doi:10.20944/preprints201901.0068.v1

A framework for the RNA-Seq based classification and prediction of disease

Naiyar Iqbal[1*], Pradeep Kumar[2]

Department of Computer Science and Information Technology[1,2]

Maulana Azad National Urdu University, Hyderabad, Telangana, India

alnaiq.rs@manuu.edu.in[1*], drpkumar1402@gmail.com[2]

Abstract. Disease classification based on biological data is an important area in bioinformatics and biomedical research. It helps the doctors and medical practitioners for the early detection of disease and support them as a computer-aided diagnostic tool for accurate diagnosis, prognosis, and treatment of disease. Earlier Microarray gene expression data have wide application for the classification of disease, but now Next-generation sequencing (NGS) has replaced the Microarray technology. From the last few years, RNA sequence (RNA-Seq) data are widely used for the transcriptomic analysis. Hence, RNA-Seq based classification of disease is in its infancy. In this article, we present a general framework for the classification of disease constructed on RNA-Seq data. This framework will guide the researchers to process RNA-Seq, extract relevant features and apply the appropriate classifier to classify any kind of disease.

Keywords: Disease classification, Read mapping, Feature selection, Machine learning.

1 Introduction

A recent breakthrough in the sequencing technologies, such as Next-Generation Sequencing (NGS) and Oxford Nanopore Technology (ONT), producing a massive amount of sequence data with comparatively better accuracy, at a faster rate and cheaper sequencing cost. At the same time, there is also exponential growth in the application of computational intelligence techniques, such as machine learning, in the field of genomics and medicine. These technologies have made the field of biology a data-rich science, with lots of scope of big data analytics. RNA sequencing (RNA-Seq), a better alternative to Microarray technology [1], enables the researchers to assess the expression degrees of genome-wide transcripts simultaneously. These gene expression profiles are nowadays considered as a rising technique for disease classification, diagnosis, and identification of potential disease biomarkers [2]. Earlier, this information from RNA-Seq data was overlooked, but by utilizing machine learning techniques, various important features can be easily identified within RNA-Seq data.

The early detection and diagnosis of disease, or disease classification, is an important task. Mostly, disease detection models are based on clinical patient data, however, which cannot give an accurate and molecular view of disease. Hence, high-throughput sequencing data, such as RNA-Seq, can be utilized for early and accurate detection of disease, which will not only give a molecular view of disease but will also identify potential disease biomarkers. For this purpose, we need both positive (disease) and negative (non-disease) RNA-Seq sample data with a sufficient number of biological replicas. Although, high dimensionality of RNA-Seq data throws diverse challenges for soft computing techniques, such as the curse of dimensionality problem, to be directly applied. Therefore, many existing classifiers cannot be directly used.

In this paper, we present a framework (i.e., a computational pipeline) for the disease classification using RNA-Seq based genes dataset from NGS platform. The article is arranged their contents as follows. Section 2 explores the review of the literature, section 3 demonstrates the proposed framework, section 4 lists software tools that can be used to implement the presented framework, and at the end section 5 concludes the paper and present forthcoming research directions.

2 Review of Literature

Literature reports that most of the gene expression-based prediction models are based on Microarray. Recently, RNA-Seq based classifications of disease are reported in the literature. Kernel Fisher Discriminant Analysis (KFDA) based on kernel machine has been proposed by Cho et al., (2004) [3] that extract an informative subset of genes form a huge amount of microarray genes and using few selected genes subset are able to classify disease. In their experiment, they use leukemia, breast cancer, and colon cancer three gene expression datasets and also they compare their model with previously developed models. The researchers verified that the proposed model is more accurate and reliable using a few informative genes. Wang et al., (2005) [4] raised the problem of existing gene ranking techniques for selection of useful genes is highly correlated. So they proposed a new hybrid model known as HykGene that works on three stages with the combination of gene ranking and clustering analysis. The first step uses feature filtering algorithm for selection of top-ranked genes, then used hierarchical clustering for generation of dendrogram at the second stage, and finally, for analysis of dendrogram a sweep-line algorithm was applied

Soft computing approach has been applied by Wang et al., (2009) [5] for the prediction of microarray-based cancer. The proposed feature selection based rough set theory which is depended on a degree for finding a few informative genes from a huge set of microarray gene expression dataset. The outcome of the proposed model found a better prediction for the selection of a single gene and a pair of genes. Also, the proposed model is simple that is based on rules via few genes and robust so that it can be able to tune the parameters for other datasets.

Raza (2014) [6] applied clustering methods which are unsupervised learning technique for analysis of microarray cancer data. The researcher used four prominent clustering techniques i.e k-means, hierarchical, density-based and expectation maximization using five distinct types of cancer datasets. In his experiment, the two-tailed t-test is applied after normalization of a huge amount of microarray gene to extract differentially expressed genes between normal and tumor two kinds of samples.

In the current generation, there is a paradigm change from microarray to RNA-Seq based for analysis of genes expression. Microarray and RNA-Seq technologies give complementary for analysis of transcriptome. The longer transcript has great potential for the detection of differential expression by the use of three published RNA-Seq high throughput data [7]. Richard et al., (2010) [8] used human RNA-Seq data for the prediction of alternative isoforms from a segment of DNA or RNA expression levels. They use RT-PCR for validation of significant fraction of the prediction. Flow Difference Matric (FDM), a graph-oriented statistical approach is applied for the identification of discriminative transcription on RNA-Seq data. Singh et al., (2011) [9] have proposed a weighted splice graph representation and a new non-parametric statistical permutation testing on ACT graphs is presented for measure the significance of discriminative transcription among pairs of sample or group of replicas.

Ning et al. (2012) [10] performed a comparative analysis for computation and correlation of MS1 and MS2 label-free mass spectrometry oriented protein on RNA-Seq gene expression data. They performed a combined analysis of two different genomic and proteomic mouse mitochondrial protein dataset and they found that the top three normalized peptide region forces from MS1 RNA-Seq gene expression data shown the best association. A new and robust algorithm Smart-Seq based mRNA-Seq is proposed by Ramsköld et al., (2012) [11] that are relevant to single cell levels. The researchers compare their proposed protocol with existing methods that show improvement of read coverage across transcripts using Smart-Seq. The authors have developed a method GSVA (Gene Set variation analysis) for microarray and RNA-Seq dataset for the analysis of variance of the gene set. The method GSVA is non-parametric and unsupervised which gives an alternative to the traditional technique of externally modeling phenotypes inside enrichment scoring model. Hänzelmann et al., (2013) [12] found that RNA-Seq is easily adopted by the GSVA method than microarray data counterpart.

A computational model, FMLNCSIM (Fuzzy measure-based long non-coding RNA functional similarity), has been developed to find the association of similar disease on the basis of assuming functional similarity of long non-coding RNA (lncRNAs) Chen et al., (2016) [13]. They found improvement in performance when they integrate previously developed model LRNSLDA (Laplacian Regularized Least Squares for lncRNA-Disease association) with FMLNCSIM. Jabeen et al., (2018b) [14] review modern methods for classification based on RNA-Seq data. They show the RNA analysis workflow pipeline with the use of machine learning techniques, for both shallow and deep learning classification.

3 The Proposed Framework

The RNA sequencing under the next-generation sequencing framework starts with sample preparation, which is one of the most important steps due to the fact that entire data analysis results are dependent on this correctness of samples it. Once the sample is prepared, it goes to sequencing step which produced a massive amount of sequence bases (e.g. A, G, C, T) in the form of short fragmented sequences, called reads, stored in a single file (.SRA or most popularly .FASTQ format) per samples. Sequencing can be done in-house, outsourced to sequencing agencies, or sometimes desired sequencing data are available in publicly available repositories such as NCBI-SRA, NCBI-GEO, etc.

The proposed RNA-Seq classification framework (Fig. 1) will work only if we have its sequences for each class with a sufficient number of samples from each class. It is recommended that we need to have at least three samples or biological replicas from each class. Fig. 1 depicts the framework for RNA-Seq based classification of disease and subsequent sections describe each step of the framework.

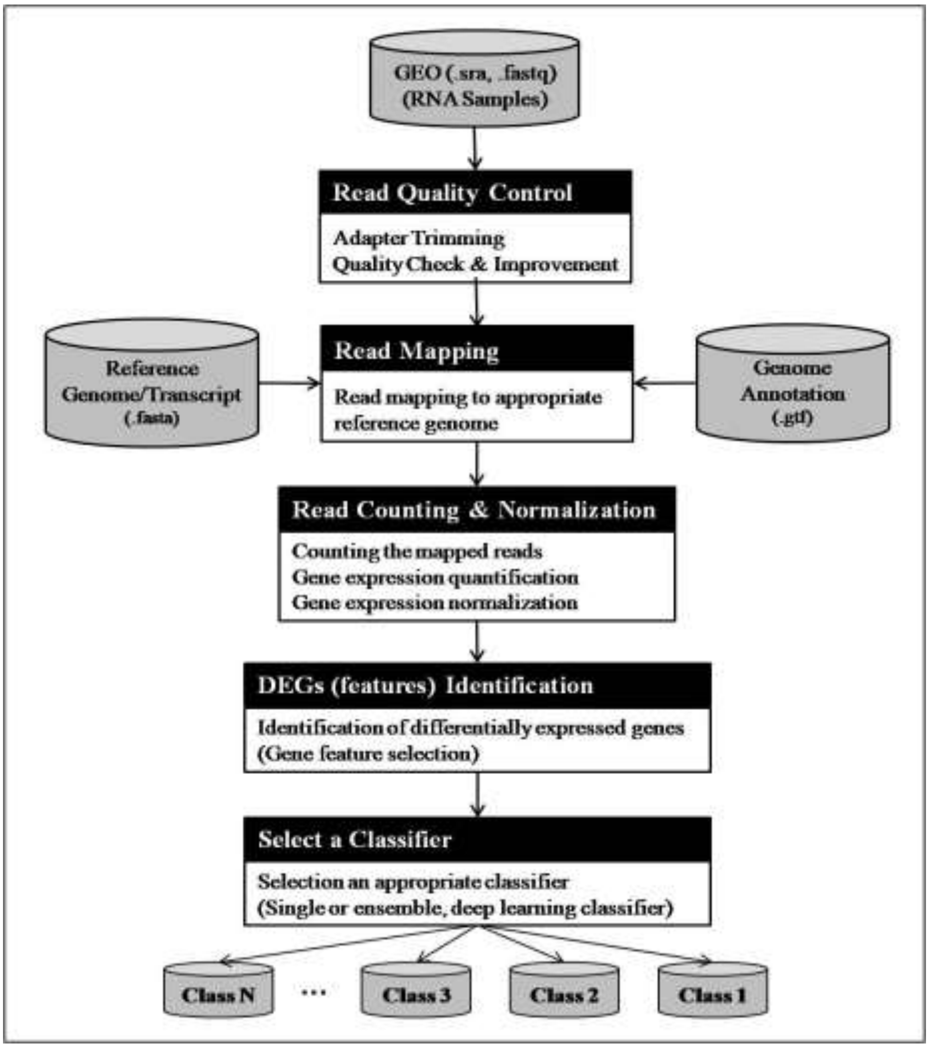
GEO (.sra, .fastq)
(RNA Samples)
Read Quality Control
Adapter Trimming
Quality Check & Improvement
Reference
Genome/Transcript
(.fasta)
Read Mapping
Read mapping to appropriate
reference genome
Genome
Annotation
(.gtf)
Read Counting & Normalization
Counting the mapped reads
Gene expression quantification
Gene expression normalization
DEGs (features) Identification
Identification of differentially expressed genes
(Gene feature selection)
Select a Classifier
Selection an appropriate classifier
(Single or ensemble, deep learning classifier)
Class N
...
Class 3
Class 2
Class 1

Fig. 1. The proposed framework for disease classification

3.1 Quality control

As quality score (Q-score) is attached to each base of the reads in the samples, therefore, we need to ensure the overall quality of reads and remove/improve the bad/corrupted reads if it's Q-score<20 [15]. For read quality assessment, FastQC tool [20] can be utilized, that assess the read quality on several quality metrics such as Q-score per base, average Q-score, percent of GC contents, and so on.

3.2 Read mapping

Read mapping refers to aligning the sequence reads to a reference genome in order to assess the numbers of reads mapped to a particular region of the genome. Read mapping algorithms are computationally expensive as millions of short sequence reads have to be mapped to the reference genome. Some of the popularly used read mapping algorithms are BWA (Burrows-Wheeler Alignment), Bowtie 2, and Tophat. A list of software tools can be found in [16].

3.3 Read count

We know that more active a gene is transcribed, the more reads we observe from it. This principle is used for gene expression identification in RNA-Seq. In RNA-Seq analysis, mapped read count are treated as gene expression. The read count is biased towards the length of the gene and sequencing depth. Hence, both gene transcript length and sequencing depth are important for read count normalization. Hence, read counts are usually performed normalization to FPKM (Fragments Per Kilobase of transcript per Million mapped reads) before downstream analysis.

3.4 Differentially expressed genes (features) identification

For the RNA-Seq based classification of disease, it is important to identify relevant features before we apply an appropriate machine learning classifiers. In the transcriptomic analysis, the identification of differentially expressed genes (DEGs) is an important task which helps to find out the gene biomarkers or gene responsible for the occurrence of a particular disease [1, 2]. Hence, DEGs can be utilized as a feature vector in order to the classification of RNA-Seq data. Some popular software tools for DEGs identification are edgeR [17], Cuffdiff [18], and DESeq2 [19].

3.5 Select an appropriate classifier

After the selection of features (e.g. DEGs), we need to choose an appropriate machine learning classifier. The selection of a classifier varies according to the complexity of the problem. For the classification of disease based on RNA-Seq normalized read count data, we can either use simple statistical learning such as Decision Tree, Naive Bayes, or sophisticated complex classifiers such as artificial neural networks with a shallow or deep architecture. In some of the cases, ensemble learning models such as Random Forest, or AdaBoost can also perform better.

4 List of available software tools to implement the framework

In order to implement the presented framework, we need a list of software that can be connected as a pipeline. Table 1 presents a list of relevant software with their descriptions for each step of the framework.

Table 1. List of software tools to implement the presented framework

Task	Tools	Descriptions	References
Quality control	FastQC	Checks the quality of the reads	[20]
Read mapping	BWA, Bowtie, TopHat,	Maps the sequence reads to the reference genome.	[16]
Read count	HT-Seq	Counts the number of reads aligned to particular transcripts on the genome	[20]
DEGs identification	edgeR, DESeq, Cuffdiff	Normalize read counts, finds DEGs, and numerous plots	[17,18,19]
Machine learning classifiers	Decision Tree, Naive Bayes, ANN, SVM, Deep neural network, etc.	Training and testing for the RNA-Seq samples classifications based on DEGs as the feature vector	[14]

5 Discussion, conclusion and future work direction

The accurate classification and prediction of disease based on genomic data is an open research problem in the field of biomedical research. Due to high throughput, low-cost, faster and deeper sequencing mechanism Next-generation sequencing (NGS) has replaced the Microarray-based classification and prediction of disease. RNA-Seq data offers better transcriptomic analysis, and hence, it can be used for the disease classification and prediction task. This paper presented a general framework for the classification of disease constructed on RNA-Seq data, which are supposed to guide the researchers to process RNA-Seq, extract relevant features and apply the appropriate classifier to classify any kind of disease.

Despite several advantages of NGS and Oxford Nanopore Technology (ONT), there is a lack of better and accurate algorithms to translate the sequence data to fruitful knowledge that can utilize in diagnosis and treatment. Also, for better analysis, we need both positive (disease) and negative (non-disease) RNA-Seq sample data with a sufficient number of biological replicas, which are still lacking behind. The high dimensionality of RNA-Seq data throws various challenges for machine learning techniques, such as the curse of dimensionality problem, to be directly applied. Hence, several existing classifiers cannot be directly used.

Some of the challenges and future work direction of RNA-Seq classification are: (i) data imbalance problem, (ii) heterogeneous nature and format of data, (iii) performance are affected due to high-density of data, (iv) high computation costs for data processing (read mapping, read counting, and training complex learning models), (v) machine learning needs large volume of data to avoid overfitting, (vi) Most of the machine learning techniques are black-box in nature, where the learning model is not easily interpretable.

Acknowledgment

The authors acknowledge Dr. Khalid Raza, Department of Computer Science, Jamia Millia Islamia for necessary discussion and suggestion on the manuscript.

References

[1] K. Raza, “Analysis of Microarray Data using Artificial Intelligence Based Techniques. In Handbook of Research on Computational Intelligence Applications in Bioinformatics” (pp. 216-239). IGI Global, 2016.

[2] A. Jabeen, N. Ahmad, and K. Raza, “Differential Expression Analysis of ZIKV Infected Human RNA Sequence Reveals Potential Biomarkers”. bioRxiv, 498295, 2018a.

[3] J. H. Cho, D. Lee, J. H. Park, and I. B. Lee, "Gene selection and classification from microarray data using kernel machine”. *FEBS letters*, *571*(1-3), 93-98, 2004.

[4] Y. Wang, F. S. Makedon, J. C. Ford, and J. Pearlman, “HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data”. *Bioinformatics*, *21*(8), 1530-1537, 2004.

[5] X. Wang, and O. Gotoh, “Microarray-based cancer prediction using soft computing approach”. *Cancer informatics*, *7*, CIN-S2655, 2009.

[6] K. Raza, “Clustering analysis of cancerous microarray data”. *Journal of Chemical and Pharmaceutical Research*, *6*(9), 488-493, 2014.

[7] A. Oshlack, and M. J. Wakefield, “Transcript length bias in RNA-seq data confounds systems biology”. *Biology direct*, *4*(1), 14, 2009.

[8] H. Richard, M. H. Schulz, M. Sultan, A. Nurnberger, S. Schrinner, D. Balzereit, ... and S. A. Haas, “Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments”. *Nucleic acids research*, *38*(10), e112-e112, 2010.

[9] D. Singh, C. F. Orellana, Y. Hu, C. D. Jones, Y. Liu, D. Y. Chiang, ... and J. F. Prins, “FDM: a graph-based statistical method to detect differential transcription using RNA-seq data”. *Bioinformatics*, *27*(19), 2633-2640, 2011.

[10] K. Ning, D. Fermin, and A. I. Nesvizhskii, “Comparative analysis of different label-free mass spectrometry-based protein abundance estimates and their correlation with RNA-Seq gene expression data”. *Journal of proteome research*, *11*(4), 2261-2271, 2012.

[11] D. Ramsköld, S. Luo, Y. C. Wang, R. Li, Q. Deng, O. R. Faridani, ... and G. P. Schroth, “Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells”. *Nature Biotechnology*, *30*(8), 777, 2012.

[12] S. Hänzelmann, R. Castelo, and J. Guinney, “GSVA: gene set variation analysis for microarray and RNA-seq data”. *BMC Bioinformatics*, *14*(1), 7, 2013.

[13] X. Chen, Y. A. Huang, X. S. Wang, Z. H. You, and K. C. Chan, “FMLNCSIM: fuzzy measure-based lncRNA functional similarity calculation model”. *Oncotarget*, *7*(29), 45948, 2016.

doi:10.20944/preprints201901.0068.v1

[14] A. Jabeen, N. Ahmad, and K. Raza, "Machine Learning-Based State-of-the-Art Methods for the Classification of RNA-Seq Data". In *Classification in BioApps* (pp. 133-172). Springer, Cham, 2018b.

[15] N. Wani, and K. Raza, "Raw Sequence to Target Gene Prediction: An Integrated Inference Pipeline for ChIP-Seq and RNA-Seq Datasets. In Applications of Artificial Intelligence Techniques in Engineering". Advances in Intelligent Systems and Computing, vol 697. Springer, Singapore, 2019.

[16] K. Raza, and S. Ahmad, "Recent advancement in Next Generation Sequencing techniques and its computational analysis". arXiv preprint arXiv:1606.05254, 2016.

[17] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". *Bioinformatics*, *26*(1), 139-140, 2010.

[18] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter, "Differential analysis of gene regulation at transcript resolution with RNA-seq". *Nature Biotechnology*, *31*(1), 46, 2013.

[19] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2", *Genome Biology*, *15*(12), 550, 2014.

[20] FactQC. A quality control tool for high throughput sequence data. (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)