Chapter 38
Identification of Candidate Genes Responsible for Age-Related Macular Degeneration Using Microarray Data

Yuhan Hao
Fordham University, USA

Gary M. Weiss
Fordham University, USA

Stuart M Brown
NYU School of Medicine, USA

ABSTRACT

A DNA microarray can measure the expression of thousands of genes simultaneously, and this enables us to study the molecular pathways underlying Age-related Macular Degeneration. Previous studies have not determined which genes are responsible for the process of AMD. The authors address this deficiency by applying modern data mining and machine learning feature selection algorithms to the AMD microarray dataset. In this paper four methods are utilized to perform feature selection: Naïve Bayes, Random Forest, Random Lasso, and Ensemble Feature Selection. Functional Annotation of 20 final selected genes suggests that most of them are responsible for signal transduction in an individual cell or between cells. The top seven genes, five protein-coding genes and two non-coding RNAs, are explored from their signaling pathways, functional interactions and associations with retinal pigment epithelium cells. The authors conclude that Pten/PI3K/Akt pathway, NF-kappaB pathway, JNK cascade, Non-canonical Wnt Pathway, and two biological processes of cilia are likely to play important roles in AMD pathogenesis.

DOI: 10.4018/978-1-5225-8903-7.ch038

1. INTRODUCTION

Age-related macular degeneration is a progressive neurodegenerative disease, and nearly 40% of people over 75 years of age have some pathological signs of AMD (Klein et al., 2011). It primarily affects retina pigmented epithelium (RPE) cells that lie beneath the retina. RPE cells help to maintain vision and usually eliminates the shedding of the outer segment of photoreceptors and promotes retinal adhesion stabilizing alignment. Dysfunction of RPE cells usually results in disruption of retinal adhesion in persistent retinal detachment or photoreceptor apoptosis (Cook et al., 1995). However, the molecular pathogenesis of AMD in RPE cells is not fully understood. Thus, our goal is to find the underlying molecular and cellular mechanism for the dysfunction of RPE cells and formation of AMD in silico.

Data mining methods have been widely applied to analyze microarray data. For example, Naïve Bayes is a commonly used generative approach. It is established on the distribution of features in each of classes and then classifies records according to the larger likelihoods for classes. Though the independence assumption is an obstacle for Naïve 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 of each feature (Demichelis et al., 2006). Logistic regression, with lasso or L1-regularization, is commonly used to handle high-dimensional data. Adaptive Lasso contains another penalty term to control lasso strength (Zou, 2006), and the elastic-net method (Zou & Hastie, 2005), combining L1 and L2 regularization, can relieve the influence of highly correlated variables. Random lasso, a random-forest-like logistic regression method, has been proposed (Wang et al., 2011). This method first applies the lasso method to bootstrap samples. Then another term, importance, is added to high-weighted variables. This method can select all highly correlated variables, whereas the normal lasso method can only select one of the highly correlated variables. Support vector machines (SVMs) have been broadly used for the analysis microarray data (Brown et al., 2000; Furey et al., 2000; Guyon et al., 2005; Statnikov et al., 2005). SVMs do not require independent variables but yield very good performance. The SVM, due to the kernel transformation, can generate complex boundary between classes. SVM's with a 'flagship' kernel have been particularly effective in many bioinformatics fields, such as DNA sequence classification and protein mass spectrometry (Noble, 2006).

More complicated methods, such as Random Forest, Artificial Neural Networks, and Deep Learning, are increasingly being utilized in the field of bioinformatics (Qi, 2012; Shen & Bax, 2013; Quang et al., 2014; Alipanahi et al., 2015). Random Forest offers several compelling benefits, since it copes well with small sample size and high dimensional or complex structure data (Yang et al., 2010). There is evidence that the performance of Random Forest is better than SVM for microarray data (Statnikov et al., 2008). Deep learning, which has advanced very quickly and has become quite popular, is extensively utilized in the bioinformatics domain (Min et al., 2016; Fakoor et al., 2013).

This study uses the microarray dataset on Gene Expression Omnibus, Series GSE29801 (Newman et al., 2012). It contains 175 RPE tissues samples from normal people and AMD patients. Although this dataset contains other 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 dataset consists of 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 dimensionality is to use stable differentially expressed genes between normal and AMD samples as features for the next-step feature selection. Stable differentially expressed genes have both large expression fold change and small p-value which is calculated from un-