.

TABLE 3
RELEVANT GENES SELECTED BY ENSEMBLE FEATURE SELECTION

Gene ID	Gene.symbol	Gene.title
58[a]	PENK	proenkephalin
79[a]	TRAF6	TNF receptor-associated factor 6
3	KATNAL2	katanin p60 subunit A-like 2
14	CLUAP1	clusterin associated protein 1
15	FAM208B	family with sequence similarity 208, member B
29	WASH1	WAS protein family homolog 1
34	TTC12	tetratricopeptide repeat domain 12
59	LINC00674	long intergenic non-protein coding
62	HIST1H3D	histone cluster 1, H3d
83	CYB5R1	cytochrome b5 reductase 1
85	CD163L1	CD163 molecule-like 1
86	MYRIP	myosin VIIA and Rab interacting protein

Those genes are selected at least by two classifiers.
[a]Only two genes are selected by all three classifiers.

4. RELATED WORK

Naive Bayes has been used in the analysis of microarray data. Though the independence assumption is an obstacle for Naive Bayes, it can be addressed by using Bayesian hierarchical models, which account for biological associations in a probabilistic framework. But for unknown interaction between genes, we still have to assume independence [12]. Logistic regression with lasso or L1-regularization is commonly used to handle high-dimensional data. Adaptive Lasso containing another term $1/w_j$ to control lasso strength [13], and elastic-net method [14] combing L1 and L2 regularization, both can relieve the influence of highly correlated variables. A random-forest-like logistic regression method is proposed, random lasso [15]. It mainly consists of two steps. First, lasso method is applied to bootstrap samples. After the first step, another term, importance is added to high-weighted variables. It can select all highly correlated variables, instead of normal lasso method which only select one of the highly correlated variables.

Support vector machines (SVMs) have already been widely used in analysis microarray data [16]–[18]. It does not require independent variables and yields very good performance. But SVM cannot directly select the most important genes for classification. The decision tree method has been used to analyze microarray data. One study indicates that decision trees can handle high-dimension data and find appropriate splitting attributes for classification of cancer [19]. However, if the cancer is caused by many genes, the decision tree may not perform well because decision trees cannot readily handle interactions between variables.

5. CONCLUSION

In this study, two genes, PENK and TRAF6, are selected by all three of our basic feature selection methods. PENK encodes a signal protein which is responsible for signal transduction in the synapse between two neurons [20] and modulates the perception of pain [21], though it has not been fully studied in other types of cells. It may also work as a signal protein between adjacent RPE cells or between

RPE cells and retina which belong to neuron cells. Also, it is regulated by two crucial genes Fos and Jun [22], which regulate cell proliferation. TRAF6 is a member of the TNF receptor associated factor (TRAF) protein family. It participates in signal transduction of both the TNF receptor and the interleukin-1 receptor [23]. Through this pathway, it regulates many signaling cascades in adaptive immunity, innate immunity and bone homeostasis. It also functions as a signal transducer in the NF-kappaB pathway in response to proinflammatory cytokines and other environmental interactions [24]. It indicates that these three pathways may play an important in AMD process.

Furthermore, PCA with 12 selected genes by any two classifiers shows a surprisingly good performance, and it suggests that those genes are vital for classification. Besides, functional clustering results from DAVID is corresponding to the function of PENK and TRAF6. The clustering results suggest that those 36 genes selected by any classifier are mainly involved in signal transduction.

This paper uses data mining methods to select relevant features for classifying AMD and in so doing provides insight into the disease. AMD is completely different from other types of cancers, so the feature selection algorithms for cancers may not be useful in this case.

Finally, we cannot find one or two genes which are responsible for classification of normal and AMD eyes status, but the selected genes suggest extracellular environment does impact the process of AMD, which also agrees with the experimental studies of AMD. Furthermore, PENK and TRAF6 have not been studied for the molecular pathogenesis of AMD in biological experiments, so further biological experiments are still needed. This paper may provide new insight into finding the AMD pathological molecular pathways.

REFERENCES

[1] Cook, Briggs, et al. "Apoptotic photoreceptor degeneration in experimental retinal detachment." Investigative ophthalmology & visual science 36.6 (1995): 990-996.

[2] Klein R, Chou CF, Klein BE, Zhang X, Meuer SM, Saaddine JB: Prevalence of age-related macular degeneration in the US population. Arch Ophthalmol 2011, 129:75-80.

[3] Newman, Aaron M., et al. "Systems-level analysis of age-related macular degeneration reveals global biomarkers and phenotype-specific functional networks." Genome Med 4.2 (2012): 16.

[4] Tsai, Chen‐An, Yi‐Ju Chen, and James J. Chen. "Testing for differentially expressed genes with microarray data." Nucleic acids research 31.9 (2003): 52.

[5] Fisher RA, Yates F: Statistical Tables for Biological, Agricultural and Medical Research. Oliver and Boyd, London, England: Hafner Press; 1948.

[6] Abeel, Thomas, et al. "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods." Bioinformatics 26.3 (2010): 392-398.

[7] Saeys, Yvan, Thomas Abeel, and Yves Van de Peer. "Robust feature selection using ensemble feature selection techniques." Machine learning and knowledge discovery in databases. Springer Berlin Heidelberg, 2008. 313-325.

[8] Dietterich, Thomas G. "Ensemble methods in machine learning." Multiple classifier systems. Springer Berlin Heidelberg, 2000. 1-15.

[9] Jiao, Xiaoli, et al. "DAVID-WS: a stateful web service to facilitate gene/protein list analysis." Bioinformatics 28.13 (2012): 1805-1806.

[10] Yeung, Ka Yee, and Walter L. Ruzzo. "Principal component analysis for clustering gene expression data." Bioinformatics 17.9 (2001): 763-774.

[11] Wang, Antai, and Edmund A. Gehan. "Gene selection for microarray data analysis using principal component analysis." Statistics in medicine 24.13 (2005): 2069-2087.

[12] Demichelis, Francesca, et al. "A hierarchical Naive Bayes Model for handling sample heterogeneity in classification problems: an application to tissue microarrays." BMC bioinformatics 7.1 (2006): 514.

[13] Zou, Hui. "The adaptive lasso and its oracle properties." Journal of the American statistical association 101.476 (2006): 1418-1429.

[14] Zou, Hui, and Trevor Hastie. "Regularization and variable selection via the elastic net." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67.2 (2005): 301-320.

[15] Wang, Sijian, et al. "Random lasso." The annals of applied statistics 5.1 (2011): 468.

[16] Brown, Michael PS, et al. "Knowledge-based analysis of microarray gene expression data by using support vector machines." Proceedings of the National Academy of Sciences 97.1 (2000): 262-267.

[17] Furey, Terrence S., et al. "Support vector machine classification and validation of cancer tissue samples using microarray expression data." Bioinformatics 16.10 (2000): 906-914.

[18] Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." Machine learning 46.1-3 (2002): 389-422.

[19] Horng, Jorng-Tzong, et al. "An expert system to classify microarray gene expression data using gene selection by decision tree." Expert Systems with Applications 36.5 (2009): 9072-9081.

[20] Comb, Michael, et al. "Proteins bound at adjacent DNA elements act synergistically to regulate human proenkephalin cAMP inducible transcription." The EMBO journal 7.12 (1988): 3793.

[21] König, Monika, et al. "Pain responses, anxiety and aggression in mice deficient in pre-proenkephalin." (1996): 535-538.

[22] Sonnenberg, June L., et al. "Regulation of proenkephalin by Fos and Jun." Science 246.4937 (1989): 1622-1625.

[23] Ye, Hong, et al. "Distinct molecular mechanism for initiating TRAF6 signalling." Nature 418.6896 (2002): 443-447.

[24] Deng, Li, et al. "Activation of the IκB kinase complex by TRAF6 requires a dimeric ubiquitin-conjugating enzyme complex and a unique polyubiquitin chain." Cell 103.2 (2000): 351-361.