Function and Pathway Analysis of Differentially Expressed Genes in Alzheimer’s Disease Dataset Using Linear Regression Model

Dr. R. Porkodi, R. Savitha
Department of Computer Science, Bharathiar University, Coimbatore, Tamil Nadu, India

Abstract
Alzheimer’s disease is an irreversible, progressive brain disorder that slowly destroys memory and thinking skills, and eventually the ability to carry out the simplest tasks. The aim of this study is to screen the potential pathways changed in Alzheimer’s disease and elucidate the mechanism of it. Published microarray data of GSE1297 series is downloaded from Gene Expression Omnibus (GEO). Significance analysis of microarray is performed using software R, and differentially expressed genes (DEGs) are harvested. The functions and pathways of DEGs are mapped in Gene Ontology and KEGG pathway database respectively. A total of 2273 genes are filtered as DEGs between normal and Alzheimer’s disease cells. This research work is to depict the molecular level changes of Alzheimer disease through microarray analysis of differentially expressed genes (DEGs), and the function and pathway enrichment of differentially expressed genes (DEGs).

Keywords
Gene Enrichment Analysis, Pathway Analysis, Differentially Expressed Genes

I. Introduction
Data mining or knowledge discovery is the process of discovering meaningful correlation patterns through large amount of data stored in repositories using pattern recognition techniques as well as statistical and mathematical techniques. Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data through which useful information and knowledge are extracted. Regression is a data mining technique used to fit an equation to a dataset. The simplest form of regression, linear regression, uses the formula of a straight line \( y = mx + b \) and determines the appropriate values for \( m \) and \( b \) to predict the value of \( y \) based upon a given value of \( x \). Advanced techniques, such as multiple regression, allow the use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation. Regression is a data mining function that predicts a number. Age, weight, distance, temperature, income, or sales could all be predicted using regression techniques.

Bioinformatics is the part of Biology which studies the properties of DNA, RNA and proteins. Bioinformatics is the science of managing, mining, and interpreting information from biological sequences and structures. Advances such as genome-sequencing initiatives, microarrays, proteomics, and functional and structural genomics have improved human knowledge. Bioinformatics can be defined as the application of computer technology to the management of biological information. Bioinformatics is the science of storing, extracting, analyzing, organizing, interpreting and utilizing information from biological molecules and sequences. Statistics is a branch of mathematics that targets on the collection, organization and interpretation of numerical data. Many researches in computational biology and bioinformatics heavily rely on the application of probabilistic models and statistical models.

II. Literature Review
Jui-Hung Hung, et. al., [4] states a central goal of biology is understanding and describing the molecular basis of plasticity: the sets of genes that are combinatorially selected by exogenous and endogenous environmental changes, and the relations among the genes. The most viable current approach to this problem consists of determining whether sets of genes are connected by some common theme, e.g., genes from the same pathway are overrepresented among those whose differential expression in response to a perturbation is most pronounced. There are many approaches to this problem, and the results they produce show a fair amount of dispersion, but they all fall within a common framework consisting of a few basic components. They critically review these components, suggest best practices for carrying out each step, and propose a voting method for meeting the challenge of assessing different methods on a large number of experimental data sets in the absence of a gold standard.

Clemens Wrzodek et. al., [7] states that the KEGG pathway database provides a widely used service for metabolic and nonmetabolic pathways. It contains manually drawn pathway maps with information about the genes, reactions and relations contained therein. To store these pathways, KEGG uses KGML, a proprietary XML-format. Parsers and translators are needed to process the pathway maps for usage in other applications and algorithms.

Christina Backes, et. al., [5] present a comprehensive and efficient gene set analysis tool, called ‘GeneTrail’ that offers a rich functionality and is easy to use. Their web-based application facilitates the statistical evaluation of high-throughput genomic or proteomic data sets with respect to enrichment of functional categories. GeneTrail covers a wide variety of biological categories and pathways, among others KEGG, TRANSPATH, TRANSFAC, and GO.

L. Krishnamurthy, et. al., [3] say during the next phase of the Human Genome Project, research will focus on functional studies of attributing functions to genes, their regulatory elements, and other DNA sequences. To facilitate the use of genomic information in such studies, a new modeling perspective is needed to examine and study genome sequences in the context of many kinds of biological information. Pathways are the logical format for modeling and presenting such information in a manner that is familiar to biological researchers. In this paper, they present an integrated system, called Pathways Database System, with a set of software tools for modeling, storing, analyzing, visualizing, and querying biological pathways data at different levels of genetic, molecular, biochemical and organismal detail.

Likun Wang, et. al., [6] say in some application, researchers may have several replicates sequenced under each condition. Current observations suggest that typically RNA-seq experiments have low technical background noise (which could be checked using DEGsseq) and the Poisson model fits data well. In such cases, users could directly pool the technical replicates together to get higher sequencing depth and detect subtle gene expression changes.
Otherwise, the methods that estimate the noise by comparing the replicates are recommended. Chen Xie, et al., [8] state high-throughput experimental technologies often identify dozens to hundreds of genes related to, or changed in, a biological or pathological process. From these genes one wants to identify biological pathways that may be involved and diseases that may be implicated. Here, they report a web server, KOBAS 2.0, which annotates an input set of genes with putative pathways and disease relationships based on mapping to genes with known annotations. It allows for both ID mapping and cross species sequence similarity mapping. It then performs statistical tests to identify statistically significantly enriched pathways and diseases. KOBAS2.0 incorporates knowledge across 1327 species from 5 pathway databases (KEGG PATHWAY, PID, Biocyc, Reactome and Panther) and 5 human disease databases (OMIM, KEGG DISEASE, FunDO, GAD and NHGRI GWAS Catalog). Jitao David, et. al., [9] demonstrate the capabilities of the KEGGgraph package, an interface between KEGG pathways and graph model in R as well as a collection of tools for these graphs. Superior to preceding approaches, KEGGgraph maintains the pathway topology and allows further analysis or dissection of pathway graphs. It parses the regularly updated KGML (KEGG XML) files into graph models maintaining all essential pathway attributes.

Seon-Young Kim, et. al., [10] say gene set enrichment analysis (GSEA) is a microarray data analysis method that uses predefined gene sets and ranks of genes to identify significant biological changes in microarray data sets. GSEA is especially useful when gene expression changes in a given microarray data set is minimal or moderate. They have developed a modified gene set enrichment analysis method based on a parametric statistical analysis model. Compared with GSEA, the parametric analysis of gene set enrichment (PAGE) detected a larger number of significantly altered gene sets and their p-values were lower than the corresponding p-values calculated by GSEA. Because PAGE uses normal distribution for statistical inference, it requires less computation than GSEA, which needs repeated computation of the permuted data set. PAGE was able to detect significantly changed gene sets from microarray data irrespective of different Affymetrix probe level analysis methods or different microarray platforms.

Liqiang Qian, et. al., [1] say squamous lung cancer (SQLC) is a common type of lung cancer, but its oncogenesis mechanism is not so clear. The aim of their study was to screen the potential pathways changed in SQLC and elucidate the mechanism of it. Published microarray data of GSE3268 series was downloaded from Gene Expression Omnibus (GEO). Significance analysis of microarrays was performed using software R, and differentially expressed genes (DEGs) were harvested. The functions and pathways of DEGs were mapped in Gene Ontology and KEGG pathway database, respectively. A total of 2961 genes were filtered as DEGs between normal and SQLC cells.

Marit Ackermann, et. al., [2] say analysis of microarray and other high-throughput data on the basis of gene sets, rather than individual genes, is becoming more important in genomic studies. Correspondingly, a large number of statistical approaches for detecting gene set enrichment have been proposed, but both the interrelations and the relative performance of the various methods are still very much unclear. They conduct an extensive survey of statistical approaches for gene set analysis and identify a common modular structure underlying most published methods. Based on this finding, they propose a general framework for detecting gene set enrichment.

### III. Objective and Methodology

The objective of the thesis is to perform gene enrichment analysis of differentially expressed genes between normal and Alzheimer’s disease cells, to identify differentially expressed genes between normal and diseased cells in Alzheimer’s disease by significance analysis of microarray using R software, to identify functions of differentially expressed genes by mapping it to Gene Ontology, identify pathways of differentially expressed genes by mapping it to KEGG pathway database respectively.

![Gene Enrichment Analysis of the Differentially Expressed Genes](image)

**Fig. 1: Gene Enrichment Analysis of the Differentially Expressed Genes**

From GEO dataset browser, a dataset of Alzheimer’s disease at various stages of severity is considered for performing gene enrichment analysis. Published microarray data of GSE1297 series is downloaded from Gene Expression Omnibus (GEO) for enrichment analysis. There are 31 samples in this dataset with four stages of severity, which consists of 8 samples at moderate stage, 7 at severe, 9 at control and 7 at incipient stages respectively. The dataset ID is GDS810. Platform is GPL96 [HG-U133A] Affymetrix Human Genome U133A Array [21].

#### A. Phase I- Pre-processing

The DEG extraction operates on Series Matrix files which contain data extracted directly from the dataset. The data has to be preprocessed to perform statistical analysis. Normalization is required in order to remove systematic variation arising from reasons other than biological differences between RNA samples. To perform normalization, soft file is read using GEOquery in R. They are usually a .text file. The GDS file is loaded and turned into an expression set object. The GEO database accepts a variety of data value types, including logged and unlogged data. R expects data values to be in log space to perform statistical analysis. To address this, a log2 transformation on values is performed. This applies the transformation on all expression values to make variation similar across orders of magnitude. By log-transforming it, we could bring it closer to normal distribution. Many statistical tests require normally-distributed data.

#### B. Phase II- Identifying Differentially Expressed Genes

T-test is used to identify differentially expressed genes between the four stages of Alzheimer’s disease condition moderate, control, severe and incipient with B-H test corrections. Then, the significance of DEGs is tested. They are corrected with Benjamini and Hochberg (B-H) test. Genes with FDR (false discovery rate) <0.05 are selected as DEGs [13]. The topTable() function in R is used to sort the differentially expressed genes in this sample.
set for all of the comparison groups. Finally, the differentially expressed genes (DEGs) with cutoff criterion FDR < 0.5 and p-value less than 0.05 were selected. At a FDR value of 0.05, a total of 2273 probes are identified to be differentially expressed in Alzheimer’s disease sample with 22283 probe IDs when compared with normal sample.

C. Function Annotation
To determine the function of DEGs, the DEGs were mapped to the GO database [20] by GOEAST tool. The biological process, cellular locations and molecular functions that are particularly over- or under-represented in DEGs are extracted and can be visualized through GOEAST [11]. This is generally referred to as GO analysis. The biological process of DEGs, such as cell division, cell cycle, signal transduction, and the metabolism of sugar, protein and fat can be viewed separately. The molecular function of DEGs, for instance, binding of proteins, oxygen transmission and NAD plus metabolism on the respiratory chain, etc., are extracted and visualized separately. Cellular components in which the DEGs are most located, such as microtubule, centromere, chromosome, cell-cell junction, and mitochondrion can be identified separately.

D. Pathway Annotation
Gene set enrichment analysis of the differentially expressed genes is conducted. About 7 pathways are identified as differently enriched pathways. They are Alzheimer’s disease pathway, Pathways in cancer pathway, Huntington’s disease pathway, Calcium signaling pathway, Apoptosis pathway, Parkinson’s disease pathway, and ECM-receptor interaction pathway. These pathways are identified as significant pathways of the harvested differentially expressed genes. Pathways of DEGs are identified by mapping the DEGs to the KEGG database. KEGG pathway enrichment analysis is done to identify the significant biological pathways related with the DEGs.

IV. Results and Discussion
A. Differentially Expressed Genes (DEGs)
Moderated t-statistics, moderated F-statistic and log-odds of differential expression are computed by empirical Bayes moderation of the standard errors towards a common value. T-test is used to identify differentially expressed genes between the four stages of Alzheimer’s disease condition, which are moderate, control, severe and incipient with B-H test corrections [18]. Significance of DEGs is tested and corrected with Benjamini and Hochberg (B-H) test. Genes with FDR (false discovery rate) < 0.05 and p-value less than 0.05 are selected as differentially expressed genes (DEGs) [16]. At a FDR value of 0.05, a total of 2273 probes are identified to be differentially expressed in Alzheimer’s disease sample with 22283 probe IDs when compared with normal sample.

B. Function Annotation
In the top 5 differentially expressed genes, i.e., DRD4, SPATA31C2, OSBP1L10, LHCGR, SEPT6, the significantly enriched biological processes are negative regulation of biological process, regulation of macromolecule metabolic process, negative regulation of biosynthetic process, etc., which are shown in dark colour. Comparatively, less significantly enriched biological processes are negative regulation of cellular catabolic process, regulation of organ growth, regulation of DNA replication, regulation of synaptic plasticity, negative regulation of fibrinolysis, etc., which is shown in lighter colour in Fig. 2. The processes which are not significantly enriched are regulation of catabolic process, regulation of cytokine biosynthetic process, regulation of proteolysis, etc., which are colourless in the graph. Similarly, significantly enriched cellular components and molecular functions of the top 5 differentially expressed genes are determined.
A. Pathway Annotation

KEGG pathway enrichment analysis is used to identify the significant biological pathways related with the DEGs. A total of 7 pathways are identified [12]. The enriched pathways are listed in Fig. 4.

<table>
<thead>
<tr>
<th>Category</th>
<th>Term</th>
<th>P-Value</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEGG_PATHWAY</td>
<td>Alzheimer's disease</td>
<td>0.00017</td>
<td>0.03</td>
</tr>
<tr>
<td>KEGG_PATHWAY</td>
<td>Pathways in cancer</td>
<td>0.00057</td>
<td>0.051</td>
</tr>
<tr>
<td>KEGG_PATHWAY</td>
<td>Huntington's disease</td>
<td>0.0007</td>
<td>0.042</td>
</tr>
<tr>
<td>KEGG_PATHWAY</td>
<td>Calcium signaling pathway</td>
<td>0.0017</td>
<td>0.074</td>
</tr>
<tr>
<td>KEGG_PATHWAY</td>
<td>Apoptosis</td>
<td>0.002</td>
<td>0.072</td>
</tr>
<tr>
<td>KEGG_PATHWAY</td>
<td>Parkinson's disease</td>
<td>0.0027</td>
<td>0.08</td>
</tr>
<tr>
<td>KEGG_PATHWAY</td>
<td>ECM-receptor interaction</td>
<td>0.003</td>
<td>0.075</td>
</tr>
</tbody>
</table>

Fig. 4: The Biological Pathways of Alzheimer’s Disease Cells Enriched in KEGG

The enriched pathway maps of the differentially expressed genes (DEGs) are generated by mapping elementary datasets (genes, proteins, small molecules, etc.) to KEGG pathway maps. All metabolic pathways are downloaded from open KEGG pathway database and analysis of DEGs is done. This is done based on hypergeometric distribution, FDR < 0.05 is chosen as cut-off criterion. In the KEGG graph, the stars represent the DEGs changed in this pathway [19]. The pathway enrichment analysis map or KEGG graph is presented for the most significant pathway, i.e., Alzheimer’s disease pathway in fig. 4.

The top enriched pathway of the differentially expressed genes is the Alzheimer’s disease pathway. In this pathway, the genes CASP3, NDUFA5, NDUVF2, UQCR10, NDUFB5, CALM3, CYCS, NDUFA4, COX4I1, COX6C, NDUFA13, APP, CDK5, ATP2A2, COX6A1, NDUFB8, APOE, COX7B, NDUFA10, PPP3CA, ADAM17, GAPDH, COX7C, UQCRQ, CALM1, LR1, UQCR1, PLCB1, SNCA, SDHB, ATP5C1, ITPR1, COX7A2L, PPP3R1, NDUFS3, PPP3CC, UQCRH, MAPK1, NDUFB7, NDUFA8, ATP5O, and COX6B1 are found to be significantly enriched.

References

[1] Liqiang Qian, Qingquan Luo, Xiaojing Zhao, Jia Huang, “Pathways Enrichment Analysis for Differentially Expressed Genes in Squamous Lung Cancer,” Shanghai JiaoTong University, Published online: 10 October 2013.


[21] Dr. R. Porkodi received the Bachelors Degree (B.Sc.) in Computer Science, Masters Degree (MCA) in Computer Applications in 1995, 1998 respectively from Madurai Kamarajar University, India. She has qualified UGC-NET for Lectureship in the year 2000. She has received Ph.D. in Computer Science in the year 2013 from Bharathiar University, Coimbatore. She has more than 15 years of teaching experience.

Her research interest includes Text Mining, Information Retrieval and Bioinformatics and she is currently working in the Department of Computer Science and Engineering, Bharathiar University, India.

Ms. R. Savitha received the Bachelors Degree (B.Sc.) in Computer Technology in 2003, Masters Degree (MCA) in Computer Applications in 2014 from Bharathiar University, Coimbatore. She is currently pursuing her M.Phil. Research at Bharathiar University, Coimbatore in the area of Microarray Analysis. Her research interest includes Microarray Analysis and Bioinformatics and Text Mining.