




Pathway-based deep clustering for molecular subtyping of prostate cancer

Ravi Teja V¹, Rema Veeranna Gowda¹, Shreeya Deshpande¹

¹ CISE, University of Florida

 These authors contributed equally to this work.

Abstract

Cancer is a complicated genetic disease that can be categorized into multiple sub types having unique molecular characteristics and clinical features. Categorizing the cancer sub types helps in choosing and building a personalized therapy for the patient as each sub type behaves and responds differently to the treatment. As the availability of data related to cancer is increasing, suitable supervised machine learning algorithms can be applied on the molecular data to identify the sub types that are clinically and genetically unique. Unfortunately several clustering based machine learning models fail to identify the sub type because of high-throughput challenges and non-linearity in the genomics data. In this paper, we are planning to implement a pathway-based deep clustering method (PACL) to categorize molecular sub types of cancer. PACL model performance is compared with several clustering based benchmark methods that have been recently proposed in this research vertical. PACL, in comparison with benchmark methods reported the lowest p-value of the log rank test. PACL interprets the model at biological pathway level and provides a solution to comprehensively identify sub types.

Keywords: Cancer sub typing, Clustering, Pathway-based analysis, Prostate cancer, TCGA

Author summary

We review and present the recent advances in the genomics understanding and advancement of human prostate cancer, with emphasis on molecular sub type classification. Using PACL we classify different sub types of prostate cancer with the goal to develop personalized therapy for each patient.

Introduction

Cancer is one of the complicated disease characterized by undesirable, uncontrolled, and uncoordinated growth of abnormal hostile body cells. Cancer can be categorized into multiple types where each type of cancer can further be classified into multiple distinct types that result in diverse response to the therapy. Cancer sub types usually progress in a single parent cell and have unique gene expression pattern, genetic identity, and protein signaling or gene regulatory network. This behaviour and study of identification of sub types based on their molecular characteristics helps in better understanding of cancer. This further helps in enhancing both diagnosis and prognosis that helps in developing or choosing a personalized therapy for cancer patients.

Over the last decade or so, significant progress has been made in understanding the genomics variations underlying Prostate Cancer (PCa) and its molecular basis. With the help of next generation sequencing, classification of PCa at different levels of molecular information, incorporating data at transcription, epigenetics, genomics and proteomic has become possible. Unique and many molecular sub types have emerged that put PCa from a poorly-understood heterogeneous disease to a disease with a collection of homogeneous molecular sub type with significant understanding.

Several machine learning models have been used to classify and identify known and unknown cancer types. For example, with the help of hierarchical clustering on gene expression data, two sub types of Diffuse Large B-Cell Lymphoma (DLBCL) were detected. K-means clustering was used to detect six sub types of Triple-Negative Breast cancer (TNBC) and five colorectal cancer (CRC) sub types were identified using Enhanced Maximum Block Improvement (eMBI) based on matrix factorization.

Dataset

We used level 3 RNASeq gene expression data of prostate cancer from The Cancer Genome Atlas (TCGA). The dataset consisted of 550 samples of 20531 genes. The pathway genesets were obtained from MiSigDB and we considered only the pathways from 4 databases - Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome, Pathway Interaction Database (PID), and BioCarta. Genes that have no association with pathways were not considered and also small pathways which include less than 15 genes were excluded to avoid redundancy with large pathways. After the preprocessing, the dataset had 1572 pathways of 10550 genes.

Goal and Approach

Kaffenberger et al. [2] identify 7 sub types of prostate cancer based on molecular sub typing. Also it is widely believed that genomics alterations follow Canonical pathways. With this method, we aim to incorporate prior pathways knowledge to find the sub types of prostate cancer. Mallavarapu et al. [1] used a similar approach for clustering Ovarian cancer and GBM sub types using a Restricted Boltzmann Machine (RBM) based architecture/model. We aim to replicate the same for Prostate cancer. In addition to RBM, we intend to use auto-encoder' as another model for clustering. We compare these two approaches with the k-means clustering method. We repeat these methods for multiple cluster sizes and compare the silhouette score to measure clustering performance.

Materials and Methods

To identify unknown cancer sub types from high-dimensional genomics data, we use Pathway-based Deep Clustering model (PACL), Auto-encoder and K-means for this project. K Means is used for comparison purposed while the main methods remain to be RBM and Auto Encoder. We will briefly discuss these methods in this paper. GitHub Repository link: <https://github.com/ShreeyaDeshpande/MLGenomics>

The dataset used for this project is the C2 curated gene dataset. The gene sets in this collection are curated from various sources, including online pathway databases and the biomedical literature. Many sets in this also contributed by individual domain experts. The gene set page for each gene set lists its source. The C2 collection data is divided into two sub-collections: Chemical and genetic perturbations (CGP) and Canonical pathways (CP). The dataset can be found here: https://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/PRAD/20160128/

for the PRAD dataset and <http://www.gsea-msigdb.org/gsea/login.jsp> for the
pathway GMT file Since C2 curated dataset is being used, there is no need for
secondary support evaluation methods.

Pathway-based Deep Clustering model

PACL uses Restricted Boltzmann Machine (RBM) which is an energy-based stochastic
model which has a hidden layer and a visible layer. The hidden units learn non-linear
transformation of the input data in a lower dimensional space whereas The visible units
correspond to input data. The two layers are connected with symmetrical weights, but
there are no intraconnections between nodes in the same layer. Hence, the hidden units,
which are conditionally independent on the visible units, represent posterior
distributions of the variables over the inputs.

K-Means Model

Originally from signal processing, k-means clustering is a method of vector quantization
that aims to partition n observations into k clusters in which each observation belongs
to the cluster with the nearest mean which serves as a prototype of the cluster. As a
result, Voronoi cells are formed in a partitioning of the data space. k-means clustering
minimizes within-cluster variances using squared Euclidean distances, but not regular
Euclidean distances, which would be the more difficult Weber problem: the mean
optimizes squared errors, whereas only the geometric median minimizes Euclidean
distances. For instance, better Euclidean solutions can be found using k-medians and
k-medoids.

Auto Encoder Model

Auto encoder is an unsupervised artificial neural network that learns how to efficiently
compress and encode data then learns how to reconstruct the data back from the
reduced encoded representation to a representation that is as close to the original input
as possible. Auto encoder, by design, reduces data dimensions by learning how to ignore
the noise in the data. Auto encoders consists of 4 main parts namely encoder,
bottleneck, decoder and reconstruction loss.

1. Encoder: In which the model learns how to reduce the input dimensions and
compress the input data into an encoded representation.
2. Bottleneck: which is the layer that contains the compressed representation of the
input data. This is the lowest possible dimensions of the input data.
3. Decoder: In which the model learns how to reconstruct the data from the encoded
representation to be as close to the original input as possible.
4. Reconstruction Loss: This is the method that measures measure how well the
decoder is performing and how close the output is to the original input. The
training then involves using back propagation in order to minimize the network's
reconstruction loss.

Silhouette score metric is used to calculate the goodness of a clustering technique for
the models. The silhouette scores for the models are described below with other results
and figures.

Table 1. Top ten ranked pathways in Prostate cancer.

Pathway name	Reference
Pathway 1	[Ref1]
Pathway 2	-
Pathway 3	-
Pathway 4	[Ref2], [Ref3]
Pathway 5	[Ref4]
Pathway 6	-
Pathway 7	[Ref5]
Pathway 8	-
Pathway 9	[Ref6]
Pathway 10	[Ref7]

Table notes List top 10 pathways found by the model and the corresponding related references from biological literature.

Results

The paper talks about two methods primarily namely RBM AutoEncoders and K-Means. These are unsupervised machine learning algorithms discussed in the paper and hence silhouette scores are calculated to calculate goodness of the clusters formed. The data clusters of AutoEncoders is depicted in Fig. 4. Similarly, the clusters of kmeans are shown in Fig. 3.

Firstly, we determined the optimal number of clusters (i.e., the number of subtypes). Silhouette scores were computed with various cluster numbers (two to ten clusters). Silhouette scores range from $[-1, 1]$, where a high score indicates better clustering performance. For all benchmark clustering methods of K-means, DKM++, HC, SC, CNMF, and CC, the original gene expression data of clusters were used to compute the silhouette score, whereas the last hidden layer node values were considered for PACL that produces high-level representations of the original data.

The silhouette scores on each clustering method are depicted in Fig. 1. Most clustering methods produced the highest silhouette scores with two clusters, which may show that two major subtypes exist in the given TCGA dataset for prostate cancer. It is worth noting that the higher silhouette score of the auto-encoder model than other methods shows that the pathway-based high-level representation of the data describes the nonlinear effects of the data. The silhouette scores are to determine the optimal number of clusters, rather than comparing the performance of the benchmark methods. The average cluster sizes in AutoEncoder were 274.5 and 144.5 with two clusters.

Issues that arose during the project and alternative approaches:

1. How will you be able to answer your original question?
 - We use unsupervised models incorporated with prior pathways knowledge to identify sub types and compare the results with current biological literature (are there 7 sub types etc.)
 - Compare the silhouette score to see if Auto-encoder performs better task at clustering.
 - Assess the hypothesis that clusters (sub types) may be associated to different survivals by log rank tests using additional data (survival)
2. What issues to you expect to arise? What future work?
 - Auto-Encoder may not be as good as standard models (RBM, k-means etc) at incorporating prior knowledge

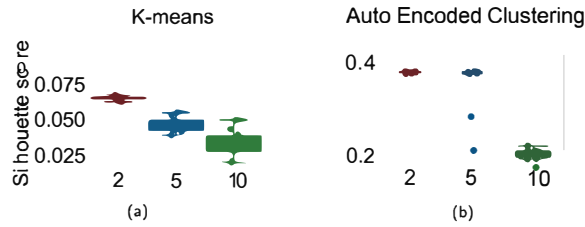


Fig 1. Silhouette scores for K-means and AutoEncoders models used in the paper

Considering the expected issues, autoEncoders worked better than the standard models like k-means which is later discussed in the results

We assessed the hypothesis that clusters (sub types) may be associated to different survivals by log rank tests. Log rank tests were performed with survival times and survival events of clusters.

Metrics Used

The silhouette Coefficient or silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

1: Means clusters are well apart from each other and clearly distinguished.

0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.

-1: Means clusters are assigned in the wrong way.

Model Interpretation

For the model interpretation, we clustered the data into two groups using RBM/Auto-encoder. The survival distributions of the two sub types are analyzed by Kaplan-Meier estimator. One cluster shows a long-term survival group (LTS), whereas another cluster indicates a short term survival group (STS).

Visualization of the nodes in the last hidden layer. The line in red separates the samples of the two clusters.

Experimental Results

Varitational AutoEncoder Loss as epochs progresses.

Top-ranked pathways by t-test between the two clusters are listed in Table 1. The ten top-ranked pathways include pathway 1, pathway 2, Most of these pathways are referred as related pathways in biological literature.

Conclusion

Blank

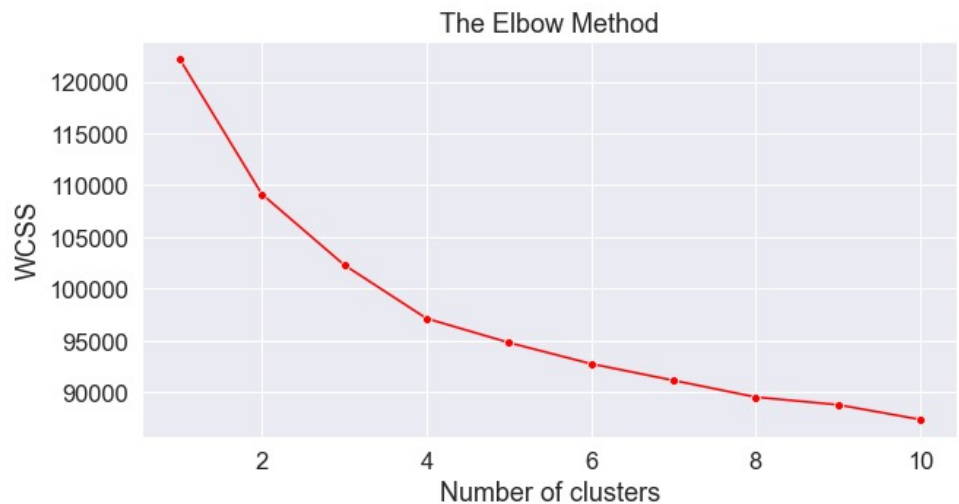


Fig 2. Optimal Cluster Decider

References

1. Mallavarapu T, Hao J, Kim Y, Oh JH, Kang M. Pathway-based deep clustering for molecular subtyping of cancer. *Methods*. 2020;173:24-31. doi:10.1016/j.ymeth.2019.06.017
2. Kaffenberger SD, Barbieri CE. M. Molecular subtyping of prostate cancer. *Curr Opin Urol*. 2016;26(3):213-218. doi:10.1097/MOU.0000000000000285
3. Lemsara, A., Ouadfel, S., Fröhlich, H. PathME: pathway based multi-modal sparse autoencoders for clustering of patient-level multi-omics data. *BMC Bioinformatics* 21, 146 (2020). <https://doi.org/10.1186/s12859-020-3465-2>
4. Chang Z, Wang Z, Ashby C, et al. eMBI: Boosting Gene Expression-based Clustering for Cancer sub types. *Cancer Inform*. 2014;13(Suppl 2):105-112. Published 2014 Oct 21. doi:10.4137/CIN.S13777
5. Chin AJ, Mirzal A, Haron H Spectral clustering on gene expression profile to identify cancer types or subtypes, *Jurnal Teknologi*. (2015).
6. Drier Y, Sheffer M, Domany E, Pathway-based personalized analysis of cancer, *Proc. Nat. Acad. Sci* (2013).
7. Koshiyama M, Matsumura N, Konishi I, Subtypes of ovarian cancer and ovarian cancer screening, *Diagnostics* (2017).
8. Lehmann BDB, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies, *J. Clinical Investigation* 121 (7) (2011) 2750–2767.
9. Mallavarapu T, Hao J, Kim Y, Oh JH, Kang M, PASCL: Pathway-based Sparse Deep Clustering for Identifying Unknown Cancer Subtypes , 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), December 2018, pp. 470–475.
10. Nidheesh N, Abdul Nazeer KA, Ameer PM, An enhanced deterministic K-Means clustering algorithm for cancer subtype prediction from gene expression data, *Comput. Biol. Med* (2017).

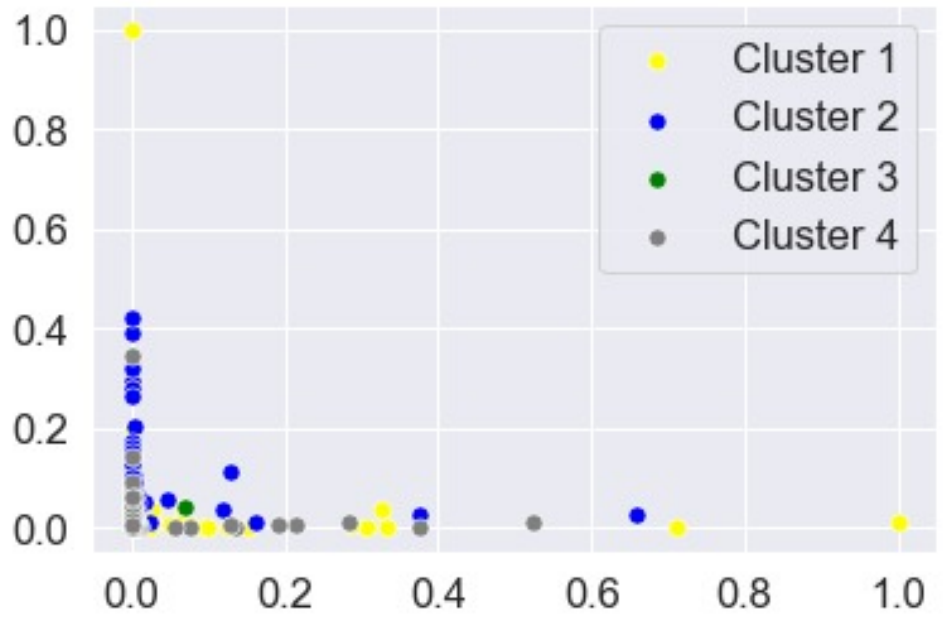


Fig 3. Kmeans Clusters

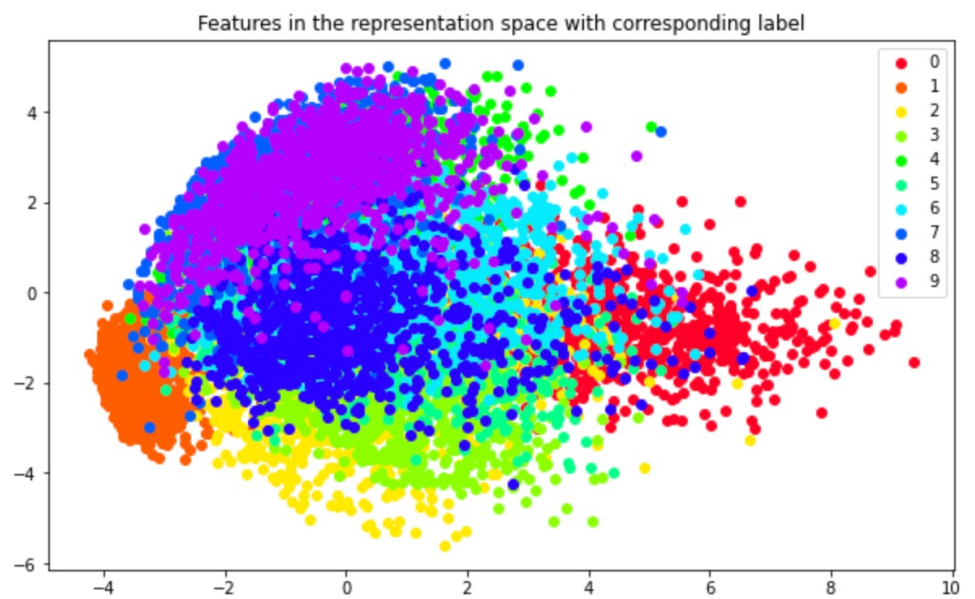


Fig 4. Clustering in space for the auto-encoder method used in the paper

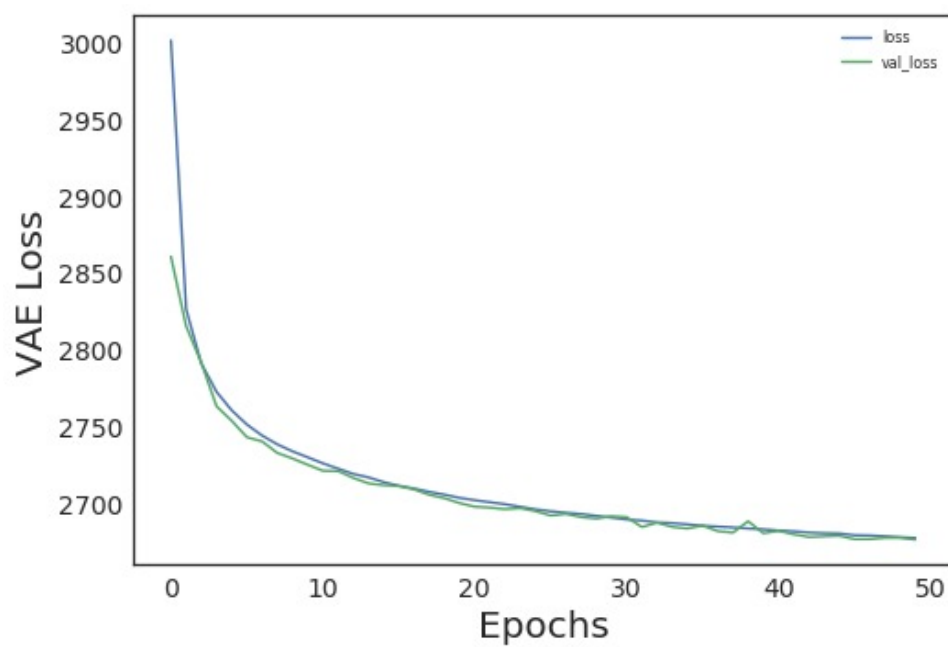


Fig 5. Varitational AutoEncoder Loss as epochs progresses