Hweyhsin Chang

CS199

12/15/2018

Principal Component Analysis of Prostate Cancer

Abstract

This Principal Component Analysis was created from the data found in "Prediction of Prostate Cancer cells Based on Principal Component" by A. Ghosh and S. Barman. It was written with the use of genomics applied to disease diagnosis, specifically prostate cancer, as it is the most common type of cancer among adult men. Normal prostate cells and their amino acids were compared to prostate cancer cells using principal component analysis, due to its ability to compare large vectors of data without much loss of information. A relatively lossless statistical analysis tool is important because human amino acid databases consist of long strings. In this paper, principal component analysis is used to generate principal components between cancer and normal prostate cells, and differentiate between the two based on maximum variance between the two. Analysis resulted in negative valued correlations, or uncorrelated data between normal and cancer cells, proving cancerous cells can be differentiated from normal cells using PCA.

Fig. 1: Data describing the correlation between normal and cancerous prostate cells.

			C	A	N	C	E	R	
	Accession no.	AAQ08976.1	AF304370.1	AF338650.1	AF455138.1	AY008445.1	FJ649644.1	NP001035756.1	NP001231873.1
	AF224278.1	-26.0477	-48.4853	-161.8877	-27.7271	-27.8535	-29.0009	-26.0555	-27.8534
N	AF331165.1	-27.7015	-51.6589	-170.885	-29.5669	-28.6867	-30.8476	-27.7092	-29.6863
O	AF462605.1	-26.4612	-49.5525	-165.2082	-28.22	-28.3435	-29.5763	-26.469	-28.3428
R	M15885.1	-27.1346	-50.3156	-167.4919	-28.6605	-28.7839	-29.8534	-27.1425	-28.7852
M	M24543.1	-25.3429	-47.4	-158.5457	-27.0901	-27.2194	-28.4354	-25.3509	-27.2198
A	M24902.1	-26.5532	-49.3816	-163.3887	-28.2532	-28.3769	-29.5725	-26.5608	-28.3768
L	NM005984.3	-26.9842	-50.3914	-166.5609	-28.832	-28.9521	-30.0249	-26.9919	-28.9508
	NM007003.2	-24.8987	-47.0019	-156.3431	-26.5466	-26.6784	27.9601	-24.9068	-26.6785

Process

I attempted to replicate the PCA results from the paper using principal component analysis on Table 3. in the paper, as seen above, which contains the correlation coefficients between 8 normal prostate cells compared to 8 cancerous prostate cells. I copied the data into a comma separated file, loaded it into a pandas dataframe with the accession numbers as column names, as well as added them as labels. I then transformed the data by applying a standard scaler to show the values onto a unit scale, and finally limited the data to 2 dimensions instead of 8. The resulting graph can be viewed below on the left, with the graphs shown in the paper juxtaposed on the right.

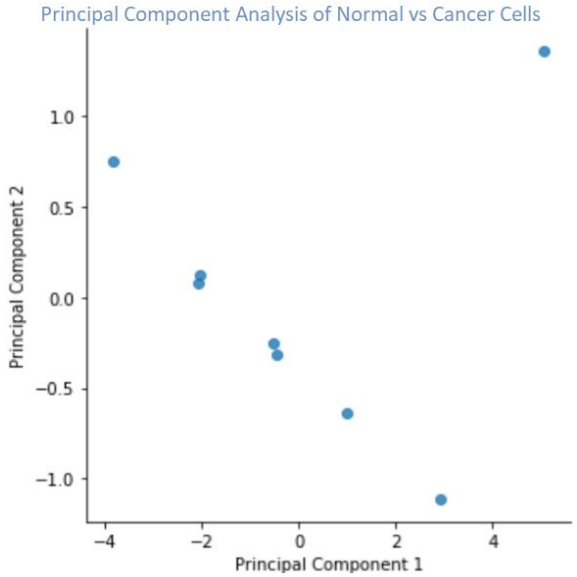

Fig. 2: Generated principal component analysis

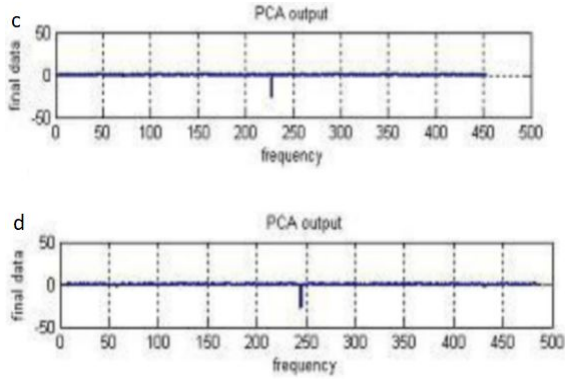

Fig. 3: Given tables:
c. AF224278.1 (normal) vs. AAQ08976.1(cancer)
d. M24543.1(normal) vs. AY008445.1(cancer)

Analysis

While this graph does not resemble the original graphs in the paper, those were generated by comparing the data from the individual data of each normal cell against

the data of each cancer cell. This resulted in multiple graphs generated for each of the combinations, instead of one graph containing all the values. The data presented in the graph correlated to the data reported in the paper, which stated, “when normal prostate cell compared with the cancerous prostate cells, out of 64 combinations, 63 shows negative value, means they are orthogonal or uncorrelated samples,” which accounts for the one outlier on the graph (Ghosh and Barman, 40). I chose this subject because cancer touches virtually everyone, be it themselves, a family member, a friend, or an acquaintance. With the amount of data on the human genome and analysis methods such as PCA, there are huge applications for analyzing and detecting cancer cells. This could lead to accurate early detection of cancer cells in patients, and for preventative screening, for all types of cancer or other diseases.

Conclusion

Principal component analysis can be used to find detect cancerous cells based on a negative correlation between normal and cancerous cells. Graphed, this data shows a strong negative linear correlation between the two sets of cells, with few outlier values, which could be misinterpreted as false-negatives. Given enough data, similar analysis can be performed on other types of cancer to be used in early detection. Coupled with a graph such as the seaborn scatterplot, the data can be presented to a patient as a simple visual to report the probability that the result of a biopsy is or is not cancerous.

Bibliograph

1. “Prediction of Prostate Cancer Cells Based on Principal Component Analysis Technique.” *Procedia Technology*, vol. 10, 28 Dec. 2013, pp. 37–44. *ScienceDirect*, www.sciencedirect.com/science/article/pii/S221201731300488X#