STAT 428 Project Report

Breast Cancer Data Analysis

Xiaodan Zhang

STAT 428-Statistical Computing

Wenxuan Zhong

University of Illinois at Urbana-Champaign

Apr.14, 2013

1. Introduction

Breast cancer is a type of cancer that originates from breast tissue, most commonly from the inner lining of milk ducts or the lobules that supply the ducts with milk. The overwhelming majority of human cases occur in women and people become curious in studying which genes play important parts in affecting the happening of the breast cancer. In this project, we will analyze a breast cancer dataset and use it to find genetic markers. For the selected genetic markers, we are going to give bio-functions corresponding to them and use the selected markers to predict the breast cancer. Then, we will discuss the limitations of the prediction model and some approaches to overcome them.

The data we use are gene expression ratio data for BRCA experiments. Gene expression ratios included in the data were derived from the fluorescent intensity from a tumor sample (BRCA1, BRCA2, or Sporadic) divided by the fluorescent intensity from a common reference ample. There are 3226 genes (rows) and 22 cell lines samples (columns) in this dataset. There are three cases of the disease: BRCA1, BRCA2 and Sporadic. And if the response is 0 or 1 where 0 represents a healthy person and 1 represents a patient with a cancer, the columns whose names contain "sporadic" can be regarded as "healthy" cases with response as 0, while others can be regarded as "breast-cancer" cases with response as 1. So, if we rearrange the columns and put all "sporadic" ones first followed by all "breast-cancer" ones, we have a corresponding "response" array with 7 0's followed by 15 1's.

2. Methodology

In order to find breast-cancer related genetic markers, I choose to do a two-sample t-test. That is, the two samples are the "healthy" group and the "breast-cancer" group. Since we need to select a limited number of genes, t-test can help us find significant genes at a specific significance level. In this way, after doing F-test for determining whether the sample variances are different or not, we can then use t-test function with different input parameters to narrow down our original 3226 genes to 60 genes at 0.01 significance level. We choose 0.01 because 0.05 only helps narrow the amount of data from 3226 to 300+, which is still too large. Moreover, the Heatmap generated with the selected genetic markers show clearer patterns of clustering than with the original data. Next, we use LDA model to predict the breast cancer. The prediction error seems low here and the LDA histogram shows two obviously isolate clusters.

3. Conclusion

In this project, by doing two-sample t-test, we choose the most significant 60 genetic markers and use them to help predict the breast cancer. The process seems reasonable and feasible. As for the bio-functions for the selected genetic markers, I choose two examples. The first marker is "HV8D9", named ARVCF armadillo repeat gene deleted in velocardiofacial syndrome [*Homo sapiens* (human)] and its bio-function is as the following: Armadillo Repeat gene deleted in Velo-Cardio-Facial syndrome

(ARVCF) is a member of the catenin family. This family plays an important role in the formation of adherens junction complexes, which are thought to facilitate communication between the inside and outside environments of a cell. The ARVCF gene was isolated in the search for the genetic defect responsible for the autosomal dominant Velo-Cardio-Facial syndrome (VCFS), a relatively common human disorder with phenotypic features including cleft palate, conotruncal heart defects and facial dysmorphology. The ARVCF gene encodes a protein containing two motifs, a coiled coil domain in the N-terminus and a 10 armadillo repeat sequence in the midregion. Since these sequences can facilitate protein-protein interactions ARVCF is thought to function in a protein complex. In addition, ARVCF contains a predicted nuclear-targeting sequence suggesting that it may have a function as a nuclear protein. [provided by RefSeq, Jun 2010]

And the second marker is "TNF1F2", named "activating transcription factor 4 (tax-responsive enhancer element B67)" and its bio-function is as the following:

This gene encodes a transcription factor that was originally identified as a widely expressed mammalian DNA binding protein that could bind a tax-responsive enhancer element in the LTR of HTLV-1. The encoded protein was also isolated and characterized as the cAMP-response element binding protein 2 (CREB-2). The protein encoded by this gene belongs to a family of DNA-binding proteins that includes the AP-1 family of transcription factors, cAMP-response element binding proteins (CREBs) and CREB-like proteins. These transcription factors share a leucine zipper region that is involved in protein-protein interactions, located C-terminal to a stretch of basic amino acids that functions as a DNA binding domain. Two alternative transcripts encoding the same protein have been described. Two pseudogenes are located on the X chromosome at q28 in a region containing a large inverted duplication. [provided by RefSeq, Sep 2011]

4. Discussion

As for the limitations of the model we choose to use, our t statistics may be affected by small or unstable variances. Moreover, as for the LDA model we use, although it maximizes class separability and keeps variance of all classes roughly constant, it implicitly not only assumes Gaussian distribution of data, but also assumes that the mean is the discriminating factor, not variance. Additionally, LDA may also underfit or overfit the data. However, in our data analysis steps, by using qqnorm function, we know that the normality assumption meets. And in order to overcome the overfitting/underfitting limitation, we use feature selection, which is basically using t-tests.