EAI Endorsed Transactions
on Energy Web

Research Article EAI.EU

Gene Expression Analysis to Mine Highly Relevant Gene Data in Chronic Diseases and Annotating its GO Terms

J. Briso Becky Bell[1,*] and S. Maria Celestin Vigila[2]

[1]Research Scholar, Department of Computer Science and Engineering, Noorul Islam Centre for Higher Education, Kumaracoil-629180, TN, India.

[2]Associate Professor, Department of Information Technology, Noorul Islam Centre for Higher Education, Kumaracoil-629180, TN, India.

E-mail: [1]brisobell30@gmail.com , [2]celesleon@yahoo.com

Abstract

Gene Expression Analysis seeks to find the highly expressive genes from a highly dimensional Microarray disease gene Database by using some statistical gene selection approaches based on supervised or unsupervised learning. Gene Ontology (GO) introduces a series of method for annotating gene function that combines semantic similarity measures by taking account on the underlying topology of gene interaction networks for structuring the graphs of the gene ontology. Initially, the genes are identified by clustering microarray disease dataset giving gene id of most expressive genes and further the genes are associated based on their biological functionalities using the gene ontology annotations taken from bioinformatics database. Also, t-test is used for finding the up-regulated genes so it can be annotated to find the most significant gene terms in hierarchical graph structure. The proposed method uses term Similarity measures to compare two or more gene ontology terms. Finally, gene functional classification and gene term association is done by forming a graph structure to be readily analysed by medical practitioner intending the nature of disease-causing genes at deeper level of understanding in chronic disorder based health care environments.

Keywords: Gene Expression Analysis, Chronic Disorder, Data Mining, Micro Array, Gene Ontology.

Received on 25 April 2020, accepted on 15 May 2020, published on 27 May 2020

doi: 10.4108/eai.13-7-2018.164821

*Corresponding author. Email: brisobell30@gmail.com

1. Introduction

Gene ontology (GO) has extensive number of techniques and methods for the assuring the biochemical qualities of genes. In order to assure the spatial structures of genes, the research area has focused on taking gene expressions [1] of Bio-Informatics analyzed along with certain roles of genes [2] Genes assume their role by connecting among them or with different macromolecules, for example nucleic acids. Generally a communication includes a contact among surfaces of two or more gene. In this way, unusual techniques have been presented for the assurance of communications among gene. Thus introducing many soft computing techniques to manage, store and compare gene terms in system biology and molecular biology data. The entire arrangement of gene interaction is additionally used as a Reference Gene Interaction Network (RGIN).

RGINs have been effectively demonstrated by utilizing undirected charts, where nodes are related to genes, and edges represented to interaction among other genes. Here we utilize a simplest undirected chart, while increasing refined models used guided and named edges to incorporate the data about the sort of biochemical affiliation and its course. In this way, the investigation of RGINs requires chart based computational techniques. RGIN information has been gathered in numerous open databases, for example, the Bio-molecular Interaction Network Database (BIND) [3] and Human Genome Gene Ontology Database (HGGODB) [4].

These databases are regularly freely accessible on the Internet offering to the client to recover information from basic query interfaces. The client can search directly through the inclusion of: (i) more or one gene identifiers, (ii) gene succession, and (iii) the organisms' name. Results may comprise of, individually, a set of genes that interact with the gene or that are space k from the seed gene, or the set of every interactions. Frequently it is difficult to detail query that is glucose synthesis is related to the every interactions.

At present there is an absence of strategies to expand the semantics of such GO fresh data with more data [5], in order to add genuine information with the data created from ontologies [2] and other sources [6]. This paper imparts on GODB information with natural data may result in more queries interfaces and in increasing gene data analysis methods [7] that may utilize such bioinformatics data in expansion to the topological data given by current GO databases(GODB) and systems.

The paper proposes a software platform for the investigation of RGIN information enhanced with gene ontology learning. Such framework could be valuable for the semantic search of information, as well as for the investigation of GODB information. The investigation of GODB systems is typically done by utilizing diagram based calculations, and partner graph properties (irregular diagrams or scale free models) to functional properties of the displayed gene. The accessibility of the information could empower the development of novel calculations lead to understand cellular biology data for physicians and genetic engineers to do further research in origin of diseases.

Initially, the different types of databases are introduced and briefly described in Section 1. Section 2 describes the relevant methods used in studying genes with databases Section 3 describes the proposed model of gene ontology analysis with clustering, annotation and semantic similarity measures. Section 4 details about the result and discussion of implemented and observed phenomenon. Finally, the system is concluded in section 5.

2. Literature Review

The molecular cell function related rich source of data is usually available as Microarrays. It consists of reports of thousands of gene expressions [8]. In order to analyse disease markers aiming drug target cure, each and every human disease are being examined and studied using microarrays experiment mostly, targeting the novelty of genes involved in diseases. Gene Expression Analysis (GEA) plays a vital role in processing the information, by embedding a large-scale expression profiling, and checking the biological interpretation of microarray data having high throughput values [9]. Yet another challenging task in microarray gene expression analysis is to identify the gene expression changes with association to particular biological condition. The essences for identifying genes involved in each biological condition must need of model with careful statistical design based on analysis and should be implemented with visual plots to infer the identified genes with precision.

It uses soft computing learning methods in order to select or cluster a set of genes from the large feature set gene data

Clustering is a technique to recognize genes that are expressed in every biological situation [10]. Clustering is one of the unsupervised learning techniques [4] for analyzing the gene expression data. The primary biological hypothesis for clustering of genes is that of like genes exhibiting similar expression patterns mostly participate the same biological process. A hierarchical clustering was performed to the sample space using correlation distance and average link, in order to select these preliminary genes. By choosing different cut points in the tree, i.e. for the different number of clusters, different prototypes can be created. At each cluster, most well represented categories are found and each new gene associated to its own category was found by taking the average of the gene expression value in same category of the genes. The categories at a fixed level were taken in order to restrict the number of categories.

Gene selection is an important step in many learning algorithms, for reducing the high dimensional data. Usually microarray-based data-sets often have thousands of genes, in order to identify Differentially Expressed Genes (DEG) [5]. One can compare the gene expression values between any two categories of data. A metric evaluates the effectiveness of each individual gene in finding the category of each sample. Here supervised learning is used by taking categorical data in to account for finding relevant scores. Researchers have developed a large number of rules to score genes linearly and maximize the speed of their calculations.

This speed is the metric's strength, but because they do not look for interactions between genes, they only select a set of strong genes. The gene selection metrics can handle binary data, but these were forced to use binary data to ensure no metric had an advantage because of a gene's structure and how it was discretised as category.

In the wide study of microarray records, one requests to interpret genes IDs for its consequence. Annotation exposes the biological consequence of genes like its molecular function, illness occupied, gene ontology, and so on [11]. Careful investigation is necessary to recognize genes that are expressed in every situation of microarray experimentation. Specific path and method implemented are crucial part of annotation; GO is an inter-related database containing gene ontology terms of every gene. Here a set of gene ontology terms are annotated to every known gene; the terms are connected as a Directed Acyclic Graph (DAG), as the levels represent specifications of each terms. For each gene terms there are three types of annotations currently available: cellular compartment, biological process and molecular function. In annotating the genes, Gene annotations of many organisms are found at different GO websites. National Centre for Biotechnology Information (NCBI) is a Gene Ontology Consortium that has a list of gene annotations which are related by their entrez identifier in GO database. Usually these annotations are mostly updated at frequent intervals and are curated by organism specific genome group members. The GO database contains field names such GO_ID, DB_Object_Symbol, GO_Name, GO_Description, Aspect_Field [12] and etc. The fields of importance are the ID and the associated gene symbol, thus creating a map for efficient search for gene ID. In this analysis you look at gene's aspect field that is annotated as molecular function to 'F'. However, to see the set of genes involved in same biologic processes ('P') or are genes located in the same compartment cell ('C').

3. Gene Annotation Technique

Gene annotation technique is utilized for knowing the function, reuse of data and content-based queries. It can be processed using Reference Gene Interaction Networks and gene ontology. The GO is selected since its capacity to give dynamic, precisely definite, structured, and controlled conditions that express genes. In addition to that, GO terms are associated with gene products and their gene sequences from different species provided by Gene Ontology Annotation (GOA). So, by using the GO one can measure the Term Similarity of any two or more GO terms. To find the relevant gene terms association and its functional classification. Here we also proposed an idea of Gene Expression Analysis such as cluster analysis to bring out the highly expressive genes using K-Means algorithm and T-Test to select up-regulated genes for finding annotated genes from GO database thus finding the number of genes annotated to terms. Also, by looking at the probabilities of each term's count one can find the most significant terms and sub-ontology is constructed using the ancestors of the top most significant terms. The proposed model is depicted in **Figure 1.**

The gene ontology [13] utilizes brought together and organized controlled vocabularies to portray quality capacity. These are the three sorts of class in which each of the GO terms: Cellular Component (CC), Molecular Function (MF) and Biological Process (BP), BP refers to a biological Process to which the gene or gene associated terms contributes. This process is accomplished by two or more ordered assemblies of biological functions. In this process it often involves a physical or chemical transformation, in the sense that something different comes out of it, when something gets into the process. The examples of biological process terms are as 'mitosis', 'purine metabolism', 'Glycolysis', 'Oxo-Nitrate Carbon Removal', etc. MF refers to the biochemical or Molecular Function of a gene or gene associated terms. These represent the capability of a gene terms whether it is potential or not. It does not specify where or when the event actually occurs but it describes what is done. The examples of functional terms are 'Enolation', 'Phosphorylation of Bis-Phosphate', 'Fructose kinesis', 'Glucose Phosphorylation', Isomerization of Phosphate', 'Clevage of Fructose Bis-Phosphate', etc.

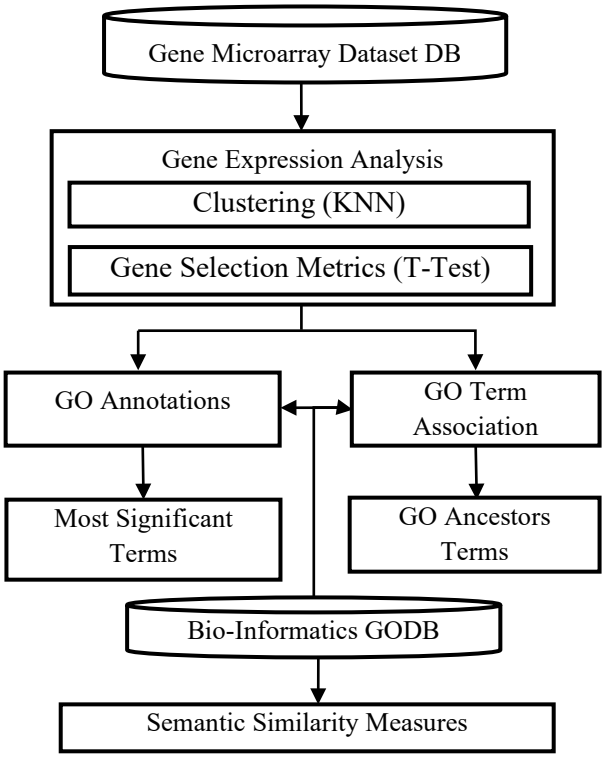


Figure 1. Proposed GO based GEA model

CC refers to the Cellular Component in which the gene or gene associated term is located i.e., the location in the cell where a gene term is active. These terms reflect our understanding of eukaryotic cell structure. The examples include such terms as 'ribosome', 'nuclear membrane', 'Golgi apparatus', 'Cytoplasm', etc.

In the GODB there are almost 12,000 GO terms [4] at present and most GO terms are progressively organized by a "part_of" relation with one or more terms, in which each GO expression is a specialization of its progressive conventional terms. There are three kinds of GO terms, Consequently, it has three DAGs by logical representing GO databases. For instance, in **Figure 2**, GO:0003674 (MF) is the base of the DAG for terms of Molecular Function, and it is the parent of GO:005488, which is thus the parent of GO:0097159 additionally, GO:005488 is a parent of GO:1901363 and both nodes GO:0097159 and GO:1901363 are parents of a new child node GO:GO003676.

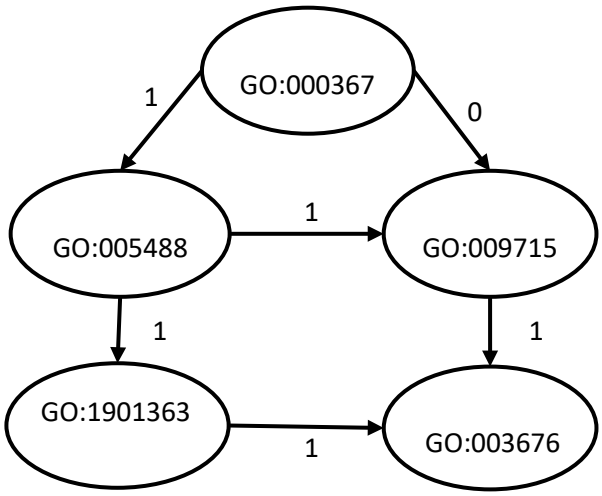


Figure 2. Sub ontology of gene ontology terms

Consider a chain of command of highlights, where each element to a gene ontology expression which is a hub in a gene ontology DAG. Each component takes a two-fold value, "0" or "1", showing whether or not an occurrence (a quality) is commented on with the relating gene ontology term. The "is-a" pecking order of the gene ontology is related with two various levels of requirements. To start with, if an element takes the esteem "1" for a given example, this suggests its progenitors in the DAG additionally take the esteem "1" for that example. For instance, if the term GO:0005488 has esteem "1" for a given quality, at that point the estimation of term GO:0003674 ought to be "1" also. Alternately, if the component takes the esteem "0" for a given occasion, this suggests its relatives in the DAG additionally take the esteem "0" for that case. For instance, in the event that the term GO:0097159 has esteem "0", term GO:0003674 ought to likewise have esteem "0".

Gene ID
8614

Gene

GO Term ID	Functional Name
GO:0007566	embryo implantation
GO:0007267	cell-cell signaling
GO:0007166	cell surface receptor
GO:00043434	response to peptide
GO:0007584	response to nutrient
GO:00046697	decidualization
GO:00033280	response to vitamin D
GO:00071456	cellular response hypoxia

GO Terms

Figure 3. Gene association with the GO terms

Bioinformatics database leukaemia contains gene level of normal and abnormal patients. The gene clustering [15] is used to identify and separate each patient's disease-causing genes. From these genes, the gene having gene ID (i.e., 8614) is searched in GODB [16]. These GODB is stored in separate sheets having GO term ID, functional description & association features. Finally, the GO database is searched for GOID by matching gene ID in **Figure 3.** Then, Semantic Similarity measures are used to test the gene terms for that chosen one ID's from the database which form a hierarchy with gene term Interaction levels of particular gene.

The idea of this approach is to assign each GO term ci occurring in gene g to its best matching partner c_0 if in gene g_0. So, gene g is associated with multiple GO terms and one GO term from gene g_0. A similarity score is computed by taking the average similarity [17] of assigned GO terms. However, while considering genes of having unequal number of GO terms the result always depend on GO terms of gene g is associated to those of gene g_0 or vice versa and in order to either take the average or the maximum of both similarity scores is decided.

Lin's measure is based on the similarity between two GO terms c_1 and c_2 is defined as the ration of the commonality of the information needed to describe c_1 and c_2 and GO terms. The commonality of c_1 and c_2 is captured again by their common ancestors as defined in equation (1)

$$\text{sim}\,(c_1,c_2) = \frac{\min\limits_{c\in S(c_1,c_2)} P(c_1,c_2)}{\min\limits_{c\in S(c_1)} P(c_1)+\min\limits_{c\in S(c_2)} P(c_2)} \quad (1)$$

This approach considers the notion of common information content of the terms in taxonomy as the basis for similarity. Here instead of p(c) a negated log of the likelihood p(c) is the relevant information gain of a term which can be quantified to terms. The lesser its information gain when higher a term appears in the ontology. The information gain that they share indicates the similarity between two nodes. By this the information gained in the set of their common ancestor nodes can be easily found by equation (2).

$$\text{sim}\,(c_1, c_2) = -\log \min_{c\in S(c_1,c_2)} P(c_1,\ c_2) \quad (2)$$

In Gene Expression Analysis K-Means clustering is a grouping approach to cluster the observations by n×p data matrix x into k clusters, so returning a vector n×1 having cluster indices for each observation. Each column corresponds to variables and each rows of x correspond to points. Clustering techniques uses a distance metrics such as Euclidean as shown in equation (3) to compare expression values of couple of genes for every research [15]. When the distance between couples of genes is little, then the two genes might be clustered. Cluster centre locations are found as a matrix having k×p rows & columns, where the centre of the cluster is j. The distances of each gene observation to each cluster centre using equation (4), and assign each gene to its closest cluster centre using equation (5).

$$d(x,c)=(x-c)(x-c)' \quad (3)$$

$$d(x_m,c_j)=\frac{d^2(x_m,c_1)}{\sum_{j=1}^{n} d^2(x_j,c_1)} \quad (4)$$

For p = 1,...,j – 1 and m = 1,...,n, select cluster centre j at random from x with less probability where the set of all observations closest to cluster centre is c_p Then, update cluster centre with a lower probability relative to the distance from the closest cluster centre that you chose already to the calculated data point.

$$d(x_m,c_j)=\frac{d^2(x_m,c_p)}{\sum_{\{h;x_{h\in c_p}}^{n} d^2(x_h,c_p)} \quad (5)$$

With the use of Statistical Significance one can compare the two categories of gene expression data values using two-sample t-test [6] for evaluating differentially expressed genes in binary categorical dataset. It performs a t-tests each gene for identifying significant changes in expression values between the two categories of samples. Conduct an unpaired t-test for differentially expressed genes with a standard two-tailed t-test on every gene as shown in equation (6) and (7), also it returns a p-value for each gene, Thus obtaining a matrix of expression values, in which each of the row corresponds to a gene, and each of the column corresponds to a sample. The x and y contain data from each of the two respective categories.

$$t=\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}} \quad (6)$$

$$s=\sqrt{\frac{(n-1)s_x^2+(m-1)s_y^2}{n+m-2}} \quad (7)$$

Where, $\bar{x}$ and $\bar{y}$ are the means of samples, s_x and s_y are the standard deviations of samples, and n and m are the sizes of samples in two categories of data. Where the two samples of data are of populations having variances equal, t distribution with n + m – 2 degrees of freedom is the test statistic under the null hypothesis. Here p is the probability of observing a test statistic as, the observed value under null hypothesis. When h = 1, which indicates the rejection of the null hypothesis at the alpha level of significance. When h = 0, which indicates a failure to reject the null hypothesis at the alpha level of significance.

In this testing method, two kinds of errors can occur, a false negative (FN) occurs when the test misidentifies a truly differentially expressed one and a false positive (FP) occurs when it is not differentially expressed; it declares the gene as differentially expressed one. Multiple hypotheses testing, test the null hypothesis of many genes simultaneously, and every test has a false positive rate (FPR).

It is defined as the fraction between the total number of positive calls and the total number of false positives in a differential expression analysis between two categories of samples. In order to determine the most differentially expressed genes a number of genes are considered having statistical significance at p-value cut-off at alpha.

FPR Estimates a positive (pFPR) value for each value in p-value. In order to estimate the test statistics null distribution, permutation methods are used, by finding the permutation of the columns in gene expression data. The whole proportion null hypothesis of truth is the quantity p_i0. By using cubic polynomial fit, y is estimated from the simulated null distribution and it is shown in equation (8).

$$y=\sum_{i=1}^{n+1} p_i\, x^{n+1-i} \quad (8)$$

Where, n is the degree of polynomial, n + 1 is the order of the polynomial and value of n ranges between $1 \le n \le 2$. The degree is the highest power of the predictor variable, and the order is the number of coefficients to be fit. When it shows both biological and statistical significance, a gene is said to be a differentially expressed one in the two categories of

samples. So, when it has higher expression in category A, it is an up-regulated gene and when it has higher expression in Category B, it is down-regulated gene.

Perform GO Annotation operation For each GO term in the gene ontology Database, count the no of annotated genes from the Microarray. The number of annotated genes to the term and the number of under or over expressed genes which are annotated to the GO term. Using this information, one can find statistically, how many GO terms are under or over represented in the MA experiment. So, you create these p-scores by including the neighbouring GO terms. To be more specific GO terms those are included as ancestor terms. When a gene is annotated to a GO term, it would increase the count of all or some of its ancestor GO terms. This takes relatives, descendants and ancestors to test different propagation schemes. Here it takes relatives to get descendants and ancestors of each term up to one level of ancestral hierarchy, in order to overcome a roughly annotated gene set.

In most significant term analysis, by considering the probabilities of each terms with counting on a chance, one can find the significant most annotated GO terms and one can calculate the statistical significance [18] How many have drawn a specific number of successes, in the number of total draws on a population, by using the probability distribution function. The p-value is calculated and the probabilities of finding such test statistical gives the total count of the GO Terms. Here it computes and assigns the p-value to associated each GO terms, and by using it one can get a list of top most significant gene ontology terms by sorting the p-values. By using the top item on the list, one can build a sub-ontology [19] which is the ancestors of the top significant terms. One can see the Terms in hierarchy, by using the calculated p-values, assign a colour to the various graph nodes. Here colour map is used to see the significant level, pale colours indicates least and red indicates most significant terms.

4. Results and Discussion

In this part, using the gene id the gene level for individual gene is displayed. This leukaemia dataset [20] has 7129 genes on the microarray and having 72 samples to each gene expressions. Using the MATLAB 2019 output is observed.

The k-means clusters the data into two groups by correlation distance measures. A number of 3242 genes are selected from a total number of 7129 genes in the first cluster category. The bottom left corner shows the first cluster of genes which are up-regulated which is observed in **Figure 4**. The genes of this cluster will be sub selected to be used for the subsequent phase of this experiment. Thus, the number of annotated terms related to MF is 17041 and a GO object with 32 Terms 38 edges is formed. By finding the p-values, colours are assigned to the graph nodes in **Figure 5.** The nodes having shade closest to green are the least significant ones and the nodes having shade red are the most significant ones. The P-value of annotated terms is shown in Table 1.

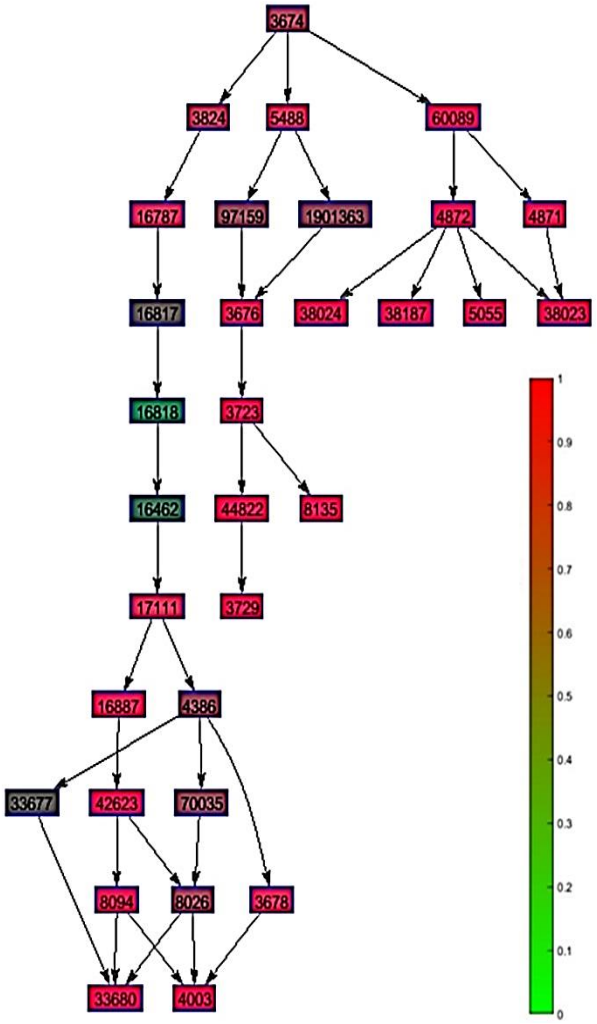


Figure 4. K-means clustering red and blue clusters

Table 1. P-Value and Count of Annotated GO Terms

GO Term	P-value	Count	GO Term Name
44822	0.0000	318 / 559	poly(A) RNA binding
03729	0.0000	277/ 492	mRNA binding
03723	0.0000	361 / 664	RNA binding
38024	0.0000	39 / 146	cargo receptor activity
04003	0.0000	41 / 52	ATP-dependent DNA receptor
38023	0.0000	142 / 403	signalling receptor activity
08135	0.0000	136 / 226	translation factor, RNA receptor
33680	0.0000	28 / 33	ATP-dependent DNA/RNA bind
05055	0.0000	35 / 129	laminin receptor activity
38187	0.0000	34 / 126	pattern recognition receptor

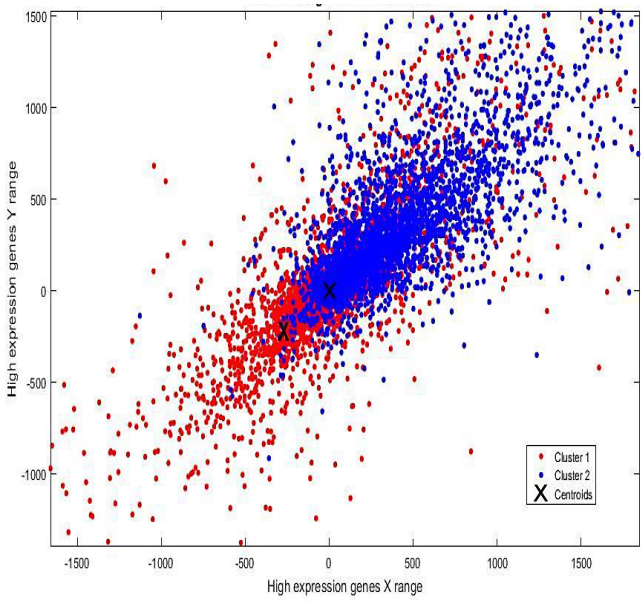
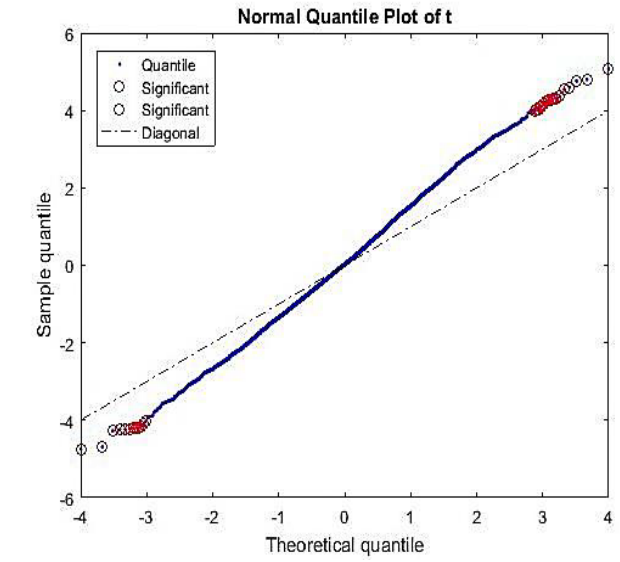


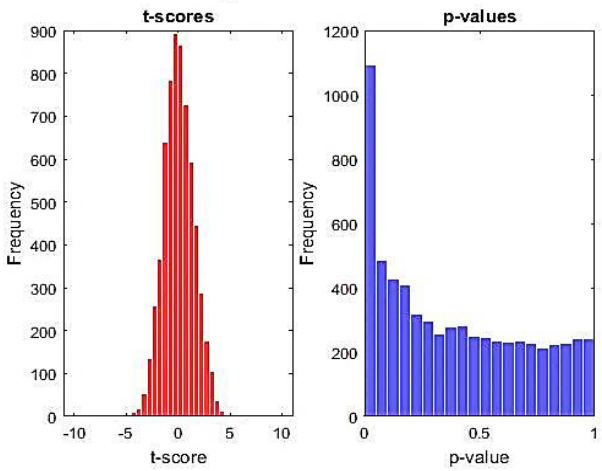
Figure 5. P-values for annotated terms in clustering

In 72 samples and 7129 genes of gene expression data in the Leukaemia dataset, the categories are of 47 ALL Acute Lymphoblast Leukaemia and 25 AML Acute Myeloid Leukaemia samples. Initially investigate a t-test on each gene for identifying significant expression change [21] in between the ALL samples and AML samples. The histograms of t-scores and p-values and the normal quantile plot of t-test are plotted for the t-tests are shown in **Figure 6**. Then select the genes which have p-values less than 0.05 the assigned cut-off value.

Likewise, while computing the q-value the observation of test using cubic polynomial fit and p-value vs q-value is displayed in **Figure 7**, which usually measures the minimum FPR which occurs while testing the test significance. When estimating FPR, it relays on the real null distribution of many tests, so the genes that have q-values less than 0.05 cut-off values are selected. Much of the genes with low FPR shows that the two categories, ALL and AML, are so biologically distinct. Then display the first 10 genes selected using P-value. The two categories of samples show both biological and statistical significance [22] between them when a gene is highly differentially expressed. So, while comparing the gene expression ratio of ALL over AML disease samples, a down-regulated gene has very high expression in AML and up-regulated gene has very high expression in ALL in this example. Plot the p-values of $-\log_{10}$ to ratio against the biological process $\log_2$ using a graph plot shown in **Figure 8**. In the plot, the genes in the list are used to label the genes. As seen in the graph plot, the up-regulated gene list is found, while genes typical down-regulated list is found from p-values. The total number of differentially expressed genes is 6415. In particular, there are the 507 up-regulated genes and 391 down-regulated genes are found and the topmost genes are displayed in Table 2.



(1)



(2) (3)

Figure 6. Quantile plot & Histogram of t-test

Table 2. P and q- values for annotated genes

Gene ID	T-score	P-values	FPR	q-value
3031	5.0750	0.0000	0.0177	0.0147
4422	4.8004	0.0000	0.0248	0.0147
4755	4.7593	0.0000	0.0194	0.0147
3815	4.7466	0.0000	0.0151	0.0147
2077	4.6890	0.0000	0.0147	0.0147
4835	4.5971	0.0000	0.0168	0.0168
1220	4.5499	0.0000	0.0170	0.0170
1179	4.2744	0.0001	0.0285	0.0210
3767	4.2423	0.0001	0.0248	0.0210
2228	4.2393	0.0001	0.0234	0.0210

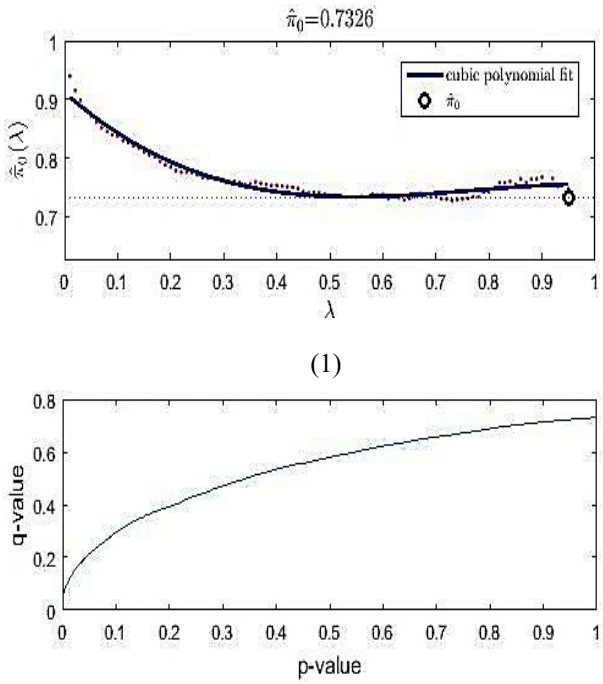


Figure 7. Testing significance by (1) Cubic polynomial fit and (2) p-value vs q-value

Table 3. P-values of annotated GO terms

GO Term	p-value	counts	GO Term Name
22857	0.00016	18 /117	Transmembrane transporter activity
46933	0.00038	7 /24	Proton-transporting ATP synthase act. RM
35605	0.00041	4 /7	Peptidyl-cysteine S-nitrosylase activity
46961	0.00049	7 / 25	Proton-transporting ATPase activity RM
03700	0.00060	15 /481	Sequence Specific DNA-binding transcription
46982	0.00060	8 /330	Transferase Activity
00981	0.00067	17 /519	sequence-specific DNA binding RNA PM II
05096	0.00069	15 /100	ATPase activator activity
03705	0.00092	1 /143	Hydrogen Exporting ATP Phase Act.
01010	0.00099	15 /467	sequence-specific DNA binding transcription factor

For finding the indices of the up-regulated genes in GO analysis one has to find the statistically significant gene ontology terms by the probability distribution function. The p-value is calculated for every gene ontology term and the probability of the number of genes annotations which are associated with each GO term can be found and it is shown in Table 3.

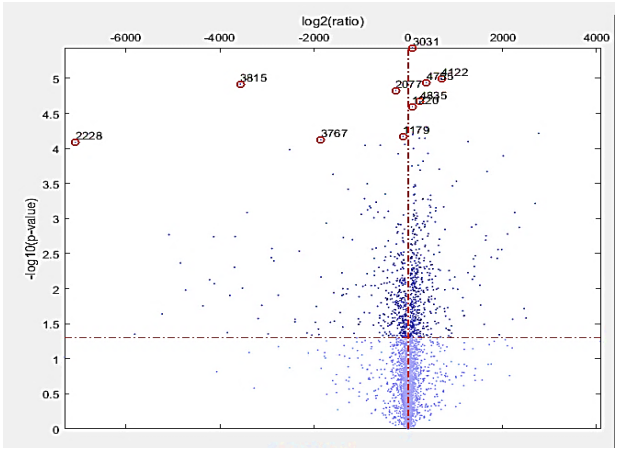


Figure 8. P-values of up and down regulated genes

Based on highest count value P-value is calculated. Then select the gene ontology terms relevant to specific molecular function thus forming a sub-ontology which includes a set of ancestors to the corresponding terms. The level of significance of p-values is shown in **Figure 9**. The nodes having shade closest to blue are the least significant ones and the nodes having shade red are the most significant ones.

Gene Ontology Term Association associates any two parent gene terms, let's consider c_0 GO:0001010 and c_1 GO:0044822, Here it represents a hierarchy of Molecular Function c and has 10 ancestor Terms c_0,1:5 & c_1,1:5 and its results is given in Table 4, where P represents the Parent Node or leaf Nodes, H is the root hierarchy of defined gene term and A corresponds to the various ancestor terms of specific parent terms. As in **Figure 10** it forms a graph of 13 nodes and 14 edges.

Table 4. GO term ancestors

GO Term Names	P/A/H	GO Term
sequence-specific DNA binding TF RTFA	P(c_0)	01010
protein binding TFA	A(c_0,1)	00988
transcription factor binding TFA	A(c_0,2)	0000989
nucleic acid binding TFA	A(c_0,3)	01071
transcription factor recruiting TFA	A(c_0,4)	01134
sequence-specific DNA binding TFA	A(c_0,5)	03700
Molecular Function	H(c)	03674
poly(A) RNA binding	P(c_1)	44822
nucleic acid binding	A(c_1,1)	03676
RNA binding	A(c_1,2)	03723
Binding	A(c_1,3)	05488
organic cyclic compound binding	A(c_1,4)	97159
heterocyclic compound binding	A(c_1,5)	1901363

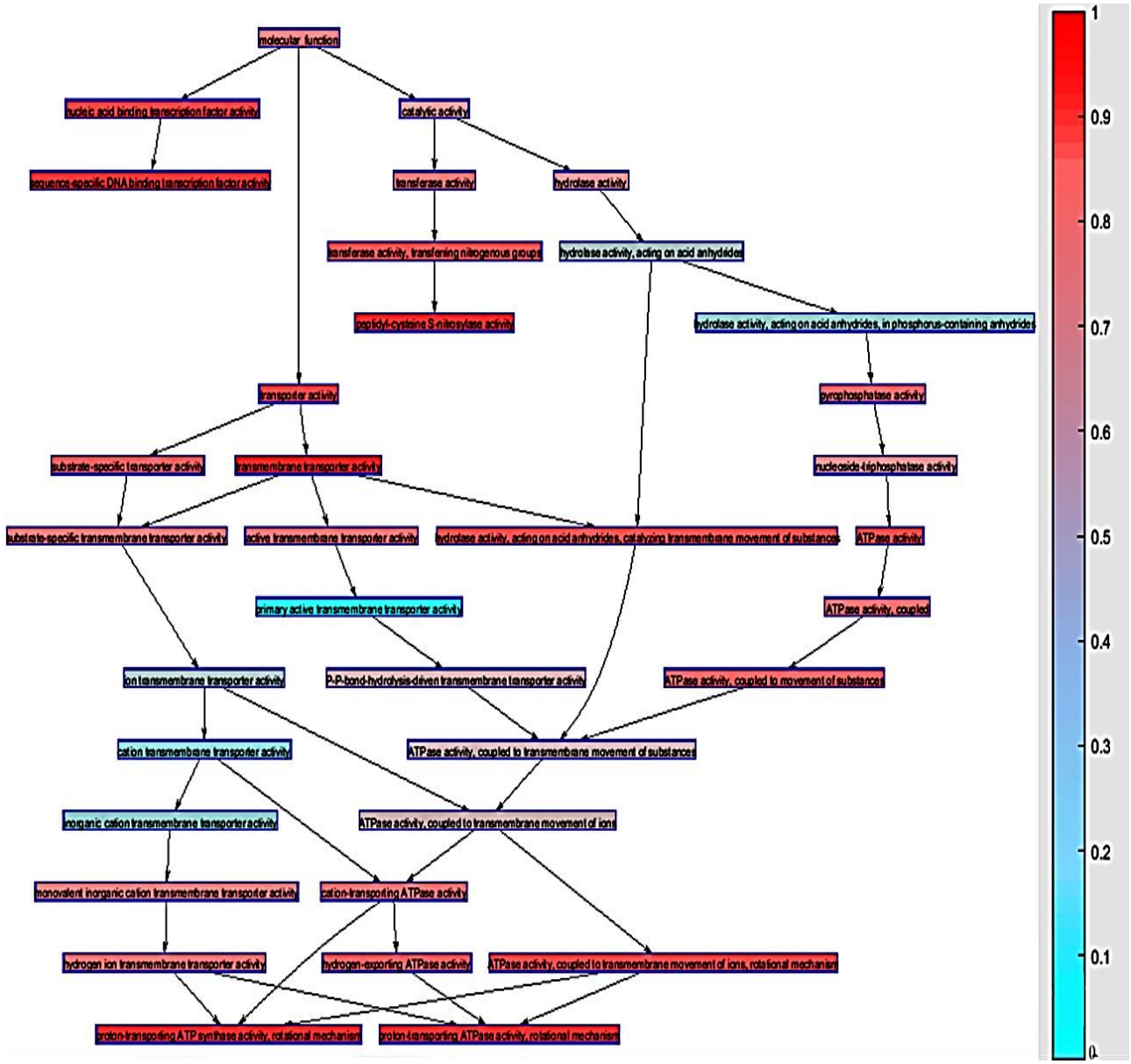


Figure 9. GO Annotation for Up-regulated genes

Table 5. Resnik's term similarity

$P(c_0,c_1)$	GO: 22857	GO: 46933	GO: 35605	GO: 46961	GO: 03700
GO: 22857	1.000	0.587	0.326	0.846	0.808
GO: 46933	0.587	1.000	0.000	0.663	0.488
GO: 35605	0.326	0.000	1.000	0.360	0.280
GO: 46961	0.846	0.663	0.360	1.000	0.665
GO: 03700	0.808	0.488	0.280	0.665	1.000
IG(c)	3.101	3.101	4.115	5.859	2.276

Table 6. Lin's term similarity

$P(c_0,c_1)$	GO: 22857	GO: 46933	GO: 35605	GO: 46961	GO: 03700
GO: 22857	0.274	0.187	0.129	0.201	0.274
GO: 46933	0.187	0.364	0.000	0.187	0.187
GO: 35605	0.129	0.000	0.519	0.129	0.129
GO: 46961	0.201	0.187	0.129	0.201	0.201
GO: 03700	0.274	0.187	0.129	0.201	0.404
IG(c)	3.117	3.117	4.044	5.774	2.286

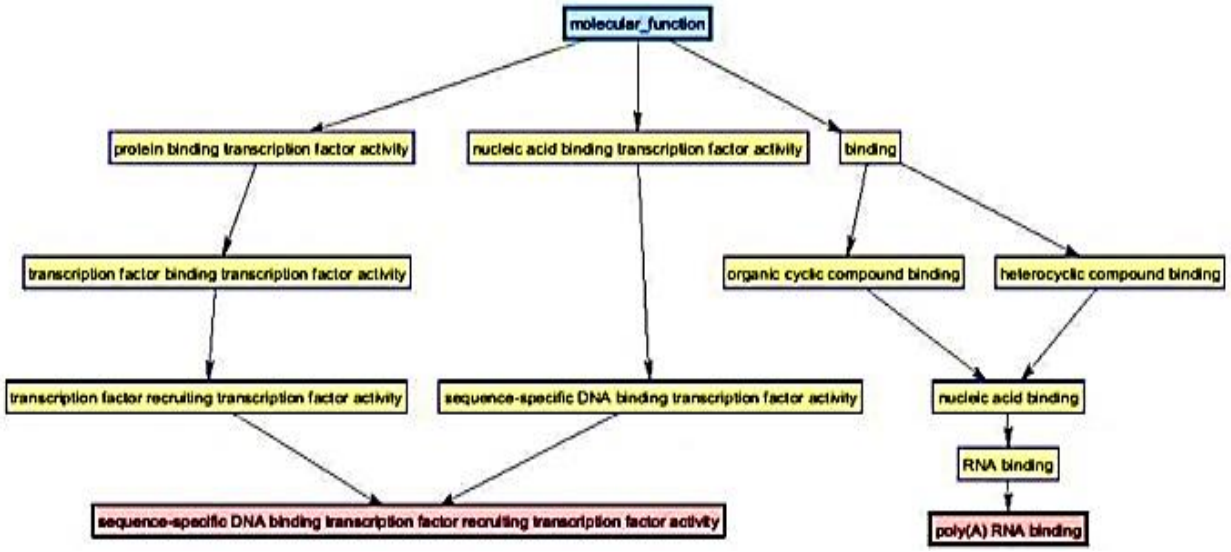


Figure 10. GO-Term Association

Term Similarity is used to validate GO analysis, the Term similarity analysis of GO terms is done, the gene Ontology terms are taken and c("GO: 0022857", "GO: 0046933", "GO: 0035605", "GO: 0046961", "GO: 0003700"), are its GO term. Here P(c_0; c_1) corresponds to the set of all common ancestors of GO terms c_0 and c_1 and IG(c) denotes the information gain of term c1. The IG of each GO term is computed for each ontological based observation and how much times a GO term or any of its indirect or direct offspring appear on annotation in GO database, as in Table 5 and 6. For term analysis let us consider the first five entries of statistical significance-based GO terms with highest P-values, and the similarity between each term and the Information Gain were significantly high for the Leukaemia genes.

5. Conclusion

Gene association of gene ontology terms is a challenging task in gene ontology analysis. The efficient implementation of statistical analysis tools using GO are analyzed and validated. In this Novel Method, The K-means Clustering Based approach and Statistical Significance based computational approach selects highly significant genes, to annotate its relevant GO terms for these significant genes using the GO database to find the most significant GO terms. Thus, the GO term association is observed and a proposed technique of annotating GO terms is practical method that is faster, easy to configure and has low computing cost, and using term similarity measures the GO Terms are validated effectively. While taking the statistical significance-based GO terms for Term similarity a high Information Gain is obtained for leukaemia genes.

References

[1] Kuznetsova, S., Irina, Artur, L., and Andreas, H. (2018) Visualisation Methods of Hierarchical Biological Data: A Survey and Review, Information Systems and Management in Creative Media, **05**(18):32-39.

[2] Rue, A., and Kévin, Paul, A. (2018) GO express: identify and visualize robust gene ontology signatures through supervised classification of gene expression data, Gene Expression Patterns, **7**(8):17-25.

[3] Andreas, S., Thomas, L., and Mario, A. (2010) Improving disease gene prioritization using the semantic similarity of Gene Ontology terms, Bio-Informatics, **26** (15):55-61.

[4] Kristian, O., Marko, L., and Sampsa, H. (2008) Fast Gene Ontology based clustering for microarray experiments, BioData Mining BioMed Central, **1**(1756):1-11.

[5] Lee, J., Shannon, C., Mithun, A., and Kumer, D. (2018) Dimension reduction of gene expression data, Statistical Theory and Practice, **12**(2):450-46.

[6] Buerger, H., and Florian, B., Jens, P., Konstantin, A., Kathrin, P., Walter, N., and Eberhard, K. (2018) Analyzing the basic principles of tissue microarray data measuring the cooperative phenomena of marker proteins in invasive breast cancer, Arxiv Preprint, **18**(3):1-28.

[7] Frank, E., and Galina, V. (2011) Pathway Analysis of Expression Data: Deciphering Functional Building Blocks of Complex Diseases, Plos Computational Biology, **7**(5):11-21.

[8] Chandra, S., Srinivasa, R., Allam, A., Srinivas, K., and Chinta, S. (2011) Gene Expression Analysis for Type-2

Diabetes Mellitus, Theotretical and Applied Information Technology, **27**(1):16-22.

[9] Vanitha, D., Devaraj, D., and Venkatesulu, M. (2015) Gene Expression Data Classification Using SVM and Mutual Information Based Gene Selection, Elsevier Computer Science, **47**(5):13 – 21.

[10] Behzadi, P., Behzadi, E., and Ranjbar, R. (2014) Microarray data analysis, Challenge, **7**(2):8-19.

[11] Briso, J., and Maria, S. (2018) Genic Disorder Identification and Protein Analysis Using Soft Computing Methods, Springer, **7**(12):73-81.

[12] Howe, E., Holton, K., Nair, S., Schlauch, D., Sinha, R., and Quackenbush, J. (2010) Multi experiment Viewer Biomedical Informatics for Cancer Research, Springer, **3**(1):267-273.

[13] Wan, C., and Alex, F. (2018) An empirical evaluation of hierarchical feature selection methods for classification in bioinformatics datasets with gene ontology-based features, Artificial Intelligence, **50**(2):201-240.

[14] Zhang, C., Wei, Z., Peter, F., and Yang, Z., (2018) MetaGO: Predicting Gene Ontology of Non-homologous Proteins Through Low-Resolution Protein Structure Prediction and Protein-Protein Network Mapping, Molecular Biology, **23**(5):22-30.

[15] Yadav, R., and Prachi S. (2018) Clustering Pathway Enrichment and Protein-Protein Interaction Analysis of Gene Expression in Neurodevelopmental Disorders, Advances in pharmacological sciences, **12**(11):32-40.

[16] Jacobson, M., Adriana, E., Sedeno, C., and Paul, P. (2018) Monitoring changes in the GO and their impact on genomic data analysis, Bio-Rxiv, **32**(8):61-69.

[17] Kimberly, A., Kylee, G., and Anjeza, P., Darby, J., Fang, Y., James, D., Chi, Z., and Lawrence, H.(2015) Genome-Wide Gene Expression in relation to Age in Large Laboratory Cohorts of Drosophila melanogaster, Genetics Research International Hindawi, **10**:1-19.

[18] Wu, G., and Feng, X. (2010) A human functional protein interaction network and its application cancer data analysis, Genome Biology, **11**(5):53-59.

[19] Werner, T. (2008) Bioinformatics applications for pathway analysis of microarray data, Current Opinion in Biotechnology, **19**(1):50-54.

[20] U.S. National Library of Medicine, (2019) NCBI-National Center for Biotechnology Information [online], http://www.ncbi.nlm.nih.gov/leukaemiadatasets.xls

[21] Malay, B., Joyshree N., and Sanghamitra B. (2016) Identifying significant microRNA–mRNA pairs associated with breast cancer subtypes, Molecular Biology Reports, **43**(1):591-599.

[22] Wan, C., (2019) Hierarchical feature selection for Knowledge discovery, Springer Science and Media LLC.