DEGnet: Identifying Differentially Expressed Genes Using Deep Neural Network from RNA-Seq Datasets

Tulika Kakati[1], Dhruba K. Bhattacharyya[1(✉)], and Jugal K. Kalita[2]

[1] Department of Computer Science and Engineering, Tezpur University, Tezpur 784028, Assam, India
tulika.kakati@gmail.com, dkb@tezu.ernet.in
[2] Department of Computer Science, University of Colorado, Colorado Springs, CO 80918, USA
jkalita@uccs.edu

Abstract. Differential expression (DE) analysis and identification of differentially expressed genes (DEGs) provide insights for discovery of therapeutic drugs and underlying mechanisms of disease. Statistical methods, such as DESeq2, edgeR, and limma-voom produce a number of false positives and false negatives and fail to differentiate between the DEGs as up-regulating (UR) and down-regulating (DR) genes linking them to disease progression. Machine learning (ML) including deep learning (DL) methods to identify DEGs from RNA-seq data face challenges due to smaller sample sizes (n) compared to number of genes (g). In this work, we propose a deep neural network (DNN) called DEGnet to predict the UR and DR genes from Parkinson's disease (PD) and breast cancer (BRCA) RNA-seq datasets. The accuracies we obtained from PD and BRCA were 100% and 87.5% respectively, higher than ML-based methods on the same datasets. However, to the best of our knowledge, we are the first to apply DNN on for classification of DEGs into UR and DR, and identify significant UR and DR genes that play role in progression of a disease. Experimental results show that DEGnet is a good performer and can be applied in other RNA-seq data, despite the $n << g$ issue.

Keywords: Deep neural network · RNA-seq · Parkinson's disease · Breast cancer

1 Introduction

Differential Expression (DE) analysis studies the variance of gene expressions across two cell conditions, such as control (or normal) and disease (or tumor). Genes with varied expressions across cell conditions have been implicated in a number of severe diseases. Therefore, DE analysis and identification of differentially expressed genes (DEGs) may provide insights into underlying mechanisms of disease and even into discovery of therapeutic drugs. Recent advances

B. Deka et al. (Eds.): PReMI 2019, LNCS 11942, pp. 130–138, 2019.
https://doi.org/10.1007/978-3-030-34872-4_15

in technologies such as next-generation sequencing have led to development of large-scale repositories of biological data, including gene expression datasets.

Recently developed statistical methods for DE analysis can be divided into two groups, parametric and non-parametric, depending upon whether the data distribution is considered a parameter. Log2 fold change (log2FC) measures the logarithmic scale in base 2 of the ratio of gene expression change in disease condition to the control condition [1]. A few methods such as, DESeq [2], DESeq2 [3], edgeR [4], and voom [5] compute variance (dispersion) in gene expression values. However, these statistical methods produce a high number of false positives and false negatives due to small biases incorporated in the estimation of dispersion for predicting DEGs from RNA-seq data. Here, we take three common methods, namely DESeq2, edgeR, and limma-voom to compare the effectiveness of our proposed model.

Later, with advances in Big Data and machine-learning (ML), ML-based DE analysis was introduced to identify DEGs [6,7], to learn from existing data and predict variations of gene expression patterns. However, application of deep learning models is a challenge for analysis of gene expression data due to smaller sample sizes (n) compared to number of genes (g), unlike image and other datasets found in usual deep learning application areas [8].

In this work, we propose a model based on deep learning which we call DEGnet to identify DEGs. The deep neural network learns from gene expressions in PD and BRCA datasets measured under control and disease conditions with log2FC change labels - 1 for up-regulation and 0 for down-regulation. The main motivation for this work is that the probability of predicting UR and DR genes using the baseline models, DESeq2, edgeR, and limma-voom from biologically validated test data based on the log2FC estimates is low. This is due to the fact the baseline methods produce high false positive rate and false negative rates due to the small biases incorporated in computing the dispersion across samples of RNA-seq data. We argue that the proposed model is generalized because it is trained on the consensus labels based on log2FC estimates of DESeq2, edgeR, and limma-voom. We also demonstrate that it predicts UR and DR genes from biologically validated test data with higher accuracy than the three baseline models. Further, we apply LR [9], KNC [10], SVM [11], GNB [12], DTC [13], and RFC [14] on PD and BRCA data and evaluate their performance in terms of accuracy, sensitivity, specificity, and precision. We found that DEGnet outperforms these traditional ML-based methods. The UR and DR genes are assessed for biological enrichment (GO enrichment and pathway analysis) using web-based tools in the ToppGene Suite [15].

Section 2 of the paper describes the datasets and the description of the strategy used by DEGnet. Section 3 gives experimental results in terms of statistical and biological validation and comparison of the performance of the proposed method with the DL based models. Finally, we conclude by presenting how the method can be developed further in the last section.

2 Method and Materials

The proposed method, DEGnet, runs on two phases. In the first phase, preprocessing, labelling and splitting of the data are done. The second phase consists of training, fine-tuning, and testing. We use log2FC estimates of statistical models (baseline models) DESeq2, edgeR, and limma-voom and prior knowledge from the literature to label the datasets. The use of log2FC estimates and knowledge of prior gene regulation with a DNN enable the capture of the non-linear patterns from biologically validated gene samples and improve the prediction performance of our model in determining UR and DR genes. Figure 1a gives the workflow of our proposed method. We use two datasets: PD (GSE68719) and BRCA (TCGA) [16] (described in the Dataset subsection).

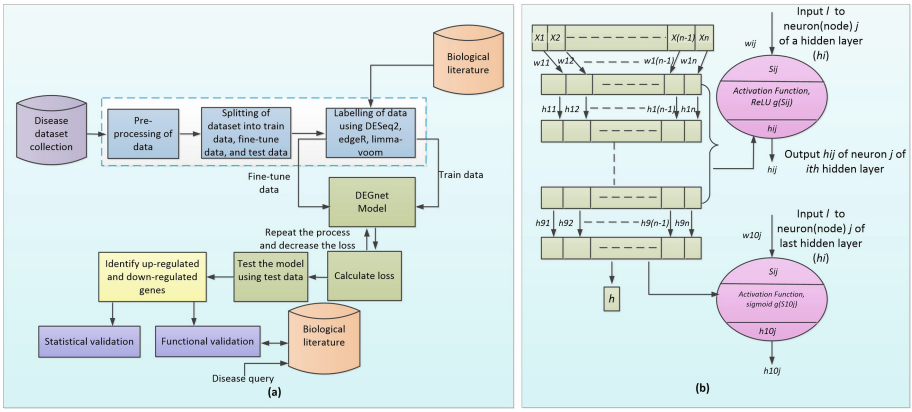

Fig. 1. (a) A representation DEGnet framework. The method has two phases. The first phase involves preprocessing, splitting, labelling of the RNA-seq data set. The second phase includes training (first level of training), fine-tuning (second level of training), testing of DEGnet model, identification of UR and DR genes and validation of identified DEGs. (b) Architecture of DNN used in DEGnet

DEGnet is a sequential deep neural network (Fig. 1b) with 1 input layer, 1 output layer, and 10 hidden layers, consisting of nodes (neurons). For batch size equal to 1, the input to the model is a vector N of size $1 \times n$, where n is the total number of control (or normal) and disease (or tumor) samples. The inputs for PD and BRCA datasets are two vectors of sizes 1×73 and 1×1215, respectively. To set the optimal number of hidden layers, we initialized the network with 2 hidden layers, and then add layers until it starts to overfit the training data and the test loss does not improve. Based on our experiment, we set the optimal number of hidden layers as 10. The most common rule to set optimal hidden layer size (number of neurons) is that the hidden layer size should be between the number of input and output size. The optimal hidden layer sizes for PD and BRCA datasets are found to be 60 and 1000, respectively based on the

different input sizes (73 and 1215) of the two datasets. In order to regularize the network, we then use dropouts rate (25%). We use rectified linear unit (ReLU) and sigmoid as activation functions, to determine whether a node should activate or not depending upon the sum of inputs-weights products of each layer. ReLU is applied at the sum h and its output is $max(0, h)$, where h is the output of each hidden layer. The sigmoid is another non-linear activation function with smooth gradient and its output range is (0,1). In our model, we predict probability of a gene to be up-regulating or down-regulating across samples. Since the probability is in the range of (0,1), sigmoid is the best choice for activation function and is used in the output layer. Also, ReLU is less computationally expensive and is most widely used activation functions. The model computes the loss that measures the error in predicting the optimal output for a given input x and updates the parameters based on the gradients. For this, we used *optim.Adam()* as an optimizer and *BSELoss()* as a loss function, which measures the binary cross entropy between the truth (y) and the predicted output (y^{pred}).

$$loss_{BCE}(y, y^{pred}) = \{l_1, l_2, \ldots, l_g\}^T,$$

$$Here \;\; l_i = -w_i[y_i^{pred} \cdot logy_i + (1 - y_i^{pred}) \cdot log(1 - y_i)] \text{ and } g \text{ is the batch size.}$$

If the $\overline{y}$ is the optimal output for test data t and if $\overline{y} < 0.5$ then $\overline{y} = 0$ otherwise 1. Once the model is trained with the consensus labelled train data; we fine tune the model using biologically validated fine-tune data. This trains the model with both log2FC estimates (sample variance) and incorporates prior knowledge of the data. Thereafter, the model is tested with the biologically validated test data. We used confusion matrix to calculate the accuracy, sensitivity (recall), specificity, and precision for evaluating the performance of our model.

2.1 Datasets and Preprocessing

We use the gene-expression datasets for PD RNA-seq (GSE68719) and BRCA (TCGA) [16] with control (or normal) and disease (tumor) samples. The first dataset contains mRNA-seq gene expression and MS3 proteomics with 29 PD and 44 control samples, profiled from human post-mortem BA9 brain tissue for PD and neurologically normal individuals. The second dataset contains the RNA-seq gene expression with 113 normal and 1102 tumor samples, profiled from breast invasive carcinoma (BRCA) expression data using an Illumina HiSeq2000 system. For preprocessing of the datasets, we remove the redundant genes and the rows with NAN values. Batch effects are removed using *removeBatchEffect()* of the edgeR package. Since the number of samples is small, we used a different approach, where we split the dataset as train data (98%), validation data or fine-tune data (1%), and test data (1%). We use three baseline methods DESeq2, edgeR, and limma-voom to calculate logFC estimates of each gene of the train data across control and disease samples. The positive logFC estimates are labelled UR and negative logFC estimates as DR. From the three baseline methods, we get three labels (of class 0 or 1) for each gene, and therefore to

remove the bias, we use the consensus labels for the genes and label them as UR (1) or DR (0) genes. The fine-tune and test data are labelled using prior knowledge acquired from the literature regarding up-regulation and down-regulation.

3 Results

Here, we show the assessment of our proposed method, DEGnet in terms GO terms enrichment, pathway enrichment and statistical metrics such as accuracy, sensitivity, specificity, and precision. We also compare the performance of DEGnet with six other ML-based methods for both PD and BRCA datasets.

3.1 Functional Validation of UR and DR Genes

In Tables 1 and 2, we show the GO enrichment and pathway enrichment in terms of p and q values for UR and DR genes extracted from the PD dataset. From Table 1, it is seen that UR and DR genes extracted using DEGnet from PD are enriched with GO terms such as activation of MAPK activity (GO:0000187), regulation of apoptotic process (GO:0042981), chemokine receptor binding (GO:0042379), CXCR chemokine receptor binding (GO:0045236), etc. which are associated with differentiation, degradation, and death of cell during pathogenesis of PD. Moreover, the PD associated pathways such as Apoptosis, Programmed Cell Death, IL-17 signaling pathway, Neurodegenerative Diseases, etc. mapped from these UR and DR genes are found to be significant with low p and q values. Similarly for BRCA, the UR and DR genes are biologically enriched. Recent studies say that there is a precise relation between Mitogen-activated protein kinase (MAPK) activation and proliferation, death, invasion of tumor during progression of cancer [17]. In Fig. 2, we show the MAPK pathway mapped from DR genes such as JUND and GADD45B of BRCA dataset identified using DEGnet.

Table 1. Analysis of GO enrichment of UR and DR genes of PD extracted using DEGnet

Disease	UR/DR genes	GO ID	p value	q value
PD	UR	Pyridine N-methyltransferase activity (GO:0030760)	6.434E–4	3.079E–2
		Activation of protein kinase (GO:0032147)	3.616E–5	2.769E–2
		Nuclease activity regulation (GO:0032069)	9.558E–5	2.769E–2
		Activation of MAPK activity (GO:0000187)	1.223E–4	2.769E–2
		Apoptotic process regulation (GO:0042981)	1.751E–4	2.769E–2
		Programmed cell death regulation (GO:0043067)	1.863E–4	2.769E–2
	DR	Cytokine activity (GO:0005125)	6.364E–7	5.761E–5
		Activity of chemokine (GO:0008009)	6.473E–7	5.761E–5
		Binding activity of chemokine receptor (GO:0042379)	1.603E–6	8.064E–5
		Binding of CXCR chemokine receptor (GO:0045236)	1.812E–6	8.064E–5
		Regulation of cell migration (GO:0030334)	2.714E–9	2.838E–6

Table 2. Analysis of pathways mapped from UR and DR genes of PD extracted using DEGnet

Disease	UR/DR genes	Pathways	p value	q value
PD	UR	Apoptosis	4.077E–4	3.319E–2
		Programmed Cell Death	4.286E–4	3.319E–2
		IGF1 pathway	2.649E–4	2.872E–2
		Cytokine Signaling in Immune system	3.263E–3	4.686E–2
		Genes regulating PIP3 signaling in cardiac myocytes	1.521E–3	4.686E–2
		IGF1 pathway	2.649E–4	2.872E–2
	DR	IL-17 signaling pathway	1.782E–8	7.825E–6
		TNF signaling pathway	4.978E–5	2.732E–3
		Neurodegenerative Diseases	1.087E–5	1.193E–3
		Chemokine signaling pathway	3.734E–4	1.091E–2
		Interleukin-4 and 13 signaling	1.315E–3	2.750E–2

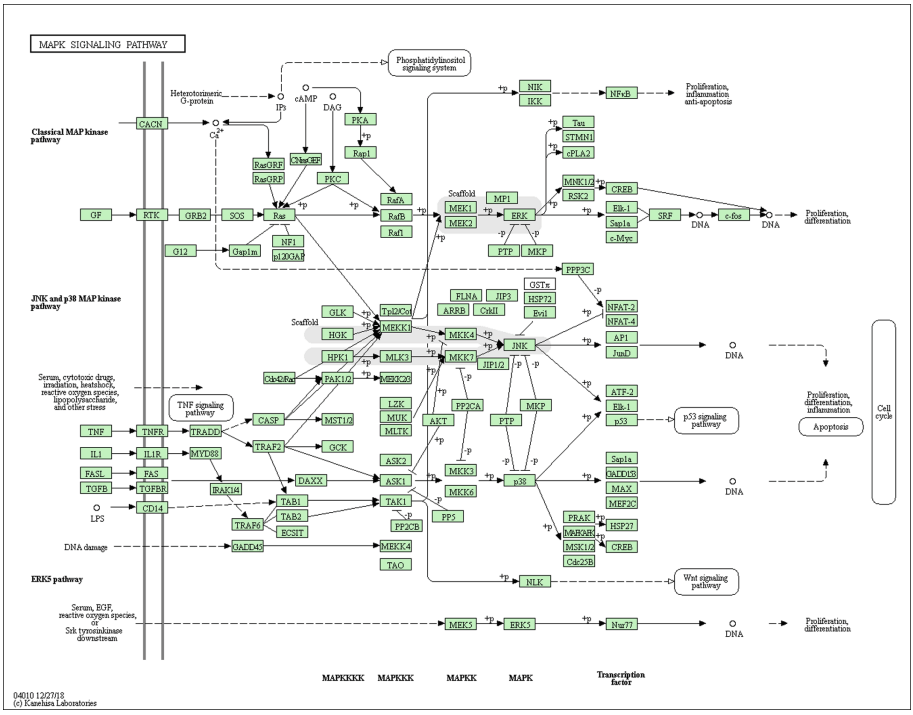

Fig. 2. MAPK pathway mapped from DR genes JUND and GADD45B of BRCA dataset. During pathogenesis of BRCA, there are perturbations in the DR genes JUND and GADD45B which cause significant disturbations in biological activities such as apoptosis, and synthesis of cells.

3.2 Statistical Analysis of UR and DR Genes

Here, we assess the performance of our proposed model with six ML based methods, namely, LR, KNC, SVM, GNB, DTC, and RFC in terms of accuracy, sensitivity, specificity, and precision on the PD and BRCA test data. For PD, except DTC and RFC, other methods find more false positives and false negatives than that of DEGnet method. Similarly, for the BRCA test data, though the DEGnet model does not obtain the best results in terms of statistical parameters, but the false positive rate and false negative rate are lower than other discussed methods. This shows that the proposed method is efficient in identifying potential disease biomarkers from a disease dataset. In Table 3, we compare the performance of DEGNet with the same six other ML-based methods in terms of accuracy, sensitivity, specificity, and precision. We see that for the PD dataset, DEGNet and DTC scores maximum accuracy, sensitivity, specificity, and precision. But for the BRCA dataset, DEGNet outperforms all six methods with 87.5% accuracy, 87.5% sensitivity, 100% specificity, and 100% precision. For DEGnet and DTC, the AUC score for the PD disease dataset is 1, which means that the model has an ideal measure of separability of true positives and false positives. The UR genes RPL3, APOD, PGK1, and PSMC1 show significant difference in functions such as protein synthesis, lipid metabolism, glycolysis pathway, catebolism and modification of proteins during pathogenesis of PD. Similarly, significant gene expression differences are seen in DR genes such as CSE1L, EEF1A, CD74, and SPP1, which are involved in transport, synthesis of protein, immune response during progression of PD [18]. The UR genes in BRCA such as SRCAP, HMGB1, PPIA, and ZNF9 are seen to play key roles in transcription, protein synthesis,

Table 3. Statistical analysis of UR and DR genes filtered from PD and BRCA

Dataset	Method	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
PD	DEGnet	100	100	100	100
	LR	53.85	62.5	50	35.71
	KNC	61.54	50	66.67	40
	SVM	50	62.5	44.44	33.33
	GNB	50	100	27.78	38.1
	DTC	100	100	100	100
	RFC	96.15	100	94.44	88.9
BRCA	DEGnet	87.5	87.5	83.74	100
	LR	40	13.04	76.47	42.86
	KNC	40	4.34	88.23	33.33
	SVM	47.5	17.39	81.03	66.67
	GNB	72.5	30.43	70.58	58.33
	DTC	40.15	21.74	82.35	62.5
	RFC	50.5	13.04	94.11	75

transport, and degradation. Similarly, significant differences are seen in the DR genes of BRCA such as TM4SF1, HMGN1, LMNA and SOD2, which participate in significant functions such as adhesion, synthesis of membrane proteins, transcription factors, and metabolism [19].

4 Discussion

The biological data has fewer samples than the number of genes and therefore the use of neural networks is challenging. In our paper, we proposed a model DEGnet, with a deep neural network of one input, multiple hidden layers, and one output layer. The trained model was used to identify UR and DR genes from PD and BRCA datasets with higher statistical and functional significances. The hallmark of the proposed method is that it can identify UR and DR with zero or minimal false positive or false negative rate. Based on the dataset size, the model may be extended later to test on other RNA-seq datasets to find potential biomarkers related to diseases by tuning the hidden layer size.

References

1. Dembélé, D., Kastner, P.: Fold change rank ordering statistics: a new method for detecting differentially expressed genes. BMC Bioinform. **15**(1), 14 (2014)
2. Anders, S., Huber, W.: Differential expression analysis for sequence count data. Genome Biol. **11**(10), R106 (2010)
3. Love, M.I., Huber, W., Anders, S.: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. **15**(12), 550 (2014)
4. Robinson, M.D., McCarthy, D.J., Smyth, G.K.: edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**(1), 139–140 (2010)
5. Law, C.W., Chen, Y., Shi, W., Smyth, G.K.: VOOM: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. **15**(2), R29 (2014)
6. Wang, L., Xi, Y., Sung, S., Qiao, H.: RNA-seq assistant: machine learning based methods to identify more transcriptional regulated genes. BMC Genom. **19**(1), 546 (2018)
7. Sekhon, A., Singh, R., Qi, Y.: DeepDiff: deep-learning for predicting differential gene expression from histone modifications. Bioinformatics **34**(17), i891–i900 (2018)
8. Kong, Y., Yu, T.: A deep neural network model using random forest to extract feature representation for gene expression data classification. Sci. Rep. **8**(1), 16477 (2018)
9. Kleinbaum, D.G., Klein, M.: Logistic Regression. Springer, New York (2002). https://doi.org/10.1007/b97379
10. Sarkar, M., Leong, T.-Y.: Application of k-nearest neighbors algorithm on breast cancer diagnosis problem. In: Proceedings of the AMIA Symposium, p. 759. American Medical Informatics Association (2000)
11. Polat, K., Güneş, S.: Breast cancer diagnosis using least square support vector machine. Digit. Signal Proc. **17**(4), 694–701 (2007)

12. Soria, D., Garibaldi, J.M., Ambrogi, F., Biganzoli, E.M., Ellis, I.O.: A 'non-parametric' version of the naive Bayes classifier. Knowl.-Based Syst. **24**(6), 775–784 (2011)
13. Singireddy, S., Alkhateeb, A., Rezaeian, I., Rueda, L., Cavallo-Medved, D., Porter, L.: Identifying differentially expressed transcripts associated with prostate cancer progression using RNA-seq and machine learning techniques. In: 2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 1–5. IEEE (2015)
14. Liaw, A., Wiener, M., et al.: Classification and regression by randomForest. R News **2**(3), 18–22 (2002)
15. Chen, J., Bardes, E.E., Aronow, B.J., Jegga, A.G.: ToppGene suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res. **37**(suppl_2), W305–W311 (2009)
16. Tomczak, K., Czerwińska, P., Wiznerowicz, M.: The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncol. **19**(1A), A68 (2015)
17. Santen, R.J., et al.: The role of mitogen-activated protein (MAP) kinase in breast cancer. J. Steroid Biochem. Mol. Biol. **80**(2), 239–256 (2002)
18. Kim, J.-M., et al.: Identification of genes related to Parkinson's disease using expressed sequence tags. DNA Res. **13**(6), 275–286 (2006)
19. Zucchi, I., et al.: Gene expression profiles of epithelial cells microscopically isolated from a breast-invasive ductal carcinoma and a nodal metastasis. Proc. Natl. Acad. Sci. **101**(52), 18147–18152 (2004)