Machine Learning Algorithms for the Identification of Potential Non-Syndromic Intellectual Disability Associated miRNAs

Julián González Betancur (julian.gonzalezbetancur@ucr.ac.cr)
University of Costa Rica School of Biology: Universidad de Costa Rica Escuela de Biologia
https://orcid.org/0000-0001-7833-368X

José A Guevara-Coto
University of Costa Rica

Adarli Romero
University of Costa Rica School of Biology: Universidad de Costa Rica Escuela de Biologia

Research article

Keywords: miRNA association, artificial intelligence, machine learning, intellectual disability, biomarker

Posted Date: June 21st, 2021

DOI: https://doi.org/10.21203/rs.3.rs-595856/v1

RESEARCH

Machine learning algorithms for the identification of potential non-syndromic intellectual disability associated miRNAs

Julián González-Betancur[1*], José A Guevara-Coto[2] and Adarli Romero[3]

*Correspondence:
julian.gonzalezbetancur@ucr.ac.cr
[1]Department of Biology, University of Costa Rica, San Pedro, Montes de Oca, Costa Rica, Waterloo Road, San José, Costa Rica
Full list of author information is available at the end of the article

Abstract

Background: Intellectual disabilities (IDs) are a group of developmental disorders with high phenotypic and genotypic heterogeneity. Association of genetic elements to IDs has typically been empirically accomplished, however recently, machine learning (ML) has proved to be an excellent instrument to elucidate these associations. miRNAs are short non-coding molecules that participate in spatiotemporal gene regulation, making them relevant for the understanding ID causality.

Methods: In this study we used the BrainSpan spatio-temporal expression database to develop a series of machine learning predictors: SVM, RF, FF-ANN, and Stochastic Gradient Descent Classifier. These models were capable of recognizing gene expression profiles. The best classifier was used to label miRNAs associated with NS-IDs using the BrainSpan expression profiles.

Results: The model with the best performance was a FF-ANN with 0.78 of F1-score, 0.78 of weighted recall and 0.78 of weighted precision. We used this model to identify miRNAs with high probability to be associated with NS-IDs using the spatio-temporal gene expression profile in the human brain. Labeled miRNAs that were annotated were associated with processes related to either IDs and-or neurodevelopmental processes.

Conclusions: The development of a machine learning framework that identified potential NS-ID miRNAs represents an interesting approach for the identification of a potential list of on genes that could be subject for further experimental validation. This study also reinforces the potential of machine learning frameworks in their discovery of potential biomarkers that could improve disease detection and management.

Keywords: miRNA association; artificial intelligence; machine learning; intellectual disability; biomarker

Background

Intellectual disabilities (ID) are a group of neurodevelopmental disorders with high phenotypic and genotypic heterogeneity[1, 2] and an estimate prevalence of 1-3% worldwide[3]. The knowledge of genetic causes for syndromic (S-IDs) and non-syndromic IDs (NS-IDs) is the result of exhaustive research in gene variants and their potential association. Such work has been mainly done using next generation sequencing[4, 1] and has provided valuable information on ID causing genetic variants. The comprehension of biological basis of IDs could help improve quality of

life in individuals with S-IDs/NS-IDs, as well as for the understanding of human cognition[4, 1, 5].

Due to their capacity to modulate expression of multiple genes, miRNAs, a class of small non-coding RNAs, play a crucial role in all tissues' development and maturation[6]. Compelling evidence exists supporting the involvement of miRNAs as important post-transcriptional regulators of encephalon development and plasticity of synaptic connections[6, 7]. Overall miRNAs' expression patterns are widely variable among brain regions and throughout the lives of individuals, time during which each miRNA can regulate the expression of many different genes, other miRNAs, and other genetic elements[8].

In the last decade miRNAs have been associated with S-IDs and with NS-IDs through the correlation of expression anomalies (overexpression and/or underexpression)[7, 2]. Similarly, variants that alter the normal pairing of specific miRNAs and their mRNA targets, have been associated with the emergence of such disorders[7, 2].

In recent years, the diagnosis of human disorders and conditions has benefited from the increased ability to convert data into knowledge. This process, defined as knowledge discovery, is primarily done using machine learning algorithms. Machine learning (ML) is a technique in the field of computation that allows a computer to learn how to recognize patterns hidden in data of any nature[9], a useful feature to work with data from complex and hard to model phenomena. Supervised learning allows us to teach the computers how to recognize classes from known data as for the computer to classify new data for us[9]. ML models are systems that apply one or several learning algorithms with the objective of analysing new data. These models have reached high popularity in recent years to study complex problems in genetics[10, 11].

Cogill and Wang[12] created a ML model that provided a list of candidates miRNAs for bio-markers in autism, utilizing a database of gene expression of the brain development. Using this same database, Gök[13] demonstrates the utility of considering different types of ML models in the search for better models with higher performance in classification problems in genetics, but focused on the association of lncRNAs with the autism spectrum.

In this paper, through the use of a computational framework that includes machine learning algorithms and functional annotations, we have presented a list of miRNAs candidates associated with NS-IDs. The list was the result of predictions generated from the highest performing machine learning model, trained with spatial-temporal gene expression data of the developing human brain. Our working hypothesis was that NS-IDs associated genes, and miRNAs that regulate them share spatial-temporal expression patterns. Based on that notion, we developed a classification system, focusing on the highest performance metrics, capable of classifying spatial-temporal expression profiles in the human brain.

Methods

Data Acquisition

For the training as well as for the miRNAs classification we utilized the gene expression profiles of the developing human brain repository BrainSpan[14]. BrainSpan

contains the expression profiles from RNA-seq in RPKM of genes and non-gene genetic elements in 16 brain areas from prenatal states to 40 y.o. with an average depth of 2 patients per area-developmental state combination. In this database each patient-area-developmental state combination is a separated column. This information was consolidated as a data matrix by the association of expression profiles data and information about spatio-temporal variables for each patient from different files available in BrainSpan.

Data Preprocessing

We selected the expression profiles of 1823 genes with no known variants associated or causing ID phenotypes. The set of 1823 genes were assigned the -1 class label to indicate their negative association with ID. In contrast, a set of 707 genes reported in the literature[1] as associated with NS-IDs were selected. These genes were labeled as NS-ID, or given the 1 class label to indicate their positive association with ID.

The amount of expression profiles for negative and positive genes were different to the number of selected genes due to BrainSpan having the mRNA profiles instead of gene profiles. We decided not to eliminate those profiles with duplicated gene names because they correspond to different isoforms with different biological activities. The pairing of the gene names and their corresponding ENCODE identification was accomplished utilizing Bioconductor's biomaRt[15] library from the programming language R version 3.5.2. We built a database (named initial DB) with gene expression profiles of positive and negative genes only.

For the initial data set the RPKM values were normalized in three ways: 1) applying base 2 logarithm, 2) applying min-max normalization, and 3) applying base 2 logarithm (log2) followed by application of min-max normalization. Using log2 and min-max is based on the notion that the logarithmic transformation can be implemented for reducing the inter and intra-sample differences or noise, whereas the min-max normalization can be used to allocate all instances in the same scale or plane, thus improving algorithm convergence. The use of the three normalization methods aimed to find the best possible performance of the ML models. All experiments were executed with the three versions of the normalized initial DB (Fig. 1).

Machine Learning model building and analysis

We used the following ML algorithms: Support Vector Machine (SVM), Random Forest (RF), and Feed Forward Artificial Neural Network (FF-ANN). SVM models were implemented in the programming language R with the library e1071[16], and in the programming language Python we implemented the SGDClassifier model, from scikit learn (sklearn)[17], which as a default uses the linear SVM. RF models were implemented in R with the library randomForest[18]. FF-ANN models were implemented as in R with the library neuralnet[19], and in Python with TensorFlow[20] and Keras[21].

For each ML model ten iterations were executed. Each iteration consisted of selecting a random seed and the creation of at least ten models of the corresponding algorithm with different combinations of its internal parameters (Fig. 1). For the creation of each model we used 70% of data for training and 30% for testing and

validation. Training data were balanced utilizing the library caret of R [22], and the library sklearn of Python. Random seeds and confusion matrix were recorded for each iteration.

SVM models were trained using "radial", "lineal", and "sigmoidal" kernels, cost and gamma (in case of sigmoidal kernel) were adjusted with the tune function of e1071 library. RF models were trained with 1500 decision trees per forest, with no replace, all other parameters were used with default values, and training data set was balanced by class using SMOTE function from DMwR[23] library of R with k equals 3 and 190% of oversampling of minority class. FF-ANN models were trained with all available solvers in Keras and MLPClassifier, with an alpha between 0.0001 and 1 in sweeps by one order of magnitude, with different combinations of activation functions for layer, and in uniform and autoencoder architectures.

Feature Selection of Important Variables

The RF model with the higher precision and F1 score was utilized for feature selection of the normalized initial DB based on the score of importance of each variable over the RF classification. Based on the feature selection we built two new databases, one with the 80% of variables with higher importance score, and another with the 60% of variables with the higher importance score. Three more iterations were executed for all algorithms using the new databases normalized with the third normalization algorithm described above.

miRNAs Classification

From the total of models created we selected the one with the higher F1 score and AUC for the classification of miRNAs. The F1 scores were calculated using the function "classification_report " from sklearn.metrics library of Python. The AUC and ROC curve were obtained using ROSE library of R[24]. The best model was retrained with the totality of the expression profiles of positive and negative genes. The miRNAs' expression profiles were obtained from the same database as the gene expression profiles, available in BrainSpan project. Once the classification of miRNAs was obtained we searched for literature that associated the positive tags resulting from the miRNAs classification with NS-IDs or with neurodevelopment. We filtered miRNAs classified as positive according to their probability to be associated with NS-IDs in order to obtain the best candidates only (those with probability greater than 0.99). The mirbase-names of the best candidates were obtained through the library biomaRt of R. The list of target genes of the best candidates was obtained with the tool STARBASE with searches of high stringency of CLIP data (greater or equals to 3 according to the tool) over the human genome.

Results

The training database consisted in the expression profiles of 27657 gene transcripts of the negative class (unassociated with NS-IDs), 11640 gene transcripts of the positive class (already reported as associated with NS-IDs), and 524 patient-brain area- developmental state variables.

The normalization algorithm that produced the best results was the third algorithm explained above (data not presented). The performance of those models

trained with the database with feature selection applied was higher than those trained with the totality of original variables (TABLE 1). However, the general performance of the models trained with the 60% of the variables was the lowest (data not presented). The best model obtained (TABLE 2) was a FF-ANN of Multilayer Perceptron Classifier type (MLPClassifier) integrated by 5 hidden layers of 400 neurons each, an alpha of 0.01, solver SGD, and random state 121. The AUC value 0.799 of the best model was the higher AUC between obtained (Fig. 2).

Once the best model was retrained the classification of 544 miRNAs was performed, and 36 of them were classified as associated with NS-IDs (TABLE 3), using their class probabilities. In this group of 36 miRNAs the probability average was 0.933 ($\pm$0.119 sd). For the miRNAs classified as negative the average probability of being positive was 0.067 ($\pm$0.119 sd). A total of 23 miRNAs were selected as best candidates due to their probability being greater or equal to 0.99 (TABLE 3). For 11 of the 23 best candidates, mirbase-names were not found in biomaRt, probably due to the original transcript names in BrainSpan being in the ENCODE 10 version. We found 7 of the best candidates already being reported as associated with brain development, cognitive development, or NS-IDs (TABLE 3). A total of 7 of the selected candidates have target genes reported in literature as associated with intellectual disorders and ND-IDs (TABLE 3) according to the results obtained from STARBASE. A total of 5 more were not found in the STARBASE system, probably due to the Gencode V10 used by BrainSpan in the identification of transcripts.

Discussion

The issue of predicting the association between genes and neurodevelopmental disorders has been successfully broached by several authors, including Cogill and Wang[12], Gök[13], and Wang and Wang[25]. However, only the best ML model obtained by Wang and Wang was an ANN type, with an Autoencoder architecture[25]. We also obtained a best ML model of ANN type but with a feed forward (FF) architecture instead of the best architecture Wang and Wang reported. The differences between the study's FF-ANN models and the remaining models can be clearly observed after the application of feature selection, suggesting that expression data patterns are better learned by FF-ANN model types, as long as the noise is filtered. We used the best model obtained with all the data (an RF type model) to perform the feature selection. This technique, which has already been reported as an embedded technique of feature selection[26] and used by Wang and Wang[25], allowed us to obtain a significant improvement in the best model, as well as all other developed models. The remaining types of model were not benefited by the feature selection as much as the best model.

In problems that requires classification in two categories random guess would obtain a performance of 0.5 correct classifications. Any higher performance indicates that information about existing patterns in the data was obtained, which is the case of this study's best model, since it obtained a considerably higher value of performance. This gained information has been of great value in other studies that use ML algorithms to associate genes with conditions, disorders and disabilities, with an effectiveness of close to 0.8, which is a similar value as the performance obtained by the best model of this study[12, 13, 25]. In biological terms, the pattern

information that was gained means that the hypothesis the study was based on, that the genes associated with NS-IDS and the miRNAs that regulate them have similar spatio-temporal expression patterns in the human brain, is correct.

For the extent of this work, it was impossible to obtain the most significant spatio-temporal variables of the brain for the performed classification due to the best model being a FF-ANN type. FF-ANN are often considered black box systems, meaning that interpretation of the importance of used entrance variables for the final prediction is either highly complex or impossible. In cases of ANNs with many layers of many neurons per layer, this is particularly true, like is the case of our best model[27].

Conclusions

It is important to note that among the miRNAs recognized as positive by the best model are miRNAs previously reported by literature as associated to NS-IDs and other neurodevelopmental disorders. According to the best ML model, all miRNAs selected as best candidates have a high probability of being associated with NS-IDs, based on their spatio-temporal expression patterns. Finding that all candidate miRNAs have target genes previously reported as associated with NS-IDs or with other neurodevelopmental disorders supports the future study of the candidates proposed by the best model, as they have a high probability of being involved in important processes related to the emergence of NS-IDs.

List of abbreviations

ML: Machine Learning. DB: Data Base. RPKM: Reads Per Kilobase of transcript, per Million mapped reads. miRNA: Micro-Ribonicleic Acid. lncRNA: long non-coding Ribonucleic Acid. IDs: Intellectual disabilities. NS-IDs: Non-Syndromic Intellectual Disabilities. S-IDs: Syndromic Intellectual Disabilities. RF: Random Forest. FF-ANN: Feed Forward Artificial Neural Network. SVM: Support Vector Machine. MLPClasiffier: Multi-Layer Perceptron Classifier. SGD: Stochastic Gradient Descent. SGDClassifier: Stochastic Gradient Descent Classifier. mRNA: Messenger Ribonucleic Acid.

Ethics approval and consent to participate
In this study we worked with human data, available in BrainSpan project's web site. No any consent to participate was required due to the absence of traceable, personal, or sensitive information.

Consent for publication
Not applicable

Availability of data and materials
The original data to replicate or findings is available in the download section of BrainSpan project's web site https://www.brainspan.org/static/download.html

Competing interests
The authors declare that they have no competing interests.

Funding
Not applicable

Author's contributions
Julián González-Betancur: Conceptualization, Methodology, Software, Formal analysis, Validation, Data Curation, Writing-Original draft.: José A. Guevara-Coto: Conceptualization, Supervision, Writing–Reviewing & Editing.: Adarli Romero: Conceptualization, Writing–Reviewing & Editing.

Acknowledgements
We would like to thank the Materials Science and Engineering Research Center (CICIMA) for allowing us to use their server cluster for the execution of our experiments. Particularly Federico Muñóz for kindly and patiently assisting us in the use of the servers. We also thank Adarly Romero Vásquez and Henriette Raventos Vorst for their advice and guidance from a biological perspective.

Author details
[1]Department of Biology, University of Costa Rica, San Pedro, Montes de Oca, Costa Rica, Waterloo Road, San José, Costa Rica. [2]Department of Computer Sciences and Informatics, University of Costa Rica, Costa Rica, San José, Costa Rica. [3]Department of Biology, University of Costa Rica, Costa Rica, San José, Costa Rica.

References

1. Vissers, L.E.L.M., Gilissen, C., Veltman, J.A.: Genetic studies in intellectual disability and related disorders. Nature Reviews Genetics **6**, 9–16 (2016)
2. Yang, J., Lui, A., He, I., Bai, Y.: Bioinformatics analysis revealed novel 3'utr variants associated with intellectual disability. Genes **11**(9), 998–1012 (2020)
3. Maulik, P.K., Mascarenhas, M.N., Mathers, C.D., Dua, T., Saxena, S.: Prevalence of intellectual disability: a meta-analysis of population-based studies. Research in Developmental Disabilities **32**, 419–436 (2011)
4. Kaufman, L., Ayub, M., Vincent, J.B.: The genetic basis of non-syndromic intellectual disability: a review. Journal of Neurodevelopmental Disorders **2**, 182–209 (2010)
5. Zedníková, I., Chylíková, B., Šeda, O., Korabečná, M., Pazourková, E., Břešťák, M., Krkavcová, M., Calda, P., Hořínek, A.: Genome-wide mirna profiling in plasma of pregnant women with down syndrome fetuses. Molecular Biology Reports **47**, 4531–4540 (2020)
6. Fiorenza, A., Barco, A.: Role of dicer and the mirna system in neuronal plasticity and brain function. Neurobiology of Learning and Memory **135**, 3–12 (2016)
7. Siew, W.H., Tan, K.L., Babaei, M.A., Cheah, P.S., Ling, K.H.: Micrornas and intellectual disability (id) in down syndrome, x-linked id, and fragile x syndrome. Frontiers in Cellular Neuroscience **7**, 1–11 (2013)
8. Ziats, M.N., Rennert, O.M.: Identification of differentially expressed micrornas across the developing human brain. Molecular Psychiatry **19**, 848–852 (2014)
9. Kotsiantis, S.B.: Supervised machine learning: A review of classification techniques. Informatica **31**, 249–268 (2007)
10. Le, D.H.: Machine learning-based approaches for disease gene prediction. Briefings in Functional Genomics **00**(00), 1–14 (2020)
11. Bracher-Smith, M., Crawford, K., Escott-Price, V.: Machine learning for genetic prediction of psychiatric disorders: a systematic review. Molecular Psychiatry **26**, 70–79 (2021)
12. Cogill, S., Wang, L.: Support vector machine model of developmental brain gene expression data for prioritization of autism risk gene candidates. Bioinformatics **32**(23), 3611–3618 (2016)
13. Gök, M.: A novel machine learning model to predict autism spectrum disorders risk gene. Neural Computing and Applications **31**, 6711–6717 (2019)
14. for Brain Science, A.I.: Home::BrainSpan: Atlas of the Developing Human Brain. https://www.brainspan.org/. Accessed: 2020-02-18
15. Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., Huber, W.: Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics **21**(16), 3439–3440 (2005). doi:10.1093/bioinformatics/bti525
16. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F.: E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. (2019). R package version 1.7-3. https://CRAN.R-project.org/package=e1071
17. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.*: Scikit-learn: Machine learning in python. Journal of machine learning research **12**(Oct), 2825–2830 (2011)
18. Liaw, M. A.and Wiener: Classification and regression by randomforest. R News **2**(3), 18–22 (2002)
19. Fritsch, F., Guenther, F., Wright, M.N.: Neuralnet: Training of Neural Networks. (2019). R package version 1.44.2. https://CRAN.R-project.org/package=neuralnet
20. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org (2015). http://tensorflow.org/
21. Chollet, F., et al.: Keras. GitHub (2015). https://github.com/fchollet/keras
22. Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., Hunt, T.: Caret: Classification and Regression Training. (2018). R package version 6.0-81. https://CRAN.R-project.org/package=caret
23. Torgo, L.: Data Mining with R, Learning with Case Studies. Chapman and Hall/CRC, ??? (2010). http://www.dcc.fc.up.pt/~ltorgo/DataMiningWithR
24. Lunardon, N., Menardi, G., Torelli, N.: Rose: a package for binary imbalanced learning. R Journal **6**(1), 82–92 (2014)
25. Wang, J., Wang, L.: Prediction and prioritization of autism-associated long non-coding rnas using gene expression and sequence features. BMC Bioinformatics **21**(505) (2020). doi:10.1186/s12859-020-03843-5
26. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics **23**, 2507–2517 (2007)

27. Géron, A.: Hands-on Machine Learning with Scikit-Learn and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd ed. edn. O'Reilly Media, Sebastopol, CA (2019)

Figures

Figure 1 Diagram of experiments execution, miRNAs classification process, and target gene recognition.

Figure 2 ROC curve of the best model obtained. AUC = 0.799.

Tables

Table 1 Performance metric F1-score by type of ML model by the size of the database used in training and testing.

Model	100,00 %[1]		80,00 %[1]	
	average	desv.est	average	desv.est
SVM	0.222	0.148	0.142	0.148
RF	**0.669**	0.006	0.594	0.005
FF-ANN	**0.669**	0.055	**0.686**	0.041
SGD	0.623	0.05	0.634	0.026

[1] 100%: All of original variables. 80%: Only variables among the 80% with higher importance according to the best RF model).Support Vector Machine (SVM), Random Forest (RF), Feed Forward Artificial Neural Network (FF-ANN), Support Vector Machine with Stochastic Gradient Descent (SGD).

Table 2 Performance metrics of the best model.

Adjust	Metric		
	Precision	Recall	F1-score
Accuracy	-	-	0.78
Macro-average	0.74	0.73	0.74
Weighted-average	0.78	0.78	0.78

Table 3 miRNAs tagged as associated with NS-IDs by the best ML model.

Ensembl_gene_id	Mirbase_id	Probability Unassociated	Probability Associated
[3]ENSG00000211575	hsa-mir-760	0.000	1.000
[2]ENSG00000221643	NF	0.000	1.000
[2]ENSG00000207789	hsa-mir-26a-2	0.000	1.000
[2,3]ENSG00000207712	hsa-mir-627	0.000	1.000
[1,2]ENSG00000207863	hsa-mir-125b-2	0.000	1.000
[1,2,3]ENSG00000207550	hsa-mir-99b	0.000	1.000
[2]ENSG00000207606	hsa-mir-554	0.000	1.000
[2]ENSG00000207945	NF	0.000	1.000
[2,3]ENSG00000207729	hsa-mir-556	0.000	1.000
[1,2]ENSG00000207991	Hsa-mir-601	0.000	1.000
[2]ENSG00000207610	NF	0.000	1.000
[1,2]ENSG00000199165	hsa-let-7a-1	0.000	1.000
[2]ENSG00000207709	NF	0.001	0.999
[2,3]ENSG00000207631	hsa-mir-641	0.001	0.999
[2]ENSG00000207690	NF	0.001	0.999
[2]ENSG00000221617	NF	0.001	0.999
[1,2]ENSG00000199017	Hsa-mir-1-1	0.002	0.998
[1,2,3]ENSG00000199179	hsa-let-7i	0.002	0.998
[2]ENSG00000253009	NF	0.003	0.997
[1,2,3]ENSG00000199043	Hsa-mir-335	0.005	0.995
[2]ENSG00000221604	hsa-mir-1293	0.005	0.995
[2]ENSG00000211995	NF	0.007	0.993
[2]ENSG00000222326	NF	0.010	0.990
[1]ENSG00000207608	Hsa-mir-127	0.016	0.984
ENSG00000215939	hsa-mir-873	0.023	0.977
[1]ENSG00000199135	Hsa-mir-101-1	0.030	0.970
ENSG00000207825	hsa-mir-519b	0.064	0.936
[1,2]ENSG00000208023	Hsa-mir-185	0.088	0.912
ENSG00000221657	NF	0.172	0.828
[1]ENSG00000207607	hsa-mir-200a	0.178	0.822
ENSG00000221541	NF	0.183	0.817
ENSG00000207819	NF	0.214	0.786
ENSG00000207626	hsa-mir-562	0.251	0.749
[1]ENSG00000211582	Hsa-mir-758	0.336	0.664
ENSG00000208022	hsa-mir-618	0.386	0.614
ENSG00000207979	hsa-mir-527	0.420	0.580

Target genes were corroborated for miRNAs with probability higher than 0.99 only. (NF) Not Found.
[1]Already associated with neurodevelopment, NS-IDs, or mental disorders.
[2]Selected as best candidates (probability higher than 0.99).
[3]Candidates with target genes associated with neurodevelopment and/or NS-IDs

Figures

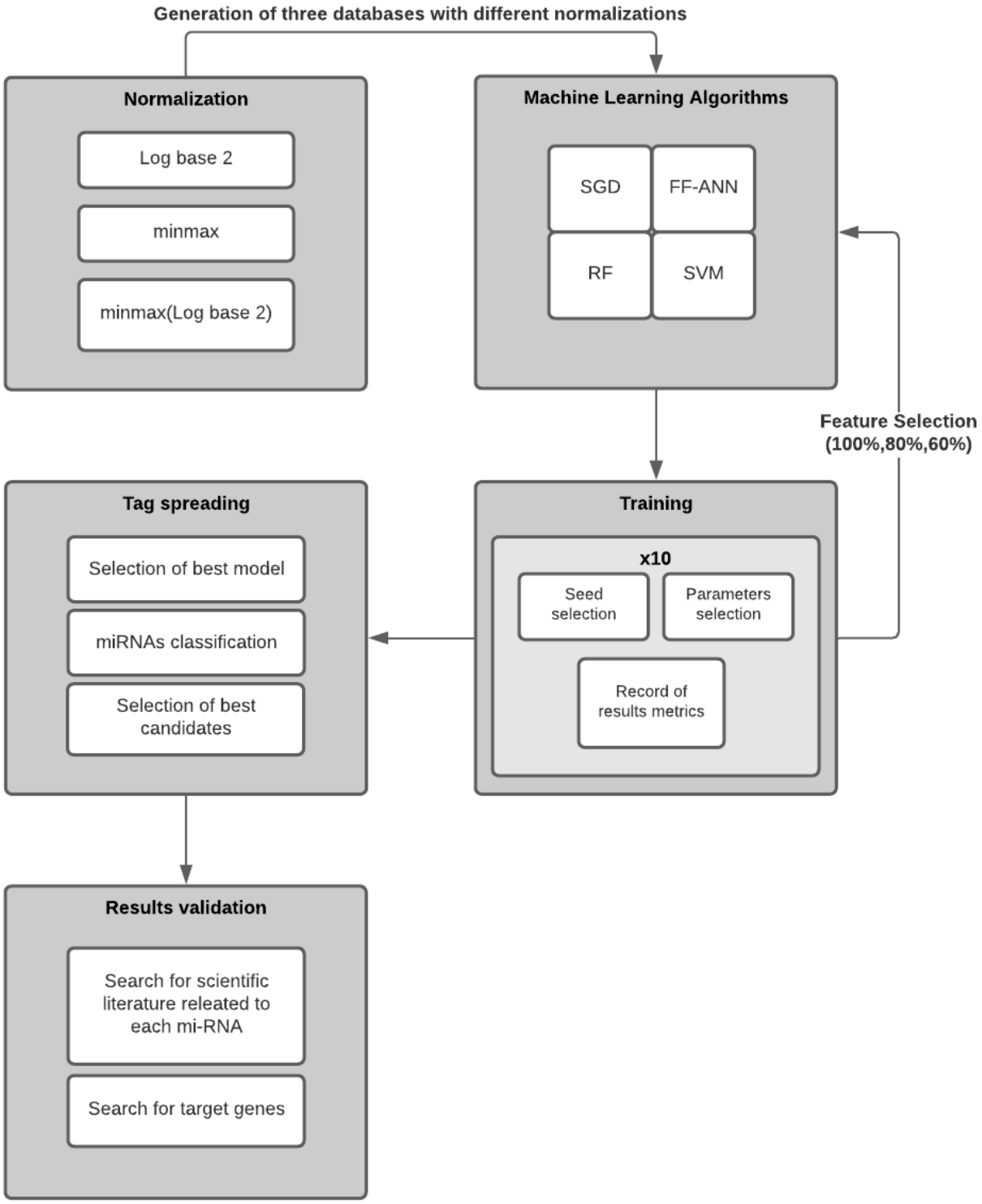

Figure 1

Diagram of experiments execution, miRNAs classication process, and target gene recognition.

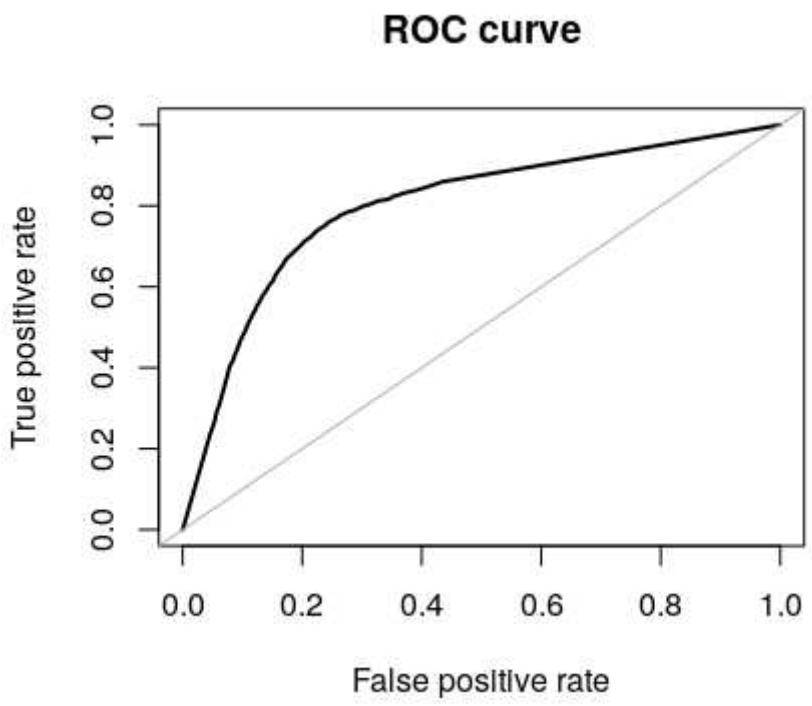

Figure 2

ROC curve of the best model obtained. AUC = 0.799.