Fabio Taddei Dalla Torre f.taddeidallatorre@studenti.unitn.it

Diagnostic model, Network and Functional enrichment analysis for Parkinson's Disease.

Abstract

This work exploit Affymetrix microarray dataset to identify a suitable classification model for the diagnosis of Parkinson's Disease. The project aims also at performing network and functional enrichment analysis on a list of significant genes identified thanks to this model. The functional enrichment analysis will be performed using DAVID and g:Profiler while EnrichNet will be used for tissue specific network analysis. Finally, active subnetwork-oriented pathways analysis will be implemented through the R package pathfindR.

Introduction

Parkinson Disease (PD) is the second most common neurological disease affecting about 1% of the population older than 60 years and 3% of people older than 80 years.[1,2] The deaths of dopamine cells in a region of the mid brain (substantia nigra) causes to suffer from rest tremor, muscular rigidity, and bradykinesia.[3]

The diagnosis of the disease relies on human expertise indeed it is based on the identification of some cardinal motor sign.[4] One of the purposes of this work is being able to identify and implement a classification model that allows to identify if an individual is subject to PD or not, starting from a gene expression profile.

In conjunction, the classification model makes it possible to extrapolate a list of significant genes. Functional enrichment analysis will be performed on said list to gain biological insight and potentially discover the effect of PD on biological functions.

Functional enrichment analysis presents some limitations. For instance, it allows only to identify functional association on overlapping genes. Besides, not considering the network of interaction between genes or protein of interest is the main drawback[5]. To overcome these weaknesses and gain the maximum knowledge out of the gene expression set, network-based analysis will be performed on the seed gene list identified by the best classification model.

This report will be structured as following: the first paragraph will present the data and the method exploited in the Data Analysis but also the tools used for the Functional Enrichment Analysis and the Network Based Analysis. The following paragraph will show the results of the previous analysis hence the performance of the various models and the biological interpretation of the results. The last paragraph will provide a summary of the analysis and will outline the peculiarity and shortcomings of the work.

Methods

Microarray dataset GSE72267 from Gene Expression Omnibus database has been investigated in this work. The dataset is composed by the blood sample of 40 patients with PD and 19 controls. GSE72267 was published by Calligaris et al. in 2015 and it was produced using GPL571 Affymetrix Human Genome U133A 2.0 Array (HG-U133A-2).

All the data processing has been carried out by using R programming language. The dataset has been downloaded by the means of the library *GEOquery* that is available via the Bioconductor package.

Data Exploratory Analysis

Visualising the distribution of the level of expression for each individual is pivotal in order to check for previous transformation or normalisation of the data and to check if these techniques are needed in an effort to reduce the non-biological noise of the data.

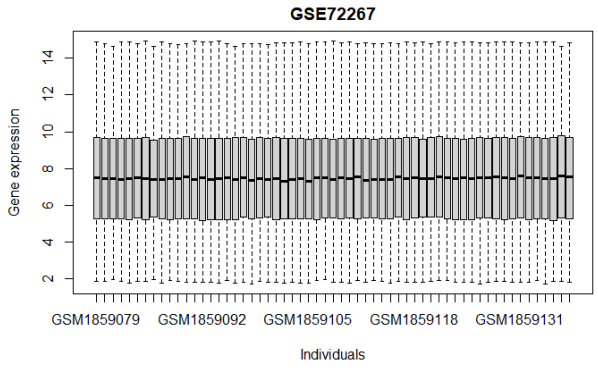


Figure 1: Boxplot of the distribution of the gene expression data for each individual.

As shown in *Figure 1*, the dataset has already been log-transformed and cleaned from the outliers.

As part of the exploratory analysis, a feature *Diagnosis* has been added to the dataset. This feature represents whether a subject was affected by PD or was a healthy control. It is possible to retrieve this feature from the *characteristics_ch1* element of the previously retrieved expression set.

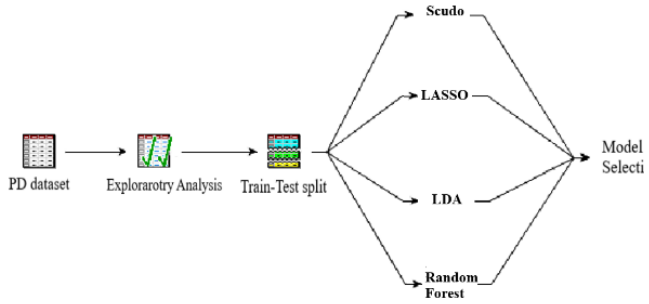


Figure 2: Applied classification methods.

Classification methods

Figure 2 shows the various classification methods that have been implemented. Hyperparameter tuning was performed for all of them to improve their performance.

Briefly, RF requires to tune the number of trees to grow (*ntree*), a value of 800 was chosen after having tested the model with other values. Another parameter is *mtry* that is the number of features chosen for each split, 10-fold CV selected an optimal number of mtry equal to 3. For LDA the parameter *prior*, that refers to the prior probability of the two classes, was set to 0,5 and 0,5 because it is desirable that the model assumes the two classes equally probable. Lasso hyperparameter *lambda*, which is a penalty parameter for considering a predictor, was tuned via CV. The optimum value of lambda was 0, this means that no penalty is introduced. 5-fold CV was performed for tuning the parameters of Scudo. This method suggested a value of 150 for *ntop* and of 150 for *nbottom* that are the number of genes considered from the ranked list of each individual. The other important parameter is *alpha* that was set to 0,7 and represent the P-value used for sampling the gens for feature selection.

For all the methods, the *caret* library has been used to perform cross validation. Moreover, as displayed in *Figure 4*, all the methods have shown significant performance improvement after having trained them on a subset of genes. Indeed, as an extreme example, it was not possible to train LDA on the full dataset at all. Only the genes showing a P-value smaller than 0.07 were chosen for the subset; P-values that were calculated via a row T-test performed thanks to the library *genefilter*. The threshold P-value was hinted while performing RF. *Figure 3* shows that only a small fraction of genes contributes to the homogeneity of nodes and leaves of the RF and so presents a higher Mean Decrease in Gini coefficient[6]. The P-value has been selected to obtain a subset of about 2 000 genes.

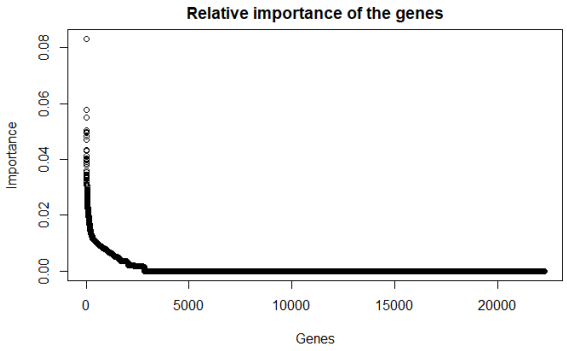


Figure 3: Relative importance of the genes calculated by Mean Decrease in Gini Coefficient

At the beginning, PCA and Clustering have also been implemented. However, these methods did not outline any interesting result. The ranked gene list produced by the classification method performed with the higher accuracy has been chosen for the following analysis.

Functional Enrichment Analysis

LDA is the method that has proven to be the most accurate. Therefore, Functional Enrichment Analysis has been performed on the 200 most important genes ranked according to LDA. GO terms and pathway enrichment analysis have been investigated, on the previously mentioned list of genes, using DAVID (https://david.ncifcrf.gov/) and g:Profiler (https://biit.cs.ut.ee/gprofiler/gost). DAVID has been selected because it provides automated solutions that allow to extrapolate biological meaning from a gene list[7]. The *Functional Annotation Chart* produced by this tool were used for the analysis. In this chart the genes from the input, that are known to have similar function, are linked together, and ranked according to their enrichment P-value. P-value lower than 0.05 means that the term is statistically significant. Also, the Benjamini-Hochberg false discovery rate is calculated[8]. For the purpose of the report this last parameter will be used to select statistically

significant terms. g:Profiler is a another widely used tool that provides a reliable service and publication ready-to-use results. Cumulative hypergeometric test is used by g:Profiler in order to produce the functional enrichment of the input gene list. This tool makes use of *g:SCS* method in order to test for false positive results. g:SCS is said to be more conservative than the previously mentioned Benjamini[9].

Biological Network Analysis

The network-based analysis has been carried out by using the web-application *EnrichNet* and the R package *pathfindR*. EnrichNet has been selected because it allows to perform tissue specific analysis in a fully automated fashion. Given that GSE72267 is compose by blood sample and PD effect mid brain cells, tissue specific analysis has been performed to come across more meaningful results. The tools make use of STRING database to produce the weighted edges of the network. KEGG, BioCarta and Reactome annotation databases were uses for the analysis. On the other hand, *pathfindR* allows active subnetwork-oriented pathways analysis. In protein-protein interaction network, active subnetwork represents a group of active interconnected genes that are mostly significant. They allow so to detect sets of interacting genes that are disease-associated. The option to cluster these pathways is also provided by pathfindR. Functionality that allows to identify representative pathways[10]. In order to extract the most relevant pathways, pathfindR make use of both the p-value of the seed gene list provided as input and information from the protein-protein interaction network. The p-value associated to the gene list is the one that has been computed, through the package *genefilter* when selecting the subset of genes for the various classification methods. Furthermore, the probe ID has bene converted into gene symbols by the mean of the *hgu133a.db* library. The analysis was conducted with the wrapper function *run_pathfindR* by keeping all the default parameters. Only the *iteration*, that refers to number of iterations for active subnetwork search and enrichment analyses was set to 1.

Results

This section will report the findings of the previously explained analysis.

Classification

For LDA, Lasso and Scudo, the performance has been calculated as the mean level of accuracy of 10-fold CV. Instead, for RF, the accuracy is 1 – mean of OOB. Indeed, for RF cross validation is not needed since already OOB predicts the response using all models that do not include that observation. Linear Discriminant Analysis (LDA) was the best model, scoring an accuracy of 93.3%, as shown by *Figure 4*. A high-level accuracy was also achieved by Lasso and Random Forest, scoring respectively 83.3% and 87.5%. To verify the performance of a model is also important to evaluate the ROC curve, displayed in *Figure 5*, that has been obtained by applying the LDA model on the test set.

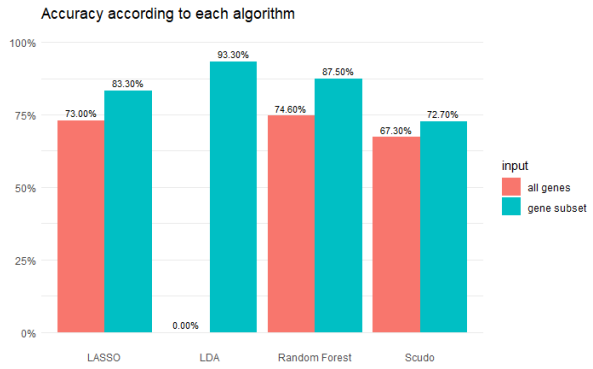


Figure 4: Level of accuracy achieved by the various method.

The ROC curve plot true positive rate against false positive rate. It is possible to summarise the plot by AUC (Area Under the Curve). This value ranges from 1, that means the model is perfectly able of distinguishing the two classes, and 0[11]. LDA produces an AUC equal to 0.8782.

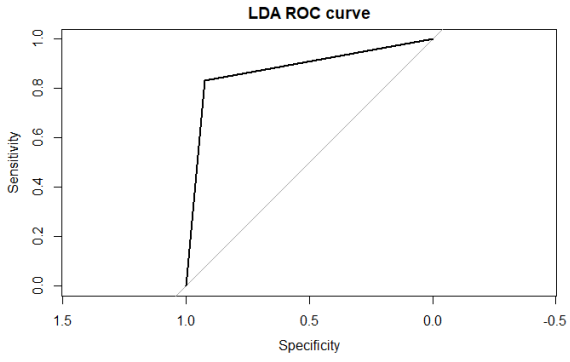


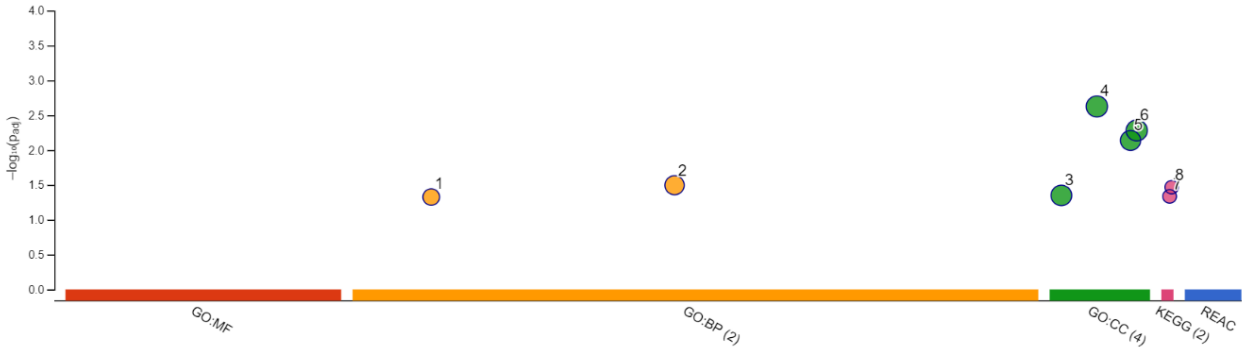
Figure 5: LDA ROC curve.

Functional Enrichment Analysis

The outcome of g:Profiler Functional Enrichment Analysis is displayed in *Figure 6* while in *Table 1* are displayed the results regarding the DAVID analysis. It is possible to see that both tools have identified, as statistically significant, two KEGG

Category	ID	Term	Benjamini
KEGG_PATHWAY	KEGG:04940	Graft-versus-host disease	$4.5 \cdot 10^{-3}$
GOTERM_BP_DIRECT	GO:0002480	Antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-independent	$6.2 \cdot 10^{-2}$
KEGG_PATHWAY	KEGG:05332	Type I diabetes mellitus	$7.4 \cdot 10^{-3}$
GOTERM_CC_DIRECT	GO:0042612	MHC class I protein complex	$2.1 \cdot 10^{-2}$
REACTOME_PATHWAY	R-HSA-1236977	R-HAS-1236977	$3.3 \cdot 10^{-2}$
KEGG_PATHWAY	KEGG:05169	Epstein-Barr virus infection	$1.6 \cdot 10^{-2}$
KEGG_PATHWAY	KEGG:05416	Viral myocarditis	$1.6 \cdot 10^{-2}$
GOTERM_BP_DIRECT	GO:0031901	Early endosome membrane	$5.4 \cdot 10^{-2}$
KEGG_PATHWAY	KEGG:05330	Allograft rejection	$2.3 \cdot 10^{-2}$

Table 1: DAVID functional enrichment analysis result.



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:BP	GO:0007626	locomotory behavior	4.721×10^{-2}
2	GO:BP	GO:0046649	lymphocyte activation	3.197×10^{-2}
3	GO:CC	GO:0005887	integral component of plasma membrane	4.477×10^{-2}
4	GO:CC	GO:0042995	cell projection	2.370×10^{-3}
5	GO:CC	GO:0098590	plasma membrane region	7.282×10^{-3}
6	GO:CC	GO:0120025	plasma membrane bounded cell projection	5.238×10^{-3}
7	KEGG	KEGG:04940	Type I diabetes mellitus	4.621×10^{-2}
8	KEGG	KEGG:05332	Graft-versus-host disease	3.413×10^{-2}

Figure 6: g:Profile functional enrichment analysis result.

pathways: *Graft-versus-host disease* and *Type I diabetes mellitus*. Different enrichment analysis tools make use of different methods, data source and identifiers type[9]. This fact gives soundness to the previously reported results. Moreover, KEGG pathways represent a very valuable finding because they allow to gain understanding about high-order biological findings and biological process[12]. DAVID has identified other pathways: *Allograft rejection*, *Viral myocarditis*, *Epstein-Barr virus infection* and a Reactome one that is *R-HAS-1236977*. Furthermore, the tool has identified three GO term: two cellular component and a biological process.

Interesting, g:Profiler reported two biological process, namely, *locomotory behaviour* and *lymphocyte activation* that were also identified by DAVID but were not considered as statistically significant according to the Benjamini corrected P-Value. The others g:Profiler results refer to four different cellular components.

Biological Network Analysis

The EnrichNet tissue specific analysis has been conducted on KEGG, BioCarta and Reactome annotation dataset, but only the Reactome one has shown worth mentioning results.

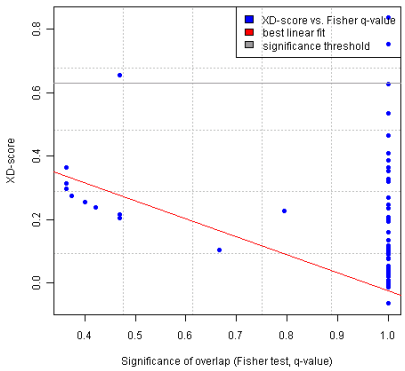


Figure 7: Regression plot: XD-score vs. Significance of overlap (Fisher test, q-value).

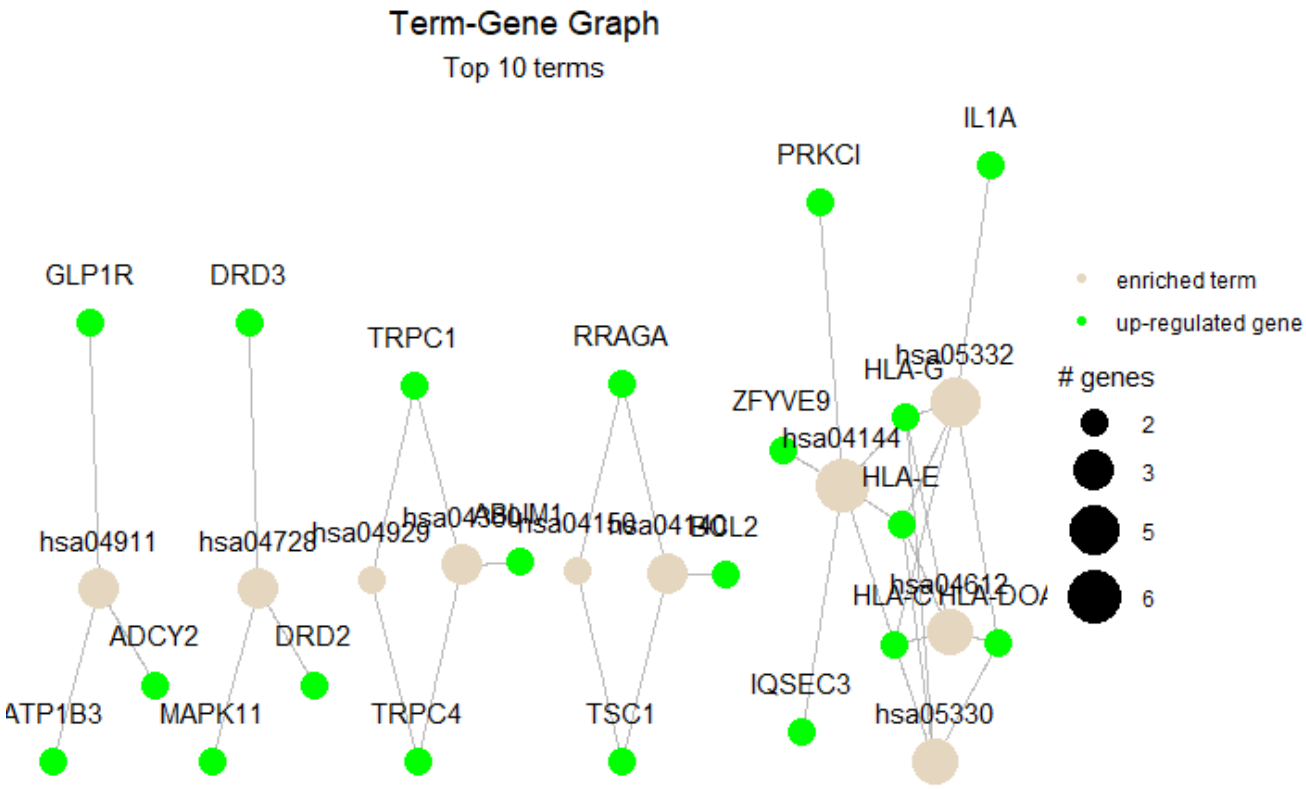


Figure 4: Network resulting from by pathfinR active-subnetwork-oriented enrichment analyses.

The first important one is the regression plot in *Figure 7*. This plot allows to choose an appropriate XD-score threshold of 0.63. The plot shows also that there is a high concentration of point on its right side; area where are reported all the non-overlapping dataset pair for which a meaningful scoring is only possible thanks to the XD-distance.

Tissue Type	Tissue XD-score
Pathway a)	
Prefrontal cortex	8.96
Globus pallidus	8.95
Amygdala	8.95
Parietal lobe	8.94
Pathway b)	
Cingulate cortex	8.98
Occipital lobe	8.98
Medulla oblongata	8.97
Prefrontal cortex	8.96
Globus pallidus	8.95
Cerebellum penducles	8.93
Caudate nucleus	8.93
Subthalamic nucleus	4.47
Pons	4.47
Parietal lobe	4.44

Table 2: Main results of the Tissue specific analysis for: Pathway a) P75NTR recruits signaling complexes and Pathway b) regulation of RHEB GTPASE activity by AMPK

Five networks have been identified as statistically significant, but in accordance with what previously stated, only their XD-score are significant while their Fischer exact test score are not. Nevertheless, two of these five pathways/processes have shown notable XD-score for tissues related to the brain. In *Table 2* are listed the most relevant tissues for these two pathways.

Figure 8 shows the network produced by pathfinR active-subnetwork-oriented enrichment analyses. Thanks to this plot is possible to see how genes interact with each other and how they are connected to different pathways/processes. *Table 3* list the most relevant KEGG pathways that are reported in the network.

KEGG ID	Pathways
hsa05332	Graft-versus-host disease
hsa04144	Endocytosis
hsa04612	Antigen processing and presentation
hsa05330	Allograft rejection
hsa04911	Insulin secretions
hsa04728	Dopaminergic synapse
hsa04929	GnRH secretion
hsa04140	Autophagy
hsa04150	mTOR signaling pathway
hsa04360	Axon guidance

Table 3: Most relevant KEGG pathways identified by pathfinR.

The chart shows that, according to the P-value calculated by genefilter, only up-regulated genes

are linked to the most relevant pathways. The plot shows that some pathways (hsa05332, hsa04144, hsa04612 and hsa05330) are highly interconnected via genes that are: *HLA-G*, *HLA-E*, *HLA-C* and *HLA-DOA*. The others displayed pathways are not to be considered less important. For instance, *hsa04728* (Dopaminergic synapse) is known to be related to *hsa05012* that is the PD pathway[13]. Moreover, the defect in *Insulin secretion* (hsa04911) is the main cause of Type 2 Diabetes Mellitus, which has been pointed out to be associated to PD[1415]. Also, *Authophagy* (hsa04140) is known to be related to the common causes of PD[16] and up-regulation and down-regulation of *mTOR signaling* have been detected in PD model, hence it is not clear whether this pathway is neuroprotective or promote PD[17]. Lastly, there are evidence that *Axon guidance pathway* is involved in PD and in other brain disorders[18].

Discussion

LDA was the model that performed the best and in view of the accuracy, ROC curve and AUC, it is possible to state that it is promising. This result is also remarkable given that PD misdiagnosis is between 10 and 25%[19]. A prominent result regard Scudo. This method has indeed proven itself to be stable with respect to parameter tuning. For instance, only slight variation in accuracy were reported while training the model with the best parameters (ntop = 150, nbottom = 150 and alpha = 0.7) and with others (ntop = 200, nbottom = 200 and alpha = 0.5).

Graft-versus-host disease (GVHD) and Type I diabetes mellitus (DM-I) were the two most significant pathways identified by the functional enrichment analysis. While there are no literature results that link GVHD with PD, it is demonstrated that PD risk increases in diabetes mellitus patients[20]. Also, the pathway that involves MHC class 1 molecules are known to play an important role in neurodegenerative disease, including PD[21]. Epstein-Barr virus infection has been identified by DAVID and is indeed known to be related with PD[22].

Functional enrichment analysis has identified the importance of MHC class 1 molecules. Thanks to the visualisation of the interaction produced by network analysis it is possible to see how classic MHC class 1 molecules (HLA-C) and non-classic MHC class 1 molecules (HLA-G and HLA-E)[23] are highly interconnected to some pathways. Among these *hsa05330* (Allograft rejection) is associated to 3D domain swapping that was observed to play an important role in Alzheimer's, Parkinson's, and prion diseases[24]. In general, the results of the network enrichment analysis are in accordance with the one of the functional analysis. Indeed, GVHD, Allograft rejections and the importance of Diabetes Mellitus are present in both analyses. On the other hand, the Tissue specific analysis has not produced any interesting network, however it has stated the importance of some genes for brain related tissue.

Summing up, LDA has shown to be a valid classification tool that also allowed to identify an effective gene list. Furthermore, network and functional enrichment analysis, performed using KEGG database have produced highly overlapping results demonstrating the importance of certain pathways and the interaction with some genes/proteins.

In conclusion, it can be stated that, starting from a gene expression database composed by blood sample, it is possible to implement a classification model with an adequate level of accuracy and to gain biological insight regarding pathways and interaction between protein that are related to the disease.

Citations

[1] Kelly et al. Molecular Brain (2019) https://doi.org/10.1186/s13041-019-0436-5
[2] Balestino R. Schapira A.H.V (2019) Parkinson disease https://doi.org/10.1111/ene.14108
[3] Lee, SA., Tsao, T.TH., Yang, KC. et al. (2011) Construction and analysis of the protein-protein interaction networks for schizophrenia, bipolar disorder, and major depression https://doi.org/10.1186/1471-2105-12-S13-S20
[4] Gelb DJ, Oliver E, Gilman S. (1999) Diagnostic criteria for Parkinson disease. Arch Neurol
[5] Glaab E., Baudot A., Krasnogor N., Schneider R., Valencia A. (2012) EnrichNet: network-based gene set enrichment analysis. doi: 10.1093/bioinformatics/bts389.
[6] Martinez T., Fernando R., Jose I. (2020) Variable importance plot (mean decrease accuracy and mean decrease Gini) https://doi.org/10.1371/journal.pone.0230799.g002
[7] Dennis, G., Sherman, B.T., Hosack, D.A. et al. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. https://doi.org/10.1186/gb-2003-4-9-r60
[8] Hui G., Yu T., Qin H., Fan S., David S. (2015) Functional enrichment analysis of three Alzheimer's disease genome-wide association studies identities DAB1 as a novel candidate liability/protective gene https://doi.org/10.1016/j.bbrc.2015.05.044.
[9] Uku R., Liis K., Ivan K., Tambet A., Priit A., Hedi P., Jaak V. (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists doi:10.1093/nar/gkz369
[10] Ege U., Ozan O., Osman U. S. (2018) pathfindR: An R Package for Pathway Enrichment Analysis Utilizing Active Subnetworks https://doi.org/10.1101/272450
[11] Kumar R, Indrayan A. Receiver operating characteristic (ROC) curve for medical researchers. (2011) https://doi.org/10.1007/s13312-011-0055-4
[12] Christian K., Falk S. (2007) Dynamic exploration and editing of KEGG pathway diagrams https://doi.org/10.1093/bioinformatics/btl611
[13] KEGG documentation: https://www.genome.jp/dbget-bin/www_bget?hsa04728
[14] Linchao D., Lei F., Xiaodong X., Jinfei F., Yandong X. (2019) Identification of core genes and pathways in type 2 diabetes mellitus by bioinformatics analysis. doi: 10.3892/mmr.2019.10522
[15] Francesca F., Giuseppe P., Ilaria C., Serena C., Francesca C. P., Claudia M., Francesco B., Pietro F., Francesco O. (2019) The Relevance of Insulin Action in the Dopaminergic System. doi: 10.3389/fnins.2019.00868
[16] Melinda A. L. D., Kai M., Ke W., Mantong Z., Daniel J. K. (2012) The Role of Autophagy in Parkinson's Disease. doi: 10.1101/cshperspect.a009357
[17] Ai-ping L., Jun C., Yuliang Z., Zhifang C., Yi H. (2017) mTOR Signaling in Parkinson's Disease. https://doi.org/10.1007/s12017-016-8417-7
[18] Timothy G. L., Spiridon P., Deborah C. M., Jarlath F., Lina S., Mariza de A., John R. H., Walter A. R., Eric A., Demetrius M. M. (2007) A Genomic Pathway Approach to a Complex Disease: Axon Guidance and Parkinson Disease. https://doi.org/10.1371/journal.pgen.0030098
[19] J. Clin, Invest Diagnosis and treatment of Parkinson disease: molecules to medicine (2006) https://doi.org/10.1172/JCI29178.
[20] Abdallah H., Rajan S. K., Rohi M., Jeevan G., Amer A., Nusrat J. (2020) Diabetes Mellitus and Parkinson's Disease: Shared Pathophysiological Links and Possible Therapeutic Implications. doi:10.7759/cureus.9853
[21] Carolina C., john D. L., David S. (2014) Neuronal MHC-I expression and its implications in synaptic function, axonal regeneration and Parkinson's and other brain diseases. https://doi.org/10.3389/fnana.2014.00114
[22] Haeman J., David A. B,, Robert G. W,, Richard J. S. (2008) Viral Parkinsonism doi: 10.1016/j.bbadis.2008.08.001
[23] Halenius, A., Gerke, C. & Hengel, H. (2015) Classical and non-classical MHC I molecule manipulation by human cytomegalovirus: so many targets—but how many arrows in the quiver? https://doi.org/10.1038/cmi.2014.105
[24] Khader S., Ramanathan S. (2012) Functional repertoire, molecular pathways and diseases associated with 3D domain swapping in the human proteome. doi: 10.1186/2043-9113-2-8