Machine learning techniques to discover genes with potential prognosis role in Alzheimer's disease using different biological sources

María Martínez-Ballesteros [a,*], José M. García-Heredia [b], Isabel A. Nepomuceno-Chamorro [a], José C. Riquelme-Santos [a]

[a] Dpto. Lenguajes y Sistemas Informáticos, Universidad de Sevilla, Seville, Spain
[b] Dpto. Bioquímica Vegetal y Biología Molecular, Universidad de Sevilla, Seville, Spain

Keywords:

Association rules

Gene expression profiles

Alzheimer's disease

Statistical significant genes

Ensemble learning

Biological knowledge integration

A B S T R A C T

Alzheimer's disease is a complex progressive neurodegenerative brain disorder, being its prevalence expected to rise over the next decades. Unconventional strategies for elucidating the genetic mechanisms are necessary due to its polygenic nature. In this work, the input information sources are five: a public DNA microarray that measures expression levels of control and patient samples, repositories of known genes associated to Alzheimer's disease, additional data, Gene Ontology and finally, a literature review or expert knowledge to validate the results. As methodology to identify genes highly related to this disease, we present the integration of three machine learning techniques: particularly, we have used decision trees, quantitative association rules and hierarchical cluster to analyze Alzheimer's disease gene expression profiles to identify genes highly linked to this neurodegenerative disease, through changes in their expression levels between control and patient samples. We propose an ensemble of decision trees and quantitative association rules to find the most suitable configurations of the multi-objective evolutionary algorithm GarNet, in order to overcome the complex parametrization intrinsic to this type of algorithms. To fulfill this goal, GarNet has been executed using multiple configuration settings and the well-known C4.5 has been used to find the minimum accuracy to be satisfied. Then, GarNet is rerun to identify dependencies between genes and their expression levels, so we are able to distinguish between healthy individuals and Alzheimer's patients using the configurations that overcome the minimum threshold of accuracy defined by C4.5 algorithm. Finally, a hierarchical cluster analysis has been used to validate the obtained gene-Alzheimer's Disease associations provided by GarNet. The results have shown that the obtained rules were able to successfully characterize the underlying information, grouping relevant genes for Alzheimer Disease. The genes reported by our approach provided two well defined groups that perfectly divided the samples between healthy and Alzheimer's Disease patients. To prove the relevance of the obtained results, a statistical test and gene expression fold-change were used. Furthermore, this relevance has been summarized in a volcano plot, showing two clearly separated and significant groups of genes that are up or down-regulated in Alzheimer's Disease patients. A biological knowledge integration phase was performed based on the information fusion of systematic literature review, enrichment Gene Ontology terms for the described genes found in the hippocampus of patients. Finally, a validation phase with additional data and a permutation test is carried out, being the results consistent with previous studies.

1. Introduction

Neurodegenerative diseases are complex syndromes characterized by a common feature: up to now, all of them progress inexorably. Although medical treatments can slow symptoms' progression, there is no cure for any of them. One of the most common neurodegenerative diseases, constituting approximately 70% of all cases [1], is Alzheimer's Disease (AD). This multifactorial and heterogeneous disorder, characterized by a progressive loss of memory and a decline in cognitive function, affects to around 10% of people between 65–85 years, increasing its risk significantly with age, reaching percentages of up to 50–60% of people over 85 [2,3]. Although AD exhibits two markedly brain histological

* Corresponding author.
E-mail addresses: mariamartinez@us.es (M. Martínez-Ballesteros), jmgheredia@us.es (J.M. García-Heredia), inepomuceno@us.es (I.A. Nepomuceno-Chamorro), riquelme@us.es (J.C. Riquelme-Santos).

characteristics, plaques and tangles accumulation [4,5], the presence of one or both symptoms are not a proof of AD development. Indeed, around a 30% of normal aged people have similar levels of plaques than AD patients of similar ages [6,7]. The main cause of AD may be genetic. In fact, up to a 5% of AD cases are due to genetic inheritance, being responsible of the appearing of an early onset AD [8]. In these cases, mutations in three different genes (*APP, PSEN1* and *PSEN2*) have been connected to pathophysiology of the disease [8–11].

Assuming its genetic background, due to changes in gene regulation or mutations accumulated through lifetime, it is important to know which genes are usually altered in AD patients, mainly in its initial states, in order to design treatments that effectively slow down -or even stop- AD progression. In this context, Data Mining and statistical techniques have been applied to find useful data patterns in Bioinformatics and Biomedicine fields to discover, for example, affected pathways in a specific syndrome or disease [12]. Specifically, several works have been published where such techniques have been focused on AD using gene expression profiles. These profiles are characterized by a low number of samples (patients) but a very high number, up to thousands, of features (gene expression) [13]. Example of data mining technique focused on AD are two types of non-supervised methods (principal component analysis and independent component analysis) that have been applied to extract and characterize the most relevant features from DNA microarray gene expression data of AD [14].

However, most of techniques used in the literature to increase AD knowledge present a low-dimensional solution. Thus, they are not sufficiently descriptive or the information provided is limited. Association Rules (AR) and particularly Quantitative Association Rules (QAR) [15] have emerged as a popular methodology to discover significant and apparently hidden relationships among attributes in a subspace of the dataset instance. The application of QAR in AD-related DNA microarray data analysis might provide a deeper knowledge into biological functions with higher relevance, since they can be used to identify gene expression patterns, helping to find which of them are associated with AD. In the context of this work, QAR are used fixing the consequent of the rule to the AD state (healthy or not) with the aim of addressing the problem as a rule-based classification.

The purpose of this work is to present an ensemble of decision trees, rules and hierarchical cluster to analyze microarray gene expression data related to AD. Usually, ensembles of classifiers are used to obtain better predictive performance. Here, we proposed an ensemble of classifiers to obtain the best configuration settings of the evolutionary algorithm GarNet [16] to identify genes highly related to AD, through changes in their expression levels between healthy and AD samples. This algorithm requires the parametrization of the objectives to be optimized and the minimum threshold of some quality measures as many other algorithms do. With the aim of obtaining the most suitable parameters of GarNet, the well-known C4.5 algorithm has been applied to the gene expression data provided in [17] with the objective of setting the minimum accuracy threshold to be satisfied by the rules obtained by GarNet. These configurations have been used to rerun GarNet to find significant genes with potential prognosis role in AD. Furthermore, a hierarchical cluster analysis has been also applied to group healthy and AD patients using the genes obtained by GarNet. It can be noted that the use of prior knowledge can outperform our approach and increase the possibilities of correcting the spurious information existing in high-throughput technology data as gene expression data. Then, a phase of biological knowledge was performed including other biological sources such as systematic review of the literature in PubMed, Gene Ontology processes and protein-protein interaction networks. Thus, the systematic search in PubMed allowed us to detect altered functions specifically in neurons of patients affected by AD. These functions include genes that encode proteins connected to metabolism, cell signaling or protein-nucleic acid interaction, suggesting an effect on overall cell functions. In addition, enrichment analysis in the context of Gene Ontology and a mapping process to network protein-protein interactions allowed us to strengthen the relationship of the discovered genes with AD. Using this ensemble learning and fusion strategy, we have found more than 90 genes whose expression is modified during AD progression, affecting processes as diverse as lipid metabolism, transcriptional regulation or protein synthesis. In addition, some of the genes are connected to cardiovascular diseases or diabetes, disorders previously related to AD. The identified genes show the complexity of AD, and could be used to design preventive treatments in healthy people before the appearance of the first AD symptoms.

The remainder of the paper is structured as follows. Section 2 presents a brief introduction about AD, summarizes the main concepts of AR and quality measures, overviews the most relevant concepts in classification and briefly presents the C4.5 algorithm. Section 3 thoroughly describes the six phases integrated in our approach to identify genes highly related to AD including the main features of GarNet algorithm. Section 4 presents and discusses the results obtained in each phase of the proposed analysis. Finally, Section 5 summarizes the conclusions drawn from the analysis conducted.

2. Preliminaries

This section provides a brief description of AD, followed by main concepts of AR, QAR and quality measures, in addition to classification, decision trees and the well-known C4.5 algorithm.

2.1. Alzheimer disease

As it has been stated in the introduction, plaques and tangles are common features in AD. The first are due to progressive extracellular deposition of amyloid $\beta(A\beta)$ peptides, due to an inadequate clearance that produces the formation of plaques. In fact, plaques accumulation can be detected in non-affected individuals even more than twenty years before the appearance of AD symptoms [18–20]. Neurofibrillary tangles are formed by the accumulation of hyperphosphorylated tau protein inside neurons, being its levels up 2-fold higher that in normal brain, affecting its normal function [21]. This post-translational alteration affects tau protein normal function. In normal conditions, tau protein interacts with tubulin, promoting the microtubule assembly. However, under hyperphosphorylated state, tau protein cannot interact with tubulin, being capable to form tau helical filaments with no activity.

Although some authors have considered that plaques and/or tangles may be considered the initiating event of AD [8,22], it is not clear if they are the cause or just symptoms of this neurodegenerative disorder, remaining its main cause elusive. It can be considered as the junction of multiple imbalances, from mitochondrial dysfunction to oxidative stress, neuroinflammation and changes in gene regulation. The last one could be considered as the hard core of the disease, due to its role in other effects. In this way, formation of $A\beta$ plaques and tau tangles are caused by the dysfunction of proteins associated to their normal processing. In addition, mitochondrial malfunction and oxidative damage can also be connected to gene regulation, due to a decrease in the synthesis of proteins from the oxidative stress pathway. This fact underlines the importance of developing tools to find common patterns in AD.

2.2. Association rule mining and quality measures

The AR learning is a popular and well-known method in the Data Mining field used to discover interesting relationships among

variables in large databases [23]. AR aim at identifying patterns that explain or summarize the data, instead of predicting the class of new data [24].

When the domain is continuous, the AR is known as QAR. In this context, let $F = \{F_1, ..., F_n\}$ be a set of features or attributes, with values in $\mathbb{R}$. Let A and C be two disjoint subsets of F, that is, $A \subset F$, $C \subset F$, and $A \cap C = \emptyset$. A QAR is a rule $X \Rightarrow Y$, in which features in A belong to the antecedent X, and features in B belong to the consequent Y, such that X and Y are formed by a conjunction of multiple boolean expressions of the form $F_i \in [v_1, v_2]$, (with $v_1, v_2 \in \mathbb{R}$). Thus, in a QAR, the features or attributes of the antecedent are related with the features of the consequent, establishing a membership value interval for each attribute involved in the rule. For example, a QAR could be numerically expressed as $EPHA10 \in [2, 2.9] \wedge TOR2A \in [1.8, 2.6] \Longrightarrow STRN4 \in [3, 3.5]$ where *EPHA*10 and *TOR2A* constitute the features appearing in the antecedent and *STRN*4 the one in the consequent.

There are several probability-based measures proposed in the literature to evaluate the generality and reliability of AR (and QAR) obtained in the mining process [25,26]. In this work, we have used the support, confidence, leverage, accuracy and gain measures to optimize and evaluate the quality of the QAR obtained by GarNet. The description and the mathematical definition can be found in [27].

Methods based on QAR have not been used to find gene associations in AD, although the technique has been used to analyze gene expression data [28] and other AD features [29]. In the first work, QAR has been used to analyze 23 genes known to be involved in arginine metabolism from yeast organism. In the second work, the authors combined a computer aided diagnosis with continuous attribute discretization and association rule mining for the early diagnosis of AD based on emission computed tomography images. Alternatively, Ponmary et al.used association rule mining to find AD patterns of amino-acid residues in the protein binding site of enzymes which has been described to have a role in AD [30].

It is noteworthy that the AD state (healthy or AD patient) has been fixed to appear in the consequent of the rules handled in this work. Thus, a rule is composed of a set of genes belonging to an interval (expression levels) in the antecedent and the AD state in the consequent. An example of the rules found is as follows: $EPHA10 \in [2.0, 2.9]$ and $STRN4 \in [3.0, 3.5] \Longrightarrow$ AD state is 1 (not healthy).

2.3. Classification and ensembles

In Machine Learning and the statistics, the classification goal is to identify to where a new observation belongs in a set of categories. It is important to consider a training set of data composed of observations or instances whose category membership is known. A wide range of classification algorithms can be found in the literature, for instance, Decision Trees, K-nearest Neighbor classifier, Bayesian Network, Neural Networks, Fuzzy Logic, Support Vector Machine, Boosting, etc.

Decision Trees [31] are one of the most frequently used in the literature. There are many specific decision-tree algorithms, the most common are, among others, ID3 [32], C4.5 [33] and CART [31]. The C4.5 algorithm [33] (and its predecessor ID3) is one of the most well-known and used decision trees.

The model overfitting in the training dataset is considered as one of the main drawbacks of classification methods. Then, the use of additional techniques such as cross-validation, regularization or pruning, among others, is necessary. Specifically, we have applied a commonly cross-validation method named k-fold cross-validation. The k-fold cross-validation is a frequently used model validation usually used to evaluate the results obtained by a Data Mining technique, specifically predictive techniques, with the aim at ensuring that the results can be generalized to an independent dataset.

A detailed explanation of this common accuracy estimation method can be found in [34].

Ensembles of classifiers enhance the performance of simple classifiers combining the outputs of several others. In the literature, we can find many examples of ensemble systems such as [35,36]. The work presented in [37] proposes a supervised learning approach to the ensemble clustering of genes using known gene-gene interaction data to improve the results for already commonly used clustering techniques. In the context of the AD problem, an ensemble based in data fusion approach for early diagnosis of AD was proposed [38]. A two stage sequential ensemble is described in [39] to perform the classification of AD based on magnetic resonance imaging features. Most of existing ensembles of the literature are devoted to improve the predictive performance of a single classifier, whereas we use an ensemble of classifiers to obtain the best configuration settings of other classifier.

3. Methodology

This section presents the main features of the techniques used to identify genes highly related to AD. The process conducted to detect genes with potential prognosis role in AD is mainly performed through the discovery of AR, in particular QAR, in six phases as can be observed in Fig. 1.

In general terms, a public microarray dataset obtained by laser capture microdissection from the entorhinal cortex was analyzed. The C4.5 algorithm has been applied to consider as minimum threshold the accuracy obtained in the model (first phase). The GarNet algorithm was applied using multiple configuration settings and a set of QAR with AD state fixed as consequent part was obtained for each experiment (second phase). The best experiments were selected in terms of accuracy measure in test sets using the accuracy obtained by C4.5 algorithm as a minimum threshold. GarNet algorithm was rerun using the selected configurations and the set of genes with potential prognosis role in AD were detected. Then, we obtained the top of frequent genes joining all gene-AD state pairs found in the rules (third phase). Gene groups were validated by hierarchical cluster analysis, in addition to statistical and biological significance validation techniques (fourth phase). A biological knowledge integration phase was performed (fifth phase). Finally, the results obtained were validated using additional data (sixth phase). The main features of each phase are detailed in the following Sections 3.1–3.6, respectively.

3.1. First phase - decision trees

A well-known Machine Learning technique based on decision trees, frequently used to tackle the AD problem, was selected as a benchmark to filter the results obtained by GarNet. Decision trees, also known as classification trees or regression trees, are commonly used as predictive modelling approaches in statistics, Data Mining and Machine Learning. A wide range of decision-trees algorithms can be found in the literature. Specifically, the well-known C4.5 algorithm [33] has been selected to classify between healthy and AD samples due to the high performance usually presented [40].

The average percentage of instances correctly classified obtained by this technique was used as a minimum threshold to select the best configurations of GarNet to perform the second phase of the process.

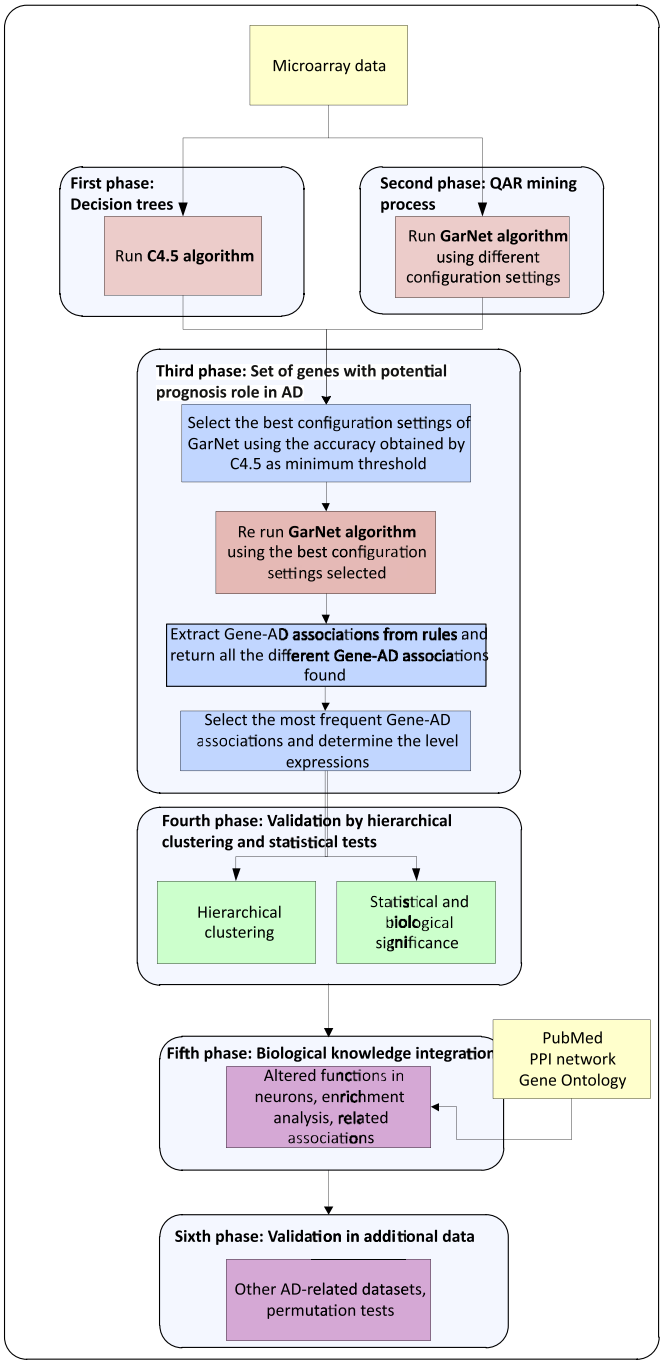

Fig. 1. Overview of general process.

3.2. Second phase - QAR mining process

GarNet [16] is a multi-objective algorithm based on the NSGA-II algorithm and it is able to find QAR in datasets with continuous attributes avoiding the discretization step. This algorithm aims at solving the main drawbacks caused by a fitness function based on a weighted objective scheme, trying to perform the best trade-off among all the measures optimized. GarNet was executed in this work to address the second phase of the analysis in order to obtain a gene subset highly related with AD. A thorough description of the algorithm can be found in the research work proposed in [16]. Additionally, the main features of GarNet are summarized below.

GarNet uses adaptive intervals instead of fixed ranges to group samples whose features share certain sets of values in continuous domain. The search for the most appropriate intervals has been carried out by means of an evolutionary process. In this process, the intervals are adjusted to find QAR with high interpretability, generality, quality and precision.

In the population, each individual constitutes a rule, being subjected to an evolutionary process, in which the genetic operators are applied. Furthermore, the instances already covered by the rules are penalized to emphasize the covering of instances not covered yet. Then, those samples covered by a few rules have a higher priority to be selected in order to generate the new population [41]. The evolutionary process ends when the number of generations is reached. The whole evolutionary process is repeated until the desired number of rules is achieved. Finally, GarNet returns the set of QAR found.

The GarNet algorithm was executed modifying the minimum QAR confidence threshold obtained and second, optimizing different group of measures following the study proposed in [27,42]. Each minimum confidence threshold combined with each group of measures comprise an experiment. Support, confidence, leverage, gain and accuracy were the selected measures optimized and different groups of 3 measures were composed.

Each experiment (each confidence minimum threshold in combination with each group of optimized measures) is executed several times using 5-fold cross-validation on the studied dataset to ensure that the performance of GarNet is stable and accurate. For each experiment, we have mined a model with a defined number of QAR in which the AD state has been fixed as consequent part of the rules and the average global accuracy (validation measure) of the model has been calculated to assess the percentage of correctly classified instances of the dataset [43].

Furthermore, we have computed the average percentage of instances incorrectly classified and the average percentage of instances not classified, that is, the instances not satisfied by any QAR of the model.

3.3. Third phase - selection process of genes with potential prognosis role in AD

After completing the first and second phase, we selected those experiments (configuration settings) in which GarNet obtains a higher accuracy value than C4.5 algorithm in the test set using 5-fold cross validation. Then, GarNet was executed again using such configuration settings in the original dataset (without using 5-fold-cross-validation) in order to obtain a set of QAR that provide information of the entire dataset. Subsequently, all the gene-AD state associations were extracted for each corresponding set of rules. Finally, we selected the most frequent gene-AD state only considering the best rules-based models in order to find potential and relevant genes highly related with AD. This process aims at discovering the most frequent and strongest associations between genes and AD avoiding those relations that occur by chance. Furthermore, the most frequent gene-gene associations among the selected AD-related genes have been extracted from the obtained rules.

To conduct this process, the best rules set obtained for each configuration setting selected were splitted into sets of attribute pairs as follows:

- First, genes belonging to the antecedent for each resulting rule were identified. Note that the AD was always fixed as a consequent of the rules. AD state can be 0 or 1, that is, 0 is used to define a healthy patient, while 1 denotes AD patients.
- Afterwards, combinations between the genes of the antecedent and the consequent (AD stage) of each rule were performed obtaining gene-AD and gene-gene pairs.
- For instance, let the following QAR be:

$$EPHA10 \in [2, 2.9] \text{ and } TOR2A \in [1.8, 2.6] \Longrightarrow \text{AD stage is } 1$$

The resulting gene-AD pairs (associations) that can be extracted from this rule are:

*EPHA*10 $\Longrightarrow$ AD stage is 1

TOR2A $\Longrightarrow$ AD stage is 1

The gene-gene associations found in this rule are:

*EPHA*10 $\Longrightarrow$ *TOR2A*

After completing the inference process of GarNet for each configuration setting selected (*K* number of configurations), the union among all the gene-AD pairs found was performed to find the most frequent gene-AD associations, hence, potential and relevant associations and, finally, gene-gene associations between these genes. Let *K* be the number of configuration settings used in the inference process. Let Π be the set of gene-AD pairs and gene-gene pairs obtained from the *k*th-configuration. The inference process output is defined as:

$$\Pi = \Pi_1 \cup \Pi_2 \cup ... \Pi_K \quad (1)$$

where Π_k, $k = 1..K$, is the set of gene-AD and gene-gene pairs obtained from the *k*th-configuration.

The final step involved the selection of the most frequent gene-AD associations from the union of the results obtained for the *K* configuration settings selected. Afterward, the most frequent gene-gene associations among these genes are found. Note that the configuration settings to perform the gene-AD extraction were selected taking into account a minimum threshold of accuracy using a benchmark method. Specifically, the well-known C.45 algorithm was applied. Finally, expression level changes between healthy and AD patients of the selected gene sets were determined through the intervals of the QAR in which the genes were involved.

An example of how the decomposition process of a QAR set into gene-AD and gene-gene pairs is defined as follows. Let the following set of QAR obtained for 2 configuration settings be:

1. **Gene-AD associations extraction**
   - **Configuration setting 1:**
     - *EPHA*10 $\in$ [2, 2.9] and *TOR2A* $\in$ [1.8, 2.6] $\Longrightarrow$ AD stage is 1
     - *FAM*158*A* $\in$ [2.7, 3.0] $\Longrightarrow$ AD stage is 1
     - *ALOX*12*B* $\in$ [2.1, 2.4] and *VILL* $\in$ [2.1, 2.3] $\Longrightarrow$ AD stage is 1
     - *PRPF*40*b* $\in$ [2.7, 3.0] and *DLGAP*2 $\in$ [2.6, 2.8] and *TAOK*2 $\in$ [2.9, 3.3] $\Longrightarrow$ AD stage is 1

     Π_1 = {*EPHA*10 - *AD*, *TOR2A* - *AD*, *FAM*158*A* - *AD*, *ALOX*12*B* - *AD*, *VILL* - *AD*, *PRPF*40*B* - *AD*, *DLGAP*2 - *AD*, *TAOK*2 - *AD*, *EPHA*10 - *TOR2A*, *ALOX*12*B* - *VILL*, *PRPF*40*b* - *DLGAP*2, *PRPF*40*b* - *TAOK*2, *DLGAP*2 - *TAOK*2}
   - **Configuration setting 2:**
     - *EPHA*10 $\in$ [2.0, 2.9] and *STRN*4 $\in$ [3.0, 3.5] $\Longrightarrow$ AD stage is 1
     - *FAM*158*A* $\in$ [2.7, 3.0] and *RPS*8 $\in$ [9.4, 11.6] $\Longrightarrow$ AD stage is 1
     - *ALOX*12*B* $\in$ [2.1, 2.4] and *PBX*4 $\in$ [2.0, 2.4] $\Longrightarrow$ AD stage is 1

     Π_2 = {*EPHA*10 - *AD*, *STRN*4 - *AD*, *FAM*158*A* - *AD*, *RPS*8 - *AD*, *ALOX*12*B* - *AD*, *PBX*4 - *AD*, *EPHA*10 - *STRN*4, *FAM*158*A* - *RPS*8, *ALOX*12*B* - *PBX*4}

     ▶**Output $\Pi = \Pi_1 \cup \Pi_2$** = {*EPHA*10 - *AD*, *TOR2A* - *AD*, *FAM*158*A* - *AD*, *ALOX*12*B* - *AD*, *VILL* - *AD*, *PRPF*40*B* - *AD*, *DLGAP*2 - *AD*, *TAOK*2 - *AD*, *STRN*4 - *AD*, *RPS*8 - *AD*, *PBX*4 - *AD EPHA*10 - *TOR2A*, *ALOX*12*B* - *VILL*, *PRPF*40*b* - *DLGAP*2, *PRPF*40*b* - *TAOK*2, *DLGAP*2 - *TAOK*2, *EPHA*10 - *STRN*4, *FAM*158*A* - *RPS*8, *ALOX*12*B* - *PBX*4}

     ▶ **Expression levels:** Expression level changes of AD patient genes were calculated considering the lower and upper bounds of the QAR obtained by GarNet and the average values of control patients. If the interval bounds of the genes in the rules are below the average value of control patients, the gene is down-regulated in AD patients. Otherwise, the gene is up-regulated. The average values of control and AD patients, in addition to the lower and upper bound of each gene in the QAR detailed in the aforementioned example and the expression level obtained, are presented in Table 1.

Table 1
Example of how the gene expression levels are calculated.

Gene symbol	Avg. expression level		QAR intervals		Regulation
	Control	AD	Lower bound	Upper bound	
EPHA10	3.8	2.3	2.0	2.9	Down-regulated
TOR2A	3.5	2.6	1.8	2.6	Down-regulated
FAM158A	4.2	2.8	2.7	3.0	Down-regulated
ALOX12B	3.1	2.3	2.1	2.4	Down-regulated
VILL	3.1	2.2	2.1	2.3	Down-regulated
PRPF40B	3.0	2.9	2.7	3.0	Down-regulated
DLGAP2	3.1	2.7	2.6	2.8	Down-regulated
TAOK2	3.8	3.1	2.9	3.3	Down-regulated
STRN4	5.0	3.4	3.0	3.5	Down-regulated
RPS8	9.0	10.6	9.4	11.6	up-regulated
PBX4	2.5	2.2	2.0	2.4	Down-regulated

3.4. Fourth phase: validation by hierarchical cluster analysis and statistical tests

The fourth phase is devoted to assess the quantitative significance of the genes related to AD obtained by GarNet from different a perspective. To fulfill this goal, hierarchical cluster analysis and statistical tests were used.

3.4.1. Hierarchical cluster analysis

A hierarchical cluster analysis was conducted to cluster patients and genes using Spearman and Pearson correlation, respectively. This method is used to assess the capability of the genes obtained by GarNet to classify between healthy and AD patients according to changes in the expression levels (down-regulated or up-regulated). Alternatively, the hierarchical cluster analysis performed for the genes reported in the third phase is used to validate expression levels changes detected by the rules intervals obtained by GarNet.

3.4.2. Statistical and biological significance

The Mann-Whitney *U*-test was applied to determine whether there is a statistically significant difference between expression levels of the reported genes in the previous phase in healthy samples and AD affected samples.

Furthermore, the bioinformatics open source software named *bioconductor*, of the well-known R software environment was used to calculate the volcano plot that visualizes the changes (fold-change) versus the statistical significance (*p*-value) of the expression levels of the genes selected by GarNet. The volcano plot has been plotted using the log-fold change and *p*-value of the top-ranked genes obtained from a linear model fit of the selected genes. The linear model fit was obtained by the *lmFit* function, and top-ranked genes were extracted by the *topTable* function, both included in the *limma* package of bioconductor (version 3.2). The *p*-value probability of a differentially expressed gene was computed through the Benjamin and Hochberg statistical tests [44]. The higher the negative log10 for each gene, the higher the probability that the gene is differentially expressed and not a false positive. The x-axis indicates the log2 value of fold-change between the two conditions.

3.5. Fifth phase: biological knowledge integration

The fifth phase consisted on the integration of the biological knowledge using different biological sources. First, the altered functions in neurons affected by AD were detected by a systematic review of the literature using the well-known PubMed, a free web literature search service developed and maintained by the National Center for biotechnology Information (NCBI) [45].

Then, an enrichment analysis of the found genes was performed in the context of Gene Ontology, a structured, controlled vocabularies and classifications that cover several domains of molecular and cellular biology, using Fatigoo tool [46] which is integrated in Babelomics 5 analysis suite.

The gene set reported by GarNet was also mapped onto the largest network of protein interactions related to Alzheimer's referred to as *AD PPI network*. This network was reported by Soler et al. in [47] where 12 well known AD associated genes were selected from OMIM Database as seed. Through yeast two-hybrid matrix screen and two-hybrid library screen, authors generated interaction core set containing all the confirmed library and matrix interactions (200 interactions between 74 nodes including seed-seed, seed-candidate and candidate-candidate interactions). This network was merged with direct interactions of seed AD genes extracted from several repositories (IntAct, DIP, MINT and HPRD). As a result, Soler et al. reported an AD network with 1704 nodes and 5881 interactions. In our analyses, the top genes were reported in official gene symbols and were converted to Uniprot accession numbers using DAVID tool [48].

Furthermore, we have analyzed which genes of those obtained by GarNet are associated to cerebral diseases using the well-known MalaCards database, in particular GeneCards [49]. This is an integrated database of all known predictable human genes, including information about diseases connected with each gene.

3.6. Sixth phase: validation using additional data

The last phase is devoted to validate the strength of the results obtained by GarNet in the previous phases.

To fulfill this goal, several AD-related datasets were used to check the significance of the subset of genes provided by GarNet. In particular, the gene regulation levels were calculated for six datasets collected from NCBI repository [45]. Then, the regulation level of each gene in the original dataset was compared with that presented for the six datasets.

Additionally, permutation tests have been applied running the same analysis performed by GarNet multiple times. Specifically, two versions of the original dataset have been generated shuffling the class labels (AD and control patients) of the instances. The percentage distribution of instances of AD and control patients was the same as the original dataset.

Then, the second and the third phases of our methodology (Sections 3.2 and 3.3) were applied in the two random versions using the best configuration settings (#8, #13 and #14) described in Section 3.2.

Finally, the results obtained from the permutation tests were compared with those reported for the original dataset in order to check if the accuracy level and the genes selected are the same.

4. Results and discussion

The dataset used to perform the analysis proposed is described in Section 4.1. Furthermore, the results obtained for the six phases are presented and discussed in the following Sections 4.2–4.7, respectively.

4.1. Alzheimer dataset

The dataset used as training data was retrieved from the gene expression data analysis described in [17]. The original dataset was provided by Dunckley et al. [50] in which 1000 neurons were collected from each of the 33 samples by laser capture microdissection from entorhinal cortex. This single-cell gene expression data is formed by 33 samples and 35,722 probesets. The data were normalized by gcRMA [51] and the resulting probesets were mapped to genes by DAVID [52]. Specifically, the 13 normal controls correspond to the Braak stages 0–II and the average age of patients was 80.1 years. Regarding the AD affected samples, they belong to Braak stages III-IV considered as 'incipient' AD and the average age of patients was 84.7. We run our approach on a subset of 1663 genes obtained from this dataset. These 1663 genes were the result of preprocessing the data set of [50] as described in [17].

4.2. First phase results - decision trees

This section details the results obtained by the selected benchmark method. The decision tree built by C4.5 using 5-fold-cross-validation over the studied dataset is shown below:

$$NDRG2 <= 2.52 : Healthy$$
$$NDRG2 > 2.52 : AD$$

The first condition, $NDRG2 <= 2.52$, corresponds to the control or healthy samples and the second condition, $NDRG2 > 2.52$ determines if the samples presents AD. The 87.87% of samples was correctly classified. In particular, the 92.3% and 85% of control and AD samples, respectively, were correctly classified. Although the rate of samples correctly classified is close to 90%, the obtained model by C4.5 provides very poor information since that only one gene appears in the final model.

The *NDRG2* gene was removed from the dataset, being the C4.5 algorithm rerun to obtain a different decision tree. In this case, the decision tree built by C4.5 removing this gene was as follows:

$$SMARCD2 <= 3.296 : Healthy$$
$$SMARCD2 > 3.296 : AD$$

It can be observed that if we desire to achieve a model including a larger set of genes, it is necessary to remove the gene appearing in the decision tree and rerun the C4.5 algorithm. In contrast, AR-based methods can obtain models comprising a high number of genes that may provide useful and relevant knowledge for experts. In order to perform the second phase of the experimentation, the accuracy obtained by C4.5 algorithm was taken as minimum threshold to select the best configurations of GarNet.

4.3. Second phase results: QAR mining process

Several experiments were carried out during the second phase to assess the performance of GarNet using multiple configuration settings with the aim at achieving the most optimal solutions for the problem at hand in this work.

This section details the sensitivity study combining multiple confidence thresholds and measures of the results achieved by GarNet. Each minimum confidence threshold in combination with each group of optimized measures comprises an experiment. Each experiment was executed 100 times using 5-fold cross-validation. The minimum thresholds for the confidence measure to be achieved by the rules were 0.8, 0.9 and 1, respectively. Alternatively, we selected the support, confidence, leverage, gain and accuracy measures and different groups of 3 measures have been used in the optimization of GarNet. These measures evaluate different features of QAR. For instance, support and confidence are

Table 2
Configuration settings used by GarNet and results obtained applying 5-fold-cross validation.

ID	Min. conf.	Groups of measures optimized					Instances (%)	
		Leverage	Confidence	Gain	Support	Accuracy	Classified	Misclassified
#1	0.8	✓	✓	✓			80.88	18.98
#2	0.8		✓	✓	✓		80.45	19.26
#3	0.8		✓	✓		✓	87.13	11.26
#4	0.8	✓	✓			✓	85.15	14.37
#5	0.8				✓	✓	77.08	22.62
#6	0.9	✓	✓	✓			82.45	17.34
#7	0.9		✓	✓	✓		85.00	14.24
#8	0.9		✓	✓		✓	**87.80**	10.40
#9	0.9	✓	✓			✓	86.92	11.83
#10	0.9				✓	✓	72.82	20.73
#11	1	✓	✓	✓			85.15	14.18
#12	1		✓	✓	✓		85.01	12.44
#13	1		✓	✓		✓	**88.80**	7.80
#14	1	✓	✓			✓	**89.03**	8.26
#15	1				✓	✓	60.21	10.35

devoted to assess the generality and the reliability of the rules, respectively. Note that confidence measure was always included in the group of measures optimized by GarNet.

Table 2 shows the different configuration settings used to evaluate the performance of GarNet and the results obtained for each one. The configuration number of each experiment is identified in the first column. Second column details the different values for the minimum confidence threshold. Third column provides the group of quality measures optimized by GarNet in each experiment. For each configuration, the objectives optimized by GarNet are marked. Finally, the last column summarizes the results achieved by GarNet according to the percentage of instances classified correctly and percentage of instances not correctly classified, respectively, after performing the 100 executions using 5-fold-cross validation. The remaining percentage of instances until 100% were not classified instances. Note that only the average results obtained in the test set are presented in this table. The best results in terms of instances correctly classified are highlighted in bold. Specifically, we have selected those configuration settings of GarNet that achieve a percentage of correctly classified instances higher or equal than 87.8% according to the results obtained by C4.5.

As can be observed, the group of measures composed of confidence, gain and accuracy optimized by GarNet is the best one, regardless the minimum confidence threshold used (configuration settings #3, #8 and #13). However, the best results in terms of instances correctly classified are those obtained by the group of measures leverage, confidence and accuracy when the minimum confidence threshold used is set to 1 (configuration setting #14). In contrast, the worst results werw obtained by the support and accuracy measures when the minimum confidence threshold used is also set to 1 (configuration setting #15).

4.4. Third phase results: genes with potential prognosis role in AD

The results obtained by GarNet using the configuration settings #8, #13 and #14 (Table 2), were selected to perform the third phase of the experimentation regarding the accuracy achieved by C4.5.

Table 3 shows the top 100 frequent genes sorted by frequency and their expression levels appearing in the rules obtained by GarNet using the aforementioned configuration settings.

The regulation of each gene was obtained according to the intervals of the rules discovered by GarNet. Green and purple colors are used to represent if the gene is downregulated or upregulated in AD, respectively. *Gene Sym.* column shows the Gene Symbol notation of each gene. *Freq. column* displays the normalized frequency of the presented top frequent genes. The value 1 represents the most frequent gene. Note that duplications were removed since some genes are frequent for both AD and control patient, thus, the final list is composed of 95 genes.

Remarkably, 19 genes are up-regulated and 84 are down-regulated. Furthermore, the group of genes obtained by C4.5, *NDRG2* (ID 21) and *SMARCD2* (ID 82) also appear in the set of genes obtained by GarNet, demonstrating, the robustness of the algorithm.

The resulting network from the most frequent gene-gene interactions extracted using the selected set of AD-related genes is shown in Fig. 2. Note that green and purple colors are used to identify the downregulated and upregulated genes, respectively, in AD.

4.5. Fourth phase results: validation by hierarchical cluster analysis and statistical tests

This section presents and discusses the results obtained in the hierarchical cluster analysis and the statistical and biological validation.

4.5.1. Hierarchical cluster analysis

Fig. 3 displays the heatmap of control and AD patients according to the top genes under study after applying the hierarchical cluster analysis. Note that data have been scaled and centered. The results are plotted as dendrograms. Spearman correlation and Pearson correlation have been used to cluster the columns (patients) and rows (genes), respectively. The resulting tree has been cut at specific height (1.5) with its corresponding clusters highlighted in the heatmap color bar.

Four groups can be observed according to the patient type and gene expression levels. It is noteworthy that set of genes obtained by GarNet provided two well defined groups that perfectly divide the samples between control and AD patients, as can be observed in the abscissa axis. Note that the labels of patients were clustered by using Pearson correlation. The down-regulated genes are grouped at the top of the heatmap. The up-regulated genes are concentrated at the bottom of the heatmap. Note that the gene regulation levels determined by the rules found by GarNet (defined in Table 3) are also consistent with those provided in the heatmap using Pearson correlation.

Additionally, the clusters obtained are displayed in Fig. 4 through the representation technique named clusplot [53] available in R software environment. It represents the principal component analysis of the two clusters of genes obtained in the heatmap using Pearson correlation. Red and blue ellipses indicate class boundaries

Table 3
Top frequent genes extracted from the rules discovered by GarNet using the best configuration settings sorted by frequency.

ID	Gene Sym.	RefSeq	Expr.	Freq.	ID	Gene Sym.	RefSeq	Expr.	Freq.
1	*EPHA10*	NM_001004338		1.00	49	*F5*	NM_000130		0.26
2	*FAM158A*	NM_016049		0.87	50	*MT1F*	NM_005949		0.26
3	*PIAS3*	NM_006099		0.81	51	*SERF2*	NM_001018108		0.24
4	*TNFSF13B*	NM_006573		0.73	52	*KLF8*	NM_007250		0.24
5	*SMYD5*	NM_006062		0.68	53	*LYZL4*	NM_144634		0.24
6	*ZNF202*	NM_003455		0.62	54	*SPTBN1*	NM_178313		0.24
7	*LIN28B*	NM_001004317		0.53	55	*KCNG1*	NM_002237		0.24
8	*CCDC43*	NM_144609		0.51	56	*ZC3H3*	NM_015117		0.23
9	*SLC26A8*	NM_052961		0.50	57	*DDX19A*	NM_018332		0.23
10	*PBX4*	NM_025245		0.46	58	*C16ORF58*	NM_022744		0.22
11	*TUBGCP6*	NM_001008658		0.46	59	*TSC22D4*	NM_030935		0.22
12	*GFRα2*	NM_001495		0.44	60	*CCBL2*	NM_001008661		0.22
13	*RPS8*	NM_001012		0.41	61	*MT1E*	NM_175617		0.22
14	*ALOX12B*	NM_001139		0.38	62	*CYP2C8*	NM_000770		0.22
15	*IGLON5*	XM_380008		0.38	63	*CAND2*	XM_371617		0.22
16	*NOC4L*	NM_024078		0.37	64	*RPS3*	NM_001005		0.22
17	*LRFN3*	NM_024509		0.37	65	*SOX11*	NM_003108		0.22
18	*GTF2E1*	NM_005513		0.36	66	*FGF12*	NM_004113		0.22
19	*SOX4*	NM_003107		0.36	67	*RPS6*	NM_001010		0.22
20	*KHK*	NM_000221		0.36	68	*SORT1*	NM_002959		0.22
21	*NDRG2*	NM_016250		0.36	69	*FAM213B*	NM_152371		0.22
22	*ST6GAL2*	NM_032528		0.36	70	*TMED4*	NM_182547		0.22
23	*KCTD13*	NM_178863		0.35	71	*GRIA1*	NM_000827		0.22
24	*PSIP1*	NM_021144		0.35	72	*TRAF7*	NM_032271		0.22
25	*VILL*	NM_015873		0.35	73	*UBAP2*	NM_018449		0.21
26	*SLC26A11*	NM_173626		0.32	74	*STRN4*	NM_013403		0.21
27	*MLIP*	NM_138569		0.32	75	*CNPPD1*	NM_015680		0.21
28	*TAOK2*	NM_016151		0.32	76	*RPL18A*	NM_000980		0.21
29	*HSD11B2*	NM_000196		0.32	77	*INPP5J*	NM_001002837		0.21
30	*GNB5*	NM_006578		0.31	78	*ST3GAL2*	NM_006927		0.21
31	*PYCRL*	NM_023078		0.31	79	*TAF1D*	NM_024116		0.21
32	*WDR60*	NM_018051		0.29	80	*IQGAP1*	NM_003870		0.19
33	*DLGAP2*	NM_004745		0.28	81	*GPHN*	NM_001024218		0.19
34	*LRRK1*	NM_024652		0.28	82	*SMARCD2*	NM_003077		0.19
35	*TRIM59*	NM_173084		0.28	83	*STOX2*	NM_020225		0.19
36	*TTC39A*	XM_375729		0.28	84	*USP9X*	NM_001039590		0.19
37	*CCDC23*	NM_199342		0.28	85	*LIN37*	NM_019104		0.19
38	*CDS1*	NM_001263		0.28	86	*TOR2A*	NM_130459		0.18
39	*MNT*	NM_020310		0.28	87	*ITM2C*	NM_001012514		0.18
40	*PTPN1*	NM_002827		0.28	88	*AQP11*	NM_173039		0.18
41	*DNASE1L2*	NM_001374		0.28	89	*PRPF40B*	NM_001031698		0.18
42	*RPL13A*	NM_012423		0.27	90	*PPWD1*	NM_015342		0.18
43	*DPY19L2*	NM_173812		0.27	91	*PUS1*	NM_001002019		0.18
44	*IFI6*	NM_002038		0.27	92	*DDA1*	NM_024050		0.18
45	*PIGW*	NM_178517		0.27	93	*RPS6KL1*	NM_031464		0.18
46	*ZNF212*	NM_012256		0.27	94	*RPL18*	NM_000979		0.18
47	*CENPT*	NM_025082		0.26	95	*ZNF583*	NM_152478		0.18
48	*TIGD3*	NM_145719		0.26					

of the genes, that is, up-regulated and down-regulated, respectively. It can be observed that the two principal components obtained explain more than 95% of the point variability and all genes, except one, are divided into two independent groups or clusters.

Thus, GarNet was able to discover a set of genes that successfully classified the samples between control and AD patients regarding their expression levels. Next section details the validation techniques applied to assess the statistical and biological relevance of the set of genes obtained by GarNet.

4.5.2. Statistical and biological significance

The Mann-Whitney *U*-test was used to determine whether there is a statistically significant difference between expression levels in healthy samples and AD affected samples in the resulting subset of genes obtained by GarNet. This non-parametric hypothesis test assess if a particular population (AD samples) tends to have larger values than the other (healthy samples). The results of Mann-Whitney *U*-test are summarized in Table 4.

Control and *AD* columns shows the average of the expression levels obtained for each group of patients (healthy and Alzheimer's patients respectively).

Diff. column shows whether the distributions of control and AD samples differed significantly (✓). It can be noted that only 6 of the 95 genes do not significantly differed according to the Mann-Whitney *U*-test and 3 of them take the last positions in the ranking. Hence, GarNet is able to find rules containing genes highly related with AD.

Alternatively, Fig. 5 shows the volcano plot that pictures the log-fold changes in base 2 on the x-axis versus the negative log of the p-value in base 10 on the y-axis. The most significantly differentially expressed genes, i.e., those that have a *p*-value lower than 0.05, are represented by a red dot. It can be observed that all the 95 genes presented a *p*-value lower than 0.05, therefore, all of them are statistically significant.

A fold-change of 1.5 cut-off (log2 threshold of 0.58) was applied to identify significant genes due to the good performance presented in previous studies [54].

Genes significantly up-regulated in AD patients are located at right in the graph, and highlighted by a purple box (16 genes). These genes have a fold-change higher than 1.5, that is, a positive log2-fold change greater than 0.58. Genes significantly down-regulated in AD patients are located at left in the graph and highlighted in a green box (61 genes). These genes have a fold-change lower than 0.66, that is, a negative log2-fold change lower than −0.58. Both significantly up-regulated and down-regulated genes are labelled with the corresponding Gene Symbol Identification. It can be noted that 77 genes (more than 80%) are biologically significant when taking into account an absolute fold-change higher than 1.5 and 100% of genes are statistically significant according to the minimum p-value considered. Again, we can state that GarNet was able to discover a set of genes highly significant both at statistical and biological level.

4.6. Fifth phase results: biological knowledge integration

This phase details the achieved results after performing the biological knowledge integration process. Specifically, the altered functions in neurons affected by AD found in the literature, the

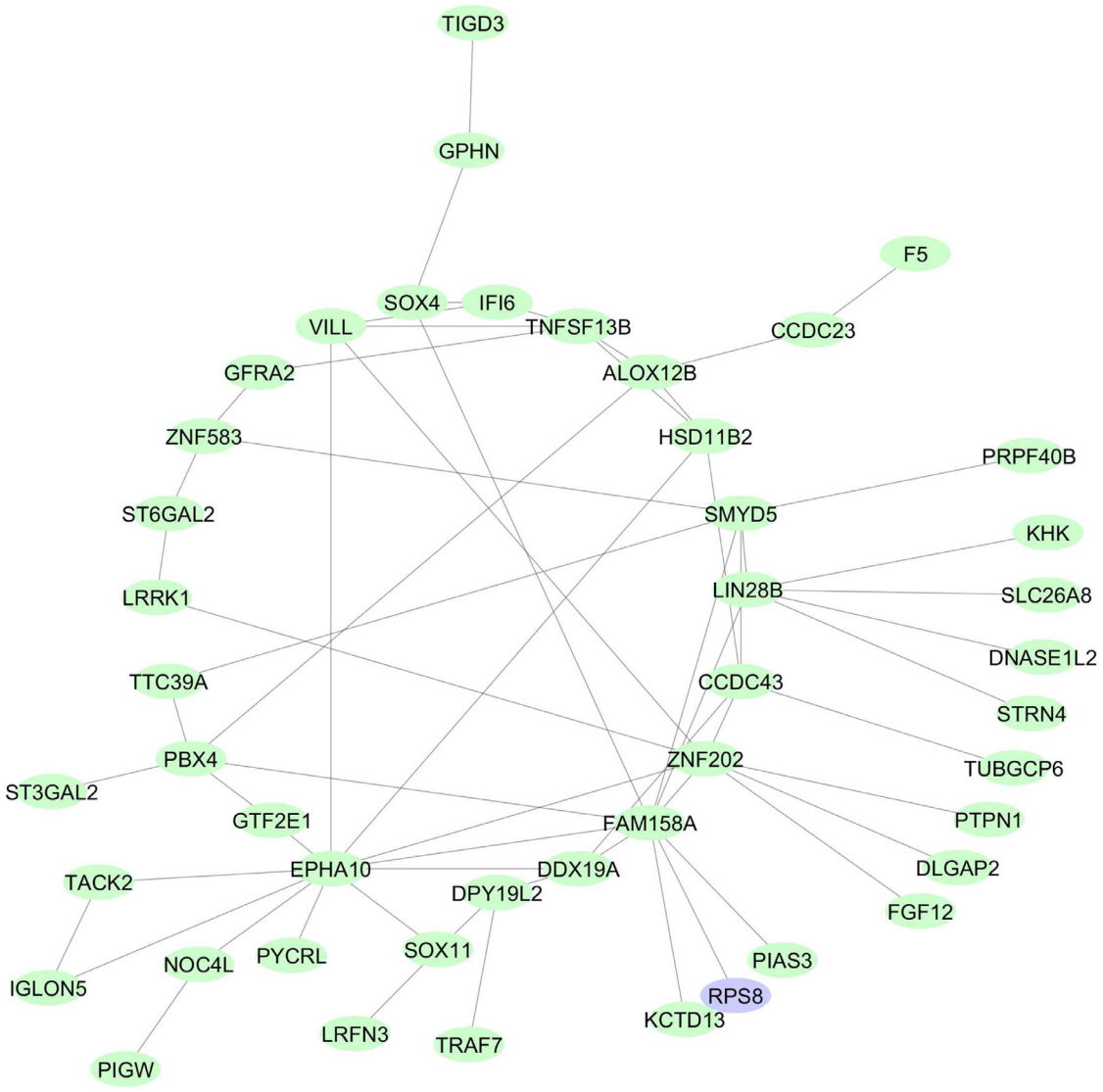

Fig. 2. Network of gene-gene interactions among AD-related genes. Green color and purple color are used to identify the downregulated and upregulated genes in AD, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

enrichment analysis, the mapping to known gene-disease associations related to AD, validation with additional datasets and permutation tests are presented in the following subsections.

4.6.1. Altered functions in neurons affected by AD

As a result of applying rule-based strategy over a dataset of AD patients, we found more than 90 genes that were significantly altered when compared to healthy people. The identified genes covered a high range of functions, both neuron specific as pleiotropic functions. These results highlight the role of AD as a multifactorial syndrome, as it has extensively stated in the literature [5,55,56].

Some of the genes (*CDS1, KLF8, SPTBN1, DDX19A, TSC22D4, GPHN, NDRG2*) found using GarNet had been previously linked to AD [17,57–62]. These genes cover different roles in neurons, including cell signaling (*CDS1, GPHN, NDRG2*), cytoskeleton structure (*SPTBN1*) or gene transcription (*KLF8, TSC22D4*) [58,63]. Although some of the found genes have been identified in other AD gene array analysis (*CDS1, SPTBN1, DDX19A*) [17,59,60], no functional information has been described yet. In addition, another additional gene, *PIAS3*, encodes a protein that acts as an inhibitor of *STAT3*, which has been connected to AD [64,65]. We also found a high number of proteins associated to neuronal processes. Amongst them, *GFRα2* expression has been connected with the development or maintenance of cognitive abilities in mice [66]. In that way, *GFRα2* altered expression in AD patients may be connected with the progressive loss of memory, one of the hallmarks of the disease.

During AD, neurons experiment a loss of its metal homeostasis [67,68]. Regarding this, we found that some genes associated to zing finger proteins are down-regulated in neurons of AD patients (i.e., *ZC3H3, ZNF202, ZNF583, Zc3h3*). Interestingly, zinc levels are lower in the brain of AD patients [69]. However, two different metallothioneins related with zinc homeostasis (MT-1E and MT-1F) were also up-regulated. These proteins are related with heavy metals detoxification and are up-regulated in order to avoid zinc neurotoxicity [68]. This apparent contradiction may be due to differences in Braak staging of neurons used for gene analysis, showing the complexity of this neurodegenerative disease. In that way, there is also ample evidence of an increment of zinc levels in AD patients, as it has been reviewed by Watt et al. [70]. The apparent opposite results obtained by González-Domínguez et al.[69] may be due to a different progression stage of the disease.

Also connected with zinc dyshomeostasis, diabetes and cardiac disease have been connected with AD [71–75]. Concerning this, we

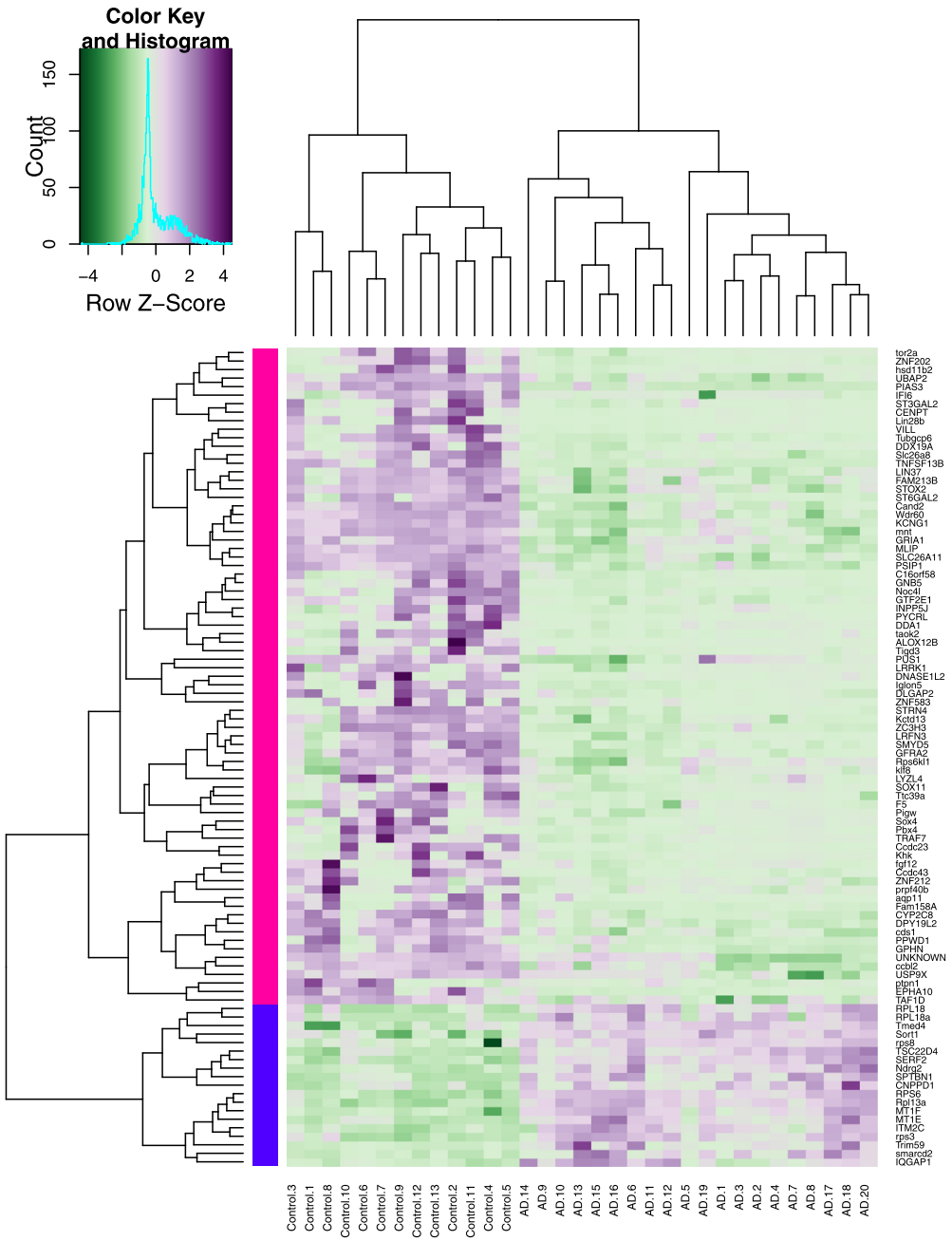

Fig. 3. Heatmap of top genes and hierarchical cluster analysis. The significant clusters bins after cutting the tree are showed in the color bar. The columns (patients) are clustered by Spearman correlation and rows (genes) are clustered by Pearson correlation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

found genes related to diabetes (i.e. *Lin28b*) or cardiac disease (i.e. *MLIP*) [76,77]. The altered expression of these genes in neurons may indicate a global deregulation of these genes in the organism.

Gene transcription and protein synthesis/degradation, are also highly affected by AD [78–80]. In that way, we found 13 transcription regulators, 7 ribosomal proteins genes and 6 genes related to protein ubiquitination (Table 5). In addition, we found other 7 genes related to RNA metabolism, like *NOC4l*, which has been connected to ribosome biogenesis. The majority of these proteins appeared to be down-regulated in AD neurons, with some exceptions. Interestingly, one of these exceptions is NDRG2, which has been detected experimentally up-regulated in AD [62]. We found up to 20 enzymes, constituting a highly diverse group with a common property: all of them were down-regulated in AD neurons. Some of them are connected to lipid or carbohydrate metabolism (*ALOX12B, CDS1, PIGW, KHK, ST3GAL2, ST6GAL2*), which suggests that affected neurons may have a lower metabolism. Altogether, we can hypothesize that, during AD progression, neurons try to compensate the general gene down-regulation with an up-regulation of ribosomal protein transcription.

4.6.2. Enrichment analysis

The top frequent genes discovered by GarNet were analyzed in the context of Gene Ontology using Fatigoo tool [46]. Detection of statistically overrepresented GO terms was done with the Fisher's exact test and multiple-testing adjustments were done with the Westfall and Young method. In Table 6, the GO biological process enriched, together with the p-value and the adjusted p-value are

Table 4
Results obtained by Mann-Whitney *U*-test

ID	Gene symbol	Control	AD	p-value	Diff.	ID	Gene symbol	Control	AD	p-value	Diff.
1	*EPHA10*	3.83	2.32	0.0001	✓	49	*F5*	6.25	4.84	0.0008	✓
2	*FAM158A*	4.20	2.82	0.0000	✓	50	*MT1F*	7.26	9.57	0.0000	✓
3	*PIAS3*	4.61	2.62	0.0000	✓	51	*SERF2*	7.97	9.28	0.0000	✓
4	*TNFSF13B*	4.31	2.52	0.0000	✓	52	*KLF8*	6.56	5.89	0.0030	✓
5	*SMYD5*	4.86	3.26	0.0000	✓	53	*LYZL4*	2.50	2.17	0.0008	✓
6	*ZNF202*	3.64	3.06	0.0000	✓	54	*SPTBN1*	3.57	4.32	0.0000	✓
7	*LIN28B*	2.87	2.40	0.0005	✓	55	*KCNG1*	7.56	5.13	0.0000	✓
8	*CCDC43*	6.95	6.26	0.0001	✓	56	*ZC3H3*	4.28	2.64	0.0004	✓
9	*SLC26A8*	4.61	3.37	0.0000	✓	57	*DDX19A*	3.37	2.93	0.0004	✓
10	*PBX4*	2.49	2.16	0.0220	✓	58	*C16ORF58*	3.80	2.59	0.0005	✓
11	*TUBGCP6*	3.97	2.40	0.0000	✓	59	*TSC22D4*	5.23	7.01	0.0000	✓
12	*GFRα2*	4.66	3.70	0.0000	✓	60	*CCBL2*	8.90	7.04	0.0000	✓
13	*RPS8*	8.99	10.58	0.0000	✓	61	*MT1E*	6.59	8.22	0.0000	✓
14	*ALOX12B*	3.14	2.30	0.0020	✓	62	*CYP2C8*	6.48	4.84	0.0000	✓
15	*IGLON5*	3.72	2.52	0.0003	✓	63	*CAND2*	10.73	9.34	0.0000	✓
16	*NOC4L*	4.72	2.24	0.0001	✓	64	*RPS3*	9.54	11.58	0.0000	✓
17	*LRFN3*	5.17	3.34	0.0001	✓	65	*SOX11*	3.45	2.98	0.0360	✓
18	*GTF2E1*	3.25	2.63	0.0000	✓	66	*FGF12*	3.16	2.81	0.0010	✓
19	*SOX4*	2.70	2.54	0.0130	✓	67	*RPS6*	9.74	11.63	0.0000	✓
20	*KHK*	3.11	2.93	0.0003	✓	68	*SORT1*	6.23	6.41	0.0000	✓
21	*NDRG2*	2.81	3.74	0.0000	✓	69	*FAM213B*	10.81	9.41	0.0000	✓
22	*ST6GAL2*	5.52	2.79	0.0003	✓	70	*TMED4*	7.89	9.36	0.0000	✓
23	*KCTD13*	9.21	6.89	0.0000	✓	71	*GRIA1*	6.28	4.55	0.0000	✓
24	*PSIP1*	9.85	8.58	0.0000	✓	72	*TRAF7*	2.63	2.45	0.0080	✓
25	*VILL*	3.13	2.25	0.0003	✓	73	*UBAP2*	9.19	7.57	0.0000	✓
26	*SLC26A11*	7.88	6.22	0.0000	✓	74	*STRN4*	4.99	3.49	0.0000	✓
27	*MLIP*	7.21	3.68	0.0000	✓	75	*CNPPD1*	3.36	3.76	0.0000	✓
28	*TAOK2*	3.84	3.11	0.0110	✓	76	*RPL18A*	3.81	7.11	0.0001	✓
29	*HSD11B2*	2.81	2.36	0.0620		77	*INPP5J*	3.90	2.70	0.0060	✓
30	*GNB5*	2.87	2.47	0.0000	✓	78	*ST3GAL2*	3.11	2.57	0.0005	✓
31	*PYCRL*	3.44	2.48	0.0240	✓	79	*TAF1D*	10.53	9.64	0.0000	✓
32	*WDR60*	7.68	6.44	0.0000	✓	80	*IQGAP1*	4.39	5.06	0.0001	✓
33	*DLGAP2*	3.09	2.78	0.0000	✓	81	*GPHN*	4.49	3.50	0.0000	✓
34	*LRRK1*	3.97	3.16	0.0070	✓	82	*SMARCD2*	2.43	2.77	0.0000	✓
35	*TRIM59*	2.89	3.57	0.0000	✓	83	*STOX2*	9.90	8.74	0.0000	✓
36	*TTC39A*	3.86	3.06	0.0000	✓	84	*USP9X*	9.52	9.00	0.0000	✓
37	*CCDC23*	6.42	5.76	0.0040	✓	85	*LIN37*	8.28	5.78	0.0000	✓
38	*CDS1*	9.44	8.43	0.0000	✓	86	*TOR2A*	3.52	2.59	0.0004	✓
39	*MNT*	10.77	9.57	0.0000	✓	87	*ITM2C*	9.07	10.80	0.0000	✓
40	*PTPN1*	2.79	2.68	0.0800		88	*AQP11*	2.36	2.13	0.3340	
41	*DNASE1L2*	2.94	2.30	0.0800		89	*PRPF40B*	3.03	2.92	0.0020	✓
42	*RPL13A*	8.90	10.97	0.0000	✓	90	*PPWD1*	3.95	3.11	0.0001	✓
43	*DPY19L2*	5.04	2.67	0.0000	✓	91	*PUS1*	5.56	4.38	0.0000	✓
44	*IFI6*	7.90	7.28	0.0004	✓	92	*DDA1*	3.03	2.84	0.5700	
45	*PIGW*	4.90	4.34	0.0060	✓	93	*RPS6KL1*	8.87	7.20	0.0003	✓
46	*ZNF212*	3.88	3.65	0.0020	✓	94	*RPL18*	10.85	12.03	0.0000	✓
47	*CENPT*	2.78	2.17	0.0030	✓	95	*ZNF583*	2.62	2.51	0.7400	
48	*TIGD3*	2.57	2.30	0.0400	✓						

Table 5
Altered functions in neurons affected by AD.

General functions		Genes found by GarNet
Protein metabolism:	(a) Gene transcription	*CAND2, GTF2E1, Klf8, LIN37, MNT, NDRG2*,* ***PIAS3****, PSIP1, SMYD5, SOX4, SOX11, TSC22D4*, ZNF202, ZNF212,* ***SMARCD2****
	(b) Ribosomal proteins	*RPS3, RPS6, RPS6Kl1*, RPS8, Rpl13A, Rpl18, Rpl18A*
	(c) Protein ubiquitination	*KCTD13,* ***PIAS3****, TRAF7, TRIM59, UBAP2, USP9X*
Enzymes		*ALOX12B, CCBL2, CDS1, CYP2C8, DPY19l2, F5, FAM213B, GNB5 INPP5J, KHK, LRRK1, LYZl4, PIGW, PPWD1, PYCRl TAOK2, TOR2A, ST3AL2, ST6GAL2,* ***ZC3H3***
Cell signaling/receptors		*DLGAP2, EPHA10, FGF12, GFRα2, GRIA1, IFI6, IQGAP1*, ITM2C** ***LRFN3****, SORT1*, STRN4, TNFSF13B*
Channels, transporters		*KCNG1,* ***SORT1****, SlC26A8, SlC26A11, TMED4*
Cytoskeletal related proteins		*GPHN, SPTBN1, TUBGCP6, VILL, WDR60*
DNA binding proteins		*CENPT, PBX4,* ***SMARCD2*, TAF1D****, TIGD3*
RNA binding proteins		*DDX19A, LIN28B, NOC4l, PRPF40B, PUS1,* ***TAF1D, ZC3H3***

Genes in bold appears in two different groups. Symbol (*) is used to represent up-regulated genes

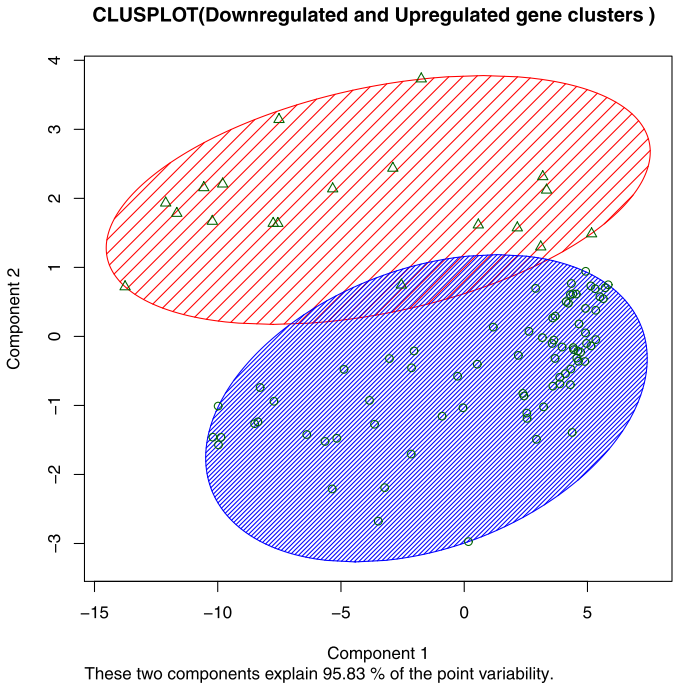

Fig. 4. Clusplot of up-regulated and down-regulated genes. Clusplot that visualizes the principal component analysis associated to the clusters of up-regulated and down-regulated genes appearing in the rules obtained by GarNet. These two components explain 95.83% of the point varibility.

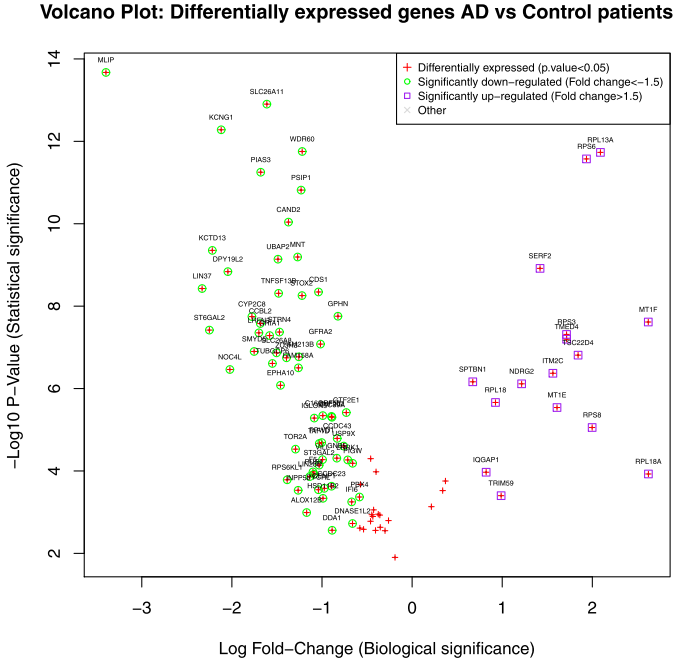

Fig. 5. Volcano plot of the expression levels of the genes selected by GarNet. The log-fold change is plotted on the x-axis and the negative log10 *p*-value is plotted on the y-axis. The most significantly differentially expressed genes are represented by a red dot. False discovery rate was used as a cut-off (adjusted *p*-value < 0.05). Genes significantly up-regulated in AD patients are located at right in the graph, represented by a purple box (fold-change > 1.5 or log2-fold threshold > 0.58). Genes significantly down-regulated in the AD patients are located at left in the graph, represented by a green circle (fold-change < 0.66 or log2-fold threshold < -0.58). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 6
Enriched GO terms by the top frequent genes discovered by GarNet.

GO Term	P value	Adj. *P* value
Positive regulation of protein sumoylation(GO:0033235)	2.2e−5	0.04
Arachidonic acid metabolic process(GO:0019369)	3e−5	0.04
Neuron projection extension(GO:1990138)	4.3e−6	0.025
Cellular response to cadmium ion(GO:0071276)	3.1e−5	0.04
Cellular response to zinc ion(GO:0071294)	3.5e−5	0.04

Table 7
Genes from top frequent genes discovered by GarNet mapped onto the AD PPI network.

Gene symbol	Uniprot	Gene description
GRIA1	P42261	Glutamate receptor, ionotropic, AMPA 1
CYP2C8	P10632	Cytochrome P450, family 2, subfamily C, polypeptide 8
PTPN1	P18031	Protein tyrosine phosphatase, non-receptor type 1
MT1F	P04733	Metallothionein 1F
PSIP1	O75475	PC4 and SFRS1 interacting protein 1
IQGAP1	P46940	IQ motif containing GTPase activating protein 1
F5	P12259	Coagulation factor V (proaccelerin, labile factor)
SPTBN1	Q01082	Spectrin, beta, non-erythrocytic 1
SORT1	Q99523	Sortilin 1
ZNF212	Q9UDV6	Zinc finger protein 212
STRN4	Q9NRL3	Striatin, calmodulin binding protein 4
TSC22D4	Q9Y3Q8	TSC22 domain family, member 4
IFI6	P09912	Interferon, alpha-inducible protein 6
ITM2C	Q9NQX7	Integral membrane protein 2C

shown. It is worth mention that GO:0071294 and GO:0071276 belong to the Gene Ontology terms of HDACi-responsive gene sets [81], being increased levels of HDAC2 described in the hippocampus of patients suffering from AD.

4.6.3. Validation on known gene-disease associations

First, the gene set obtained by GarNet was mapped onto the network reported by Soller et al. where 12 well known AD associated genes were selected from OMIM Database as seed [47]. Protein-protein interactions network is referred as “AD PPI network”. This network is reported using UNIPROT identifiers and we have used DAVID tool to convert the top frequent genes to UNIPROT identifiers. The top frequent genes discovered by GarNet using the best configuration settings were mapped onto the AD PPI network for the creation of a module with the first neighbors of mapped genes. In this module, we found 14 genes from the top frequent genes (see Table 7), containing this module 90 genes and 178 interactions as can be observed in Fig. 6.

Furthermore, the gene set reported by GarNet was also mapped into the well-known database MalaCards and in particular, GeneCards [49]. Table 8 shows the subset of genes found by GarNet that are associated to cerebral, diabetes and heart diseases in GeneCards. These genes have been related to syndromes like Parkinson’s disease, ataxia, dementia and Alzheimer’s disease, among others. Since these syndromes present a numerous and overlapping effects, the genes affected by them may be common.

After the study performed, we have seen that many of the genes found by our methodology are related to brain diseases as demonstrated some existing repositories.

4.7. Sixth phase results: validation in additional data

This section details the validation of the results obtained by GarNet checking the significance of the gene subset in other AD-related datasets. Additionally, random versions of the original dataset were generated to quantify the strength of the provided results.

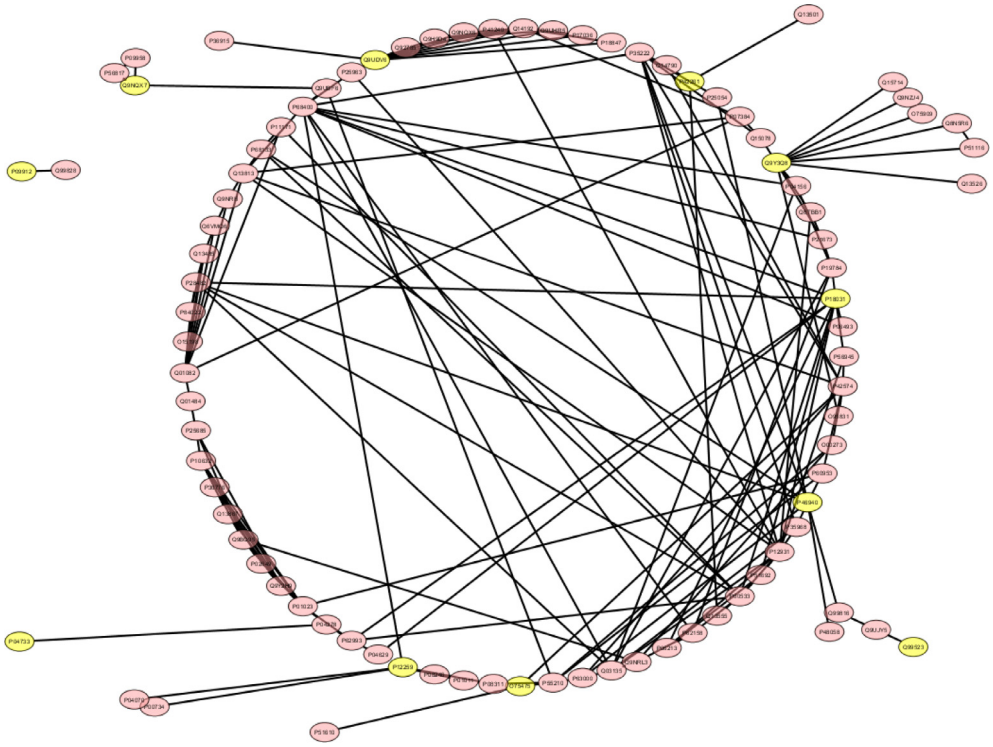

Fig. 6. Alzheimer's disease PPI network. The top frequent genes discovered by GarNet using the best configuration setting were mapped onto the AD PPI Network to create a module with the first neighbors of mapped genes highlighted in yellow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 8
Genes found in MalaCards associated to cerebral diseases.

Gene symbol	Decription in Malacards
TUBGCP6	Microcephaly, Mental retardation
IGLON5	Dementia, Ataxia
SOX4	Ataxia, Neuronitis, Cerebritis
KHK	Mental Depression
ST6GAL2	Neuronitis, Cerebritis
TAOK2	Autism Spectrum Disorder,
LRRK1	Parkinson Disease, Nervous system cancer
CDS1	Nervous system cancer, Werner Syndrome
MNT	Miller-Dieker Syndrome, Diabetes Mellitus
PTPN1	Diabetes Mellitus, Werner Syndrome
PIGW	Hyperphosphatasia with Mental Retardation, West syndrome
ZC3H3	Werner Syndrome
CYP2C8	Myocardial infaction
SOX11	Mental retardation, neuronitis, cerebritis
GRIA1	Alzheimer Disease, Schizophrenia, Neuronitis
INPP5J	Lowe Syndrome
TAF1D	Huntington Disease
IQGAP1	Werner Syndrome
GPHN	Autism Spectrum Disorder, neuronitis
USP9X	Mental retardation, neuronitis
PRPF40B	Rett syndrome
PUS1	Myopathy, Werner syndrome

4.7.1. Significance in other AD-related data

The set of genes obtained by GarNet was checked with other AD-related datasets with the aim of evaluating the relevance of them.

In particular, the following datasets collected from NCBI repository [45] have been analyzed:

- **D1 - GSE39420**: Control patients against AD patients with the Psen gene mutated.
- **D2 - GSE39420**: Control patients against an early onset AD.
- **D3 - GSE4757**: Healthy neurons against diseased neurons in AD patients.
- **D4 - GSE5281**: Entorhinal cortex of healthy patients against AD patients.
- **D5 - GSE28146**: Control patients against moderate AD patients.
- **D6 - GSE28146**: Control patients against severe AD patients.

Table 9 reports the validation performed in the six described datasets. Column *gene* indicates the name of gene in Gene Symbol format. Column *Ref.* refers to the regulation level of the genes obtained by GarNet in the dataset used in this paper. Columns *D1* to *D6* show the regulation level of each selected gene, respectively, in the external datasets previously described. The regulation level of each gene in each dataset was calculated taking into account the average of AD patients regarding the average of control patients using the fold change measure previously described. It can be noted that green color denotes that the gene is down-regulated, purple color refers to up-regulated gene, N.D. stands for genes without significant differences between healthy and AD patients. White color indicates that the gene has not been found in the analyzed dataset.

As it can be observed, a 65% of the genes presents the same behavior if we consider 4 of the 6 datasets as success. This fact indicates that most of the analyzed genes appear commonly in the same way that in the analyzed dataset by GarNet. This result shows that this gene subset could be used as a general biomarker of AD.

4.7.2. Permutation tests

To quantify the result strengths obtained by GarNet, we compared the predictive capability of our approach using the original dataset and other two versions of the dataset where the class labels of the instances are shuffled, henceforth named dataset random 1 and dataset random 2, respectively. Note that the proportion of control and AD patients has been maintained in the random datasets.

Table 9
Validation of the found genes using other AD-related datasets.

Gene	Ref.	D1	D2	D3	D4	D5	D6	Gene	Ref.	D1	D2	D3	D4	D5	D6
C16ORF58								GPHN		ND	ND				
CDS1								TOR2A							
CYP2C8								ZNF583							
DDX19A								LRFN3							
DLGAP2								PUS1							
FGF12								ST3GAL2							
GRIA1								VILL							
KCTD13								AQP11							
SMYD5								DNASE1L2		ND	ND				
TAF1D								PIGW							
TTC39A								TAOK2		ND	ND				
CCDC43								TNFSF13B							
DPY19L2								CCDC23							
EPHA10								CENPT							
FAM158A								INPP5J							
HSD11B2								GTF2E1		ND	ND				
PBX4								ZNF212			ND			ND	ND
RPS6KL1								PRPF40B		ND	ND				
SLC26A11								TUBGCP6		ND	ND				
ST6GAL2								UBAP2		ND	ND				
STOX2								PPWD1							
LIN28B								SOX11							
WDR60			ND					ZNF202							
CAND2								PIAS3							
DDA1								ZC3H3		ND	ND				
GNB5								F5							
IFI6								NOC4L							
KCNG1								PTPN1							
KHK								ITM2C		ND	ND				
KLF8								SERF2							
MLIP								RPS8						ND	ND
MNT								TMED4							
SLC26A8								TRIM59							
CCBL2		ND	ND					NDRG2		ND	ND				
GFRA2						ND	ND	RPL13A							
STRN4						ND	ND	SMARCD2						ND	ND
LRRK1								SPTBN1		ND	ND				
TRAF7								RPS3							
USP9X			ND					RPL18A							
ALOX12B								SORT1							
LIN37								RPL18							
PSIP1								RPS6							
PYCRL								MT1E							
SOX4								MT1F							
FAM213B						ND	ND	CNPPD1							
TIGD3		ND	ND					IQGAP1							
IGLON5						ND	ND	TSC22D4							
LYZL4			ND												

*green color: down-regulated, purple color: up-regulated, N.D.: no significant differences, white color: not found

Table 10
Results obtained by GarNet using different versions of the dataset where the class labels of the instances are shuffled.

ID. configuration	Control classified (%)	AD classified (%)	Not covered (%)
Original dataset			
#8	39.39	60.61	0
#13	39.39	60.61	0
#14	39.39	60.61	0
Dataset random 1			
#8	30.39	0.55	69.06
#13	33.94	5.88	60.18
#14	31.88	4.76	63.36
Dataset random 2			
#8	25.70	0	74.3
#13	32.39	1.85	65.76
#14	29.48	1.70	68.82

*39.39% of instances are control patients and 60.61% of instances are AD patients in all datasets. The percentage of control and AD patients matches with the distribution of both type of patients in the dataset.

To fulfill this goal, we applied the methodology proposed in both random datasets using the best configuration settings of GarNet (#8, #13 and #14) described in Section 4.3.

Table 10 summarizes the percentage of control and AD instances correctly classified in addition to the uncovered instances by the rules obtained by GarNet for each dataset (original, random 1, random 2) and each configuration setting.

Note that the 39.30% of instances are control patients and the 60.61% are AD patients.

As can be observed, the rules obtained by GarNet correctly classify the 100% of instances when is applied in the original dataset. Regarding the random datasets, although all the control patients are correctly classified, the percentage of AD patients correctly classified is very low with values close to 0. The percentage of uncovered instances is very high, achieving values greater than the 60%.

It can be noted that GarNet algorithm is not able to detect significant rules for AD patients when the class labels of instances are shuffled.

Additionally, the top of 100 frequent genes that appear in the rules found by GarNet in the random datasets were selected. It is noteworthy that the top of 100 frequent genes indicated for random datasets are associated with control patients instead of AD patients. As described Table 10, GarNet is not able to mine rules associated for AD patients from the random datasets. Only 5 genes obtained in the original dataset have been also found in the random dataset 1, and only 2 in the random dataset 2. This demonstrates that the genes found by GarNet do not have the same level of accuracy and the same variables when the permutation datasets are used.

5. Conclusions

In this work, we have presented the integration of three machine learning methods: decision trees, quantitative rules and hierarchical cluster analysis in AD gene expression profiles. We aim at providing gene expression patterns and a deeper knowledge into biological functions with higher relevance. To fulfil this purpose, we used different external sources of information such as PubMed,

GO and PPI network. From the best of our knowledge, QAR-based methods have emerged as a popular methodology to discover hidden relationships among variables in a subspace of a dataset, but it has not been used in AD-related dataset.

The evolutionary GarNet algorithm has been applied to select a set of genes highly related with the AD according to their expression level in a preprocessed dataset composed initially by 33 samples and 35,722 probesets collected by laser capture microdissection from entorhinal cortex. To fulfill this goal, six phases were performed. Both first and second phases constitute an ensemble of classifiers to establish the best configurations settings of another classifier. The first one applies the well-known C4.5 algorithm that was used to define the minimum threshold to select the best configurations of GarNet in terms of instances correctly classified. The second one was devoted to apply the GarNet algorithm using multiple configuration settings with the aim of obtaining QAR from the studied dataset. The third phase is applied to rerun GarNet with the configurations that overcome the accuracy obtained by C4.5 algorithm. Then, we identify the set of genes with potential prognosis role in AD. The fourth phase tackled both the validation by hierarchical cluster analysis, fold change and statistical tests. Biological knowledge integration based on the information fusion of prior biological knowledge, Gene Ontology enrichment analysis and mapping of the genes obtained to AD PPI network was performed in the fifth phase. Finally, the sixth phase performed the results validation obtained using additional datasets and permutation tests. The results provided by GarNet could be used to perfectly divide the expression profiles between control and AD as the heatmap and clusplot shown. The absolute fold-change of most of the genes was higher than 1.5 as displayed the volcano plot presented. The Mann-Whitney *U*-test was used to prove that GarNet has been able to discover a set of genes highly significant at statistical level.

The results have shown that the obtained rules successfully characterize the underlying information, grouping relevant genes for the problem under study and agreeing with prior biological knowledge. We found 90 genes that were significantly altered in AD patients. Some of them were previously linked to AD and neuronal process as GFRα2, which has been connected with the development or maintenance of cognitive abilities in mice. Our method provided genes associated to diabetes and cardiac disease, e.g. *LIN28B* and *MLIP*, supporting the hypothesis that there may be an association between these two diseases and AD. Finally, GO enrichment analysis found two enriched terms previously connected to AD, being an additional result that validates our QAR strategy. Altogether, QAR can be used to find significant altered genes not only in AD, but in other diseases, using similar approaches.

Acknowledgments

The financial support from the Spanish Ministry of Science and Technology, projects TIN2011-28956-C02-02 and TIN2014-55894-C2-1-R, and from the Junta de Andalucia, P11-TIC-7528, is acknowledged. We thank Lourdes Riquelme for assistance with the review of the text and for her comments that greatly improved the manuscript.

References

[1] C. Ferri, M. Prince, C. Brayne, H. Brodaty, L. Fratiglioni, M. Ganguli, K. Hall, K. Hasegawa, H. Hendrie, Y. Huang, A. Jorm, C. Mathers, P. Menezes, E. Rimmer, M. Scazufca, Global prevalence of dementia: a delphi consensus study, Lancet 366 (9503) (2005) 2112–2117.

[2] D. Evans, H. Funkenstein, M. Albert, P. Scherr, N. Cook, M. Chown, L. Hebert, C. Hennekens, J. Taylor, Prevalence of alzheimer's disease in a community population of older persons: higher than previously reported, J. Am. Med. Assoc. 262 (18) (1989) 2551–2556.

[3] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, E.M. Stadlan, Clinical diagnosis of alzheimer's disease: report of the nincds-adrda work group under the auspices of department of health and human services task force on alzheimer's disease., Neurology 34 (7) (1984) 939–944.

[4] M. Goedert, M.G. Spillantini, A century of Alzheimer's disease, Science 314 (5800) (2006) 777–781.

[5] K. Iqbal, F. Liu, C.-X. Gong, Alzheimer disease therapeutics: focus on the disease and not just plaques and tangles, Biochem. Pharmacol. 88 (4) (2014) 631–639.

[6] B. Tomlinson, G. Blessed, M. Roth, Observations on the brains of demented old people, J. Neurol. Sci. 11 (3) (1970) 205–242.

[7] D.W. Dickson, H.A. Crystal, L.A. Mattiace, D.M. Masur, A.D. Blau, P. Davies, S.H. Yen, M.K. Aronson, Identification of normal and pathological aging in prospectively studied nondemented elderly humans., Neurobiol. Aging 13 (1) (1992) 179–189.

[8] C. Jack, D. Holtzman, Biomarker modeling of alzheimer's disease, Neuron 80 (6) (2013) 1347–1358.

[9] D. Campion, C. Dumanchin, D. Hannequin, B. Dubois, S. Belliard, M. Puel, C. Thomas-Anterion, A. Michon, C. Martin, F. Charbonnier, G. Raux, A. Camuzat, C. Penet, V. Mesnage, M. Martinez, F. Clerget-Darpoux, A. Brice, T. Frebourg, Early-onset autosomal dominant alzheimer disease: prevalence, genetic heterogeneity, and mutation spectrum, Am. J. Human Genetics 65 (3) (1999) 664–670.

[10] B. Frank, S. Gupta, A review of antioxidants and Alzheimer's disease., Ann. Clin. Psychiatry 17 (4) (2005) 269–286.

[11] M.A. Wollmer, Cholesterol-related genes in alzheimer's disease, Biochimica et Biophysica Acta (BBA) - Molecul. Cell Biol. Lipids 1801 (8) (2010) 762–773. Lipids and Alzheimer's Disease.

[12] A. Holzinger, M. Dehmer, I. Jurisica, Knowledge discovery and interactive data mining in bioinformatics - state-of-the-art, future challenges and research directions, BMC Bioinform. 15 (Suppl 6) (2014). I1.

[13] A review of microarray datasets and applied feature selection methods, Inf. Sci. 282 (2014) 111–135.

[14] W. Kong, X. Mou, Q. Liu, Z. Chen, C. Vanderburg, J. Rogers, X. Huang, Independent component analysis of alzheimer's dna microarray gene expression data, Mol. Neurodegener. 4 (1) (2009) 1–14.

[15] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, J. Riquelme, An evolutionary algorithm to discover quantitative association rules in multidimensional time series, Soft Comput. 15 (10) (2011) 2065–2084.

[16] M. Martínez-Ballesteros, I. Nepomuceno-Chamorro, J. Riquelme, Discovering gene association networks by multi-objective evolutionary quantitative association rules, J. Comput. Syst. Sci. 80 (1) (2014) 118–136.

[17] M. Ray, J. Ruan, W. Zhang, Variations in the transcriptome of alzheimer's disease reveal molecular networks involved in cardiovascular diseases, Genome Biol. 9 (10) (2008). R148.

[18] R.J. Bateman, C. Xiong, T.L.S. Benzinger, A.M. Fagan, A. Goate, N.C. Fox, D.S. Marcus, N.J. Cairns, X. Xie, T.M. Blazey, D.M. Holtzman, A. Santacruz, V. Buckles, A. Oliver, K. Moulder, P.S. Aisen, B. Ghetti, W.E. Klunk, E. McDade, R.N. Martins, C.L. Masters, R. Mayeux, J.M. Ringman, M.N. Rossor, P.R. Schofield, R.A. Sperling, S. Salloway, J.C. Morris, Clinical and biomarker changes in dominantly inherited Alzheimer's disease, N. Engl. J. Med. 367 (9) (2012) 795–804.

[19] A.S. Fleisher, K. Chen, Y.T. Quiroz, L.J. Jakimovich, M.G. Gomez, C.M. Langois, J.B. Langbaum, N. Ayutyanont, A. Roontiva, P. Thiyyagura, W. Lee, H. Mo, L. Lopez, S. Moreno, N. Acosta-Baena, M. Giraldo, G. Garcia, R.A. Reiman, M.J. Huentelman, K.S. Kosik, P.N. Tariot, F. Lopera, E.M. Reiman, Florbetapir PET analysis of amyloid-ßdeposition in the presenilin 1 E280A autosomal dominant alzheimer's disease kindred: a cross-sectional study, Lancet Neurol. 11 (12) (2012) 1057–1065.

[20] V.L. Villemagne, S. Burnham, P. Bourgeat, B. Brown, K.A. Ellis, O. Salvado, C. Szoeke, S.L. Macaulay, R. Martins, P. Maruff, D. Ames, C.C. Rowe, C.L. Masters, Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic alzheimer's disease: a prospective cohort study, Lancet Neurol. 12 (2013) 357–367.

[21] E. Kopke, Y.C. Tung, S. Shaikh, A.C. Alonso, K. Iqbal, I. Grundke-Iqbal, Microtubule-associated protein tau. abnormal phosphorylation of a non-paired helical filament pool in alzheimer disease., J. Biol. Chem. 268 (32) (1993) 24374–24384.

[22] J.L. Price, J.C. Morris, Tangles and plaques in nondemented aging and preclinical alzheimer's disease, Ann. Neurol. 45 (3) (1999) 358–368.

[23] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, in: Proceedings of the International Conference on Very Large Databases, 1994, pp. 478–499.

[24] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006.

[25] L. Geng, H. Hamilton, Interestingness measures for data mining: a survey, ACM Comput. Surv. 38 (3) (2006) 1–42.

[26] G. Piatetsky-Shapiro, Discovery, analysis and presentation of strong rules, in: Knowledge Discovery in Databases, 1991, pp. 229–248.

[27] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, J. Riquelme, Selecting the best measures to discover quantitative association rules, Neurocomputing 126 (2014) 3–14.

[28] E. Georgii, L. Richter, U. Rückert, S. Kramer, Analyzing microarray data using quantitative association rules, Bioinformatics 21 (suppl 2) (2005) ii123–ii129.

[29] R. Chaves, J. Ramírez, J.M. Górriz, Integrating discretization and association rule-based classification for alzheimer's disease diagnosis, Expert Syst. Appl. 40 (5) (2013) 1571–1578.

[30] D. Ponmary Pushpa Latha, D. Joseph Pushpa Raj, Measuring interesting amino acid patterns for alzheimer's disease related studies targets on the binding site using association rule mining, J. Appl. Pharm. Sci. 3 (7) (2013) 25–30.
[31] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, CRC press, 1984.
[32] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.
[33] J.R. Quinlan, C4.5: programs for machine learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.
[34] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, Morgan Kaufmann, 1995, pp. 1137–1143.
[35] J. Abellán, Ensembles of decision trees based on imprecise probabilities and uncertainty measures, Inform. Fusion 14 (4) (2013) 423–430.
[36] L.I. Kuncheva, Combining pattern classifiers: methods and algorithms, Wiley-Interscience, 2004.
[37] A.K. Rider, G. Siwo, S.J. Emrich, M.T. Ferdig, N.V. Chawla, A supervised learning approach to the ensemble clustering of genes, Int. J. Data Min. Bioinform. 9 (2) (2014) 199–219.
[38] R. Polikar, A. Topalis, D. Parikh, D. Green, J. Frymiare, J. Kounios, C.M. Clark, An ensemble based data fusion approach for early diagnosis of alzheimer's disease, Inf. Fusion 9 (1) (2008) 83–95.
[39] M. Termenon, M. Graña, A two stage sequential ensemble applied to the classification of alzheimer's disease based on mri features, Neural Process. Lett. 35 (1) (2012) 1–12.
[40] S. Mestizo Gutiérrez, M. Herrera Rivero, N. Cruz Ramírez, E. Hernández, G.E. Aranda-Abreu, Decision trees for the analysis of genes involved in alzheimer's disease pathology, J. Theor. Biol. 357 (0) (2014) 21–25.
[41] M. Martínez-Ballesteros, S. Salcedo-Sanz, J. Riquelme, C. Casanova-Mateo, J. Camacho, Evolutionary association rules for total ozone content modeling from satellite observations, Chemomet. Intell. Lab. Syst. 109 (2) (2011) 217–227.
[42] M. Martínez-Ballesteros, J. Riquelme, Analysis of measures of quantitative association rules, in: Proceedings of the International Conference on Hybrid Artificial Intelligent Systems, in: Lecture Notes in Computer Science, 6679, 2011, pp. 319–326.
[43] I.H. Witten, E. Frank, Data mining: practical machine learning tools and techniques, Morgan Kaufmann Series in Data Management Systems, 2nd, Morgan Kaufmann, 2005.
[44] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Series B (Methodol.) 57 (1) (1995) 289–300.
[45] Pubmed resource, 2015, (http://www.ncbi.nlm.nih.gov/pubmed/). [Online; accessed in October 2015].
[46] F. Al-Shahrour, P. Minguez, J. Tárraga, et al., FatiGO: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments, Nucleic Acids Res. 35 (2007) W91–W96.
[47] M. Soler-López, A. Zanzoni, R. Lluís, U. Stelzl, P. Aloy, Interactome mapping suggests new mechanistic details underlying Alzheimer's disease, Genome Res. 21 (3) (2011) 364–376.
[48] David tools, 2016, (https://david-d.ncifcrf.gov/). [Online; accessed in May 2016].
[49] N. Rappaport, N. Nativ, G. Stelzer, M. Twik, Y. Guan-Golan, T.I. Stein, I. Bahir, F. Belinky, C.P. Morrey, M. Safran, D. Lancet, Malacards: an integrated compendium for diseases and their annotation, Database 2013 (2013), doi:10.1093/database/bat018.
[50] T. Dunckley, T. Beach, K. Ramsey, A. Grover, D. Mastroeni, D. Walker, B. LaFleur, K. Coon, K. Brown, R. Caselli, W. Kukull, R. Higdon, D. McKeel, J. Morris, C. Hulette, D. Schmechel, E. Reiman, J. Rogers, D. Stephan, Gene expression correlates of neurofibrillary tangles in alzheimer's disease., Neurobiol. Aging 27 (2006) 1359–1371.
[51] R. Irizarry, Z. Wu, H. Jaffee, Comparison of affymetrix genechip expression measures., Bioinformatics 22 (2006) 789–794.
[52] G. Dennis, B. Sherman, D. Hosack, J. Yang, W. Gao, H. Lane, R. Lempicki, David: database for annotation, visualization, and integrated discovery., Genome Biol. 4 (2003) P3.
[53] G. Pison, A. Struyf, P.J. Rousseeuw, Displaying a clustering with clusplot, Comput. Stat. Data Anal. 30 (4) (1999) 381–392.
[54] A.A. Margolin, S.-E. Ong, M. Schenone, R. Gould, S.L. Schreiber, S.A. Carr, T.R. Golub, Empirical bayes analysis of quantitative proteomics experiments, PLoS ONE 4 (10) (2009) e7454.
[55] K. Iqbal, I. Grundke-Iqbal, Alzheimer's disease, a multifactorial disorder seeking multitherapies, Alzheimer's Dementia 6 (5) (2010) 420–424.
[56] M. Storandt, D. Head, A.M. Fagan, D.M. Holtzman, J.C. Morris, Toward a multifactorial model of alzheimer disease, Neurobiol. Aging 33 (10) (2012) 2262–2271.
[57] L. Zhang, X. Ju, Y. Cheng, X. Guo, T. Wen, Identifying tmem59 related gene regulatory network of mouse neural stem cell from a compendium of expression profiles, BMC Syst. Biol. 5 (1) (2011) 1–12.
[58] R. Yi, B. Chen, J. Zhao, X. Zhan, L. Zhang, X. Liu, Q. Dong, Kruppel-like factor 8 ameliorates alzheimer's disease by activating β-catenin, J. Mol. Neurosci. 52 (2) (2014) 231–241.
[59] M. Ray, W. Zhang, Analysis of alzheimer's disease severity across brain regions by topological analysis of gene co-expression networks, BMC Syst. Biol. 4 (1) (2010) 1–11.
[60] M.-G. Hong, A. Alexeyenko, J.-C. Lambert, P. Amouyel, J.A. Prince, Genome-wide pathway analysis implicates intracellular transmembrane protein transport in alzheimer disease, J. Hum. Genet. 55 (10) (2010) 707–709.
[61] C.M. Hales, H. Rees, N.T. Seyfried, E.B. Dammer, D.M. Duong, M. Gearing, T.J. Montine, J.C. Troncoso, M. Thambisetty, A.I. Levey, J.J. Lah, T.S. Wingo, Abnormal gephyrin immunoreactivity associated with alzheimer disease pathologic changes, J. Neuropathol. Exp. Neurol. 72 (11) (2013) 1009–1015.
[62] C. Mitchelmore, S. Buchmann-Moller, L. Rask, M.J. West, J.C. Troncoso, N.A. Jensen, Ndrg2: a novel alzheimer's disease associated protein, Neurobiol. Dis. 16 (1) (2004) 48–58.
[63] A. Jones, K. Friedrich, M. Rohm, M. Schafer, C. Algire, P. Kulozik, O. Seibert, K. Muller-Decker, T. Sijmonsma, D. Strzoda, C. Sticht, N. Gretz, G.M. Dallinga-Thie, B. Leuchs, M. Kogl, W. Stremmel, M.B. Diaz, S. Herzig, Tsc22d4 is a molecular output of hepatic wasting metabolism, EMBO Mol. Med. 5 (2) (2013) 294–308.
[64] C. Chung, J. Liao, B. Liu, X. Rao, P. Jay, P. Berta, K. Shuai, Specific inhibition of stat3 signal transduction by pias3, Science (New York, N.Y.) 278 (5344) (1997) 1803–1805.
[65] C. Chung, J. Liao, B. Liu, X. Rao, P. Jay, P. Berta, K. Shuai, Tyk2/stat3 signaling mediates β-amyloid-induced neuronal cell death: implications in alzheimer's disease, J. Neurosci. 30 (20) (2010) 6873–6881.
[66] V. Voikar, J. Rossi, H. Rauvala, M.S. Airaksinen, Impaired behavioural flexibility and memory in mice lacking gdnf family receptor α2, European. J. Neurosci. 20 (1) (2004) 308–312.
[67] M.C. Carreiras, E. Mendes, M.J. Perry, A.P. Francisco, J. Marco-Contelles, The multifactorial nature of alzheimer's disease for developing potential therapeutics., Curr. Top. Med. Chem. 13 (15) (2013) 1745–1770.
[68] D. Mizuno, M. Kawahara, The molecular mechanisms of zinc neurotoxicity and the pathogenesis of vascular type senile dementia, Int. J. Mol. Sci. 14 (11) (2013) 22067–22081.
[69] R. González-Domínguez, T. García-Barrera, J.L. Gómez-Ariza, Homeostasis of metals in the progression of alzheimer's disease, BioMetals 27 (3) (2014) 539–549.
[70] I.J.W. Nicole T. Watt, N.M. Hooper, The role of zinc in alzheimer's disease, Int. J. Alzheimers Dis. 2011 (2011) 1–10.
[71] C. Rosendorff, M. Beeri, J. Silverman, Cardiovascular risk factors for Alzheimer's disease, Am. J. Geriatr. Cardiol. 16 (3) (2007) 143–149.
[72] R. Stewart, Cardiovascular factors in Alzheimer's disease, J. Neurol., Neurosurg. Psychiatry 65 (2) (1998) 143–147.
[73] S. Craft, S. Watson, Insulin and neurodegenerative disease: shared and specific mechanisms, Lancet Neurol. 3 (3) (2004) 169–178.
[74] C. MacKnight, K. Rockwood, E. Awalt, I. McDowell, Diabetes mellitus and the risk of dementia, alzheimer's disease and vascular cognitive impairment in the canadian study of health and aging, Dement Geriatr Cogn Disord. 14 (2002) 77–83.
[75] J. Janson, T. Laedtke, J. Parisi, P. O'Brien, Petersen, R.C., P. Butler, Increased risk of type 2 diabetes in alzheimer disease, Diabetes 53 (2) (2004) 474–481.
[76] R. Graf, M. Munschauer, G. Mastrobuoni, F. Mayr, U. Heinemann, S. Kempa, N. Rajewsky, M. Landthaler, Identification of lin28b-bound mrnas reveals features of target recognition and regulation, RNA Biol. 10 (7) (2013) 1146–1159.
[77] E. Ahmady, S.A. Deeke, S. Rabaa, L. Kouri, L. Kenney, A.F.R. Stewart, P.G. Burgon, Identification of a novel muscle a-type lamin-interacting protein (mlip), J. Biol. Chem. 286 (22) (2011) 19702–19713.
[78] S. Sekar, J. McDonald, L. Cuyugan, J. Aldrich, A. Kurdoglu, J. Adkins, G. Serrano, T.G. Beach, D.W. Craig, J. Valla, E.M. Reiman, W.S. Liang, Alzheimer's disease is associated with altered expression of genes involved in immune response and mitochondrial processes in astrocytes, Neurobiol. Aging 36 (2) (2015) 583–591.
[79] X.-F. Chen, Y.-w. Zhang, H. Xu, G. Bu, Transcriptional regulation and its misregulation in alzheimer's disease, Mol. Brain 6 (1) (2013) 1–9.
[80] W.S. Liang, T. Dunckley, T.G. Beach, A. Grover, D. Mastroeni, K. Ramsey, R.J. Caselli, W.A. Kukull, D. McKeel, J.C. Morris, C.M. Hulette, D. Schmechel, E.M. Reiman, J. Rogers, D.A. Stephan, Altered neuronal gene expression in brain regions differentially affected by alzheimer's disease: a reference data set, Physiol. Genomics 33 (2) (2008) 240–256.
[81] B.E.L. Lauffer, R. Mintzer, R. Fong, S. Mukund, C. Tam, I. Zilberleyb, B. Flicke, A. Ritscher, G. Fedorowicz, R. Vallero, D.F. Ortwine, J. Gunzner, Z. Modrusan, L. Neumann, C.M. Koth, P.J. Lupardus, J.S. Kaminker, C.E. Heise, P. Steiner, Histone deacetylase (hdac) inhibitor kinetic rate constants correlate with cellular histone acetylation but not transcription and cell viability, J. Biol. Chem. 288 (37) (2013) 26926–26943.