Proceedings of the 26th Annual International Conference of the IEEE EMBS
San Francisco, CA, USA • September 1-5, 2004

A DNA-based Pattern Recognition Technique for Cancer Detection

David Peterson and Charles H. Lee
Department of Mathematics
California State University, Fullerton

Abstract—The proper orthogonal decomposition (POD) technique (also known as the Karhunen-Loève transform) has been used as a model reduction tool for many applications in engineering and science. In principle, one begins with an ensemble of data, called *snapshots*, collected from an experiment or laboratory results. The POD technique is then used to produce a set of basis elements that can span the original snapshot collection using the fewest possible degrees of freedom. It is such capability that allows us to extract the representative characteristics of a cancer from a collection of DNA microarray samples known to be cancerous. The resulting few POD elements can be regarded as dominant cancerous patterns, which can be used to determine whether an arbitrary DNA microarray sample is cancerous. In our study, we consider two types of cancers, liver and bladder. DNA microarray data are downloaded from the Stanford Microarray Database. Our findings indicate that the POD method can successfully detect both cancer types, although our approach can be applied to other types of disease or cancer.

TABLE OF CONTENTS

1. INTRODUCTION

Our objective in this paper is to use a pattern recognition technique on DNA data expressed in microarrays to detect cancer. We start with a collection of DNA microarray data of individuals who suffer a specific type of cancer. Each DNA microarray contains the expressions of thousands of individual genes on a single surface that is about the size of a microscope slide. Such image allows one to see genes that are induced or repressed in an experiment. Thus signatures of a cancer may be encrypted in the DNA microarrays, and once found, can be used for detection. To extract such representative patterns out of an ensemble of cancerous samples, we employ the *proper orthogonal decomposition* (POD) method. The detail of the POD method can be found in Section 2. The DNA microarray data we use in this paper are from the liver cancer [1] and the gastric cancer [2] studies. Both sets are downloaded from the Stanford Microarray Database, genome-www5.stanford.edu. In each case, we have both the cancerous (100+) and normal (70+) samples. Only 85% of the cancerous samples are randomly selected for the POD method use, the rest and the normal samples are reserved for identifying and detecting purposes. The study is repeated 100 times with a different set of 85% cancerous samples each time. Our results indicate that we can positively identify the cancerous samples at all times. The details of our studies are described in Section 3. It is noteworthy to mention that although our study focuses on liver and bladder cancers, the method is not necessarily restricted to these types of diseases.

2. MATHEMATICAL FORMULATION FOR POD

The POD method has received much attention in recent years as a tool to reduce the complexity and dimensions of dynamical models in engineering and science [3]-[5]. In principle, one begins with an ensemble of data $\{V_i(\vec{x})\}_{i=1}^{n_s}$, called *snapshots*, collected from an experiment or laboratory results. The POD technique is then used to produce a basis $\{\Phi_i(\vec{x})\}_{i=1}^{n_s}$ whose first few elements contain all the dominant features of the entire snapshots collection. In other words, the primary component Φ_1 captures *most* of the essential features of the original ensemble, while subsequent basis elements capture more of the smaller and finer variability between the snapshots. As a result, we wish to choose the primary component Φ_1 such that the quantity

$$\sum_{i=1}^{n_s} |\langle V_i, \Phi_1 \rangle|^2 \tag{1}$$

is as large as possible with $\langle \cdot, \cdot \rangle$ denotes the inner product. By assuming Φ_1 is a linear combination of the snapshots,

$$\Phi_1(\vec{x}) = \sum_{j=1}^{n_s} w_j V_j(\vec{x}), \tag{2}$$

where $\vec{w} = [\, w_1 \quad w_2 \quad \cdots \quad w_{n_s} \,]^T$ is the weighting vector assigned to the snapshots. Thus maximizing the quantity in (1) is equivalent to maximizing the following

$$\| \theta \vec{w} \|^2 = \vec{w}^T \theta^2 \vec{w}, \tag{3}$$

where θ is the covariance matrix of the snapshots with its (i, j) component, $\theta_{i,j}$, defined by

$$\theta_{i,j} = \langle V_i, V_j \rangle, \quad i = 1, \cdots, n_s, \, j = 1, \cdots, n_s. \tag{4}$$

Note that with distinct snapshots, the covariance matrix θ is symmetric positive definite and thus the weighting vector that maximizes (3) will also maximize

$$J(\vec{w}) = \vec{w}^T \theta \vec{w}, \tag{5}$$

In this process, the weighting vector for the primary component is exactly the dominant eigenvector of θ

corresponding to the largest eigenvalue. Let us denote the eigenpairs of θ by $\left\{\left(\lambda_j, \vec{u}^{(j)}\right)\right\}_{j=1}^{n_s}$ and sort the eigenvalues in decreasing order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n_s} \geq 0$. It follows that

$$\Phi_j(\vec{x}) = \sum_{i=1}^{n_s} u_i^{(j)} V_i(\vec{x}), \text{ and} \quad (6)$$

$$\sum_{i=1}^{n_s} \left|\langle V_i, \Phi_1 \rangle\right|^2 \geq \sum_{i=1}^{n_s} \left|\langle V_i, \Phi_2 \rangle\right|^2 \geq \cdots \geq \sum_{i=1}^{n_s} \left|\langle V_i, \Phi_{n_s} \rangle\right|^2. \quad (7)$$

Once the POD basis elements are found, they can be used for comparison with other images. Projections form a simple way of implementing the comparison. If V is an arbitrary image to be tested, then

$$P_\Phi(V) = \sum_{i=1}^{n_s} \frac{\langle V, \Phi_i \rangle}{\langle \Phi_i, \Phi_i \rangle}, \quad (8)$$

measures of the correlation between V with POD elements. The larger the magnitude of $P_\Phi(V)$ the greater the correlation there is between the image V and the original set of images. Due to the dominance and optimality of the POD basis only a first few elements are needed in the projection (8). In our study, we seek solely Φ_1 as

$$\lambda_1 \approx \sum_{i=1}^{n_s} \lambda_i. \quad (9)$$

3. Case Studies

As an application of the method, we examined DNA microarray data from references [1] and [2]. The data were obtained from the Stanford Microarray Database at *genome-www5.stanford.edu*. This analysis used the log(base 2) of the R/G normalized ratio (mean). Data for each of these references contain normal tissue samples in addition to the samples from tumorous tissue. Genes were only included in the analysis if good data were present in over 80% of the samples. For samples which were missing data for a particular gene, the missing value was imputed with the average of the values for that gene from the other samples. After imputing the missing data, the average value for each gene was removed. If this is not done, the method requires more than just the principal component to distinguish between the normal and cancerous tissue samples.

The principal component was determined using a random selection of the tumorous tissue samples. Projections onto this principal component were performed for all the tumorous samples, as well as all of the normal tissue samples. We then compare the projections for the normal and tumorous samples. If the principal component derived from the tumorous samples is significantly different from that for a normal tissue sample, the normal tissue projections should differ considerably from the tumorous tissue projections.

3.1 Chen Liver Cancer Data

Reference [1] contained data from 76 normal tissue samples and 105 primary liver tumor samples. We find the principal component of the tumorous samples using the POD analysis. The analysis was performed 100 times, each time using a different set of 85 of the tumor samples selected at random. The remaining 20 tumor samples were reserved for testing purposes.

The projections for all of the samples onto the principal components for each case are summarized in Figure 1. In this figure, the horizontal axis is the case number and the vertical axis represents the projections for the samples onto the principal component. Samples 1 through 105 were the tumorous samples whereas samples 106 through 181 were the normal tissue samples. Figure 1 shows that a very large percentage of the normal tissue samples (samples 106 through 181) have negative projections onto the principal component. The tumorous samples (samples 1 through 105) show more variability, but about 75% of them show positive projections.

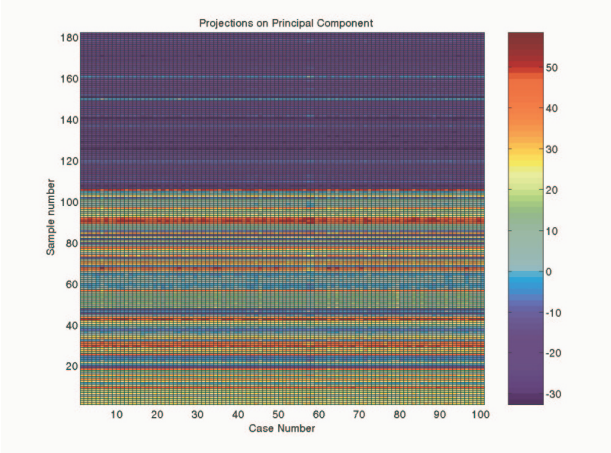

Figure 1 – Projections onto Principal Component – Chen Liver Cancer Data

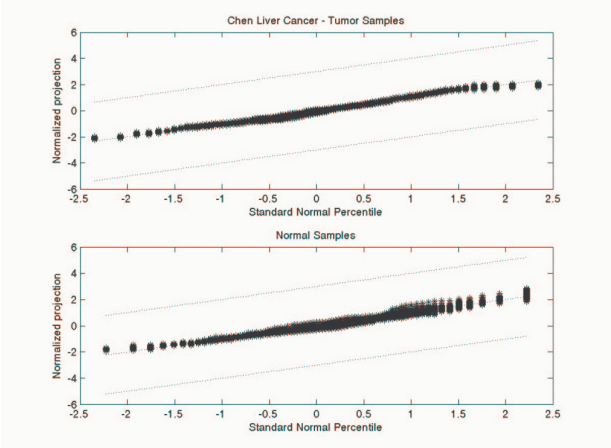

Figure 2 – Percentile Limits vs. Standard Normal Distribution – Chen Liver Cancer Data

We generated statistics for the projections of the tumorous samples to find the sample mean and standard deviation for each of the 100 cases. We then averaged these values to determine an average mean and standard deviation value for tumorous samples. The projections for the tumorous samples tend to be normally distributed. To show this, we 'normalized' the projections by subtracting the mean and dividing the result by the standard deviation. The projections were then sorted into ascending order, and the percentile values were plotted against those from a standard normal distribution. The results are shown in the top plot of Figure 2. If the projections are normally distributed, the percentile values should fall close to the middle line shown on the Figure. The top and bottom line on the Figure show the mean plus and minus three sigma values. The percentile values for the tumorous sample projections line up fairly well with those from the standard normal. Thus, it is a reasonable to assume that these projections are normally distributed.

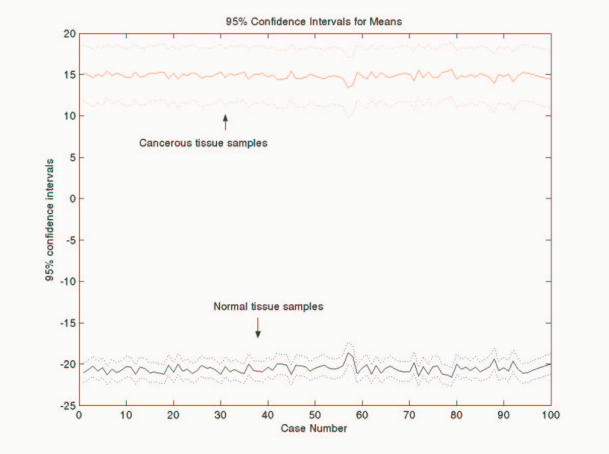

Figure 3 – 95% Confidence Intervals for Mean– Chen Liver Cancer Data

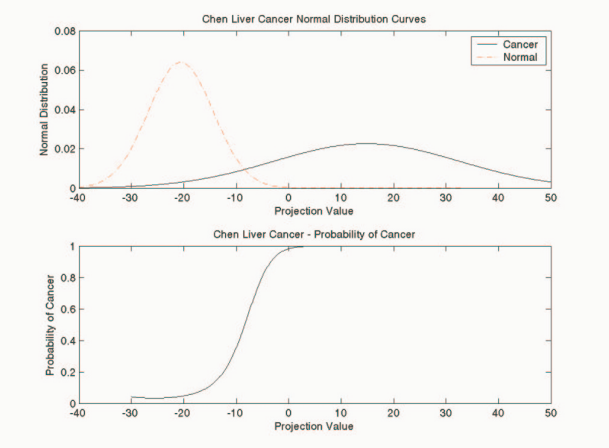

Figure 4 – Normal Density Functions and Probability of Cancer– Chen Liver Cancer Data

Similar statistics were generated for the projections from the normal tissue samples, with the percentile values plotted against those from a standard normal distribution in the bottom plot of Figure 2. From this figure, it also seems reasonable to consider the projections from the normal tissue samples as being normally distributed.

The 95% confidence intervals for the means are shown in Figure 3 as a function of the case number. This figure shows the consistency of the mean values for the projections, regardless of which tumor samples were used in determining the principal component.

The normal probability density functions for the projections are shown in the top plot of Figure 4. This figure shows that if the projection of a sample is positive, the sample is almost certainly tumorous. There is about a 25% probability of a tumorous sample having a negative projection. The probability of cancer as a function of the projection value is shown as the bottom curve of Figure 4. The figure shows that as the projection value becomes more negative, it becomes less likely for the sample to be cancerous.

3.2 Chen Bladder Cancer Data

A similar analysis was performed using the Chen bladder cancer data from Reference [2]. The data used in this analysis consisted of 103 cancerous tissue samples and 21 normal tissue samples. Similar to the analysis in section 3.1, the principal component for the tumorous samples was performed 100 times, each time using 83 randomly selected samples for the principal component analysis. The resulting projections onto the principal components are shown in Figure 5. Examination of the figure shows that there is much more variability for the projections. About 40% of the projections from tumorous samples are negative.

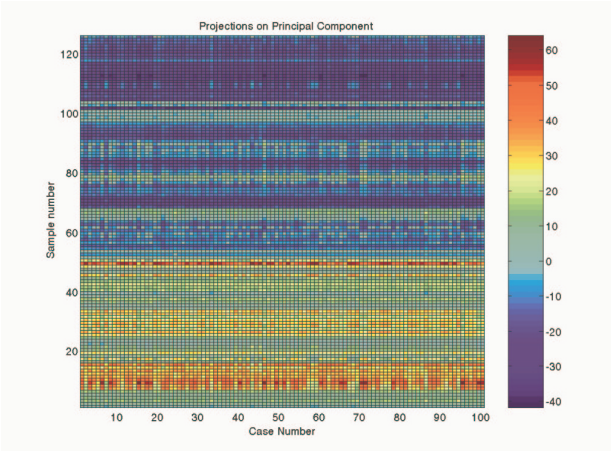

Figure 5 – Projections onto Principal Component – Chen Bladder Cancer Data

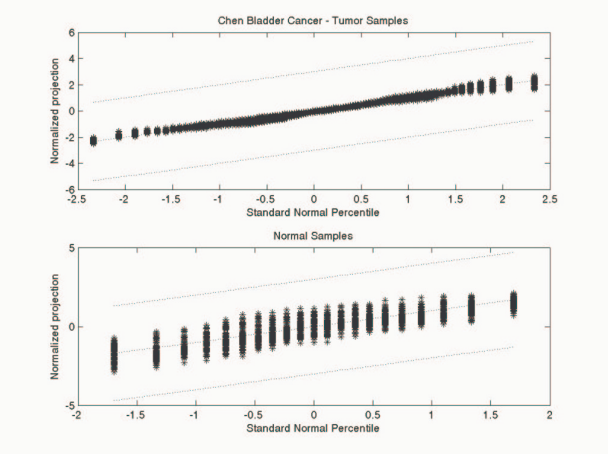

Figure 6 - Percentile Limits vs. Standard Normal Distribution – Chen Bladder Cancer Data

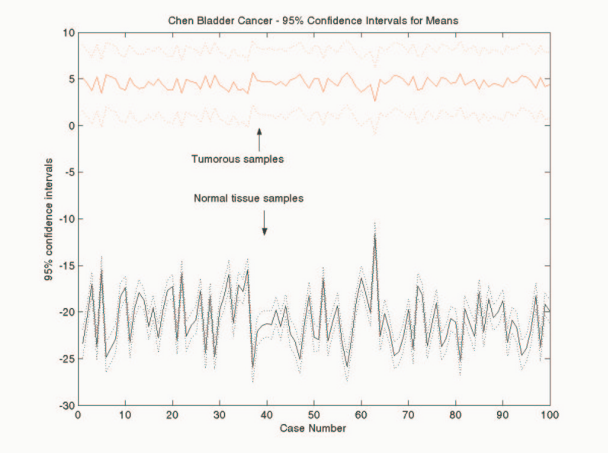

Figure 7 – 95% Confidence Intervals for Means – Chen Bladder Cancer Data

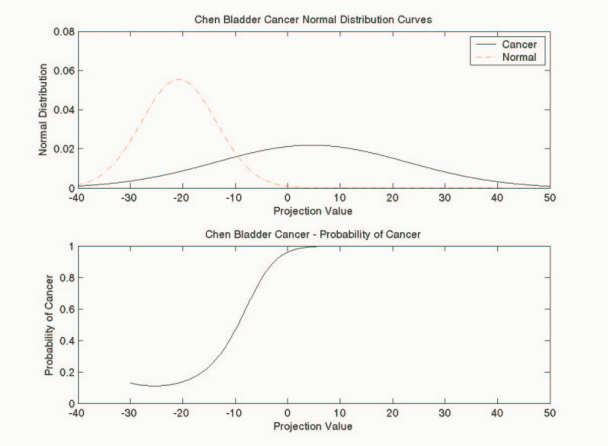

Figure 8 – Normal Density Functions – Chen Bladder Cancer Data

We plotted percentile values against those for a standard normal distribution (Figure 6). Once again, we can see that the normal distribution assumption is not unreasonable, although there is more variability in the normal tissue sample data. 95% Confidence intervals for the mean are shown in Figure 7. This figure shows the sensitivity of the results with regard to which samples were used for the POD process. The normal probability distribution functions are plotted in the top plot of Figure 8, with the probability of a sample being cancerous plotted as a function of the projection value in the bottom plot of Figure 8. Once again, the Figure shows that if a sample has a positive projection, it is almost certainly tumorous. If the projection is negative, however, there is about a 40% probability that the sample is tumorous.

4. Summary and Conclusions

The above study showed an example of how the Proper Orthogonal Decomposition method can be used for a simple pattern recognition application. The principal component of a set of images is found. The magnitude of the projection of an arbitrary image onto the principal component is a measure of the correlation of an arbitrary image with the original set of data.

As a practical application, the process was used to form the principal components for a set of DNA microarray data for tumorous samples. Then projections were made for normal tissue samples, as well as other tumorous samples, against the principal components.

The analysis was performed using data from two different studies. In both cases, positive projections indicate tumorous samples. However, the method is prone to false negatives; in the liver cancer study 25% of the tumorous samples had negative projections, while 40% of the tumorous samples in the bladder cancer study had negative projections.

References

[1] Chen, X., et. al., "*Variation in Gene Expression Patterns in Human Liver Cancers*", Mol Biol Cell. 2002 Jun; 13(6): 1929-39.

[2] Chen, X., et. al., "*Variation in Gene Expression Patterns in Human Gastric Cancers*", Mol Biol Cell. 2003 Aug; 14(8): 3208-15. Epub 2003 Apr 17.

[3] H.V. Ly and H.T. Tran, "*Modeling and Control of Physical Processes using Proper Orthogonal Decomposition*" Computers and Mathematics with Applications, vol. 33 (2001) pp. 223-236.

[4] H.V. Ly and H.T. Tran, *"Proper Orthogonal Decomposition for Flow Calculations and Optimal Control in a Horizontal CVD Reactor,"* Quarterly of Applied Mathematics, Vol. 60 No. 4 (2002) pp. 631-656

[5] C.H. Lee and H.T. Tran, "*Reduced-Order Feedback Control for Thin Film Flows*," Journal of Computational and Applied Mathematics (2004), to appear.