**Molecular Classification of Cancer by Gene Expression Monitoring**

Namrata Ray, Jugal Marfatia, Md Muhtasim Billah

**Abstract**

Although cancer classification has improved over the last few decades, yet there are only a few approaches reported in literature for detecting new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). For this project, one of those existing studies was taken as inspiration where a generic approach to cancer classification, based on gene expression monitoring by DNA microarrays, was described and applied to human acute leukemias as a test case. In that work, a suggested class discovery technique was able to spontaneously distinguish between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) without previous knowledge of these classes. For this project, as the classification strategy, principal component analysis and clustering has been performed which was successfully able to not only classify the different cancer cases (ALL or AML) but also demonstrated the explained variance in terms of the number of principal components. The derived results, in line with the literature of interest, demonstrate the feasibility of cancer classification based solely on gene expression monitoring and propose a general tactic for predicting cancer classes for other types of cancer, irrespective of any previous information.

Keywords: gene expression, PCA, clustering, DNA microarrays, leukemia.

**Introduction**

To distinguish among various pathogenetically different tumor types and targeting specific therapies to them has been one of the major challenges for cancer treatment. Improvement in cancer classification has thus been vital to achieve progresses in cancer therapeutics. Cancer classification, which was primarily based on morphological appearance of the tumor, has some limitations as tumors with similar histopathological manifestation can take considerably different clinical pathways and display dissimilar responses to therapy. Another difficulty arises for cancer classification because of its reliance on particular biological traits, instead of systematic and unbiased approaches for identifying tumor subtypes.

In the literature, based on which this project has been carried out, such an approach has been described depending on global gene expression analysis. They divided cancer classification into two major tasks: class discovery and class prediction. Class discovery refers to defining previously unrecognized tumor subtypes and class prediction refers to the assignment of particular tumor samples to already-defined classes. Human acute leukemias was chosen as a test case for this. Although the distinction between AML and ALL had been well established already, no single test was sufficient to establish the diagnosis till then. Rather, the clinical practice involved an experienced hemato-pathologist’s interpretation of the tumor’s morphology, histochemistry, immune-phenotyping, and cytogenetic analysis, each performed in a separate, highly specialized laboratory. Although usually accurate, leukemia classification remained imperfect and errors did occur. But distinguishing ALL from AML is critical for successful treatment.

Though microarray studies have primarily been descriptive rather than analytical, it has been suggested (*10*) that such microarrays could provide a tool for cancer classification. In the literature being discussed, a systematic method was developed to tackle cancer classification based on the simultaneous expression monitoring of thousands of genes using DNA microarrays (*9*). They began with class prediction: How could one use an initial collection of samples belonging to known classes (such as AML and ALL) to create a “class predictor” to classify new, unknown samples? They developed an analytical method and first tested it on distinctions that are easily made at the morphological level, and then turned to the more challenging problem of distinguishing acute leukemias, whose appearance is highly similar.

For the project, the principal component analysis (PCA) and clustering has been used to classify ALL and AML cancer types using the data provided by Golub *et al*. Explained variance in terms of the principal components has also been reported. The statistical analytical model was able to successfully distinguish between the cancer types which remained as the sole purpose of this project.

**Data Set**

The dataset for this project comes from a proof-of-concept study published in 1999 by Golub *et al*. It showed how new cases of cancer could be classified by gene expression monitoring (via DNA microarray) and thereby provided a general approach for identifying new cancer classes and assigning tumors to known classes. These data were used to classify patients with acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL).

The initial leukemia data set consisted of 38 bone marrow samples (27 ALL, 11 AML) obtained from acute leukemia patients at the time of diagnosis (13). RNA prepared from bone marrow mononuclear cells was hybridized to high-density oligonucleotide microarrays, produced by Affymetrix and containing probes for 6817 human genes (14). For each gene, a quantitative expression level was obtained. Samples were subjected to a priori quality control standards regarding the amount of labeled RNA and the quality of the scanned microarray image (15). Intensity values have been rescaled such that overall intensities for each chip are equivalent. Also, there exists an independent dataset (for test, 34 samples) used in the paper.

**Methods**

The first issue was to explore whether there were genes whose expression pattern was strongly correlated with the class distinction to be predicted. The 6817 genes were sorted by their degree of correlation (16). To establish whether the observed correlations were stronger than would be expected by chance, we developed a method called “neighborhood analysis”

The second issue was how to use a collection of known samples to create a “class predictor” capable of assigning a new sample to one of two classes.

Step 1: Standardize 7123 genes of 72 individuals in the study.

Step 2: Reduce the dimensionality by conducting PCA (n\_components = 30).

Step 3: Visually see the variation explained by each component.

Step 4: Conducted Clustering using kmeans (n\_clusters = 3)

Step 5: Visually check the proportion of people with ALL vs AML in each cluster.

Step 6: Conclude that clustering via PCA can help classify cancer types.

**Results**

**Conclusion**