An improved catalogue of putative synaptic genes defined by their temporal transcription profiles through an ensemble machine learning approach

Flavio Pazos (✉ fpazos@iibce.edu.uy)

Instituto de Investigaciones Biológicas Clemente Estable <https://orcid.org/0000-0003-4980-1071>

Pablo Soto

Instituto de Investigaciones Biológicas Clemente Estable

Martín Palazzo

Instituto de Investigaciones Biomédicas de Buenos Aires (Max Plank Partner)

Gustavo Guerberoфф

Instituto de Matemática y Estadística "Rafael Laguarda", Facultad de Ingeniería, UDELAR

Patricio Yankilevich

Instituto de Investigaciones Biomédicas de Buenos Aires (Max Plank Partner)

Rafael Cantera

Instituto de Investigaciones Biológicas Clemente Estable & Zoology Department, Stockholm University

Research article

Keywords: Synapse, Machine learning, Temporal transcription profiles, Gene function prediction

Posted Date: May 14th, 2019

DOI: <https://doi.org/10.21203/rs.2.9628/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background. Assembly and function of neuronal synapses require the coordinated expression of a yet undetermined set of genes. Previously, we had trained an ensemble machine learning model to assign a probability of having synaptic function to every protein-coding gene in *Drosophila melanogaster*. This approach resulted in the publication of a catalogue of 893 genes that was postulated to be very enriched in genes with still undocumented synaptic functions. Since then, the scientific community has experimentally identified 79 new synaptic genes. Here we used these new empirical data to evaluate the predictive power of the catalogue. Then we implemented a series of improvements to the training scheme and the ensemble rules of our model and added the new synaptic genes to the training set, to obtain a new, enhanced catalogue of putative synaptic genes. **Results.** The retrospective analysis demonstrated that our original catalogue was indeed highly enriched in genes with unknown synaptic function. The changes to the training scheme and the ensemble rules resulted in a catalogue with better predictive power. Finally, training this improved model with an updated training set, that includes all the new synaptic genes, we obtained a new, enhanced catalogue of putative synaptic genes, which we present here announcing a regularly updated version that will be available online at: <http://synapticgenes.bnd.edu.uy> **Conclusions.** We show that training a machine learning model solely with the whole-body temporal transcription profiles of known synaptic genes resulted in a catalogue with a significant enrichment in undiscovered synaptic genes. Using new empirical data, we validated our original approach, improved our model and obtained a better catalogue. The utility of this approach is that it reduces the number of genes to be tested through hypothesis-driven experimentation.

Background

The synapse, a specialized contact between neurons, is currently of fundamental importance for our understanding of learning, memory and other brain functions. Assembly and function of neuronal synapses require the coordinated expression of a yet undetermined set of genes (1). Since the biological roles of the vast majority of known amino acid sequences remain partly or completely unknown (2), computational prediction of gene function is an open research problem of much relevance. In recent years diverse methodologies have been assayed, with a strong prevalence of machine learning approaches, with the top-performing algorithms, architectures and training schemes remaining function-specific and context-dependent (3).

In a previous study, (4), we implemented an ensemble machine learning model that assigned a probability of being “synaptic” to each protein-coding gene of *Drosophila melanogaster*. Instead of relying on GO annotations, to construct the training set we used an *ad hoc* definition of “synaptic gene”. The features from which the synaptic function was inferred were the whole-body transcription levels of all *Drosophila* genes at 24 developmental stages, published by the modENCODE project (5). As far as we know, this is the only study that predicts gene function relying solely on temporal transcriptions profiles obtained through NGS technologies.

Our model combined three learning algorithms: k-NN, Random Forests and SVM to obtain a catalogue which was greatly enriched with genes mostly expressed in the central nervous system, genes for which a synaptic function had not been discovered in *Drosophila* but whose homologous have a known synaptic function in *Homo sapiens*, and genes that already had some synapse related GO annotation in *Drosophila* but that did not fulfill our definition of synaptic gene. After excluding these already annotated genes we obtained a final catalogue, which we postulated it was highly enriched in genes for which a synaptic function was going to be discovered.

Since the publication of our catalogue (4) researchers around the world have identified and experimentally validated 79 “new synaptic genes”, allowing us to analyze

the predictive power of our approach *a posteriori*. As a significant proportion of the new synaptic genes were in our catalogue, we felt encouraged to improve it. In addition to expand our training set with the new synaptic genes, we implemented some changes to our training scheme and to the ensemble criteria to generate a new model. To analyze if these changes really improve the prediction performance, we used the original training set (Supplementary Table 2) to train our new model and test if this results in a catalogue with higher enrichment in new synaptic genes compared with that published in 2015.

Finally, once we demonstrated that the changes resulted in a better catalogue, we incorporated all the new synaptic genes to the training set, retrained the model and obtained a new, enhanced catalogue of putative synaptic genes that we publish here. A monthly updated version of this catalogue will be available online at: <http://synapticgenes.bnd.edu.uy>.

Results

Evaluation of the original catalogue

Since the publication of our catalogue (4), 79 new synaptic genes (NSG) have been experimentally identified in *Drosophila*. Additional file 1 lists these genes as well as the references reporting their synaptic functions. The table also shows the probability of being synaptic that had been assigned to each of these genes by the three algorithms in our original model and whether or not the gene was included in our original catalogue. Almost a third of the NSG identified after the publication of our work were included in our catalogue. In terms of list enrichment, the catalogue has an enrichment in NSG of 4.45 with a p-value $< 10^{-9}$ (See Table 1).

Comparison between the original and the new model when both are trained with the original training set. The enrichment in new synaptic genes found in the catalogue obtained with the new model is 20% higher than that found in the catalogue obtained with the original model. (p-values are calculated assuming a hyper-geometric distribution).

Improved training scheme and ensemble rules

We implemented a number of changes to the training scheme and the ensemble rules of our approach (see Fig. 1, and Methods). To test whether these changes really improved the performance of our model we trained a new model with the original training set and obtained an alternative catalogue of putative synaptic genes. Thereafter, we compared the enrichment in NSG found in the original catalogue with that found in this catalogue. As shown in Table 1, the implemented changes gave the model a better predictive power measured as enrichment in NSG.

Fig. 1 – Training scheme & ensemble rules

Beginning with an unbalanced training set, we constructed five smaller, slightly differing and balanced training sets. To construct each of these balanced training sets we picked out four fifths of the positives examples and randomly selected the same number of negative examples. We repeated this procedure five times, leaving out a different fifth of the positive examples each time. Using the positive and negative examples left out each iteration, we defined five test sets, used to independently calculate the accuracy and the AUC of the ROC for each classifier. Then we trained three models with each training set, obtaining 15 different classifiers. The hyper parameters of each classifier were set by grid search combined with 10-fold cross validation. Then we considered the intersection of the 15 classifications, conformed by those genes classified as synaptic by all the classifiers. Finally, after ranking these genes by the mean probability of being synaptic assigned by the classifiers, we set a probability threshold maximizing the enrichment of the resulting catalogue in new synaptic genes.

A new catalogue of putative synaptic genes

After confirming that the changes implemented to the training scheme and ensemble criteria resulted in a catalogue more enriched in NSG, we incorporated these genes to the training set and then repeated the whole procedure described above. We obtained a catalogue of 1.257 *Drosophila* genes, all of which received a probability of being synaptic above 0,95 from each of the 15 classifiers. After excluding the 201 genes already annotated with some synapse related GO term (6), we obtained a final catalogue of 1.171 putative synaptic genes (Additional file 3).

Monthly updated online catalogue

The model we are presenting here will be re-trained on a regular basis, incorporating to the training set each new gene identified as having synaptic function. This will result in a continuously updated catalogue that will be available here: <http://synapticgenes.bnd.edu.uy>. At this site visitors will find the updated list of synaptic genes used to train the model.

Discussion

A catalogue of genes postulated to have high probability of being important for the function of neuronal synapses was published four years ago (4). Since then, 79 “new synaptic genes” (NSG) were experimentally identified by others. These NSG, identified with a variety of experimental methods, offered a great opportunity to test the predictive power of our machine learning approach and to test whether it could be improved. The original catalogue was obtained by training an ensemble machine-learning model that assigned each protein-coding gene present in the data set a probability of having synaptic function, only considering its temporal transcription profile during development. We hypothesized that the catalogue was highly enriched in genes of relevance for *Drosophila* synapse assembly and function still not recognized as such.

The main goal here was to use these new data to perform *a posteriori* evaluation of the predictive power of our original approach, to evaluate a series of changes to the architecture of the model that could enhance its predictive power and to obtain a new, enhanced catalogue of putative synaptic genes by including to the training set all the NSG.

We found here that almost a third of the genes for which a synaptic function had been experimentally identified by other colleagues between 2015 and 2019 were present in our catalogue. This represents a good experimental validation of the predictive power of our machine learning approach and demonstrates that it is possible to make predictions on gene function using machine learning based entirely on temporal transcription data.

On the other hand, our original model assigned very low probabilities to some genes that were later proven to have synaptic functions. One possible explanation to this relies in that our model is trained only with temporal transcription profiles. Assembly and function of synapses most probably requires the coordinated expression of hundreds of genes, including genes encoding proteins that repress other genes, which transcription profiles will probably be specular. It is quite probable that any model exclusively trained with transcription profiles will fail to recognize genes with specular transcription profiles. It is also worth noting that none of the new synaptic genes belonged to the list of “non-synaptic genes” with which we had used to train the algorithms, thus providing conclusive validation for the biological criteria used to construct the training set.

Conclusions

We show here that a catalogue of *Drosophila* putative synaptic genes obtained by an ensemble machine learning model four years ago has a significant enrichment in genes whose synaptic function was discovered after the publication of the catalogue. It is worth noting that the model was trained exclusively with temporal transcription profiles from a whole-body developmental transcriptome. We implemented a number of changes to the training scheme and the ensemble rules and thereafter trained this new model with the original training set. This generated a catalogue that was even more enriched in new synaptic genes than the original catalogue, indicating that the adjusted model had better performance. Finally, we

included all the new synaptic genes in the training set and trained a model that incorporates the proposed changes, obtaining a new, enhanced catalogue of putative synaptic genes.

We are making this catalogue available to the scientific community, firmly believing that this will facilitate the identification of genes important for the assembly and function of synapses, by means of gene silencing, mutant analysis, electrophysiology, neuroanatomy, behavioral assays and other traditional protocols, all of which will most likely lead to a better understanding of the function of the nervous system. The catalogue is available at: <http://synapticgenes.bnd.edu.uy>

Methods

Evaluation of the predictive power of our original model

Using the same *ad hoc* definition for “synaptic gene” we used in our previous work, we performed a bibliographic revision and identified 79 genes that gathered strong experimental evidence about its importance for neuronal synapses after the publication of our catalogue. Then we analyzed the enrichment of our original catalogue in these new synaptic genes.

A new training scheme

To obtain our original catalogue we trained three learning algorithms (k-NN, Random Forest and SVM) with an unbalanced training set, in which there were many more negative than positive examples. After an exhaustive bibliographic revision, aimed to construct a list of genes whose importance for neuronal synapses had gathered strong experimental evidence, 92 positive examples were selected (Additional file 1). On the other hand, 397 genes were selected as negatives examples, based on two biological criteria: genes with no expression during developmental stages when massive synaptogenesis is known to take place, or genes with no expression during adult life in one of the two sexes (5).

With the aim of improving the predictive power of our model, here we use the original unbalanced training set to construct five smaller, slightly differing and balanced training sets (See Fig. 1). To construct each of these balanced training set we picked out four fifths of the original positives examples and randomly selected the same number of negative examples. We repeated this procedure five times, leaving out a different fifth of the positive examples each time. Each of these training sets includes a slightly different portion of the original positive examples and a completely different set of the negative examples. Using the positive and negative examples left out when constructing each balanced training sets we also defined five test sets, used to independently calculate the accuracy and the AUC of the ROC for each classifier.

Ensemble rules

Our original catalogue comprised all those genes for which the probability of being synaptic assigned by each of the three learning algorithms was above a certain threshold. This threshold was set to obtain a catalogue of a given size, following a *rationale* about how many undiscovered synaptic genes could probably exist in *Drosophila*. Now we followed a different approach (See Fig. 1). We trained the same three algorithms with five training sets, which resulted in 15 different classifiers. The hyper parameters of each classifier were set by grid search. Each classifier assigned a different probability of being synaptic to each gene. Then we set the classification threshold maximizing the enrichment of the resulting catalogue in new synaptic genes.

All calculations were performed using Jupyter Notebooks and Sklearn (7).

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of Data and Materials

All data generated or analyzed during this study are included in this published article [and its supplementary information files].

Competing interests

The authors declare that they have no competing interests

Funding

FPO received financial assistance from PEDECIBA (Uruguay) and received funds from the Agencia Nacional de Investigación e Innovación (ANII, Uruguay). RC received funds from the Sistema Nacional de Investigadores (Uruguay).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

FPO conceived the study, performed the experiments and data analysis and wrote the manuscript. PS discussed the experiments prepared the web site and corrected the manuscript. MP discussed the experiments and corrected the manuscript. GG supervised the project and corrected the manuscript. PY supervised the project and corrected the manuscript. RC conceived the study, supervised the project and corrected the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by grant FSDA_1_2017_1_142427 from Agencia Nacional de Investigación e Innovación. FPO is the recipient of a PhD fellowship from Agencia Nacional de Investigación e Innovación and GG and RC were supported by Sistema Nacional de Investigadores (Uruguay).

References

1 - Lassek M, Weingarten J, Volkandt W. The synaptic proteome. *Cell Tissue Res.* 2015; 359:255–65. doi: 10.1007/s00441-014-1943-4

2 - UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 20118; 46: 2699

3 - Jiang Y., Oron T. R., Clark W. T., Bankapur A. R., D’Andrea D., et al., 2016 An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol.* 2017; 184. doi:10.1186/s13059-016-1037-6

4 - Pazos Obregón F, Papalardo C, Castro S, Guerberoff G, Cantera R. Putative synaptic genes defined from a Drosophila whole body developmental transcriptome by a machine learning approach. *BMC Genomics.* 2015; 16:694–707. doi: 10.1186/s12864-015-1888-3

5 - Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, Brown JB, Cherbas L, Davis CA, Dobin A, Li R, Lin W, Malone JH, Mattiuzzo NR, Miller D, Sturgill D, Tuch BB, Zaleski C, Zhang D, Blanchette M, Dudoit S, Eads B, Green RE, Hammonds A, Jiang L, Kapranov P, et al. The developmental transcriptome of Drosophila melanogaster. *Nature.* 2011; 471:473–9. doi: 10.1038/nature09715.

6 - Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000; 25:25–9.

7 - Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res.* 2011; 12:2825–2830

Tables

Table 1 – Comparison between the original and the new model

Model	Original	New	New
Training set	original	original	new
Classification threshold	0.9	0.95	0.95
Genes above threshold	988	152	1,371
Genes annotated with any synapse related GO term	95	29	201
Genes in final catalogue	893	123	1,170
Enrichment in NSG	4.45	5.48	
p-value	1.07E-09	1.09E-03	

Figures

Figure 1

Training scheme & ensemble rules

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplement1.xls](#)
- [supplement2.xls](#)
- [supplement2.xls](#)