Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230344

Differential Gene Expression Data Analysis of ASD Using Random Forest

Pragya[a,1], Praveen Kumar GOVARTHAN[a], Kshitij SINHA [b], Sudip MUKHERJEE[a,1] and Jac Fredo AGASTINOSE RONICKOM[a,1]
[a] *School of Biomedical Engineering, Indian Institute of Technology (BHU), Varanasi, Uttar Pradesh, India*
[b] *School of Biochemical Engineering, Indian Institute of Technology (BHU), Varanasi, Uttar Pradesh, India*
ORCiD ID: Pragya https://orcid.org/0000-0003-3815-3232

Abstract. Autism spectrum disorder (ASD) is a developmental disability caused by differences in the brain regions. Analysis of differential expression (DE) of transcriptomic data allows for genome-wide analysis of gene expression changes related to ASD. De-novo mutations may play a vital role in ASD, but the list of genes involved is still far from complete. Differentially expressed genes (DEGs) are treated as candidate biomarkers and a small set of DEGs might be identified as biomarkers using either biological knowledge or data-driven approaches like machine learning and statistical analysis. In this study, we employed a machine learning-based approach to identify the differential gene expression between ASD and Typical Development (TD). The gene expression data of 15 ASD and 15 TD were obtained from the NCBI GEO database. Initially, we extracted the data and used a standard pipeline to pre-process the data. Further, Random Forest (RF) was used to discriminate genes between ASD and TD. We identified the top 10 prominent differential genes and compared them with the statistical test results. Our results show that the proposed RF model yields 5-fold cross-validation accuracy, sensitivity and specificity of 96.67%. Further, we obtained precision and F-measure scores of 97.5% and 96.57%, respectively. Moreover, we found 34 unique DEG chromosomal locations having influential contributions in identifying ASD from TD. We have also identified chr3:113322718-113322659 as the most significant contributing chromosomal location in discriminating ASD and TD. Our machine learning-based method of refining DE analysis is promising for finding biomarkers from gene expression profiles and prioritizing DEGs. Moreover, our study reported top 10 gene signatures for ASD may facilitate the development of reliable diagnosis and prognosis biomarkers for screening ASD.

Keywords. Gene expression data, NCBI, Autism Spectrum Disorder, Random Forest, Statistical test

1. Introduction

Autism spectrum disorder (ASD) is a developmental disability characterized by social communication, interaction, and restricted or repetitive behaviors or interests [1,2]. It is caused by environmental and genetic factors. Studies on gene expression can help us to

[1] Corresponding Authors: Pragya, E-mail: pragya.rs.bme22@itbhu.ac.in, Dr. SM, E-mail: sudip.bme@iitbhu.ac.in, E-mail: Dr. JFAR, jack.bme@iitbhu.ac.in.

identify the protein that is primarily responsible for ASD. Moreover, Differential gene expression study helps to understand the biological differences at the genetic level between typical and diseased conditions [3]. Many technologies can capture gene expression from the DNA or RNA such as Microarray DNA, qPCR, and RNAseq [4]. However, these methods have the disadvantage of high cost and time-consuming. Although the list of risk genes implicated by de-novo mutations is growing, it is still very likely far from complete, with an estimated full set of ASD genes ranging from several hundred to more than 30,000. In the search for additional de-novo mutations, sequencing studies continue to be an important approach, but the current sequencing cost is still very high, especially for large samples [5]. As an alternative strategy, advanced analytical approaches like machine learning and statistical methods, which leverage previously implicated genes and prior knowledge, have the potential to enhance risk gene discovery in an efficient and cost-effective manner [6,7,8].

In this study, we have proposed a machine learning-based process pipeline for efeffectively identifying the candidate chromosomal locations (Hereafter will be referred tas genes for simplicity) in ASD. Random Forest (RF), an ensemble-based classifier, is trained with the gene expression data of ASD and TD from the NCBI GEO database. We have chosen RF, as it's a widely-used and established machine-learning algorithm for classifying datasets with many features. RF is an ensemble method that combines multiple decision trees, handling complex relationships between features and the target variable. It can also estimate feature importance, identifying informative genes for classification. The built machine learning model is validated using 5-fold cross-validation and candidate DEGs responsible for ASD were found.

2. Methods

The processing pipeline adopted in this study is shown in Figure 1. The gene expression data were downloaded from the NCBI GEO database of GSE7329 [9]. The gene expression data provides information about the expression level of the 43,932 genes of 30 samples (ASD=15 and TD=15) [10]. The gene expression data was preprocessed using an algorithm implemented in the R programming language.

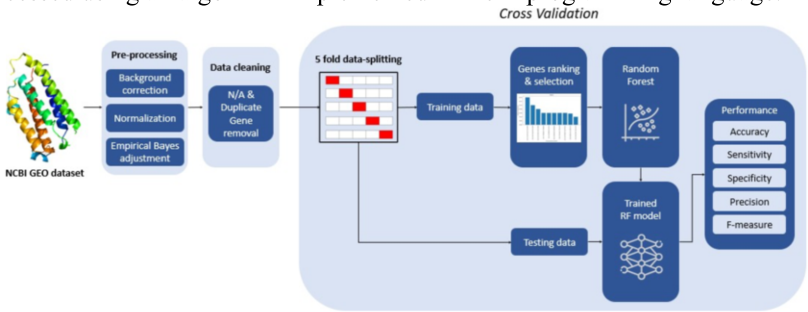

Figure 1. Flowchart of the proposed pipeline

Our study used multiple software packages including GEOquery, LIMMA and Tidyverse to predict disease outcomes based on gene expression data. In addition, the Python programming language package Sci-kit learn 1.0.1 was also used for building and training machine learning models to predict disease outcomes based on gene expression data. Initially, background correction was performed to subtract the background intensity from the foreground intensity for each spot. It attempts to adjust the data for ambient intensity surrounding each feature. We normalized the gene data and subsequently applied empirical Bayes statistics to differential expression to rank genes in order of evidence for differential expression. The computed linear model fit was then applied to generate the moderated t-statistics, moderated F-statistic, and log-odds of differential expression by empirical Bayes moderation of the standard errors towards a global value. We excluded 2,686 genes from the analysis, as their chromosomal locations were not available in the data set. Furthermore, 2 duplicates of 15 genes and 10 duplicates of 75 genes were observed and we computed the average for these genes. This involved taking the sum of the expression values for each gene and dividing it by the number of times that gene occurred in the dataset. To remove these redundant and missing gene names, we sorted and filtered the data in Microsoft Excel 2016, which led to the exclusion of these genes. Our final analysis included a total gene count of 40,556. We per- formed extensive 5-fold cross-validation to evaluate the performance of the RF machine learning model. Further, we optimized the number of features (Chromosomal locations) for the training model using the feature ranking method (Top 10 Chromosomal locations) of RF. During training, the genes were ranked according to their importance and the top 10 genes were only included in building the model in each fold. We computed the performance metrics like accuracy, sensitivity, specificity, precision and F-measure to evaluate the performance of the classification.

3. Results and Discussions

Figure 2 shows the occurrence of significant genes across 5-fold cross-validation by RF. There are 10 most important genes per fold (50 in total), but due to the overlapping of certain genes in the 5 folds, we ended up with 34 unique genes whose number of occurrences add up to 50. It can be observed that ‘chr3:113322718-113322659’ (Gene name-germinal centre expressed transcript) was the significant DEG present in every fold. So, we conclude that this gene plays a significant role in ASD. Other genes such as ‘chr3:197078850-197078791 and chr4:147533543147533484’ occurred in 3 out of 5 folds indicating influential contributions to the classifier. Furthermore, 8 chromosomal locations were present in 2 folds and the rest were present in a single fold.

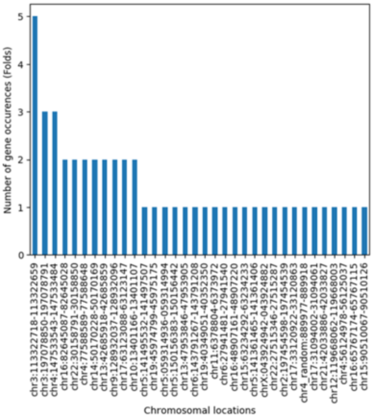

Figure 2. Histogram plot of the number of occurrences concerning chromosomal locations of genes in the 5-fold cross-validation

We obtained 96.67% for the accuracy metrics and the same for the sensitivity and specificity metrics. Similarly, the model also achieved a high precision of 97.5% and an F-measure score of 96.57%. The results showed that our proposed pipeline was able to recognize ASD from TD with high average classification accuracy.

Table 1 depicts the top 10 DEGs identified based on the RF model and statistical method. Our results show that chr3:113322718-113322659 and chrX:72826272-72826213 were the most significant genes by RF and t-test respectively. However, the top 10 genes identified by the RF and t-test were dissimilar.

Table 1. Comparison of top 10 DEGs of ASD identified through RF and t-test methods

S.No.	RF		t-test	
	Chromosomal location	Score	Chromosomal location	p-value
1	chr3:113322718-113322659	0.011	chrX:72826272-72826213	9.66E-33
2	chr3:197078850-197078791	0.006	chr11:61161623-61161682	1.08E-25
3	chr4:147533543-147533484	0.007	chr12:54797596-54797655	4.33E-25
4	chr16:82645087-82645028	0.006	chr6:74284574-74284515	9.88E-25
5	chr22:30158791-30158850	0.008	chr6:33351778-33351946	1.22E-24
6	chr4:77588589-77588648	0.006	chr1:44913389-44913448	1.55E-24
7	chr14:50170228-50170169	0.008	chr2:232403578-232403637	2.6E-24
8	chr13:42685918-42685859	0.007	chr17:34259949-34259890	3.62E-24
9	chr9:128932037-128932096	0.006	chr16:1952138-1952082	4.56E-24
10	chr17:63123088-63123147	0.006	chr7:5340087-5340028	5.17E-24

4. Limitations and Future work

Our process pipeline has produced a high classification accuracy of 96.7% to discriminate between ASD and TD. However, it has a few limitations. We ranked and selected the top 10 genes but failed to find similar genes between RF and t-test analysis. We can select the top 20, 30, or 40 genes and then re-validate the results. We can also use other preprocessing pipelines, like low-level preprocessing or high-level preprocessing methods [11]. In addition, we have considered only a single gene expression dataset and can use more datasets for improved performance. We built the models using machine learning classifiers but never attempted deep learning algorithms. We used only RF; however, we can use other classifiers such as support vector machine, logistic regression, XGBoost etc.

5. Conclusions

We have proposed a machine learning framework for identifying the genes responsible for causing ASD in an individual. We achieved an accuracy of 96.67% and the top performing gene was the germinal center expressed transcript 2 (chromosomal location chr3:113322718-113322659) by the RF algorithm. However, with the statistical analysis, we found the chromosomal location chrX:72826272-72826213 was responsible for ASD. The chromosomal locations were found different in both approaches. However, the RF model produced high classification accuracy. Our study shows the possibility of utilizing the proposed model in a potential application for screening ASD and TD in a clinical environment.

References

[1] Ansel A, Rosenzweig JP, Zisman PD, Melamed M, Gesundheit B. Variation in gene expression in autism spectrum disorders: an extensive review of transcriptomic studies. Frontiers in neuroscience. 2017 Jan 5;10:601.

[2] Ronicko JF, Thomas J, Thangavel P, Koneru V, Langs G, Dauwels J. Diagnostic classification of autism using resting-state fMRI data improves with full correlation functional brain connectivity compared to partial correlation. Journal of Neuroscience Methods. 2020 Nov 1;345:108884.

[3] Rylaarsdam L, Guemez-Gamboa A. Genetic causes and modifiers of autism spectrum disorder. Frontiers in cellular neuroscience. 2019:385.

[4] Nagy A´ , La´nczky A, Menyha´rt O, Gyo˝rffy B. Validation of miRNA prognostic power in hepatocellular

[5] Carcinoma using expression data of independent datasets. Scientific reports. 2018 Jun 15;8(1):1-9.

[6] Lin Y, Afshar S, Rajadhyaksha AM, Potash JB, Han S. A machine learning approach to predicting autism risk genes: Validation of known genes and discovery of new candidates. Frontiers in genetics. 2020 Sep 10;11:500064.

[7] Asif M, Martiniano HF, Vicente AM, Couto FM. Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology. PloS one. 2018 Dec 10;13(12):e0208626.

[8] Go¨k M. A novel machine learning model to predict autism spectrum disorders risk gene. Neural Computing and Applications. 2019 Oct;31(10):6711-7.

[9] Brueggeman L, Koomar T, Michaelson JJ. Forecasting risk gene discovery in autism with machine learning and genome-scale data. Scientific reports. 2020 Mar 12;10(1):1-1.

[10] Agastheeswaramoorthy K, Sevilimedu A. Drug REpurposing using AI/ML tools-for Rare Diseases (DREAM-RD): A case study with Fragile X Syndrome (FXS). bioRxiv. 2020 Jan 1.

[11] Nishimura Y, Martin CL, Vazquez-Lopez A, Spence SJ, Alvarez-Retuerto AI, Sigman M, Steindler C, Pellegrini S, Schanen NC, Warren ST, Geschwind DH. Genome-wide expression profiling of lymphoblastoid cell lines distinguishes different forms of autism and reveals shared pathways. Human molecular genetics. 2007 Jul 15;16(14):1682-98.

[12] Dziuda DM. Data mining for genomics and proteomics: analysis of gene and protein expression data. John Wiley & Sons; 2010 Jul 16.