Ensemble transcript interaction networks: A case study on Alzheimer's disease

Rubén Armañanzas, Pedro Larrañaga, Concha Bielza*

Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Madrid, Spain

A B S T R A C T

Systems biology techniques are a topic of recent interest within the neurological field. Computational intelligence (CI) addresses this holistic perspective by means of consensus or ensemble techniques ultimately capable of uncovering new and relevant findings. In this paper, we propose the application of a CI approach based on ensemble Bayesian network classifiers and multivariate feature subset selection to induce probabilistic dependences that could match or unveil biological relationships. The research focuses on the analysis of high-throughput Alzheimer's disease (AD) transcript profiling. The analysis is conducted from two perspectives. First, we compare the expression profiles of hippocampus subregion entorhinal cortex (EC) samples of AD patients and controls. Second, we use the ensemble approach to study four types of samples: EC and dentate gyrus (DG) samples from both patients and controls. Results disclose transcript interaction networks with remarkable structures and genes not directly related to AD by previous studies. The ensemble is able to identify a variety of transcripts that play key roles in other neurological pathologies. Classical statistical assessment by means of non-parametric tests confirms the relevance of the majority of the transcripts. The ensemble approach pinpoints key metabolic mechanisms that could lead to new findings in the pathogenesis and development of AD.

Keywords:
Bayesian network classifiers
Interaction networks
Alzheimer's disease
High-throughput data

1. Introduction

Computational intelligence (CI) techniques have proven able to help physicians to analyze gene activities within complex diseases. Following this breakthrough research, CI-driven systems biology has recently gained interest within the neurological community as a tool for unveiling new findings and/or proposing new working hypotheses [1]. Up to now, one of the biggest challenges in this field has been to look for key genetic mechanisms and compounds in complex neurodegenerative pathologies. In actual fact, tools to address biological relationships and interactions are currently at the cutting edge [2].

Many approaches have been put on trial [3] in order to understand such complex relations, ranging from pure Bayesian networks [4] to statistical validations by multiple random simulation [5], new graphical models to match gene interactions [6] or biological validation of previously reported interactions [7]. The main thrust of all this research is to assume that a gene transcript behaves like a random variable and that the behavior of the entire system can be represented by a joint probability distribution. The regulatory interactions between the transcripts across that distribution are expected to produce corresponding probabilistic dependences within their expression levels [8].

Within this framework, most research looks for differentially expressed genes to build models. However, fewer papers

explicitly focus on the statistical information provided by comparing different sample types. In such a supervised-class experimental design, the phenotype statistical distribution may report interactions among genes based on their behavior across the different conditions [3]. It is possible to assign confidence levels to interactions based on the frequency of their appearance in an induced pool of Bayesian classifiers. Depending on the confidence level, the expert can set up different interaction networks, ranging from very simple to dense forest-like structures.

The techniques to produce this hierarchy of ensemble networks are borrowed from the field of machine learning and statistics. First, Bayesian classifiers use no *a priori* biological information and consider only the phenotype distribution. Second, a feature subset selection procedure is used to reduce the dimensionality from thousands to only hundreds of candidate genes [9]. And, third, results produced by *non-parametric bootstrap* – dataset sampled with replacement – are conservative. In scenarios where the number of samples is very low, it is crucial to aim for a low ratio of false positive findings. Last, ensemble or consensus techniques reinforce the search of robust and reliable gene interactions [10,11].

Throughout this paper, we use this ensemble approach to investigate a gene expression dataset of Alzheimer's disease (AD). The analysis focuses on gaining an understanding of dysregulation in the hippocampal entorhinal cortex (EC), as well as on the multiple comparison of the hippocampal entorhinal cortex and dentate gyrus (DG). The aim behind this research is to formulate working hypotheses about why there is such a big difference in the extent of the damage to the above hippocampal structures between elderly AD patients and healthy controls.

The paper is organized as follows. Section 2 presents the dataset, the experimental design and the induction of ensemble Bayesian networks, respectively. Section 3 shows the experimental parameters and running results for both analytical comparisons. Section 4 gives an in-depth biological discussion of the most important findings, corroborating previous knowledge and stating new hypotheses based on the reported results. Lastly, Section 5 explains the conclusions and ideas for future work. Supplementary content is available with extended information on all the results and interaction networks.

2. Materials and methods

2.1. *Microarray data*

The available data contain gene expression profiles from six AD and six control brain samples. The samples were obtained at autopsy, and there are two different cohorts: one from the DG region and another from the EC subregion of the hippocampus.

The microarray technology used to retrieve gene activity was an Affymetrix HG-U133A genechip array. A single sample is hybridized for each array, thereby outputting a total of 24 hybridized arrays. The acquired microarray dataset was scaled to a value of 500, and probes with a 3′/5′ ratio in the GADPH and actin gene greater than 7 were excluded from the study. A total of 7610 probes that passed the last Affymetrix detection algorithm filter were retained as valid probes for the subsequent data analysis (MAS 5 and GeneSpring 5.0.3 were the tools used for the process). For more details on each of these steps and the actual dataset, see the original paper [12]. Throughout this paper, we will use the term transcript as the product measured by each probe of the microarray. Similarly, the term gene will refer to the associated gene from which that transcript is synthesized. The terms were matched by the *Ensembl BioMart* tool, using the Ensembl Genes Release 62 dataset.

2.2. *Experimental design*

Experimental design refers to the way that the different gene profiles are configured in a supervised classification problem. To get all the possible options, we considered a list of biological facts supported by the original study:

- EC appears to be a prime target for AD, as it is highly vulnerable to the effects of ischemia and anoxia;
- DG is the neighboring subregion of the EC most resistant to AD;
- differences in entorhinal function between controls and AD patients are age and time independent.

All these statements reveal two different biologically important scenarios: (a) the comparison of the EC gene profiles between AD patients and controls; and (b) a multiclass study with other combinations of samples and individuals.

Therefore, two different data mining analyses were conducted: EC-AD *vs* EC-Control (see Section 3.2.1) with 12 samples in a dichotomic supervised classification problem; and EC-AD *vs* EC-Control *vs* DG-AD *vs* DG-Control (see Section 3.2.2) with all 24 samples in a four-class (or multiclass) supervised classification problem. This experimental design substantially differs from the design used by [12]. The analysis pipelines were also different in terms of running scheme and mathematical approaches. Therefore, final results showed a limited degree of coincidence between both studies.

2.3. *Ensemble Bayesian networks of highly reliable dependences*

The data analysis methodology combines a resampling method with an inner feature selection technique and a Bayesian k-dependence classifier to output a gene interaction network formed by arcs above a set confidence level. This chained process can be used as a tool to unveil or corroborate biological hypotheses [13].

The method for building the ensemble Bayesian networks was originally proposed in [10]. It is based on searching robust arcs from the whole set of arcs configured by a pool of Bayesian networks classifiers (BNC). Briefly, we can define the number of occurrences o_{ij} as the number of times a given arc l_{ij} – where X_i and X_j are head and tail nodes, respectively – has been induced across B datasets. These B datasets correspond to the B bootstrapped datasets from the original dataset. For each resampled dataset an intermediate feature subset selection process is also tackled to select the most relevant genes.

After the induction of all the BNCs, it is possible to define an occurrence threshold t of reliability. Using that threshold, arcs with occurrences equal to or greater than t are retained. This set of retained arcs is denoted as L_t. By inspecting L_t, it is also possible to state which set of variables or features is covered by all arcs. This feature set comprises the relevant features subset and is denoted by $S(L_t)$.

Interestingly, by changing t, we can build a hierarchy of models, ranging from a model with just one arc and hence two features, to a model that includes almost all the detected arcs or conditional dependences. Since the method looks for BNCs, arcs that form cycles are removed. Cycles of more than two variables are unfeasible due to the formulation of the BNC in use. Finally, given a t value, the expert is reported with the correspondent network structure. It is therefore possible to control the scope of the study and to isolate findings that could constitute future working hypotheses.

2.4. *Differential expression measures*

To supplement the results of the ensemble networks, two different definitions of fold-change (FC) were used to compare the expressions of the identified relevant transcripts. These two FCs are the classical expression ratio or FC_r and the expression difference or FC_d. Thus, FC_r is defined as the ratio between the median value of the transcript expressions within the disease and the control samples, whereas FC_d compares the same values but removing the median control expression from the median disease expression. We here propose the use of two different univariate hypothesis tests to check the significance of the transcripts previously detected as relevant by the ensemble of networks.

In dichotomic studies, it is common to assess significance using p-values in a t-test. Unfortunately, this is not such a good approach when there are few samples and normal distribution assumptions cannot be guaranteed. The alternative proposed here is to use a non-parametric test: the Wilcoxon rank sum test for equal medians.

The Wilcoxon test is only applicable with two samples, but we require a four-factor test to analyze the results of the multiclass experimental design. In this case, we used Friedman test for multiple treatments of a series of subjects [14]. The Friedman test is able to jointly compare the activity of each transcript evaluated in the four different class configurations. This is the best test for investigating the significance of the differences between the four phenotypes in the multiclass problem.

3. Results

3.1. *Running parameters*

The data analysis method introduced in Section 2.3 includes a set of running parameters to be fixed: the feature subset selection, the BNC to be induced and the number of times that the bootstrap loop is performed, B. Also, and especially in the gene expression context, all these parameters are expected to setup a scenario with an affordable runtime.

For the subset selection step, we used correlation-based feature subset selection [9] or CFS. CFS finds low redundancy feature subsets, which are also highly correlated to the supervised class variable. The CFS search strategy was configured as a forward greedy hill-climbing search that starts from an empty set of features. This search strategy guarantees that the cardinality dimension of the output subsets is not high.

After reducing the dataset using CFS, a kDB classifier is induced using a k value of 4 [15]. Thanks to this Bayesian classifier and the fixed k value the graphical models are both flexible (capable of inducing many diverse structures) and not sparse when inducing the dependence structures. Since the number of available instances is particularly low, experiments had to be as robust as possible. Therefore, the number of times the main bootstrap loop is performed was set to $B = 10{,}000$ times for both data mining analyses. In addition, this high sampling rate prevents trapping in local optima.

Bayesian classifiers typically deal with discrete variables. Hence, a process was run to discretize the original continuous data. On the basis of its biological activity, we assume that a gene does not have many different activity states. A general criterion in microarray analysis is that this number of states is three: an up-regulated, a down-regulated and a baseline or null activity. Taking up this idea, we considered equal width discretization [16] in three different bins to be the best method for parsing the continuous values into discrete states. Any bias included by the discretization is not expected to affect the real gene profiling behavior [17,18].

3.2. *Ensemble interaction networks*

As discussed in Section 2.3, the practitioner must set an occurrence threshold t to output both the list of interaction networks and the associated list of highly relevant features. The number of times two variables are jointly selected and included as head and tail of the same conditional dependence is strongly influenced by the running parameters in use. Hence, the empirical distribution of the computed values is completely unknown. This distribution is formed by a X-axis reflecting the number of occurrences an arc was included, whereas Y-axis shows how many arcs reach such threshold. Experimentally, the distribution of the results is right-skewed with extreme values on the right tail of the plot. These are precisely the most relevant values: a small set of arcs with very high occurrence values. In order to retain only these extreme links, we retrieve the 0.999 quantile from the empirical distribution. The associated threshold levels for such quantile are detailed for both tackled analyses (see Sections 3.2.1 and 3.2.2). Both full transcript networks are available as supplementary content.

3.2.1. *EC-AD vs EC-Control*

The total number of different arcs was 135,880. The most frequent dependence identified from the above arcs was between probes 200099_s_at and 201358_s_at (transcripts RPS3A and COPB1), with 2666 occurrences. Filtering these values to the 0.999 quantile, we got a threshold level of 577 ($t = 576.44$). A total number of 63 probes and 136 conditional dependences were retained for this level (see Supplementary Tables 1 and 2 for the full probe/gene list).

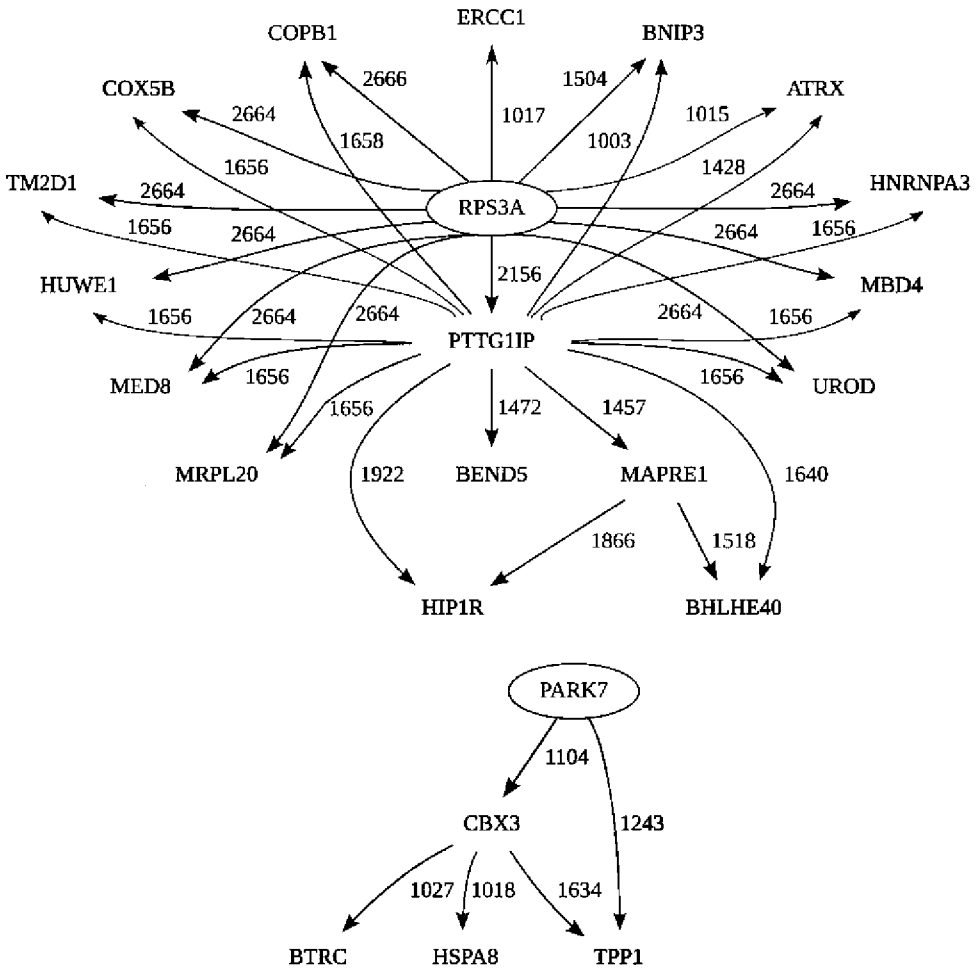

Fig. 1 – Results for the EC comparison between controls and AD patients. The reliability threshold is set to 1000 out of 10,000 main iterations. The network includes 23 transcripts and 35 conditional dependences. Arc labels represent the number of times the associated arc and nodes have been found together across the 10,000 models. Unfilled nodes map root nodes in the structure, whereas shaded nodes denote child nodes.

The full interaction network described by the 136 arcs is available as supplementary content. To illustrate the core network and the most important transcripts, Fig. 1 shows the subnetwork that corresponds to a t level of 1000 – with 35 conditional dependences and 23 probes. Notice that Fig. 1 illustrates two disconnected graphs for this threshold. These two structures are connected in the original $t=577$ network.

3.2.2. *EC-AD vs EC-Control vs DG-AD vs DG-Control*

More than 500,000 arcs were configured. Of these, 28,355 arcs were repeatedly presented in more than five runs. The dependence between 200872_at and 201170_s_at (transcripts S100A10 and BHLHE40) was present in 1875 out of all 10,000 models. Following the extreme quantile selection criterion (0.999), the threshold level was set into 703 ($t=702.87$). This retrieved a total of 28 arcs comprising 22 probes (listed in Supplementary Table 3). Fig. 2 shows the most robust interaction network structure. As in the previous analysis, the confidence threshold for Fig. 2 has been raised to 1000, including 9 transcripts and 12 highly reliable conditional dependences.

4. Discussion

Throughout this section we will discuss the findings reported by the ensemble approach for both analyses. The discussion focuses on transcripts whose relevance within the neurodegenerative domain has been previously proven. Similarly, new possibly relevant transcripts are also explored in search of new working hypotheses.

4.1. *EC-AD vs EC-Control*

This comparison focuses on checking how neuronal death in the EC of AD patients changes the transcript profiling with respect to control samples. To do this, we firstly discuss the

Table 1 – Of the 63 probes reported by the ensemble in the EC comparison, 17 map disease-related genes and 10 out of the 17 are related to neurological pathologies (upper part).

Gene	Disease
BHLHE40	Bipolar disorder
ERCC1	Cerebral oculofacioskeletal syndrome 4
HUWE1	X-linked mental retardation
ATRX	Alpha-thalassemia myelodysplasia syndrome
PARK7	Parkinson's disease
RAB3GAP1	Warburg micro syndrome 1
VPS13B	Cohen syndrome
HIP1R	Huntington's disease
COMT	Panic disorder, susceptibility to schizophrenia
TPP1	Late-infantile neuronal ceroid lipofuscinosis
BIN1	Centronuclear myopathy
TPI1	Hemolytic anemia
IL6R	Multiple myeloma
KRT10	Epidermolytic hyperkeratosis
HLA-DPB1	Beryllium disease
PPARGC1A	Familiar lipodystrophy
UROD	Porphyria cutanea tarda

transcripts that are somehow linked with human diseases. Significantly, 17 out of the 63 probes found by the network ensemble are related to or are triggers of several diseases. More interestingly, 10 out of these 17 probes are probes related to neurological disorders. Table 1 lists both sets of associated genes.

To discuss the possible rationale behind the presence of these genes, Table 2 presents the p-values obtained from the Wilcoxon rank sum test comparing the expression profiles. Notice that only transcripts illustrated in Fig. 1 are listed. A total of 18 out of 23 transcripts exhibit statistically significant differences at $\alpha = 0.05$. Moreover, the first 10 transcripts have a p-value lower than 0.01. These results corroborate the key role played by these transcripts, previously reported by the multivariate ensemble structure. The first ten transcripts are sorted in three groups with increasing p-values. In addition, by using the Genotator text-mining database [19], 7 out of the 23 transcripts are traced to have key roles in the reported literature of AD: RPS3A, BTRC, TM2D1, PARK7, COX5B, TPP1 and HSPA8. There follows a brief biological discussion of some important links between the transcripts and the condition.

4.1.1. *AD pathogenesis and/or metabolism*

One of the most relevant transcripts in terms of both network connections and p-value is RPS3A. This gene has highly reliable dependences on almost all the genes reported by the network at a confidence level of 1000. This topological position may have a direct correspondence in biology. Ribosomes consist of a small 40S subunit and a large 60S subunit that, together, are composed of 80 structurally distinct proteins. The RPS3A gene encodes a S3AE ribosomal protein that is a component of the 40S subunit. Grupe et al. [20] performed an association study of candidate genes on chromosome 10 for triggering late-onset Alzheimer's disease (LOAD). They conducted two rounds of analyses with a total of 779 LOAD samples and 629 controls. The study analyzed 1412 SNPs with 677 putative functional mutations. Results reported just five relevant mutations. Of these, marker rs498055, located in a gene homologous to RPS3A, was significantly associated with AD with an allelic p-value of 0.0001. This study implicates RPS3A gene in the pathogenesis of this disorder. Looking at the fold-change values in Table 2, we can corroborate that its activity in the AD entorhinal cortex is greatly underexpressed compared with the control samples: a 4.9521 logRatio decrease. Also there is a significant variance in its expression level across the control samples, whereas AD samples have a low constant expression (see Supplementary Fig. 1).

MED8 transcript is part of the 20 subunits of the mediator complex, which is required to activate mRNA production. MED8 presents interactions with proteins of key relevance in the central nervous system, like ARRB2, which plays a role in the regulation of synaptic receptors by inhibiting beta-adrenergic receptor function. Another target protein is CCNC or cyclin C, whose expression has been proven to be involved in the pathogenesis of Alzheimer's disease [21].

TM2D1 is a beta-amyloid peptide-binding protein. Beta-amyloid peptide has a toxic effect on neurons, including death, morphological and physiological alterations (among others) and the consequent loss of cognitive abilities observed in AD [22]. TM2D1 interacts with APP amyloid beta (A4) precursor protein, which is a cell surface receptor and transmembrane precursor protein that is cleaved by secretases to form a number of peptides. Some of these peptides form the amyloid protein plaques found in the brains of patients with Alzheimer disease.

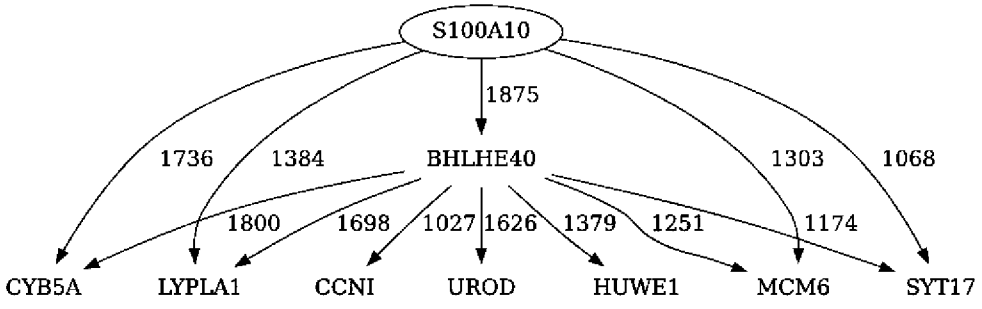

Fig. 2 – Core results for the four-class supervised problem data analysis: EC-AD *vs* EC-Control *vs* DG-AD *vs* DG-Control. Individual labels represent the occurrence level of each arc (t is set to 1000 times out of the 10,000 bootstrap samplings). Unfilled nodes map root nodes in the structure, whereas shaded nodes denote child nodes.

Table 2 – Fold-change expressions of the transcripts found to have highest relevancies in the entorhinal cortex comparison between AD patients and controls. Fold-change is expressed in three complementary ways, namely, ratio (FC_r), logarithmic transformation of the ratio ($\log_2 FC_r$) and difference (FC_d). The p-value in the last column shows the output of a Wilcoxon rank sum test to compare the expression levels. Transcripts are listed in increasing p-value order.

	FC_r	$\log_2 FC_r$	FC_d	p-Value
UROD	0.2884	−1.7938	−860.7	0.0022
HIP1R	0.3416	−1.5497	−777.7	0.0022
BEND5	0.7262	−0.4615	−125.3	0.0022
ATRX	0.7919	−0.3366	−492.8	0.0022
RPS3A	0.0323	−4.9521	−8889	0.0043
BHLHE40	0.5713	−0.8077	−846.4	0.0043
BTRC	1.3652	0.4491	63.6	0.0043
MED8	0.0431	−4.5350	−515.8	0.0087
TM2D1	0.2419	−2.0476	−622.0	0.0087
PARK7	0.7135	−0.4871	−3116	0.0087
HNRNPA3	0.0592	−4.0774	−628.9	0.0152
MBD4	0.1782	−2.4884	−378.9	0.0152
MRPL20	0.0105	−6.5684	−1300.6	0.0260
HUWE1	0.0624	−4.0013	−2427.8	0.0260
ERCC1	0.1833	−2.4479	−409.3	0.0260
COX5B	0.8077	−0.3081	−1026.6	0.0260
PTTG1IP	1.5275	0.6112	951.5	0.0260
TPP1	1.4906	0.5759	603.1	0.0411
COPB1	0.3177	−1.6544	−1475.3	0.0649
BNIP3	0.5631	−0.8286	−1861.2	0.0649
MAPRE1	1.1517	0.2037	67.7	0.2403
HSPA8	0.6991	−0.5164	−1783.3	0.3095
CBX3	0.8795	−0.1852	−140.9	0.3939

4.1.2. *Circadian rhythm*

One of the symptoms in severe AD is the asynchronity of the circadian rhythm [23]. Of our list of genes, BHLHE40, also known as DEC1, plays a role in the finer regulation and robustness of the molecular clock components CLOCK/BMAL1 [24]. Previous research pointed to CLOCK gene regulation in the correct setup of the circadian rhythm of the metabolism. In our experiments, the gene profiling of BHLHE40 shows an underexpression in AD samples, a fact that may be linked with the dysfunction of CLOCK and, therefore, of the circadian rhythm. Another transcript find to be relevant, BTRC, associates with beta-catenin destruction motifs by activating the NF-kappaB pathway and inhibiting the beta-catenin pathway. Such inhibition has already been studied as an alteration of the circadian clock gene expression in mice [25].

These findings match the hypotheses of [26], stating that amyloid beta production follows the circadian rhythm, rising when a person is awake and falling during sleep. [26] also suggests that excessive sleep debt could cause a chronic build-up of amyloid beta protein, which could hypothetically lead to AD.

4.1.3. *Other central nervous system diseases*

To the best of our knowledge the other 5 transcripts have not been previously related to AD. However, they might play roles in the central nervous system metabolism since they are all related (or have domain interactions) to several other neurological diseases. Among them, HIP1R is named after the Huntingtin-interacting protein, or, mutations in ATRX are associated with the X-linked mental retardation syndrome. For PARK7, it has been widely studied because its defects are the cause of autosomal recessive early-onset Parkinson's disease.

4.2. *EC-AD vs EC-Control vs DG-AD vs DG-Control*

The aim of this second analysis is to locate relevant transcripts and/or relationships that are differentially expressed in all four tissues under study. Results in Section 3.2.2 reported a simpler network structure (see Fig. 2) than in the case of single EC comparison. Even so, there are some similarities: the presence of the HUWE1, UROD and BHLHE40 transcripts.

As in the previous analysis, Table 3 lists the p-values for three sets of hypothesis tests applied to the set of transcripts found in Fig. 2. The first column, labeled as *PATvsCON*, contrasts the expression profiles of patient and control samples no matters what the type of tissue is. The second column, *ECvsDG*, groups the profiles by tissue type. In both cases, the Wilcoxon test is used to retrieve the associated p-values. The column labeled *MULTICLASS* lists the p-value output by the Friedman hypothesis test, comparing all four types of tissue-patient profiles as in the ensemble network analysis.

Let us examine the values of the Friedman test in Table 3, as they were computed in a similar manner to the ensemble network. Hence, there are a total of 7 out of 9 transcripts that show statistical significance at $\alpha = 0.05$. This supports the findings from the networks and adds even more robustness to the data analysis. We also used Genotator text-mining database [19] to check the already published relation between these genes and AD: 3 out of the 9 transcripts seem to have key roles in AD: S100A10, CYB5A and CCNI. There follows a brief discussion of the biological foundations between some of these transcripts and the condition.

4.2.1. *AD pathogenesis and/or metabolism*

In the network topology of Fig. 2, S100A10 forms a central node. This is the only transcript that is not conditionally

Table 3 – Log_2 fold-change expressions and p-values of the transcripts found to have highest relevancies in the multiple class comparison. The first two columns list values output by grouping the EC and DG expressions and computing the changes comparing the activities of patients against controls (PATvsCON). The third and fourth columns list values output by grouping the patient and control expressions and computing the changes comparing the activities of both classes in EC and DG (ECvsDG). Finally, the multiclass column presents the p-values of a Friedman test for multiple comparison of all four expressions. Transcripts are listed in increasing order of Friedman test p-value.

	PATvsCON		ECvsDG		MULTICLASS
	$\log_2 FC_r$	p-Value	$\log_2 FC_r$	p-Value	p-Value
S100A10	1.4880	<0.0002	−0.2879	0.0120	<0.0008
MCM6	−0.3244	0.0051	0.0509	0.5444	0.0074
LYPLA1	−0.5938	0.0006	0.0996	0.5444	0.0081
UROD	−1.8375	<0.0001	−0.3531	0.0226	0.0169
HUWE1	−4.6466	<0.0002	−0.5201	0.8852	0.0186
BHLHE40	−0.6219	0.2602	0.4674	0.1124	0.0203
CYB5A	−0.1370	0.1410	−0.1907	0.0262	0.0293
SYT17	0.0253	0.8852	−0.7766	0.0783	0.1447
CCNI	−0.1761	0.4357	0.3873	0.7950	0.2214

dependent on any other, and its links are within the highest level occurrences set. S100A10 plays a pivotal role in the dynamic modulation of serotonergic 1B receptor function, and is involved in the pathogenesis of major depressive disorder (MDD) and the therapeutic mechanisms of antidepressant action [27]. In our data, the expression profile shows a clear dysregulation between AD and control samples in both DG and EC. A recent study [28] identifies another S100 family member (S100A7) as a novel biomarker of AD involved in the attenuation of beta amyloid neuropathology in mice. The findings of [28] suggest that S100A7 expression in the brain promotes α-secretase activity in the AD brain, precluding the generation of amyloidogenic β-amyloid peptides. This overexpression matches our expression profiles for S100A10 (see Supplementary Fig. 2). The quantitative p-value reported by the Friedman test also corroborates the key role identified by the ensemble network.

4.2.2. *Circadian rhythm*

BHLHE40 has previously been proposed as a key transcript in the circadian dysregulation symptom of AD patients, and hence in the beta-amyloid production. Its expression profile shows that only in the EC tissue is a differential expression detected. For the DG, the expression levels are similar in both AD and control samples, where variance is higher in the control tissue (see Supplementary Fig. 2). This expression profile makes full sense in the light of recent research on the importance of EC only in the regulation of the circadian rhythm of the hypothalamic–pituitary–adrenal (HPA) axis [29].

4.2.3. *Other central nervous system diseases*

UROD shows a clear underexpression in both the DG and EC of AD samples. The decrease in UROD concentration alters the production of heme and, hence, hemoglobin. Hemoglobin is distributed in AD patients in a brain region-dependent manner, where the highest levels are to be found in the hippocampus [30]. Our finding may corroborate previous hypotheses that hemoglobin levels are inversely related to the AD condition [31–33]. We also find the HUWE1 transcript, which is directly related to the X-linked mental retardation syndrome [34]. It regulates neural differentiation and proliferation.

5. Conclusions

Systems biology is breaking new ground in search of answers to the complex and devastating neurodegenerative disease domain. Of these diseases, Alzheimer's disease is one of the best known, affecting millions of elderly people worldwide. Computational intelligence techniques are now developing promising approaches and reporting results that build bridges between disciplines [35].

In this study, we focused on a data mining approach that is able to retrieve key transcripts in high-throughput gene expression analysis. Since the number of samples in this kind of analysis is still very low, reliable approaches are required to dispel the so-called *curse of dimensionality*. To do this, we tackled two different supervised classification problems based on Alzheimer's disease microarray data. First, we compared the gene expression profiles of patient and control samples collected from the entorhinal cortex. Second, we compared all these samples with the respective dentate gyrus hippocampal subregions.

The computational intelligence approach on trial makes use of three different and complementary machine learning approaches: bootstrap resampling, multivariate filter subset selection and a Bayesian network classifier. As the output, this combination provides the researcher with a highly reliable ensemble gene-interaction network. Precisely, the aim of this research is to propose these findings as possible targets for deeper and more detailed studies.

The reported results suggest interesting findings. New potential relationships have been pinpointed, including the role of BHLHE40 in the regulation of the circadian rhythm in AD patients. Several other findings are a potential source of working hypotheses. Of these, we have discussed AD pathogenesis, transcription regulation and products related to other neurological conditions. In actual fact, ensemble network findings corroborate previous findings in AD: the importance of the RPS3A gene within AD pathogenesis or the inverse relationship of hemoglobin levels to the AD condition.

Practitioners in the field are aware that some of these relationships may be numerical artifacts. The need for multiple and independent validations of these statistical findings is crucial. However, the production cost of all these hypotheses

is really low, whereas their contributions could be enormous in terms of key insights for future research. Hence, more monitoring and validation will, as in any kind of biomedical research, be needed.

Conflict of interest

None declared.

Acknowledgements

We would like to thank Dr. Lidia Alonso-Nanclares for her review on the neuropathology discussion. This work has been partially supported by projects TIN2010-20900-C04-04, Consolider Ingenio 2010-CSD2007-00018 and Cajal Blue Brain of the Spanish Ministry of Science and Innovation (MICINN). R.A. is supported by a Juan de la Cierva postdoctoral fellowship (MICINN).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2011.11.011.

REFERENCES

[1] P. Villoslada, L. Steinman, S.E. Baranzini, Systems biology and its application to the understanding of neurological diseases, Annals of Neurology 65 (2) (2009) 124–139.
[2] F. Noorbakhsh, C.M. Overall, C. Power, Deciphering complex mechanisms in neurodegenerative diseases: the advent of systems biology, Trends in Neurosciences 32 (2) (2009) 88–100.
[3] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J.A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, V. Robles, Machine learning in bioinformatics, Briefings in Bioinformatics 7 (1) (2006) 86–112.
[4] J.M. Peña, J. Björkegren, J. Tegnér, Growing Bayesian network models of gene networks from seed genes, Bioinformatics 21 (Suppl. 2) (2005) ii224–ii229.
[5] S.G. Baker, B.S. Kramer, Identifying genes that contribute most to good classification in microarrays, BMC Bioinformatics 7 (2006) 407.
[6] K.-C. Liang, X. Wang, Gene regulatory network reconstruction using conditional mutual information, EURASIP Journal on Bioinformatics and Systems Biology 2008 (2008) 253894.
[7] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, J. Vert, Classification of microarray data using gene networks, BMC Bioinformatics 8 (1) (2007) 35.
[8] D. Pe'er, A. Tanay, A. Regev, Minreg: a scalable algorithm for learning parsimonious regulatory networks in yeast and mammals, Journal of Machine Learning Research 7 (2006) 167–189.
[9] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.
[10] R. Armañanzas, I. Inza, P. Larrañaga, Detecting reliable gene interactions by a hierarchy of Bayesian network classifiers, Computer Methods and Programs in Biomedicine 91 (2) (2008) 110–121.
[11] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, Y. Saeys, Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, Bioinformatics 26 (3) (2010) 392–398.
[12] S.A. Small, K. Kent, A. Pierce, C. Leung, M. Suk Kang, H. Okada, L. Honig, J.-P. Vonsattel, T.-W. Kim, Model-guided microarray implicates the retromer complex in Alzheimer's disease, Annals of Neurology 58 (6) (2005) 909–919.
[13] D. Otaegui, S. Baranzini, R. Armañanzas, B. Calvo, M. Muñoz-Culla, P. Khankhanian, I. Inza, J.A. Lozano, T. Castillo-Triviño, A. Asensio, J. Olascoaga, A. López de Munain, Differential micro RNA expression in PBMC from multiple sclerosis patients, PLoS One 4 (7) (2009) e6309.
[14] M. Friedman, A comparison of alternative tests of significance for the problem of m rankings, Annals of Mathematical Statistics 11 (1) (1940) 86–92.
[15] M. Sahami, Learning limited dependence Bayesian classifiers, in: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, The AAAI Press, California, 1996, pp. 335–338.
[16] R. Kerber, Chimerge: discretization for numeric attributes, in: National Conference on Artificial Intelligence, San Jose, California, 1992, pp. 123–128.
[17] A. Sáenz, M. Azpitarte, R. Armañanzas, F. Leturcq, A. Alzualde, I. Inza, F. García-Bragado, G. De la Herran, J. Corcuera, A. Cabello, C. Navarro, C. De la Torre, E. Gallardo, I. Illa, A. López de Munain, Gene expression profiling in limb-girdle muscular dystrophy 2A, PLoS One 3 (11) (2008) e3750.
[18] R. Armañanzas, B. Calvo, I. Inza, M. López-Hoyos, V. Martínez-Taboada, E. Ucar, I. Bernales, A. Fullaondo, P. Larrañaga, A.M. Zubiaga, Microarray analysis of autoimmune diseases by machine learning procedures, IEEE Transactions on Information Technology in Biomedicine 13 (3) (2009) 341–350.
[19] D.P. Wall, R. Pivovarov, M. Tong, J.-Y. Jung, V.A. Fusaro, T.F. DeLuca, P.J. Tonellato, Genotator: a disease-agnostic tool for genetic annotation of disease, BMC Medical Genomics 3 (2010) 50.
[20] A. Grupe, Y. Li, C. Rowland, P. Nowotny, A.L. Hinrichs, S. Smemo, J.S. Kauwe, T.J. Maxwell, S. Cherny, L. Doil, K. Tacey, R. van Luchene, A. Myers, F. Wavrant-De Vrièze, M. Kaleem, P. Hollingworth, L. Jehu, C. Foy, N. Archer, G. Hamilton, P. Holmans, C.M. Morris, J. Catanese, J. Sninsky, T.J. White, J. Powell, J. Hardy, M. O'Donovan, S. Lovestone, L. Jones, J.C. Morris, L. Thal, M. Owen, J. Williams, A. Goate, A scan of chromosome 10 identifies a novel locus showing strong association with late-onset Alzheimer disease, American Journal of Human Genetics 78 (1) (2006) 78–88.
[21] U. Ueberham, A. Hessel, T. Arendt, Cyclin C expression is involved in the pathogenesis of Alzheimer's disease, Neurobioloy of Aging 24 (3) (2003) 427–435.
[22] J.L. Price, J.C. Morris, So what if tangles precede plaques? Neurobiology of Aging 25 (6) (2004) 721–723.
[23] D.G. Harper, L. Volicer, E.G. Stopa, A.C. McKee, M. Nitta, A. Satlin, Disturbance of endogenous circadian rhythm in aging and Alzheimer disease, American Journal of Geriatric Psychiatry 13 (5) (2005) 359–368.
[24] A. Nakashima, T. Kawamoto, K.K. Honda, T. Ueshima, M. Noshiro, T. Iwata, K. Fujimoto, H. Kubo, S. Honma, N. Yorioka, N. Kohno, Y. Kato, Dec1 modulates the circadian phase of clock gene expression, Molecular and Cellular Biology 28 (12) (2008) 4080–4092.
[25] X. Yang, P.A. Wood, C.M. Ansell, M. Ohmori, E.Y. Oh, Y. Xiong, F.G. Berger, M.M. Peña, W.J. Hrushesky, Beta-catenin induces beta-TrCP-mediated PER2 degradation altering circadian clock gene expression in intestinal mucosa of ApcMin/+mice, Journal Biochemistry 145 (3) (2009) 289–297.

[26] J.-E. Kang, M.M. Lim, R.J. Bateman, J.J. Lee, L.P. Smyth, J.R. Cirrito, N. Fujiki, S. Nishino, D.M. Holtzman, Amyloid-beta dynamics are regulated by orexin and the sleep–wake cycle, Science 326 (5955) (2009) 1005–1007.
[27] R.F. Tzang, C.J. Hong, Y.J. Liou, Y.W. Yu, T.J. Chen, S.J. Tsai, Association study of p11 gene with major depressive disorder, suicidal behaviors and treatment response, Neuroscience Letters 447 (1) (2008) 92–95.
[28] L. Ho, H. Fivecoat, J. Wang, G.M. Pasinetti, Alzheimer's disease biomarker discovery in symptomatic and asymptomatic patients: experimental approaches and future clinical applications, Experimental Gerontology 45 (2010) 15–22.
[29] W. Zhu, R. Zhang, C. Hu, H. Umegaki, Effect of the entorhinal cortex on diurnal ACTH and corticosterone release in rats, Neuro Endocrinology Letters 29 (1) (2008) 159–162.
[30] C.-W. Wu, P.-C. Liao, L. Yu, S.-T. Wang, S.-T. Chen, C.-M. Wu, Y.-M. Kuo, Hemoglobin promotes a-beta oligomer formation and localizes in neurons and amyloid deposits, Neurobiology of Disease 17 (3) (2004) 367–377.
[31] R.S. Pandav, V. Chandra, H.H. Dodge, S.T. DeKosky, M. Ganguli, Hemoglobin levels and Alzheimer disease: an epidemiologic study in India, American Journal of Geriatric Psychiatry 12 (5) (2004) 523–526.
[32] C. Hock, K. Villringer, F. Müller-Spahn, M. Hofmann, S. Schuh-Hofer, H. Heekeren, R. Wenzel, U. Dirnagl, A. Villringer, Near infrared spectroscopy in the diagnosis of Alzheimer's disease, Annals of the New York Academy of Sciences 777 (2006) 22–29.
[33] H. Arai, M. Takano, K. Miyakawa, T. Ota, T. Takahashi, H. Asaka, T. Kawaguchi, A quantitative near-infrared spectroscopy study: a decrease in cerebral hemoglobin oxygenation in Alzheimer's disease and mild cognitive impairment, Brain and Cognition 61 (2) (2006) 189–194.
[34] G. Froyen, M. Corbett, J. Vandewalle, I. Jarvela, O. Lawrence, C. Meldrum, M. Bauters, K. Govaerts, L. Vandeleur, H. Van Esch, J. Chelly, D. Sanlaville, H. van Bokhoven, H.H. Ropers, F. Laumonnier, E. Ranieri, C.E. Schwartz, F. Abidi, P.S. Tarpey, P.A. Futreal, A. Whibley, F.L. Raymond, M.R. Stratton, J.P. Fryns, R. Scott, M. Peippo, M. Sipponen, M. Partington, D. Mowat, M. Field, A. Hackett, P. Marynen, G. Turner, J. Gécz, Submicroscopic duplications of the hydroxysteroid dehydrogenase HSD17B10 and the E3 ubiquitin ligase HUWE1 are associated with mental retardation, American Journal of Human Genetics 82 (2) (2008) 432–443.
[35] J.A. Miller, M.C. Oldham, D.H. Geschwind, A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging, The Journal of Neuroscience 28 (6) (2008) 1410–1420.