Detecting Genes Responsible for Cancer Development

1 Problem Statement

Due to the advancement of micro-array and RNA-sequencing technology, it is possible to measure expression profiles of thousands of genes across a set of samples during disease progression, cellular development. Computational analysis of such transcriptome datasets has been found to be useful in understanding mechanisms underlying disease progression and identifying key drivers which trigger disease progression [1, 2, 3]. Although a lot of effort have been made to analyze gene expression data in order to distinguish normal cells from abnormal cells over the last two decades, it is a challenging task due to the high dimensionality and complexity of these data. Moreover, the presence of a large number of genes and typically a small number of samples lead to the phenomena called curse of dimensionality. Hence, there remains a critical need to improve accuracy and identify genes which play instrumental roles during cancer progression.

2 Background Work

Many approaches have been proposed for the classification of cancer cells and healthy cells using gene expression profiles [3, 4, 5, 6]. For instance, the self-organizing map (SOM) was used to analyze leukemia cancer dataset. A support vector machine (SVM) with a dot product kernel has been applied to the diagnosis of ovarian, leukemia, and colon cancer. SVMs with nonlinear kernels (polynomial and Gaussian) were also used for the classification of breast cancer tissues from microarray gene expression data. Due to the large number of genes, high amount of noise in the gene expression data and the complexity of biological networks, there is a need to deeply analyze the transcriptome data for identifying the genes which play key roles during cancer progression. To deal with some of the aforementioned challenges, principal component analysis (PCA) has been proposed for dimensionality reduction of expression profiles. However, PCA reduces the dimensionality of the data linearly and it may not extract some nonlinear relationships in the data, whereas other approaches such as Kernel PCA (KPCA) may be capable of uncovering these nonlinear relationships. K-nearest neighbors (KNN) unsupervised learning also has been applied to breast cancer data. Recently, researchers have applied PCA with a combination of autoencoder to capture non-linear relationships in data. But using a single autoencoder may not extract all the useful representations from the noisy, complex, and high-dimensional expression data. One way to deal with this challenge is reducing the dimensionality incrementally which can be achieved by the multi-layered architecture of an stacked denoising auto-encoder (SDAE)[7] with reduced loss of information.

3 Materials and Methods

3.1 Dataset

RNA-seq expression data can be obtained from The Cancer Genome Atlas (TCGA) database [8] for both healthy and cancer cells.

3.2 Methods

Figure 1 shows the general framework which can be used to analyze transcriptome data.

3.2.1 Data preprocessing: After obtaining the raw dataset, quality assessment can be performed. This step includes removal of the low-quality sequences, exclusion of the poor-quality reads with more than a certain number of unknown bases and trimming the sequencing adapters and primers [6].Once we obtain the preprocessed data, we can get the feature count matrix by mapping to the reference genome.Subsequently, Fragments-Per Kilobase of transcript per Million mapped reads (FPKM) or Reads-Per Kilobase of transcript per Million mapped reads (RPKM) normalization me-

thods can be used to normalize the expression data. Afterwards, Synthetic Minority over-sampling TEchnique (SMOTE) [9] can be used to transform data into a more balanced representation.

3.2.2 Training & Testing: This is the most important step in classification in which of the SDAE [7], a dropout regularization factor can be used in order to randomly exclude fractions of hidden units in the training procedure by setting them to zero. This method prevents nodes overfitting from co-adapting too much and consequently avoids overfitting.

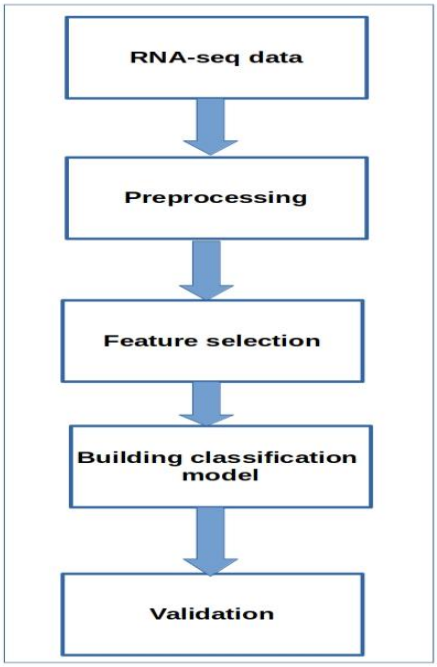

Figure 1: Framework of gene detection system during cancer progression

3.2.3 Evaluation Measures: Accuracy, sensitivity, specificity, precision and F-measure can be used to evaluate the performance of the classifier.

3.3 Experimental Design

3.3.1 Software and Hardware Requirements: For this implementation, GPU can be used with Keras library with Theano backend running on an Nvidia Tesla K80.

References

1.Gov, E et al.. "Differential co-expression analysis reveals a novel prognostic gene module in ovarian cancer". Scientific Reports. 7(1): 4996. 2017.

2. Filteau, M et al. "Gene coexpression networks reveal key drivers of phenotypic divergence in lake whitefish". Molecular Biology and Evaluation. 30(6): 1384-96. 2013.

3. Kim, BJ et al.. "Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method". Proceedings of the National Academy of Sciences of the Inited States of America. 115(6): 1322-1327, 2018.

4. Danaee, P et al.."A deep learning approach for cancer detection and relevant gene identification". Pacific Symposium on Biocomputing. 22: 219-229. 2017.

5. Zhou, Y et al.."An artificial neural network method for combining gene prediction based on equitable weights". Neurocomputing. 71(4-6): 538-543, 2008.

6. Zaraslz, G et al. “A comprehensive simulation study on classification of RNA-Seq data”. PloS One. 12(8): e0182507, 2017.
7. Vincent, P et al.. “Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a local denoising criterion”. Journal of Machine Learning Researc. 11 (2010) 3371-3408. 2010.
8. Saleem, M et al.. “Linked cancer genome atlas database”. I-SEMANTICS '13 - Proceedings of the 9th International Conference on Semantic Systems: 04-06 September 2013 – Graz. 2013.
9. Chawla, NV. “SMOTE: Synthetic Minority Over-sampling Technique”. Journal of Artificial Intelligence Research. 16 (2002), 321 – 357.