Predicting Life Expectancy of Acute Myeloid Leukemia (AML) Patients Based on Gene Expression of Cancer Cells

Max Drach

Abstract— **Doctors must often decide whether to offer a cancer patient a risky treatment or let the patient continue to live with cancer. Trialing a variety of dimension-reduction techniques, we determine one that best clusters cancer patients based on survival data. Then we predict the life expectancy of a cancer patient using a Kaplan-Meier estimator on reduced-dimension data. PCA was the most effective dimension-reducing algorithm based on our data visualization metric, and was used to show that reduced data can be used to predict the survival rate of cancer patients.**

I. INTRODUCTION

A. *Weighing the Risks of Cancer Treatment*

Acute Myeloid Leukemia (AML) kills over 10,000 people in the United States each year [1]. However, many of these deaths are not due to the cancer itself, but to complications that occur during treatment. Cancer treatments themselves often pose significant risks. For example, bone marrow transplants, which are common last-resort treatment options for AML patients, have a survival rate of less than 30% [2]. Thus, doctors are often faced with the difficult decision to either give a patient a risky, painful treatment, or to let the patient survive as long as possible without the treatment.

No two cases of AML are alike, and specifics of a patient's particular instance of cancer can indicate how long the patient will survive. In particular, cancer cell gene expression frequencies indicate specific characteristics within a cell. Gene expression frequency data, coupled with patient survival data, can estimate a particular cancer patient's odds of survival.

B. *Project Goals*

This project had two primary objectives. Our first objective was to determine a reliable way to reduce our high-dimensional data set while preserving meaningful structure. Genes often interact in complex ways; many genes solely exist to regulate the frequencies of others. To determine the diemnsion-reduction algorithm that created the best low-dimension visualization of our data, we applied a variety of both linear and non-linear dimension-reducing algorithms to find one that best grouped patients based on chance of survival, best maintaining the variances within the data relevant to survival diagnosis.

Our second objective was to predict a patient's chance of survival based on the patient's gene frequency data. While it is difficult to appropriately divide high-dimensional data into groups based on survival probabilities, we used our reduced-dimension data that best captured relevant cell variances to achieve high prediction accuracy. These predictions would provide a practical use to our dimension-reduction analysis, demonstrating a technique that could aid patients with a wide range of diseases beyond AML.

II. DATA

A. *Procurement*

DNA microarray technology is increasingly used to produce gene expression frequency profiles of patients. These microarrays contain hundreds of short DNA segments attached to small regions upon a slide. To determine gene frequencies from a sample, large quantities of DNA or RNA from the sample cell are tagged with fluorescent molecules, then injected onto the slide. DNA or RNA sequences that match those attached to the slide bind to these molecules, and those that do not are washed off. When exposed to the correct wavelength of light, the molecules bound to the slide glow. The intensity of the hue of a particular region of the slide indicates the relative frequency of the gene in the original sample. Figure 1 from [5] shows this process.

B. *Gene Frequency Data Structure*

Our gene expression frequency data came from a collection established by researchers from the Stanford Center for Cancer Systems Biology. The data is available at precog.stanford.edu.

Our visualization data, selected from the Precog collection, was originally procured in 1999 by the AML

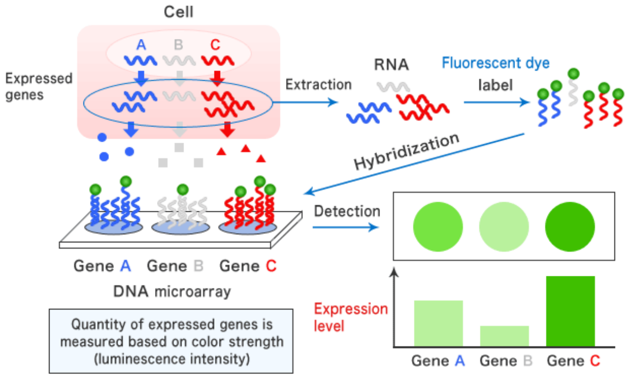

Fig. 1. DNA Microarray

Cooperative Group [3]. Data was collected from cancer cells of 159 patients with AML and contained expression frequencies for 11,979 genes. Our gene expression data was in the form of a matrix containing positive numbers indicating the frequency of a particular gene – row m, column n corresponded to gene m, patient n. We had a separate matrix containing survival data for each patient. This matrix contained a row containing an arbitrary number between 1 and 40, which indicated the time that the patient's status was checked, and a 1 or 0 indicating whether the patient was dead or alive, respectively [3].

C. *Feature Selection*

The types of genes in our frequency data set varied widely. Genes ranged from cell membrane protein sequences to chromosomal reading frames. Because these genes were from a random sampling, only a few genes from this random selection were likely to actually be responsible for decreased life expectancy of a patient.

To determine which genes were most relevant to predicting patient survival rates, we used a Cox Regression model to relate change in gene frequency to the length of time before the individual either passed away or was verified to be alive. We then sorted the genes with the greatest absolute z-score, which specifies how strongly the change in a particular gene correlates to a change in patient life expectancy [6]. We used the 50 genes with the greatest absolute z-score values as the dimensions of each individual.

We also reduced the data to have 0 mean and variance of 1 for our dimension-reduction algorithms that required normalized data.

III. Dimension-Reducing Methods

We reduced our data with 50 selected features using a variety of linear and nonlinear methods.

A. *PCA*

PCA is frequent first choice for dimension reduction. However, it requires that data be linearly correlated. Although a linear algorithm like PCA may not be able to capture non-linear relationships between data dimensions, we apply PCA as a first step.

B. *LLE*

Locally Linear Embedding is a clustering algorithm. It takes into account the relationships of data in multiple dimensions, preserving relative distances between these points in the lower-dimensional space. Since multiple gene frequencies of those that we selected may correspond with increased probability of death, a clustering algorithm may be appropriate to ensure that reduced data maintains the relative distances of influential dimensions.

C. *T-SNE*

Similar to LLE, T-SNE relies on the proximity of data points in the high-dimensional space. However, it models distances with a probability distribution that tends to emphasize local clusters. Again, because dimensions of our data could be related be related, T-SNE is a reasonable choice.

D. *Diffusion Maps*

Diffusion maps is an algorithm group points based on their connectivity – the probability of moving from one point to another during a random walk. Diffusion maps have been used for a wide variety of applications, from speaker identification to image compression [9]. This algorithm has the potential to to capture structure in the data that other algorithms cannot.

E. *Determining Visualization Quality*

To determine effectiveness of our dimension-reducing algorithms, we again used Cox Regression to relate each of the reduced dimensions to the survival time of each patient. Since our data gives the alive-dead status of each patient after an arbitrary time period, and this period differs for each patient, so we cannot simply gauge algorithm effectiveness by seeing if our algorithm correctly classifies patients as alive or dead.

Thus, we use the resulting absolute value of the Cox z-value on each dimension as a confidence metric for how well our algorithm preserved desired structure related to patient life expectancy in our data.

IV. Visualization Results

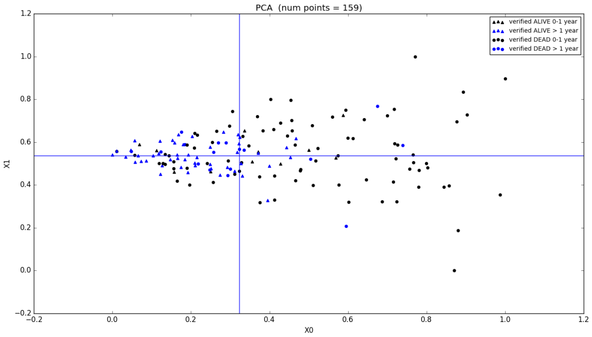

Fig. 2. PCA — z-index of x-axis=7.18

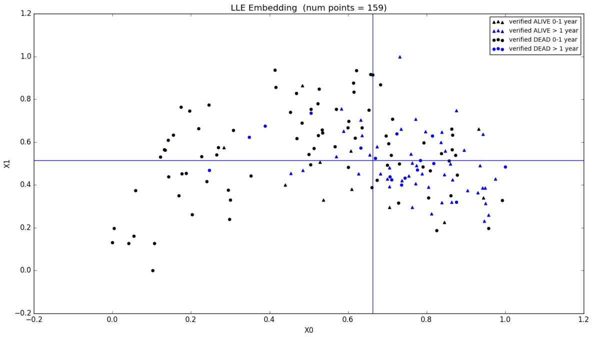

Fig. 3. LLE — z-index of x-axis=-6.91

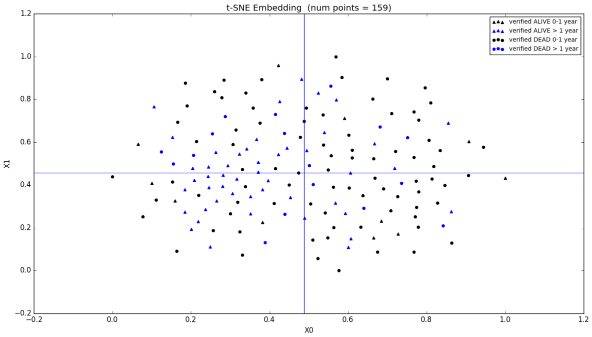

Fig. 4. T-SNE — z-index of x-axis=3.29

The black circles represent patients verified dead between 1-2 years and the blue triangles correspond to patients found to be alive after 2 years. The vertical and horizontal lines mark the median points of the data for either dimension. Only the horizontal dimension Cox z-indexes are displayed.

A. PCA

Of these three methods, PCA performed the best out of all of these algorithms with a z-index value of 7.18. Since this linear algorithm performed so well, capturing nonlinear interactions between the different gene dimensions does not appear to be important in order to model the gene's impact on survival. However, the way that we chose our features may also have favored PCA. We chose features that we were most confident had an impact on the lifespans of our patients – this may have favored data with high-variance, which is often preserved well with a linear algorithm such as PCA.

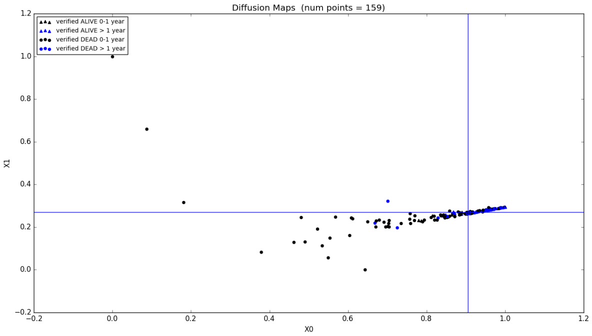

Fig. 5. Diffusion Maps — z-index of x-axis=-7.08

B. LLE

Locally Linear Embedding performed pretty well with an absolute z-index of 6.91, but not nearly as well as PCA. Since LLE is a clustering algorithm, its lower performance suggests that the proximity of points across multiple dimensions does not correlate that strongly with change in life expectancy.

C. T-SNE

T-SNE did not perform well, with a z-index of 3.29. This also suggests that relationships between data dimensions is not that important for predicting life expectancy. Maintaining local groups from the high-dimensional space does not translate to groups with similar survival rates in a low-dimensional space.

D. Diffusion Maps

Diffusion maps acquired the high z-index of 7.08, but did not represent the data in a visually useful way. Diffusion maps was strongly affected by points that were particularly far from the mean of the data, spreading these few points across the graph while clustering others too close together to provide useful intuition. Diffusion Maps does appear to divide the data appropriately, but is too sensitive to outliers to usefully represent this kind of genetic frequency data.

V. Gene-Prediction Methods

Gene expression data mapped to a low-dimensional space can help us predict how long a patient will live. From our whole data set, we can build a Kaplan-Meier plot to determine how long a patient with AML is expected to live. Then, by grouping data based on low-dimension location, we can provide estimates for how long a patient will live. After training a parametric algorithm like PCA on one data set, we can use our learned PCA parameters to map a patient into a portion of a low-dimensional space correlated with a specific specific survival curve.

A. Determining Accurate Survival Probability

The Kaplan-Meier survival estimator creates a survival curve based on survival data in which the status of any given individual is only known at an arbitrary time after a certain period. The function takes the following form [7]:

$$S(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

At any time t, the probability that any individual is still alive is given by the the difference of the number of alive and the number dead $(n_i - d_i)$ divided by the number alive (n_i), multiplied at all intervals from $0-t$.

For our full set of training data, our Kaplan-Meier plot is given by Figure 6. At 10 months, only 50% of patients are expected to still be alive, making a bone marrow transplant appear to be worth the potential 30% risk.

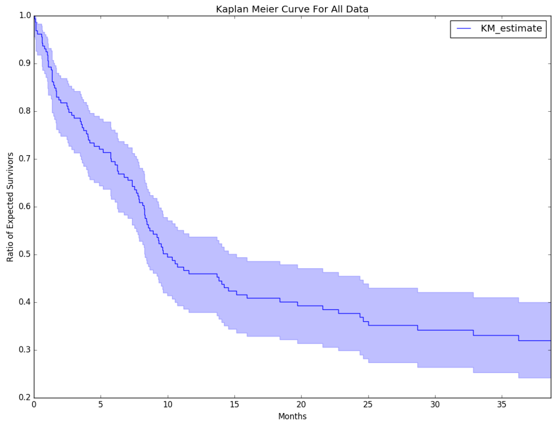

Fig. 6. Kaplan-Meier — All Training Data

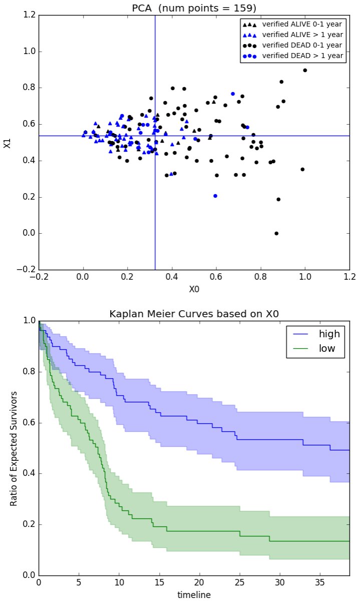

Fig. 7. PCA Train Data — Divided by X0 dimension median

B. Data Sets

We used the same set from the visualization portion of this paper for our training set [3]. Our test data set came from a study that procured data from microarrays in Santa Clara, California, which was also studying gene expression of AML patients [4].

C. Dividing Data And Predicting

To predict the survival rate of a patient, we first reduced the dimensions of our training data with PCA. We split data along the dimension corresponding to the largest Cox z-index, using the median of these points projected on the x-axis. We can then create a separate Kaplan-Meier plot for each of these two groups of our training data to see how well we separated the data as in Figure 7.

Now can use using the PCA parameters basd on our training set to reduce data from our test set. Once reduced, we divide our test set along the same dimension using the same median values from our training set. We map the points in each of these sets to the appropriate curve that we created using our training data.

VI. Gene-Prediction Results

As is clear in Figure 7, dividing the training data across the median of the reduced dimension divides the our cancer patients into very distinct groups. 80% of the patients in the upper group are alive after 10 months, a significant increase from the 50% in our original graph, Figure 6. In the bottom group, 30% of patients are alive after 10 months. This gene expression data divides cancer patients into groups with drastically different survival ratings, highly useful information for doctors deciding whether to recommend a risky treatment.

To determine the accuracy of our prediction, we compared the high/low patient divide determined by our training PCA parameters median divider to the high/low divide from when we reduced our test set with a PCA and median determined from the test data itself. The groupings of patients were fairly consistent, with only 5% of patients grouped differently in the training data set than in the test data set.

To determine whether the two Kaplan-Meier curves determined by our training accurately mirrored those of our test data set, we plotted the test group's own Kaplan-Meier curves produced by our training data set. Unfortunately, the curves were quite different, as can be seen by comparing Figure 7 with Figure 8. This is likely due to differences in the time and location of the two studies from which we procured our data.

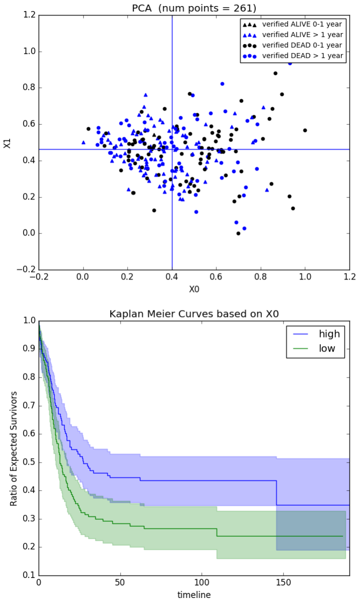

Fig. 8. PCA Test Data Mapped Using Own PCA Parameters — Divided by X0 dimension median

VII. Conclusion

Dimension reduction techniques can help researchers make sense of high-dimensional gene expression frequency data. PCA, although a linear algorithm, is highly effective at preserving features in genetic data related to chance of survival. Data partitioning on reduced gene expression data, with additional research, may accurately predict the survival rate of cancer patients and can help doctors make difficult treatment decisions.

VIII. Acknowledgments

We would like to thank Andrew Gentles (Stanford University) for his support and for providing the data for this project.

References

[1] ”Leukemia – Acute Myeloid.” American Cancer Society. 6 Jun. 2016.

[2] Bunim, Juliana. ”Survival Rates for Pediatric Bone Marrow Transplants One of the Top in Nation.” University of San Francisco, 12 Jan. 2012. 6 Jun. 2016.

[3] Metzeler, K.H. et al. ”An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia.” (2008) US National Library of Medicine. Pubmed. Web. Jun 6 2016.

[4] Wouters, BJ. et al. ”Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome.” US National Library of Medicine (2009) Pubmed. Web. Jun 6 2016.

[5] ”Outline of detection method of genes by DNA microarrays.” Toray. Jun 26 2016.

[6] Cox, John and Fischer Black. ”Valuing Corporate Securities: Some Effects of Bond Indenture Provisions.” The Journal of Finance. 31.2 (1976). Web. Jun 26 2016.

[7] Kaplan, E.L. and Paul Meier. ”Nonparametric Estimation from INcompelte Observations.” Journal of the American Statistical Association. 53.282 (2008). Web. Jun 26 2016.

[8] Gentles, Andrew J. et al. ”The Prognostic Landscape of Genes and Infiltrating Immune Cells Across Human Cancers.” Nature Medicine. 21 (2015): 938-945. Web. Jun 6 2016.

[9] de la Porte, J. et al. ”An Introduction to Diffusion Maps.” University of Stellenbosch, South Africa. (2008). Web. Jun 6 2016.