436r An Improved Pca Approach for Microarray Data Analysis

Derrick Rollins, Dongmei Zhai, and Ramon Gonzalez

A critical challenge in bioinformatics is the extraction of concrete information from global gene expression data that can accurately associate critical genes with the experimental conditions (i.e., assay). The difficulty lies in the fact that the number of genes, typically in the order of thousands, is much greater that the number of assays (e.g., treatments, subjects, conditions, etc.), typically less than twenty. Thus, efficient information extraction and dimensionality reduction methods are needed to analyze this type of data. Two promising and popular approaches are singular value decomposition (SVD) and principal component analysis (PCA). SVD and PCA are related and Berrar et al. [1] describe their relationship in detail. However, the focus of this talk is only PCA as a tool in analyzing microarray data. Moreover, in this talk, we present a new and powerful way to apply PCA to microarray data that overcomes several limitations of the current PCA approach. The limitations of the current PCA approach include an inability to verify (or deny) association of a principal component vector to the known experimental conditions and then to accurately identify the critical genes.

In our computationally more intensive proposed approach, we are able to obtain a new type of principal components and demonstrate their high ability to scrutinized different experimental conditions. We will demonstrate this ability using real and simulated microarray data. The study with real data consists of expression data with 4290 identified genes and fourteen experimental conditions, with four treated with ethanol and ten untreated conditions [2]. This talk will demonstrate prefect identification of treated and untreated cases by the first principal component. Thus, we are able to conclude that this component physically represents ethanol treatment. Next we split the ethanol group from the non-ethanol group and determine the cumulative contributions for each gene for these two groups using this principal component. We next ordered the genes by contribution for both groups and then determine the genes with the strongest and weakest contributions to the ethanol cases and the non-ethanol cases.

To assess the ability of our method to correctly identify genes associated with experimental conditions we generated artificial gene expression data following the procedure in [1]. In this study we were able to assess the proposed method along with the current approach. Since, in a simulation study, one knows the truth, we were able to evaluate the ability of each approach to obtain this truth. In one study, we created artificial data exactly following a study in [1] and verified that we were able to obtain the same results they reported. In this study, all genes were given random noise, certain genes were also given sinusoidal behavior, and certain other genes were given exponentially decaying behavior. In this study we demonstrate the weaknesses of the current approach and the strengths of the proposed method as described above. More specifically, we show that the proposed method was able to obtain and identify the principal component that represented sinusoidal behavior and to identify all the genes that were given sinusoidal or exponentially decaying behavior. We also show that this was not possible with the current PCA approach. Finally, this talk also presents results of other studies where we imposed specific types of functions on certain genes and give additional support for the use of the proposed PCA method.

[1] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, Singular value decomposition and principal component analysis, in "A Practical Approach to Microarray Data Analysis (D. P. Berrar, W. Dubitzky, M. Granzow, eds.) Kluwer: Norwell, MA, 2003. pp. 91-109. LANL LA-UR.

[2] Gonzalez, R., H. Tao, J. E. Purvis, K.T. Shanmugam, S.W. York, and L.O. Ingram. (2003). Gene Array-Based Identification of Changes that Contribute to Ethanol Tolerance in Ethanologenic Escherichia coli: Comparison of KO11 (Parent) to LYO1 (Resistant Mutant). Biotechnol. Prog. 19 (2): 612-623.