Identification of differentially co-expressed genes in Alzheimer's disease related data

Irene Teinemaa
Supervisor: Elena Sügis

Abstract. To the date large amount of studies address the basis of brain ageing by studying the pathways involved in this process and by identifying the interactions through which the ageing phenotype that develops in normal and in disease conditions. These studies have depicted association of various genes and proteins with the potential cause of the disease. However, underlying mechanisms of Alzheimer's disease are poorly understood and using only standard methods such as differential expression analysis and various clustering techniques in microarray data studies is not enough. In this work we apply differential co-expression (DC) analysis that is designed to identify changes in the correlation patterns between disease conditions. As a result, differentially co-expressed gene pairs in the Alzheimer's disease specific microarray data are identified. During the analysis of the DC pairs, UCHL1-CDK5 and CDK5-CCNI were identified as the corresponding protein-protein interactions in the BioGrid database. Identified gene pairs have evidence to be connected with the genes that are strongly associated with Alzheimer's disease.

1 Introduction

Ageing is inevitably a complex process that is natural to all human beings. It causes social, psychological and physical changes. Most of the changes affect the quality of life in a negative way. Many studies have been devoted to improving or prolonging the quality of human's life and reducing the dementia in older generation. One of the examples of dementia is Alzheimer's disease, which affects 50% of adults over 85 years old [1].

Alzheimer's disease is a complex multifactorial neurodegenerative disease and the leading cause of dementia among the elderly people. It has been shown in previous studies [2] that both environmental as well as genetic factors influence the risk of AD. Despite the discovery of a few genes influential to AD, the whole process is not yet well understood.

In order to understand a disease, it is essential to study the differences between healthy and affected tissues. One method is differential expression (DE), which enables to identify genes of which mean expression level varies across the two conditions. However, the function and regulatory activities of a gene can be affected without affecting its expression levels [3] and therefore relevant genes might be missed by DE analysis. Another approach is based on the correlation between gene expression patterns to obtain groups of genes with similar expression profiles using clustering methods. The disadvantage of this approach

is that co-expression-based methods assume that the expression patterns of the discovered groups of genes are correlated in all studied conditions [4].

Although these methods provide valuable insights, in real life the diseases are more complex and can not be understood with DE and clustering analysis alone. Another type of technique is differential co-expression (DC) which aims to find pairs of genes that are correlated in one condition and not correlated or negatively correlated in the other.

In this project, differential co-expression analysis is applied to Alzheimer's disease specific data. Experiments are carried out with various set-ups, using the whole data set as well as considering three brain regions separately. Over three thousand differentially co-expressed gene pars were detected. Further analysis showed that DC pairs of UCHL1-CDK5 and CDK5-CCNI were found to have corresponding protein-protein interactions in the BioGrid database. Also, there is evidence that the mentioned gene pairs are connected with genes that are strongly associated with Alzheimers disease.

The next section describes briefly the used method for identifying DC pairs. Section 3 describes the data set used in the analysis. The results of the findings are described in section 4. Section 5 summarizes the findings.

2 Methods

The association between two genes is usually calculated as Pearson correlation. These correlations are calculated between all pairs of genes for each condition. For n genes there are n*(n-1)/2 pairs of genes, therefore the computation of the correlation matrix and methods for the DC analysis are computationally very expensive.

In this project, the DC analysis is performed with an empirical Bayesian approach [5]. After the correlation matrix for each condition is built, a modified EM algorithm is used to choose the best-fitting model. The output of the method is a list of gene pairs with corresponding false discovery rate (FDR) values for estimating their significance.

3 Data

The data used in the analysis is gene expression data set E-GEOD-36980 from ArrayExpress database. The data was normalized and filtered using Factor Analysis for Robust Microarray Summarization (FARMS) [6]. The normalized and filtered data contains 71 samples of 1559 genes. 27 of the samples are from Alzheimer (AD) instances and 44 from non-AD instances. The samples are from 3 different regions of brain: frontal cortex (31 samples), temporal cortex (28) and hippocampus (12).

For further analysis of DC pairs we used publicly available protein-protein interaction network that was obtained from BioGrid database.

4 Results

4.1 Differential co-expression

Differential co-expression analysis was performed in four parts: over all data and over the data divided into three groups according to brain region.

The EBcoexpress [7] package provides three variations of the EM method for co-expression. The first one, zero-step, is fast, but depends highly on initialization of parameters. The one-step variation performs one iteration of EM-calculation and usually its quality is as good as the full EM-calculation, but is faster. The third function performs full EM-calculation over the data. The calculations were performed with first and second method for each of the data groups. For computational feasibility, the full EM was calculated only on the whole data set and Hippocampus, as the fit for the latter with first and second function was much worse than for other groups of data. However, the fit with full EM was not significantly better for Hippocampus subset of data.

A hard threshold of 0.95 was used as the significance threshold. The resulting number of gene pairs is presented in table 1.

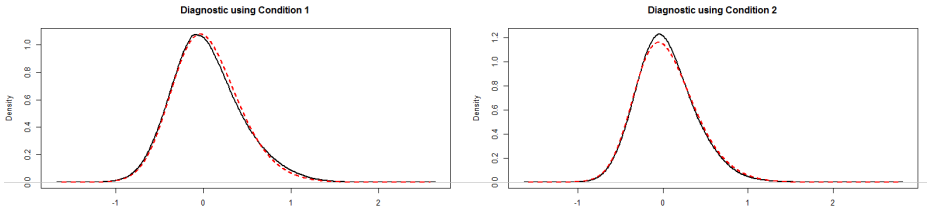
Table 1: Number of found DC pairs

	Zero-step	One-step	Full EM
Whole data set	4510	3352	3172
Frontal cortex	23	21	-
Temporal cortex	4	0	-
Hippocampus	2	0	0

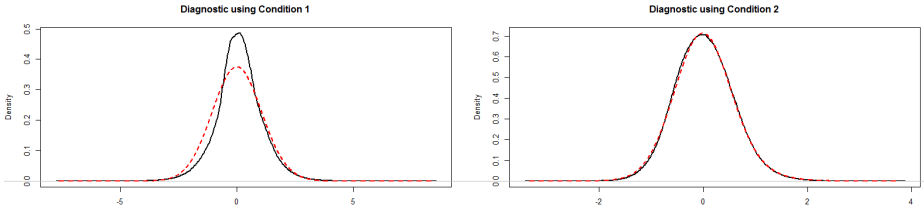
The comparison of the empirical and theoretical distributions of the whole data set with one-step function and Hippocampus data set with full EM function are visualized in figure 1. The chosen model fits the whole data set well, but the model for Hippocampus fits the first condition relatively badly.

4.2 Mapping DC pairs to PPI network

The found DC pairs were mapped to protein-protein interaction network from BioGrid database. The mapping showed that there is almost no overlap between the resulting pairs from any of the groups and the BioGrid network. Among all the results, two DC pairs were present in BioGrid network. These were ENSG00000154277 (UCHL1) - ENSG00000164885 (CDK5) and ENSG00000164885 (CDK5) - ENSG00000118816 (CCNI). The first pair was found as differentially co-expressed by analysis over the whole data set using the one-step function. However, full EM did not mark this pair as DC. Conversely, the second pair was identified as DC by the full EM function, but not the one-step variation. The expression data of the two gene pairs are shown on figure 2.



(a) Whole data set, one-step variation



(b) Hippocampus data set, full EM

Fig. 1: Comparison of the empirical (black) and theoretical (red) distributions. The theoretical form of the distribution is based on the parameters that were chosen by EM. Left figures show the fit across AD instances and right across non-AD instances.

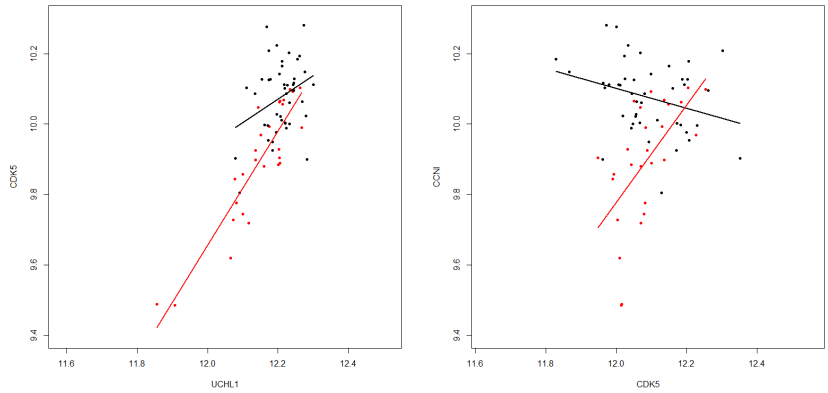
To identify common interacting partners of UCHL1, CDK5 and CCNI, we extracted and merged 1-hop neighborhoods of each gene in the resulting DC network calculated over full data set. In the one-step DC network, UCHL1 has 132 neighbors and CDK5 has 76. The merged neighborhood network contains 162 nodes and 706 edges. In the full EM DC network, CDK5 has 72 neighbors and CCNI 8.

To identify known protein-protein interactions for the given genes, we extracted and merged 1-hop neighborhoods of the corresponding genes from the String database. The resulting network contains well-known Alzheimer's disease related genes MAPT and APP. The merged network of first neighbors of UCHL1, CDK5 and CCNI genes is shown on figure 3. In the BioGrid network, UCHL1 has 74 neighbors, CDK5 has 222 and CCNI 10.

4.3 Analysis of DC network

The DC network for the whole data set using second method has 997 nodes and 3352 edges. Average node degree is 6.7. Table 2 shows 7 largest hubs with the number of nodes and edges in the modules based on that hub. Both UCHL1 and CDK5 are among those hubs.

There are two 5-cliques in the DC network, both of which contain UCHL1 and one of them also contains CDK5. Both cliques also contain genes LRRTM3, GOT1 and MLLT11. Pairwise investigation of expression data of the clique members shows that the pairs are positively correlated in AD cases and not correlated



(a) UCHL1 and CDK5 (b) CCNI and CDK5

Fig. 2: Expression scatterplot of gene pairs UCHL1-CDK5 and CCNI-CDK5. Red indicates AD and black non-AD.

Table 2: Hub-based modules in the DC network. Analysis was performed across the whole data set with two groups Alzheimer vs. Control.

Gene	Nodes	Edges
UCHL1	133	313
ST6GALNAC5	125	287
PLD3	97	109
ATP6V0D1	92	119
AP1S1	80	128
CDK5	77	156
PDE2A	76	77

(or weakly correlated) in non-AD samples. The two cliques are shown on figure 4.

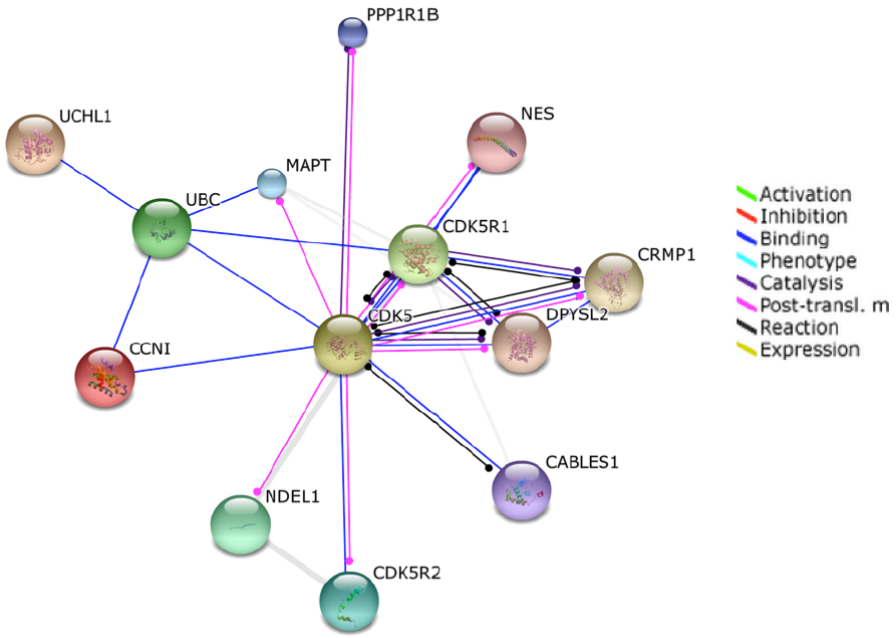
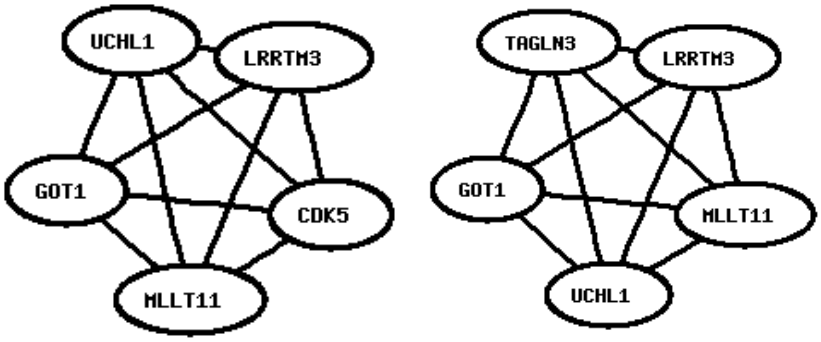


Fig. 3: Merged network of first neighbors of UCHL1, CDK5 and CCNI genes from String database. The resulting network contains well-known Alzheimer's disease related genes MAPT and APP.



(a) (b)

Fig. 4: 5-cliques in the DC network. Analysis was performed across the whole data set with two groups Alzheimer vs. Control.

MCL clustering resulted in 96 clusters. Some statistics of the largest clusters is presented in table 3. UCHL1 belongs to the largest cluster and CDK5 to the second largest cluster. Statistical enrichment analysis of GO anntotations shows that the two largest clusters are involved in the following biological processes: neurological system process, transmission of nerve impulse, synaptic transmission, cell-cell signaling and neurotransmitter release cycle.

Table 3: MCL clusters in the DC network. Analysis was performed across the whole data set with two groups Alzheimer vs. Control.

Cluster	Nodes	Edges	Density
1	118	351	5.1%
2	97	175	3.8%
3	63	16	0.8%
4	56	92	6.0%
5	38	47	6.7%

5 Conclusions

The DC analysis on Alzheimer's data set resulted in identification of two gene pairs UCHL1-CDK5 and CDK5-CCNI that have corresponding protein-protein interactions in the BioGrid database. The 1-hop PPI neighborhood network of the mentioned genes contains genes MAPT and APP which are well-known to be associated with Alzheimer's disease.

In addition, UCHL1, CDK5 and CCNI have a common neighbor Ubiquitin C (UBC). The ubiquitin pathway is involved in the pathogenesis of several diseases including neurodegenerative disorders.

Some further analysis of the genes UCHL1, CDK5, CCNI and UBC is necessary to understand whether and how they are involved in the regulatory mechanism of the Alzheimer's disease.

References

1. Bishop, N.a., Lu, T., Yankner, B.a.: Neural mechanisms of ageing and cognitive decline. Nature **464** (2010) 529–35
2. Kamboh, M.I., Demirci, F.Y., Wang, X., Minster, R.L., Carrasquillo, M.M., Pankratz, V.S., Younkin, S.G., Saykin, a.J., Jun, G., Baldwin, C., Logue, M.W., Buros, J., Farrer, L., Pericak-Vance, M.a., Haines, J.L., Sweet, R.a., Ganguli, M., Feingold, E., Dekosky, S.T., Lopez, O.L., Barmada, M.M.: Genome-wide association study of Alzheimer's disease. Translational psychiatry **2** (2012) e117
3. de la Fuente, A.: From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. Trends in genetics : TIG **26** (2010) 326–33
4. Amar, D., Safer, H., Shamir, R.: Dissection of regulatory networks that are altered in disease via differential co-expression. PLoS computational biology **9** (2013) e1002955

5. Dawson, J.A., Kendziorski, C.: An Empirical Bayesian Approach for Identifying Differential Coexpression in High-Throughput Experiments. Biometrics **68** (2012) 455–465
6. Hochreiter, S., Clevert, D.A., Obermayer, K.: A new summarization method for affymetrix probe level data. Bioinformatics **22** (2006) 943–949
7. Dawson, J.A.: EBcoexpress: EBcoexpress for Differential Co-Expression Analysis. (2012) R package version 1.4.0.