Gene Cell Tissue. 2016 July; 3(3):e38415. doi: 10.17795/gct-38415.

Published online 2016 July 12. **Research Article**

Identification of Genes Involved in the Early Stages of Alzheimer Disease Using a Neural Network Algorithm

Marie Barati,[1] and Mansour Ebrahimi[2,*]

[1]University of Applied Science and Technology Centre of Nehbandan, Nehbandan, IR Iran
[2]Department of Biology, School of Basic Sciences, University of Qom, Qom, IR Iran

[*]*Corresponding author*: Mansour Ebrahimi, Department of Biology, School of Basic Sciences, University of Qom, Qom, IR Iran. E-mail: mebrahimi14@gmail.com

Received 2016 April 13; **Revised** 2016 June 20; **Accepted** 2016 June 26.

Abstract

Alzheimer disease is one form of dementia in old age. Alzheimer disease, the incurable disease, which is usually in the seventh decade of human life, shows its symptoms. The disease may be present for years without clinical symptoms. The current study identified the genes with altered expression in patients with Alzheimer disease. The important sequence of each gene in Alzheimer disease was found and introduced as a biomarker of this disease. The present study used microarray libraries related to Alzheimer disease. Finally, the data were weighted using 10 data mining methods, including methods such as support vector machine (SVM), deviation, information gain ratio and the Gini coefficient. Sequences with least two algorithm weights above 0.5 were selected as the most important sequences. Then, a neural network algorithm (neural net, auto multilayer perceptron and perceptron) was run on 11 data bases from the weighted perceptron algorithm, resulting in a careful 97% best performance.

Keywords: Alzheimer Disease, Genes, Data Mining, Neural Network Algorithm

1. Context

In recent years, extensive links between various fields of knowledge are observed, especially in the fields of data mining and different sciences. Data mining is related to different sciences in a wide variety of areas. Fraud detection, pattern recognition, natural language processing, data mining and bioinformatics are among some issues that have a wide link with data mining. In the field of medical sciences, diagnosis, prognostic and treatment response rate of treatment procedures are evident examples of using data mining in medicine. Given the importance of Alzheimer disease and the absence of a definite treatment for this disease, it was decided to use data mining techniques to identify hidden patterns in effective genes involved in the incidence of the disease. In the current study, data mining techniques had an important role in identifying and extracting the repetitive patterns and their importance in the incidence of this disease. The aim of the research sample was to help understand better and prevent the development and spread of diseases such as Alzheimer disease, which has no definitive method of treatment.

1.1. Alzheimer

This disease was described for the first time in 1906 by Alois Alzheimer, a German psychiatrist and neuropathologist, by observing the pathological features of the disease in the brain pathology of a patient diagnosed with verbal and behavioral disorders (1). These disorders are often observed in people over 65 years (2). But the rare early Alzheimer disease may be happen much sooner (3). Although, the progression of the disease is different among people, there are general symptoms for this disease, (4) early stages of Alzheimer are confused with the symptoms of aging and even stress (5). The usual symptom in the early stages is barely remembering of the most recent events, if someone is suspected of Alzheimer disease the person is checked with tests which assess the behavior and abilities of the person and after that with the person is checked with brain scan (6). With the progression of the disease, patient has signs such as confusion, touchiness, attack, mood swings, speech problems and loss of the long term memory. By decreasing patient's tolerance, he avoids family and society (5, 7). Gradually, physical performance is lost and it causes death (8). Since outburst of the disease is different from person to person, predicting the effects is difficult. Before this disease becomes apparent, it progresses for long time and it may even progress for years without any recognition. The average life span after diagnosis is seven years (9). The current treatments decrease the signs of the disease and there is not a treatment to stop the process of the disease (10). Since it does not have a definite cure the patients rely on the others and the disease has effects on the structure of the society (11). Alzheimer in the devel-

oped countries is one of the most expensive diseases of the society (12, 13).

1.2. Review of the Literature

The conducted researches on the data mining and Alzheimer with the classification method started in 1990 and still persist. The current study refers to the conducted researches such as the classification based on the neural network of perceptual, natural condition, insane, Alzheimer and vascular dementia from single photon emission by tomography data of computed image of the brain conducted in 1995. The current study aimed to classify and identify the three groups of patients with Alzheimer disease, patients with mental decline and the healthy elderly people. The obtained results showed that the artificial neural network method had effects on the diagnosis of brain images and other areas (14).

2. Evidence Acquisition

2.1. Methods

First, experiments relevant to the subject matter conducted by the microarray method were collected to build the corresponding database, since the databases for data mining models should be prepared based on the microarray method. Then, the appropriate microarray libraries to the subject under study were searched. For this purpose the national certification board for Alzheimer care (NCBAC) website was visited. Alzheimer disease was the subject of the study.

Among the libraries found in the search, five libraries had the desired conditions. Each of them is described in Table 1 respectively.

To normalize the data, it was necessary to transfer them into Expression Console software. The reason for normalization is that normally these data contain some noise, the data should be normalized to remove the noise resulted from the light intensity during the test. Remote method invocation (RMI) algorithm was used for normalization. RMI algorithm (robust multi-array average: RMA) is a powerful linear model in probe levels to minimize the effect of affinity difference between specific probes. This approach increases the sensitivity to small changes in test and control samples. RMI is a multi-chip and parallel approach; therefore, all intended arrays for comparison should be included together in the summarization stage (last stage). Then all of the five test data were normalized and prepared to compare sick and healthy samples. In the next stage, the test was designed. In this stage, Bayes T. was conducted on each sample vs control sample and the algorithms can be categorized, compared and performed according to the method proposed in the paper. Bayes test is conducted in a way that a comparison should be made between two groups. Here, two groups of sick and healthy were compared to identify the gene changes after being diagnosed with Alzheimer disease. Then, the information related to the annotation of each gene taken from the valid data bases was collected. The information was prepared to transfer into the RapidMiner software by obtaining the sequence of each gene. In this step, the 10-feature selection (weighting) algorithm was applied on the data, the implementation of these algorithms was done in the RapidMiner software and as a result, the feature selection method was briefly represented in each; for more information refer to the following networks (15).

2.2. Neural Network

An artificial neural network (ANN), usually called neural network (NN), is a mathematical or computational model inspired by the structure and functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation (the central connectionist principle is that mental phenomena can be described by interconnected networks of simple and often uniform units). In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are usually used to model complex relationships between inputs and outputs or to find patterns in data (16).

2.3. Perceptron

The perceptron is a type of artificial neural network invented in 1957 by Frank Rosenblatt. It can be seen as the simplest kind of feed-forward neural network: a linear classifier. Besides all biological analogies, the single layer perceptron is simply a linear classifier which is efficiently trained by a simple update rule: for all wrongly classified data points, the weight vector is either increased or decreased by the corresponding example values. The coming paragraphs explain the basic ideas about neural networks and feed-forward neural networks (16).

2.4. MLp

A multilayer perceptron (MLp) is a feed-forward artificial neural network model that maps sets of input data onto a set of appropriate output. An MLp consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLp utilizes back propagation

Table 1. Downloaded Micro-Array Libraries

The Sampled Region	Sample Number		Study Item	Year	Series
	Sick	Control			
Hippocampal genes	22	9	Alive human	2004	GSE1297
Entorhinal cortex	10	10	Alive human	2006	GSE4757
Hippocampal genes	22	88	Alive human	2011	GSE28146
Hippocampal genes	8	8	Alive mice	2012	GSE32536
Temporal cortex - Frontal cortex - Hippocampal	32	47	Alive human	2013	GSE6980

for training the network. This class of networks consists of multiple layers of computational units, usually interconnected in a feed-forward way. In many applications the units of these networks apply a sigmoid function as an activation function (16).

2.5. Neural Net

A feed-forward neural network is an artificial neural network where connections between the units do not form a directed cycle. In this network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes (if any) to the output nodes. There are no cycles or loops in the network. To use a more sophisticated net, the neural net should be used (16).

3. Results

By applying the feature selection algorithms on the data, more important sequences were identified. By applying the 10-feature selection algorithms on 1092 sequences, 23 sequences which had more than 5/0 weight in two algorithms were identified to be more important than other sequences. The results of each weighting algorithm or feature selection were as follows:

After performing 10-feature selection algorithms, the number of sequences with 5.0 more weight in at least two algorithms was identified as the most important sequences. Among these sequences, TGCCCC, AGCCTG, AATTG, GAATAT and AAATTG sequences with eight times repetition were identified to be the most important sequences.

3.1. Output of Neural Net Algorithms

Table 2 reviewed performance of the neural net algorithm; the best performance related to perceptron algorithm on uncertainty database; the worst performance related to the neural net algorithm with 0.7 value on deviation database; the performance of auto MLP algorithm on all of the databases were identical and good.

The output of the neural network algorithms in Table 3 reviewed the performance of the algorithms of the neural networks. Perceptron algorithm with 0.97 had the best performance. Neural net algorithm with 0.74 among three algorithms had the lowest performance. On average, auto MLp algorithm had the best performance and neural net had the lowest performance.

4. Discussion

Alzheimer disease is the most important degenerative brain disease which can be equally observed in both genders. It appears both hereditary and accidentally. The accidental kind is most common and the most common cause for it is the effect of unknown environmental and metabolic factors (17). The increasing rates of Alzheimer incidence is warning, which is becoming a social concern of many countries [6, 8]. In America, this disease is considered as one of the ten causes of mortality. In a study recently conducted in Hopkins university, it was estimated that in 2050 from each 85 individuals from the earth population, one will be diagnosed with Alzheimer disease (18). Micro-array technology is a very powerful method providing the opportunity to study the expression of thousands of genes simultaneously and identify thousands of protein interactions (19). This technology has two major subcategories: DNA possibility and protein. Using DNA microarray provides the micro-array to study the expression of thousands of genes simultaneously. The aims of such gene analyses are as follows:

First: How to explain the effect of any single gene on other genes; second: How the gene is expressed in healthy and diseased cells. For example, different types of cancer have similar morphologic symptoms, therefore, direct diagnosis methods can be provided by using gene expression data.

On the other hand, protein array is a kind of measuring method to help medical experts with measuring and existence of proteins in the biological samples including

Table 2. Output of Neural Net Algorithms

Data Base	Auto MLp	Neural Net	Perceptron	
Uncertainty	0.9	0.9	1.0	0.9
SVM	0.9	0.9	0.9	0.9
Rule	0.9	0.9	0.9	0.9
Relief	0.9	0.9	0.9	0.9
PCA	0.9	0.9	0.8	0.8
Info gain ratio	0.9	0.9	0.8	0.8
Info gain	0.9	0.9	0.9	0.9
Gini index	0.9	0.9	0.9	0.9
Deviation	0.9	0.7	0.8	0.8

Table 3. Output of Neural Net Algorithms

Algorithms	Average	Best Performance	Lowest Performance
Auto MLp*	0.89	0.94	0.86
Perceptron	0.88	0.97	0.8
Neural net	0.86	0.89	0.74

blood (19). Analyzing gene expression and its changes are affected by factors such as: treatment, pathogens and cell damage (20). Since 1998, researches are conducted by micro-arrays on Alzheimer disease and as an example, a study entitled "gene expression profiles of multiple genes in single neurons of Alzheimer disease (9)", is the genotype of a cell's genetic makeup, an organism or an individual with an indication to a certain preferred features. In this research Northern blot, dot-blot and transcriptase along with analyzing reverse polymerase chain reaction showed altered levels of expression in several messages in the brain of individuals diagnosed with Alzheimer disease. Since all the cells are not affected by this disease; the method can clearly study the changes regarding the disease in the individual's cells. The current study aimed to identify the biomarker, first, using the pre-prepared samples taken from different areas of the brain of those affected by Alzheimer disease, using the RMA algorithm, the samples were normalized and then by designing tests to compare healthy samples with control samples, genes were identified with change in their expression. These tests were designed with Flex Array software and Bayes test, a powerful algorithm to compare the two groups, was used to identify the genes with change in their expression. Due to the large number of obtained results, and to increase the accuracy of the test, results were limited to values with change in expression more than 5.1 and less than 5.1. Identifying these biomarkers is important in the way that they accelerate disease diagnosis process and in many cases inhibit disease incidence in the individual and its progress with preventive actions.

The important effect of amyloid beta precursor protein, PAZ, kinase and microglobulin genes in Alzheimer disease was validated in the previous researches too. In a research conducted in 2001 in Saint Mateu, California, on Alzheimer disease, 31 genes had an over expression and 87 genes had a down expression which was completely consistent with the current study samples and some of the unknown genes were also added to the list of genes that previously had a role in the incidence of Alzheimer disease (21).

In this research, some Xist genes were observed between the genes with over expression or down expression, which did not observe in the results of the previous studies and it showed the effect of these genes on the incidence of this disease. According to the differences in the samples and the disease intensity between the used samples in the current study and other studies, differences in the intensity of expression are shown. In a research conducted in 2001, researchers identified Alzheimer biomarkers in a few mice (22), and again it had common results with the current study to an extent. However, it was necessary to find more accurate and more reliable results to conduct the current study on human samples and with sampling of different brain segments. In the current study, as already expressed, samples included humans and mice with different disease intensity and male and female gen-

der and also different regions in the brain including, temporal and frontal and hippocampal brain cortexes which presented comprehensive results which no study had conducted to this extent. Early studies similar to the current study worked on a certain cortex of the brain or a certain sample, such as dead or alive human and/or mouse. That is the reason the current study results were more comprehensive than those of other studies, and genes besides the previously identified samples were added to the list of biomarkers. In 2004, a study entitled: "micro-array analysis in Alzheimer disease and normal aging" identified genes with expression change by sampling brain cortex. The results of this experiment was consistent with the obtained results of the current study; for example, beta, actin, L21 ribosomal protein, eukaryotic translation initiation factor 5A, neuronal thread protein genes in the current study were introduced with the highest (21), but not the low expression; therefore, only slight differences existed in gene expression level which was normal according to the differences in the samples. The current study examined five libraries regarding the Alzheimer disease; in each of these tests, the specific regions of the brain were tested. This shows that extensive results were obtained. After conducting the experiment on each of the five libraries and comparing normal and defective genes in different cortical layers of the brain, between samples with different disease intensity, it was observed that a few genes in these tests were common. After removing all the iterations and after obtaining all the test results from the library, it was found that genes with over expression had the most changes, as shown in the tables below. After integrating all the test results and removing the repeated results, five genes were identified with over expression, shown in the Table 4. In the table, the over expression of Xist gene at X-inactive specific transcript can be observed with 8.68652 change in value.

Table 4. List of Genes With the Highest Over Expression

Probe Set ID	Gene Name	Fold Change
235446-at	XIST X-inactive specific transcript	8.68652
221728-x-at	XIST X-inactive specific transcript	5.361924
224588-at	XIST X-inactive specific transcript	4.702068
243712-at	XIST X-inactive specific transcript	4.330641
33323-r-at	SFN	3.823175
214218-s-at	XIST X-inactive specific transcript	3.73422
1565483-at	EGFR	3.626728
224589-at	XIST X-inactive specific transcript	3.181633

This method was followed to find the highest levels of decrease in expression. They were identified after the integration of the results of all tests and removing duplicate results of five genes with the highest decrease in expression, shown in Table 5. In the table it can be observed that the highest decrease in expression belonged to 2A5COL gene with low rates of 4.561804 negative changes.

Table 5. Genes with the Lowest Expression

Probe Set ID	Gene Name	Entrez Gene	Fold Change
221730-at	COL5A2	1290	-4.561804
221805-at	NEFL	4747	-4.432566
221729-at	COL5A2	1290	-4.353842
203798-s-at	VSNL1	7447	-4.080669
213436-at	CNR1	2159	-3.972829

Acknowledgments

This paper presented a final report from a student master's degree thesis in information technology with the tracking code of 2172733. The project was approved by the faculty of engineering at the Qom University, Qom, Iran.

Footnotes

Authors' Contribution: All authors had equal contribution in design, work, statistical analysis and writing of the manuscript.

Conflict of Interest: The authors declared no conflict of interest.

References

1. Berchtold NC, Cotman CW. Evolution in the conceptualization of dementia and Alzheimer's disease: Greco-Roman period to the 1960s. *Neurobiol Aging.* 1998;**19**(3):173–89. [PubMed: 9661992].
2. Brookmeyer R, Gray S, Kawas C. Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. *Am J Public Health.* 1998;**88**(9):1337–42. [PubMed: 9736873].
3. Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM. Forecasting the global burden of Alzheimer's disease. *Alzheimers Dement.* 2007;**3**(3):186–91. doi: 10.1016/j.jalz.2007.04.381. [PubMed: 19595937].
4. Perl DP. Neuropathology of Alzheimer's disease. *Mt Sinai J Med.* 2010;**77**(1):32–42. doi: 10.1002/msj.20157. [PubMed: 20101720].
5. Waldemar G, Dubois B, Emre M, Georges J, McKeith IG, Rossor M, et al. Recommendations for the diagnosis and management of Alzheimer's disease and other disorders associated with dementia: EFNS guideline. *Eur J Neurol.* 2007;**14**(1):1–26. doi: 10.1111/j.1468-1331.2006.01605.x. [PubMed: 17222085].
6. Alzheimer's A. 2011 Alzheimer's disease facts and figures. *Alzheimers Dement.* 2011;**7**(2):208–44. doi: 10.1016/j.jalz.2011.02.004. [PubMed: 21414557].

7. Tabert MH, Liu X, Doty RL, Serby M, Zamora D, Pelton GH, et al. A 10-item smell identification scale related to risk for Alzheimer's disease. *Ann Neurol.* 2005;**58**(1):155–60. doi: 10.1002/ana.20533. [PubMed: 15984022].
8. Holtzman DM, Herz J, Bu G. Apolipoprotein E and apolipoprotein E receptors: normal biology and roles in Alzheimer disease. *Cold Spring Harb Perspect Med.* 2012;**2**(3):006312. doi: 10.1101/cshperspect.a006312. [PubMed: 22393530].
9. Molsa PK, Marttila RJ, Rinne UK. Survival and cause of death in Alzheimer's disease and multi-infarct dementia. *Acta Neurol Scand.* 1986;**74**(2):103–7. [PubMed: 3776457].
10. Lin CH, Chen PK, Chang YC, Chuo LJ, Chen YS, Tsai GE, et al. Benzoate, a D-amino acid oxidase inhibitor, for the treatment of early-phase Alzheimer disease: a randomized, double-blind, placebo-controlled trial. *Biol Psychiatry.* 2014;**75**(9):678–85. doi: 10.1016/j.biopsych.2013.08.010. [PubMed: 24074637].
11. Ory MG, Hoffman R, Yee JL, Tennstedt S, Schulz R. Prevalence and impact of caregiving: a detailed comparison between dementia and nondementia caregivers. *Gerontologist.* 1999;**39**(2):177–85. [PubMed: 10224714].
12. Meek PD, McKeithan K, Schumock GT. Economic considerations in Alzheimer's disease. *Pharmacotherapy.* 1998;**18**(2 Pt 2):68–73. [PubMed: 9543467] discussion 79-82.
13. Zhu CW, Sano M. Economic considerations in the management of Alzheimer's disease. *Clin Interv Aging.* 2006;**1**(2):143–54. [PubMed: 18044111].
14. deFigueiredo RJ, Shankle WR, Maccato A, Dick MB, Mundkur P, Mena I, et al. Neural-network-based classification of cognitively normal, demented, Alzheimer disease and vascular dementia from single photon emission with computed tomography image data from brain. *Proc Natl Acad Sci U S A.* 1995;**92**(12):5530–4. [PubMed: 7777543].
15. Ebrahimie E, Ebrahimi M, Sarvestani NR, Ebrahimi M. Protein attributes contribute to halo-stability, bioinformatics approach. *Saline Systems.* 2011;**7**(1):1. doi: 10.1186/1746-1448-7-1. [PubMed: 21592393].
16. RapidMiner Technical Support Available from: http://docs.rapidminer.com/.
17. Hardman JGLL, Gilman AG. Goodman & Gilman's The Pharmacological Basis of Therapeutics. 10 ed. New York: McGraw-Hill; 2001. p. 2045.
18. Chilukoti N, Early K, Sandhu S, Riley-Doucet C, Debnath D, editors. Assistive technology for promoting physical and mental exercise to delay progression of cognitive degeneration in patients with dementia. 2007 ieee biomedical circuits and systems conference. 2007; IEEE; pp. 235–8.
19. Adomas A, Heller G, Olson A, Osborne J, Karlsson M, Nahalkova J, et al. Comparative analysis of transcript abundance in Pinus sylvestris after challenge with a saprotrophic, pathogenic or mutualistic fungus. *Tree Physiol.* 2008;**28**(6):885–97. [PubMed: 18381269].
20. Loring JF, Wen X, Lee JM, Seilhamer J, Somogyi R. A gene expression profile of Alzheimer's disease. *DNA Cell Biol.* 2001;**20**(11):683–95. doi: 10.1089/10445490152717541. [PubMed: 11788046].
21. Arisi I, D'Onofrio M, Brandi R, Felsani A, Capsoni S, Drovandi G, et al. Gene expression biomarkers in the brain of a mouse model for Alzheimer's disease: mining of microarray data by logic classification and feature selection. *J Alzheimers Dis.* 2011;**24**(4):721–38. doi: 10.3233/JAD-2011-101881. [PubMed: 21321390].
22. Ricciarelli R, d'Abramo C, Massone S, Marinari U, Pronzato M, Tabaton M. Microarray analysis in Alzheimer's disease and normal aging. *IUBMB Life.* 2004;**56**(6):349–54. doi: 10.1080/15216540412331286002. [PubMed: 15370883].