Finding Informative Genes Related to Alzheimer Disease using Supervised Independent Component Analysis

Faezeh Dorri1
faezeh.dorri@gmail.com

Houshang Hassibi1
hassibi@eetd.kntu.ac.ir

Seyed Hassan Paylakhi2
hpaylakhi@gmail.com

Elahe Elahi3
elahe.elahi@gmail.com

1 Biomedical Engineering Department, K. N. Toosi University of Technology, Tehran, Iran
2 Department of Genetics, Faculty of Basic Science, Tarbiat Modares University, Tehran, Iran
3 College of Science, University of Tehran, Tehran, Iran

Keywords: Alzheimer disease, Supervised independent component analysis, Microarray data

1 Introduction

Alzheimer’s Disease (AD) is the most common neurodegenerative disorder among people age 65 and older. Because of increasing average survival age throughout the world, AD is considered an important public health issue and considerable efforts are devoted to AD related research. One of the objectives of current research is identification of genes and molecular pathways involved in the etiology of AD. For this aim, the most powerful approach being used involves analysis of gene expression using microarrays. State of the art microarrays essentially allow assessment of expression of all human genes simultaneously in one experiment. Microarray data is analyzed with various outlooks, and one of these involves application of mathematical tools to best retrieve and reveal information hidden in the vast amount of data produced by the microarray protocols. In the present study, a supervised independent component analysis (SICA) is applied to two publically available gene expression microarray data sets with the objective of identifying the most informative genes with regards to AD. Independent Component Analysis (ICA) is a computational method for decomposing a multi-variant signal into additive components, that assumes the source signals are non-gaussian and mutually independent [1]. In a supervised ICA, information about classes to which samples belong is also provided.

2 Method

AD gene expression data set analyzed in this study was the result of a recent publication [2]. In this study the arrays were used to assess the relative level of expression of more than 20,000 human genes, and both elderly healthy individuals and AD affected individuals were tested. Because of various strategic impediments, the numbers of samples tested in such experiments are usually few. In this case, there were always less than 20 individuals in each class in each of the studies. Various filtering processes were applied, the most important being removal of genes with equal average expression level among patients and controls. Subsequently, a $n \times m$ matrix was created where the individuals and genes were represented, respectively, in rows and columns. A final column was added to the matrix that identified the class to which each sample belonged; the two possible classes were affected or non-affected status with respect to AD. The resulting matrix is $n \times (m + 1)$, also called matrix $X$. An ICA algorithm was applied to decompose this matrix into two matrices $A$ and $S$, where $A$ is $n \times l$
and $S$ is $l \times (m + 1)$. Row of $A$ reveal the weight with which independent components contribute to observed profiles, while rows of $S$ reveal the contribution of each gene and class information to each of the independent components. The independent components can be considered hidden factors. Finally, the last column of matrix $S$ which related to the class information was used to identify the independent components most relevant to AD status. The important independent components were identified based on the following property: $S_c > m + v$ or $S_c < m - v$, where $S_c$ is the $(m + 1)$-th column of $S$, and $m$ and $v$ are, respectively, the mean and variance of the corresponding column. For each independent component, genes were identified whose expression levels are notably different from other genes with respect to that component. A threshold level of difference was set for selection of the genes. The genes identified are considered candidate genes relevant to the AD phenotype. The algorithm was run 100 times in order to obtain robust results. The greater the number of runs in which a particular gene was identified, the greater the confidence that the gene does in fact have a role in AD. A graph was drawn to show the number of runs in which each gene was found to be relevant to one or more independent component. Twenty seven genes that were most often reported were selected to evaluate possible relevance to AD by other biological criteria. Seven biological parameters were used, and biological confirmation of relevance of the gene was considered to exist if the gene qualified at least one of the parameters. Again, the larger the number of biological parameters a gene qualified, the greater the confidence that the gene does in fact have a role in AD.

3 Results

The filtering process applied reduced the number of genes to be analyzed to approximately 5000. The SICA protocol results identified approximately 100 of these genes to be relevant to one or more of the independent components. That is, approximately 100 genes passed the threshold value of difference in expression as compared to expression of other genes for at least one independent component. The graph was sharply asymptotic in this respect. Of the 27 genes with most often reported, 14 showed biological relevance as assessed by qualification for at least one biological parameter. Notably, 1 qualified for four biological parameters and 1 qualified for five parameters. The latter genes were GFAP and CRYAB.

4 Conclusion

We believe the SICA protocol used in this study was fruitful for the case of Alzheimer’s Disease. One of its favorable qualities is that it ultimately identified a fairly small number of genes (100 of the 20,000 originally analyzed for expression level). Although this feature may mean that some relevant genes were not identified, the limited number realistically allows further wet lab study on the role of these genes the results of which are the ultimate hallmark of medical research. Our statistical approach also provides information that can be used to prioritize the genes for further study. The algorithm used needs to be applied to other complex disease for which microarray gene expression data is available in order to assess its value as a general approach for identification of genes relevant to diseases.

References
