

A REVIEW OF VARIOUS METHODS USED IN THE ANALYSIS OF FUNCTIONAL GENE EXPRESSION DATA

Houda Fyad, Fatiha Barigou, Karim Bouamrane, Baghdad Atmani

LIO Laboratory, Department of Computer Science, Faculty of Exact and Applied Sciences University of Oran 1 Ahmed Ben Bella
BP 1524, 31000 El M'naouer Oran, Algeria

ABSTRACT

Sequencing projects arising from high-throughput technologies including those of sequencing DNA microarray allowed measuring simultaneously the expression levels of millions of genes of a biological sample as well as to annotate and to identify the role (function) of those genes. Consequently, to better manage and organize this significant amount of information, bioinformatics approaches have been developed. These approaches provide a representation and a more 'relevant' integration of data in order to test and validate the researchers' hypothesis. In this context, this article describes and discusses some techniques used for the functional analysis of gene expression data.

KEYWORDS

Microarray, Genome Annotation, Functional Analysis, Genes Expression Data, Data Mining, Clustering, Classification, Gene Ontology.

1. INTRODUCTION

This paper is an extension of the previous work [1]. In this paper, we add the description of the bioinformatics approaches that have been developed to improve the quality of annotation. Also, we discuss some hybrid methods that performing well in analysis of functional gene expression data.

The successful developments of high-throughput sequencing technology including those of sequencing DNA microarray generated a large volume of genomic data. This technology allows measuring the simultaneous expression of a large number of genes, or even all the genes contained in the genome under many and varied conditions. Also, it identifies the rate of gene expression (over or under expressed); characterization of genes differentially expressed; the establishment of a characteristic profile of a given biological state. Therefore, it provides to researchers the opportunity to study the coordinated behaviour of genes and so better understanding the function of a gene in an experimental situation.

The massive data produced presents a significant challenge for data storage and analysis. In this case, bioinformatics tools are essential for data management. Thus, the transition of the genome sequencing to the annotated genome gave rise to methods, tools, and bioinformatics platforms, to help many areas of biology to manage and organize this mass of data. Some of these approaches using data mining have been developed to find the similar expression profiles of genomic data. Others have used controlled vocabularies or ontologies to capture the semantics of biological concepts describing biological objects such as genomic sequences, genes or gene products. And

some have joined the two above-mentioned approaches. All of this, providing biologists with a more "relevant" representation and data integration allowing them to analyze their genomic data, test and confirm their assumption throughout the experimental cycle.

This article gives a comprehensive overview of the different bioinformatics and data mining approaches employed in the functional analysis of gene expression data. The rest of the paper is organized as follows: section two introduces the concept of the genome annotation with its three levels of complexity. Section three describes the different data mining techniques used in the functional analysis of gene expression. The fourth section deals with the use of gene ontology to build a gene expression profile. And, finally, the fifth section provides some concluding remarks and gives an outlook for future works.

2. GENOME ANNOTATION

DEFINITION AND STRATEGIES FOR GENOME ANNOTATION

As underlined by several authors' genome annotation (or DNA annotation) is extraction, definition, and interpretation of features on the genome sequence derived by integrating computational tools and biological knowledge [2, 3, 4].

An annotation is a note added by way of explanation or commentary. Once a genome is sequenced, it needs to be annotated to make sense of it. Annotation could be:

- gene products names
- functional characteristics of gene products
- physical characteristics of gene/protein/genome
- overall metabolic profile of the organism

For example, genome annotation is notably used by biologists for identification of different genes expressed in plants organs (root, leaf,...) during a cycle of development like the *Arabidopsis thaliana* plant [5], also it is used for identification of genes involved in the rice tolerance to salinity [5, 6] and possibly for the discovery of new functions by the association of genes with "known" genes based on the co-expressed and co-regulation in coral [7]. For the *Drosophila*, it was to determine the present/absent genes in neural flow and synaptic transmission routing [8]. For the mouse, the study consists of analyzing the over or under expression of genes across different genetic manipulation of embryos and adults and the effects of environmental conditions [9]. In medicine [10], it allows distinguishing and classifying types of tumors, knowing the genes expressed on a large number of patients to observe the effect of a drug (e.g. anti-cancer), examine the effect of a treatment on the expression of genes, to compare healthy tissue from diseased tissue, treated against untreated.

The process of annotation can be divided into three levels [11]:

- The syntactic or structural annotation: it identifies sequences presenting a biological relevance (genes, signals, repetitions, etc).
- The functional annotation: it predicts the potential functions of the previously identified genes (similarities of sequences, patterns, structures, etc) and collects any experimental information (literature, big data sets, etc).
- Relational or contextual annotation: it determines the interactions between the biological objects (families of genes, regulatory networks, metabolic networks, etc).

In summary, genome annotation involves finding the relationship between a new sequence of genes and other known sequences that often reveal its structure, function, and evolution [12]. Moreover, these different levels of annotation are not separated, rather intermingled, and are very closely related. The genome annotation is a precisely interconnection of these three levels mentioned above [11]. Many bioinformatics approaches have been developed to improve the quality of annotation. They are: (i) sequence similarity searching, (ii) searching for patterns in sequence.

2.1. SEQUENCE SIMILARITY SEARCHING METHODS

Sequence similarity searching, compares two sequences or multiple sequences in order to identify similar regions of the gene that can indicate structural and/or functional relationships between these biological sequences [13]. This comparison uses either the global alignment such as the Needleman-Wunsch algorithm [14] or the local alignment as the Smith-Waterman algorithm [15]. The Needleman-Wunsch algorithm is based on dynamic programming. It performs the optimal global alignment of two sequences [13, 16]. It comprises three steps: (i) Initialization of the score and the traceback matrices, (ii) Calculation of scores and filling in the score and traceback matrices and (iii) Inferring the alignment from the traceback matrix [13]. It guarantees to find the maximum score alignment. This was the first application of dynamic programming for the comparison of biological sequences.

The Smith-Waterman algorithm achieves the optimal local alignment and discovers the best alignment of the conserved domain of two small sequences [13, 16]. This algorithm has similar steps of the Needleman-Wunsch algorithm but with some differences: in the initialization step, the first row and the column of the matrix are filled with zeros [17]. In the scoring step, the negative scores are not included and are replaced by zero [17]. Finally, in traceback step, the algorithm starts with the cell that has the highest score and works back until a cell with a score of zero is reached [17]. The difference with the Needleman-Wunsch algorithm is that while the latter searches for global sequence alignments (involving the entire length of the two sequences to be aligned), the Smith-Waterman algorithm also looks for local alignments, involving only regions or segments of the two sequences analyzed. It therefore makes it possible, for example, to identify proteins which possess a domain in common among other different domains.

2.2. METHODS FOR PATTERNS SEARCHING IN SEQUENCE

A pattern is an element that has multiple occurrences in a given set of biological sequences [18]. The aim of methods for searching patterns in sequences is to identify an unknown pattern in a given set of a sequence [18]. To perform this pattern search, several methods are proposed and they are based on various algorithms: (i) the exhaustive search algorithm, (ii) the probabilistic optimization algorithm, (iii) the deterministic optimization algorithm.

The principle of the exhaustive search algorithm is to determine all users' possible patterns with for each of them a score or statistical significance. Enumerative method using exhaustive search algorithm works as follows: all possible patterns meeting the constraints provided by the user are listed. For each pattern, the program finds its occurrences in the input data sequences and, according to these occurrences, assigns a score or a statistical significance to the pattern. Long patterns are more difficult to identify because the running time increases exponentially with the length of the [18].

Assuming that a pattern occurs in all sequences, the purpose of the probabilistic optimization algorithm is to look for the best pattern of a fixed length W of sequences. The algorithm is carried out in iterations. The result of each iteration is a set of sub-sequences of length W minus one from each sequence. This set of subsequences represents the occurrences of the pattern. We can compute a position weight matrix characterizing the pattern from this set of occurrences [18].

Sometimes a pattern cannot be described well by a simple deterministic pattern and one may wish to express it in a form of a stochastic model, such as Hidden Markov model or position weight matrix (which is a simpler version of HMM). This kind of pattern is discovered using iterative expectation maximization techniques that do not necessarily converge to the global maximum [18].

In addition to the genome annotation by alignment of sequences, which allowed the discovery of functions of various organisms, methods, and techniques of data mining have appeared for grouping and classifying genes to similar behaviour. Also controlled vocabularies to organize and standardize this information and avoid the errors of annotation due in particular to this deluge of information coming from high-speed technologies.

In the next sections, we present methods and techniques using (i) data mining for identification of genes co-expressed in an analysis of expression data. (ii) Ontology (Gene Ontology (GO)) for data annotation and (iii) approaches that combine data mining and ontologies for functional analysis of gene expression data.

3. FUNCTIONAL GENE EXPRESSION DATA ANALYSIS USING DATA MINING

To answer the questions of biologists such as: are there clusters according to the genes expression profiles? What distinguishes these samples, these genes? Can we predict clusters, classifications? Data mining methods have been used to classify, aggregate and visualize these expression data.

Data mining is a process that is used to search through a large amount of data in order to find useful information. Several data mining methodologies have been proposed to analyze large amounts of gene expression data. Most of these techniques can be broadly classified as cluster analysis and classification techniques. These techniques have been widely used to identify patterns expressions and co-expressed genes and to construct models able of predicting the behavior of genes. In this paper, we focus on clustering, classification and association rule.

3.1. CLUSTERING TECHNIQUES

Clustering has for objective to describe data independent of any a priori knowledge and to reduce the amount of data by categorizing or grouping similar data items together. To categories genes with similar functionality, various clustering methods are used:

- Hierarchical methods like agglomerative hierarchical clustering (AHC)
- Partitioned methods like K-means and C-fuzzy means,
- Model-based methods like self-organizing map (SOM)

Several works are considered to be the pioneers in this field [19, 20, 21]. Clustering was used on pharmacovigilance data [22] and in the diagnosis of cancer [23]. Many comparative studies have been conducted to determine the most efficient clustering algorithm [24, 25, 26, 27] but currently, no consensus is established.

K-means method is used in various applications such as time-series yeast gene expression analysis [28] and the classification breast cancer subtypes [29]. However, in the real nature of biological data, a gene may be involved in several biological processes at once. Hence the use of the Fuzzy C-Means method [30] to give the possibility to a gene belonging to more than one expression profile at a time. This method has been improved for a better classification of the expression data studied. For example, in medical diagnostic, the Possibilistic Fuzzy C-Means (PFCM) clustering technique performed well (with accuracy value equal to 78,13%) for classifying the patients into different clusters of thyroid diseases, compared to Fuzzy C-means (with accuracy value equal to 66,97%) [31]. The same observation was confirmed on lung diseases. The PFCM clustering technique gives the highest accuracy 94.95% compared with k-means and fuzzy C-Means clustering algorithm [32].

To conclude, firstly, we can say that clustering can work well when there is already a wealth of knowledge about the pathway in question, but it works less well when this knowledge is sparse [33]. And secondly, several clustering algorithms have been proposed to analyze gene expression data. In general, there are no best clustering methods. They focus on models and characteristics of various data. Table 1, shows a comparison of these techniques.

Table 1. Clustering methods comparison

Method	Principle	Example of use	Advantages	Disadvantages
K-Means	Decomposes the data set into a set of disjoint clusters; identifies subsets of genes with similar behaviour.	Used by [34] to determine the expression profile during seven periods of the cell cycle in yeast.	Relatively efficient Easy implementation Allows to obtain a mean profile for each class Well suited to large data sets.	Need to specify K, the number of clusters in advance Sensitive to noisy data and outliers Sensitive to start point Genes are forced to belong only to 1 cluster may not converge.
Hierarchical clustering	Proceeds successively by either merging smaller clusters into larger ones, or by splitting larger clusters: it offers an intuitive visual distribution of the data.	Used by [35] to classify and visualize dataset resulting from a proteomic analysis on species of pathogenic bacteria food-derived: <i>Listeria</i> monocytogenes and also used with the proteomics dataset to identify genes differentially expressed on sarcopenia in accordance with rat age.	Does not require the number of clusters to be known in advance No input parameters (besides the choice of the similarity) Computes a complete hierarchy of clusters.	Not scale well: runtime No explicit clusters No automatic discovering of optimal clusters.
Self organizing maps (SOM)	Partitioning experiments genes into a known number groups by association to nodes.	Used by [36] to find groups of genes primarily involved in the differentiation mechanisms of enterocytes.	The position of the groups space reflects the degree similarity between data. Data projected in a same neighborhood have close profiles expression Insensitive to missing values.	Need to specify the number of expected groups The results depend on the chosen distance.
Fuzzy C-mean	Identifies genes pertaining to different regulatory clusters.	In [37], it provides a more interesting distribution of gene clusters compared to "ordinary" clustering methods when tested with melanoma and leukemia dataset.	Each gene can belong to multiple clusters.	No "natural" visualization of the data "Outlier" genes forced to belong to some cluster.

3.2. CLASSIFICATION TECHNIQUES

Classification employs a set of pre-classified data (training set) to develop a model that can classify the population of records at large. Among the most used methods are distinguished:

- **K-Nearest Neighbors (kNN):** this method is very requested by biologists for its simplicity of interpretation. The classifier searches the k nearest neighbors of an unknown sample based on a distance measure. The most common metric used in Bioinformatics is the absolute Pearson coefficient. For clinical end points and controls from breast cancer, neuroblastoma, and multiple myeloma, authors in [38] generated 463,320 kNN models by varying feature ranking method, a number of features, distance metric, a number of neighbors, vote weighting and decision threshold. They identified factors that contribute to the MAQC-II project performance variation.
- **Support Vector Machines (SVM):** its principle is to search a hyperplane of optimal separation between two classes of sample space characteristics. This method was applied to the tumor classification from biochips. The SVM [39] or SVM combined with other techniques such as LDA [40] discriminate against non-linearly separable data and some of these approaches offer the possibility to define several classes. Other works have applied SVM with MI (Mutual Information) for the classification of colon cancer and Lymphoma [41]. Others authors have proposed SVM with GA (Genetic algorithm) for also the classification of cancer. This hybrid method is tested on two gene expression of Tumor datasets. The experimental results indicate that this method is efficient, robust and provides important information for cancer diagnosis in the future [42]. But the disadvantage of SVM technique is to find the optimal separator border, from a set of learning in order to deal with cases where the data are not linearly separable. Another inconvenient is the principle of an SVM is only applied to a problem with two classes. The generalization to multiple classes involves decomposition of the original problem into a set of sub binary problems between a particular class to the aggregation of all of the other classes ("one vs. all") or all classes "one versus one".
- **Decision trees (DT):** it is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. The model built is in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. Some authors, working on leukemia data (acute myeloid leukemia, acute lymphoblastic leukemia and chronic lymphoblastic), compared the performance of DT with the Subgroup Discovery Algorithms and SVM method [43]. According to the authors, DT gives good results. Other authors have combined a meta-heuristic called Particle Swarm Optimization (PSO) with DT (C4.5) and use it for patients' cancer data. They evaluated the performance of the proposed method (PSODT) and compare it with other algorithms of classification, such as SOM, DT (C4.5), neural networks, SVM, and Naive Bayes. The results have shown that PSODT provides better than the others methods [44].
- **Association rule (AR):** An association rule is an expression of the form $X \Rightarrow Y$, where X and Y are sets of items. Association is usually to find frequent item set findings among large data sets. Association Rule algorithm generate rules with confidence values. A study has been done in this regard by defining three different semantics addressing different biological goals: (i) similar expression levels between genes, (ii) similar variations in expression levels of genes, (iii) evolution in levels of gene expression. These rules have been applied to tumors breast and integrated into database software named

MeV of the TIGR environment dedicated to the interpretation of microarray data [45]. The same authors made an improvement by adding rules for building regulatory networks from gene expression data filtered based on the five quality indices: support, confidence, lift, leverage and conviction [46]. In following study [47], authors developed the DAR (Dynamic association rules) algorithm. They modified association rule mining technique by determining the minimum support and minimum confidence based on the concept of a confidence interval and hypothesis testing.

3.3. TOOLS FOR ANALYSIS OF GENE EXPRESSION

There is an important range of tools for the application of classification methods and gene grouping. They include implementation of the main methods of clustering (Hierarchical clustering, k-Means, SOM, etc.), accompanied by various graphical representations (heat maps, three-dimensional chart) facilitating the interpretation of the obtained results [48]. In table 2, we present examples of (software) tools for the classification and grouping of gene expression data.

Software	URL reference
Weka	http://www.cs.waikato.ac.nz/ml/weka/
SAS Artificial	http://www.sas.com/technologies/analytics/data mining/miner/
IBM/SPSS Clementine	http://www.spss.com/software/modeling/modeler-pro/
SVMlight	http://svmlight.joachims.org
LIBSVM	http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/
Cluster and Treeview	http://rana.lbl.gov/EisenSoftware.htm
dChip	http://biosun1.harvard.edu/complab/dchip/
MeV	http://www.tm4.org/mev/
MAGIC Tools	http://www.bio.davidson.edu/projects/magic/ magic.html

Table 2. Tools/Environnement for gene classification and clustering.

4. FUNCTIONAL GENE EXPRESSION DATA ANALYSIS USING ONTOLOGIES

The role of controlled vocabularies or ontologies is to capture the biological concepts describing biological objects such as genomic sequences, genes or gene products. These concepts are derived from publications of the results of the sequencing of genomes and their annotations. Therefore, the use of bio-ontologies becomes essential to deal with the heterogeneity of data and sources. It unifies the different definitions to improve the quality of data and facilitate the sharing and exchange of data.

4.1. BIOLOGICAL AND BIOINFORMATIC ONTOLOGIES

The Gene Ontology (GO) project [49] aims to provide a structured vocabulary to specific biological fields for describing gene products (protein or mRNA) function in the cellular context. It includes three parallel ontologies which are increasingly used by the bioinformatics community: (i) molecular functions, (ii) biological processes and (iii) cellular components. Terms are interconnected by relationship (is a, part of, regulates, etc).

GO is considered as the essential resource for the annotation. It is thus used by many portals (RefSeq, UniProt, KEGG, PDB, TAIR, etc.). Gene Ontology Annotation [50] is a portal dedicated to the data annotation of various interest organisms by using GO. AmiGO [51] is a portal that provides access to GO website, it contains many cross-references with other information systems. The Open Biomedical Ontology project (OBO) [52] is designed to create reference ontologies in biology and biomedical. The platform National Center for Biomedical Ontology (NCBO) [53] develops and maintains a web application called BioPortal which allows researchers to access and use biomedical ontologies.

The Sequence Ontology (SO) project [54] was initially developed by the Gene Ontology Consortium for the definition of the characteristics of sequences that should be used in the annotation. It includes databases of model organisms such as WormBase, FlyBase, Mouse Genome Informatics group, and institutes such as the Sanger Institute and EBI. Other resources such as ArrayExpress at the EBI [55], GEO at NCBI [56], for the filing of data, expression of genes also contain information on the annotation of various organisms.

4.2. ONTOLOGIES OF THE MICROARRAY EXPERIMENTS

A formal description of experiences is extremely important for the organization and execution of experiments in biology. For example, the DNA chips for Microarray Gene Expression Data project (MGED) [57] provide terms to annotate all aspects of an experience of DNA chips of its design with the definition of hybridization, to the preparation of the biological sample and the protocols used for hybridization on the chip and the analysis of data.

The terms MGED are organized in the form of ontology. It was built for the description of biological samples and their use in microarray experiments. This description focuses on biological material (biomaterials) and some treatments used during the experiment, thus, the ontology will be used directly by users to annotate their experiences on microarrays as well as developers of software and databases through structured queries experiences [58].

4.3. SEMANTIC SIMILARITY MEASURES

When biological entities are described using a common ontology, they can be compared by means of their annotations. This type of comparison is called semantic similarity. Several studies have been published describing and evaluating diverse semantic similarity measures. Semantic similarity has become a valuable tool for validating the results drawn from biomedical studies such as gene clustering, gene expression data analysis, prediction and validation of molecular interaction, etc.

The adoption of ontologies for annotation provides a means to compare entities on aspects that would otherwise not be comparable. For instance, if two gene products are annotated within the same schema, they can be compared by comparing the terms with which they are annotated [59]. The Gene Ontology is the main focus of investigation of semantic similarity in molecular biology because comparing gene products at the functional level is crucial for a variety of applications.

The authors in [59] give an interesting survey of semantic similarity measures applied to biomedical ontologies and describe examples of applications to biomedical research. As outlined by the authors, this survey will clarify how biomedical researchers can benefit from semantic similarity measures and help them choose the approach most suitable for their studies.

Several semantic similarity measures have been developed for use with GO. According to the strategies they employ, we distinguish:

4.3.1. MEASURES FOR COMPARING TERM

- **Node-based** [60, 61]: determines the information shared by two terms. A constraint of these measures is that they look only at a single common ancestor despite the fact that GO terms can have several disjoint common ancestors.
- **Edge-based** [62, 63, 64]: use the directed graph topology to compute distances between the terms to compare.
- **Hybrid** [65, 66]: combine different aspects of node-based and edge-based methods.

4.3.2. MEASURES FOR COMPARING GENE PRODUCTS: TO ASSESS THE FUNCTIONAL SIMILARITY BETWEEN GENE PRODUCTS:

It is necessary to compare sets of terms rather than single terms. Several strategies have been proposed, they are grouped into two categories:

- **Pairwise** [59, 67]: measure functional similarity between two gene products by combining the semantic similarities between their terms.
- **Groupwise** [59, 68]: calculates directly similarity by one of three approaches: set, graph, or vector.

An early work was to measure the information content of the terms of the Gene Ontology (GO) [69]. Then it was evaluating some similarity measures such as Resnik, Lin, and Jiang which are node-based measures on these annotated terms. Then, the same authors have investigated semantic similarity measures and their application to ontological annotations of the SWISS-PROT database. They found a correlation between the semantic similarity of GO terms and the sequence similarity of the same genes aligned by BLAST [70].

In [71] controlled vocabularies containing medical concepts such as MeSH and SNOMED-CT were evaluated by a new measure based cross-modified path length feature between the concept nodes [71]. Afterward, measures have been developed to take into account the fact that both terms can have several disjoint common ancestors (DCA) [72].

To overcome the weaknesses of the existing Gene Ontology browsers which use a conventional approach based on keyword matching, a genetic similarity measure is introduced in [73] to find a group of semantically similar Gene Ontology terms. The proposed approach combines semantic similarity measure with a parallel genetic algorithm. The semantic similarity measure is used to compute the similitude strength between the Gene Ontology terms. Then, the parallel genetic algorithm is employed to perform batch retrieval and to accelerate the search in large search space of the Gene Ontology graph.

In [74] authors have attempted to improve existing measures such as the Wu-Palmer measure by adding metadata by taking into account codes of evidence (codes that specify the quality of the annotation), the types of relationships between the GO terms deriving the metabolic pathways of different organisms (regulates, positively regulates, negatively regulates) and the qualifier NOT. This measure was applied to the metabolic pathways between species: human, mouse and the chicken [74].

However, although the Gene ontology, which is the reference for describing biological objects such as genome sequence, genes or gene products, it has only a static view of these biological objects and does not allow visualization that could express these concepts in space and time.

Hence a combination of data mining to group similar expression profiles (static or temporal) and ontologies as additional annotation resources is desirable for the functional analysis of genes.

5. FUNCTIONAL GENE EXPRESSION DATA ANALYSIS COMBINING ONTOLOGY AND BY DATA MINING

Generally, the data analysis of expression takes place in two main steps: (i) identification of the groups of genes co-expressed, for example, by using clustering algorithms (ii) functional analysis of these groups by using a controlled vocabulary such as the Gene Ontology (GO).

The following work [75] associates the first step to the second one. A transversal approach was developed based on the parallel grouping of the genes according to the biological annotations (vocabulary Gene Ontology), medical (UMLS terminology), genomic (characteristics of sequences) and experimental results (expression data). This approach has proved to be as powerful as a classical approach functioning in two phases. Others authors have suggested an approach based on fuzzy modelisation of differential expression profiles combined with data from GO, KEGG, and Pfam [76]. An improvement of this approach was added by the same author by using the Formal Concepts Analysis method in upstream to get genes that have same expression profiles and same functional 'behaviour', and in downstream, it visualizes the results by Lattice [77]. Another work proposed a Bayesian join modelling approach for the analysis of expression data in parallel steps, which incorporates the information of functional annotations into expression data analysis of experimental data obtained for *Escherichia coli*. Simulation experiments show improved results for the integrated approach in both differentially expressed genes and altered genes sets when compared to conventional methods [78]. In this following study, the authors combined GO terms with gene expression data by using affinity propagation clustering algorithm. Next, the neighbourhood rough set was used to select representative genes for each cluster. In the end, a similarity distance was used as heuristic information and quantitative analysis was applied to get the proper biological similarity weight for each dataset for the classification of plant stress response [79].

6. DISCUSSION AND CONCLUSION

This article outlines various methods used in the functional analysis of gene expression data. Bioinformatics' approaches were the first to contribute to a fine analysis of genes (at the level of the sequence), their expressions and their regulation, as well as, the discovery of mutations or changes in the origin of diseases. Analysis deals with the identification of similar regions between several sequences, prediction of the structural properties of genes, etc. In addition to this, data mining methods, besides their diversity, appeared like a simple and obvious solution for determining expression profiles and the grouping or classification of genes with similar behaviour. However, to ensure a complete analysis, we must give an annotation and a meaning to the results. That is to say bring semantics that could be achieved through controlled vocabularies such as GO and other sources of knowledge such as UniProt, KEGG, etc. Consequently, for better representation of co-expressed genes groups and a more "relevant" integration of genomic data supporting researchers in their experiments, recent works has been realized with both approaches. As perspective, it would be interesting to do inter-species annotation on plants such as *Solanum lycopersicum* (Tomato) because it contains a lot of anti-oxidants which protects from the aging and certain cancers or on *Medicago truncatula* for its fixation of nitrogen in the soil with some model plants like *Arabidopsis thaliana*. The approach which will be used is the third one which employs data mining and ontologies for functional analysis of the expression data by accessing profile data of expression and annotation via NCBI GEO, ArrayExpress sequence databases, using the Gene Ontology (GO) and Plant Ontology (PO) which includes terms of growth and stages of development of the plant and terms on the morphological and anatomical structures

(tissues and cell types) of plants. The study will be on the aspect of space-time of terms by using Gene Ontology Annotation (GOA) as a resource.

REFERENCES

- [1] Fyad, H. Barigou, F & Bouamrane, K. (2016), "Computational methods for functional analysis of gene expression", Second International Conference on Advances in Computer Science and Information Technology (ACSTY 2016), Chennai, India, November 26~27, 2016, Vol. 6, No.14, pp. 1– 14. DOI : <http://dx.doi.org/10.5121/csit.2016.61401>.
- [2] Borah, P. (2011). "Biological databases with emphasis on biodiversity and conservation".
- [3] Médigue, C. & Moszer, I. (2007). "Annotation, comparison and databases for hundreds of bacterial genomes". *Research in Microbiology*, Vol. 158, No. 10, pp. 724-736. DOI: <http://doi.org/10.1016/j.resmic.2007.09.009>.
- [4] Yandell, M. & Ence, D. (2012). "A beginner's guide to eukaryotic genome annotation". *Nature Reviews Genetics*, Vol. 13, No. 5, pp. 329-342. DOI: <http://dx.doi.org/10.1038/nrg3174>.
- [5] Aharoni, A. & Vorst, O. (2002). "DNA microarrays for functional plant genomics". *Plant Molecular Biology*, Vol. 48, pp. 99-118. DOI: <http://dx.doi.org/10.1023/A:1013734019946>.
- [6] Rensink, W. A. & Buell, C. R. (2005). "Microarray expression profiling resources for plant genomics". *Trends in plant science*, Vol. 10, No. 12, pp. 603-609. DOI: <http://dx.doi.org/10.1016/j.tplants.2005.10.003>.
- [7] Grasso, L. C. Maindonald, J. Rudd, S. Hayward, D. C. Saint, R. Miller, D. J. & Ball, E. E. (2008). "Microarray analysis identifies candidate genes for key roles in coral development". *BMC genomics*, Vol. 9, No. 1, pp.1-18. DOI: <http://dx.doi.org/10.1186/1471-2164-9-540>.
- [8] Guenin, L. Raharijaona, M. Houlgatte, R. & Baba-Aissa, F. (2010). "Expression profiling of prospero in the Drosophila larval chemosensory organ: Between growth and outgrowth". *BMC genomics*, Vol. 11, No. 1, pp. 1-15. DOI: <http://dx.doi.org/10.1186/1471-2164-11-47>.
- [9] Sharov, A.A. Piao, Y. Ko, M.S. (2010). "Gene expression profiling of mouse embryos with microarrays", *Methods in Enzymology*, Vol. 477, pp. 511–541. DOI : [https://dx.doi.org/10.1016/S0076-6879\(10\)77025-7](https://dx.doi.org/10.1016/S0076-6879(10)77025-7).
- [10] Govindarajan, R. Duraiyan, J. Kaliyappan, K. & Palanisamy, M. (2012). "Microarray and its applications", *Journal of Pharmacy & Bioallied Sciences*, Vol. 4, No. 6, pp. 310-312. DOI: <http://doi.org/10.4103/0975-7406.100283>.
- [11] Médigue, C. Bocs, S. Labarre, L. Mathé, C. Vallenet, D. (2002). "The annotation in silico of genome sequences", *MEDECINE/SCIENCES*, Vol. 18, pp. 237-250. [original reference in French].
- [12] Xu, D. (2009). "Computational methods for protein sequence comparison and search", *Current protocols in protein science*, pp. 2.1. 1-2.1. 27. DOI: <http://doi.org/10.1002/0471140864.ps0201s56>.
- [13] Abhilash, C B. & Rohitaksha, K. (2014). "A Comparative Study on Global and Local Alignment Algorithm Methods", *International Journal of Emerging Technology and Advanced Engineering*, Vol 4, No. 1, pp. 1-10. ISSN 2250-2459.
- [14] Needleman, S. B. & Wunsch, C. D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins", *Journal of molecular biology*, Vol. 48, No 3, pp. 443-453. DOI: [https://doi.org/10.1016/0022-2836\(70\)90057-4](https://doi.org/10.1016/0022-2836(70)90057-4).
- [15] Smith, T. F. & Waterman, M. S. (1981). "Identification of common molecular subsequences", *Journal of molecular biology*, Vol. 147, No 1, pp. 195-197. DOI: [https://doi.org/10.1016/0022-2836\(81\)90087-5](https://doi.org/10.1016/0022-2836(81)90087-5).
- [16] Nigro, C. (2009). "Bioinformatics: Sequence Alignment". In *Conference Proceedings*.
- [17] Meghna Mathur & Geetika. (2013). "Multiple Sequence Alignment Using MATLAB", *International Journal of Information and Computation Technology*, Vol. 3, No. 6, pp. 497-504.
- [18] Brejová, B. Vinar, T. & Li, M. (2003). "Pattern Discovery: Methods and Software". In *introduction to Bioinformatics*, chapter 29, pp. 491-522, Humana Press.

- [19] Eisen, M. B. Spellman, P. T. Brown, P. O. & Botstein, D. (1998). "Cluster analysis and display of genome-wide expression patterns", In Proceedings of the National Academy of Sciences, Vol. 95, No. 25, pp. 14863-14868.
- [20] Tamayo, P. Slonim, D. Mesirov, J. Zhu, Q. Kitareewan, S. Dmitrovsky, E. & al., (1999). "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation". In Proceedings of the National Academy of Sciences, Vol. 96, No. 6, pp.2907-2912. DOI: <https://doi.org/10.1073/pnas.96.6.2907>.
- [21] Pan, W. Lin, J. & Le, C. T. (2002). "Model-based cluster analysis of microarray gene-expression data", Genome Biology, Vol. 3, Research0009. DOI: <http://dx.doi.org/10.1186/gb-2002-3-2-research0009>.
- [22] Shannon, W. Culverhouse, R. & Duncan, J. (2003). "Analyzing microarray data using cluster analysis", Pharmacogenomics, Vol. 4, No. 1, pp.41–52. DOI: <http://dx.doi.org/10.1517/phgs.4.1.41.22581>.
- [23] Smolkin, M. & Ghosh, D. (2003). "Cluster stability scores for microarray data in cancer studies", BMC bioinformatics, Vol. 4, No. 1, pp. 1-7. DOI: <http://dx.doi.org/10.1186/1471-2105-4-36>.
- [24] Yeung, K. Y. Haynor, D. R. & Ruzzo, W. L. (2001). "Validating clustering for gene expression data", Bioinformatics, Vol. 17, pp. 309-318. DOI: <http://dx.doi.org/10.1093/bioinformatics/17.4.309>
- [25] Dudoit, S. & Fridlyand, J. (2002), "A prediction-based resampling method for estimating the number of clusters in a dataset", Genome biology, Vol. 3, No. 7, pp. RESEARCH0036-1-RESEARCH0036-21. DOI: <http://dx.doi.org/10.1186/gb-2002-3-7-research0036>.
- [26] Romualdi, C. Campanaro, S. Campagna, D. Celegato, B. Cannata, N. Toppo, S. Valle, G. Lanfranchi, G. (2003). "Pattern recognition in gene expression profiling using DNA array: a comparative study of different statistical methods applied to cancer classification", Human Molecular Genetics, Vol. 12, No. 8, pp. 823-836. DOI: <http://dx.doi.org/10.1093/hmg/ddg093>.
- [27] Priness, I. Maimon, O. & Ben-Gal, I. (2007). "Evaluation of gene-expression clustering via mutual information distance measure", BMC Bioinformatics, Vol. 8, No. 1, pp. 1-12. DOI: <http://dx.doi.org/10.1186/1471-2105-8-111>.
- [28] Tavazoie, S. Hughes, J. D. Campbell, M. J. Cho, R. J. & Church, G. M. (1999). "Systematic determination of genetic network architecture", Nature genetics, Vol. 22, No. 3, pp. 281-285. DOI: <http://dx.doi.org/10.1038/10343M3>.
- [29] Masuda, H. Baggerly, K. A. Wang, Y. Zhang, Y. Gonzalez-Angulo, A.M. & al., (2013). "Differential response to neoadjuvant chemotherapy among 7 triple-negative breast cancer molecular subtypes", Clinical Cancer Research, Vol 19, No.19, pp.5533-5540. DOI: <http://dx.doi.org/10.1158/1078-0432.CCR-13-0799>.
- [30] Maji, P. & Paul, S. "Clustering Rough Sets Fuzzy Sets Microarray", (2012). Book "Perception and Machine Intelligence", Vol. 7143, pp. 203-210, Springer.
- [31] Simhachalam, B. & Ganesan, G. (2014). "Possibilistic fuzzy c-means clustering on medical diagnostic systems". In Contemporary Computing and Informatics (IC3I), 2014 International Conference on (pp. 1125-1129). DOI: <http://dx.doi.org/10.1109/IC3I.2014.7019729>.
- [32] Scaria, T. Stephen, G. & Mathew, J. (2016). "Gene Expression Data Analysis using Fuzzy C-means Clustering Technique", Gene Expression, Vol. 135, No.8, pp. 33-36. DOI: <http://dx.doi.org/10.5120/ijca2016908470>
- [33] Fiehn, O. Kloska, S. & Altmann, T. (2001). "Integrated studies on plant biology using multiparallel techniques", Current Opinion in Biotechnology, Vol. 12, No.1, pp.82-86. DOI: [http://dx.doi.org/10.1016/S0958-1669\(00\)00165-8](http://dx.doi.org/10.1016/S0958-1669(00)00165-8).
- [34] Anusuya, S. Bhanu, D. N. U. & Kasthuri, E. (2015). "yeast gene expression analysis using k means and FCM", International Journal of Pharma and Bio Sciences, Vol. 6, No.3: B, pp. 395 – 400.
- [35] Meunier, B. Dumas, E. Piec, I. Bechet, D. Hebraud, M. & Hocquette, J. F. (2007). "Assessment of Hierarchical Clustering Methodologies for Proteomic Data Mining", Journal of Proteome Research, Vol. 6, No. 1, pp. 358-366. DOI: <http://dx.doi.org/10.1021/pr060343h>.

- [36] Bédérine-Ferran, H. Le Meur, N. Gicquel, I. Le Cunff, M. Soriano, N. Guisle, I. & al., (2004). "Transcriptome variations in human CaCo-2 cells: a model for enterocyte differentiation and its link to iron absorption", *Genomics*, Vol. 83, No. 5, pp. 772-789.
DOI: <http://dx.doi.org/10.1016/j.ygeno.2003.11.014>.
- [37] Kim, S. Y. Choi, T. M. & Bae, J. S. (2006). "Fuzzy Types Clustering for Microarray Data", *International Journal of Computational Intelligence*, Vol. 2, No. 1, pp. 12-15.
- [38] Parry, R. M. Jones, W. Stokes, T. H. Phan, J. H. Moffitt, R. A. Fang, H. Shi, L. Oberthuer, A. Fischer, M. Tong, W. & Wang, M. D. (2010). "K-Nearest Neighbor Models for Microarray Gene Expression Analysis And Clinical Outcome Prediction", *Pharmacogenomics Journal*, Vol. 10. No.4, pp. 292-309.
DOI: <http://dx.doi.org/10.1038/tpj.2010.56>.
- [39] Li, F. & Yang, Y. (2005). "Analysis of recursive gene selection approaches from microarray data". *Bioinformatics*, Vol. 21, No. 19, pp. 3741-3747.
DOI: <http://dx.doi.org/10.1093/bioinformatics/bti618>.
- [40] Nijima, S. & Kuhara, S. (2006). "Recursive gene selection based on maximum margin criterion: a comparison with SVM-RFE", *Biomedcentral*, Vol.7. pp. 1-18.
DOI: <http://dx.doi.org/doi:10.1186/1471-2105-7-543>.
- [41] Vanitha, C. D. A. Devaraj, D. & Venkatesulu, M. (2015). "Gene Expression Data Classification using Support Vector Machine and Mutual Information-based Gene Selection", *Procedia Computer Science*, Vol. 47, pp. 13-21.
DOI: <http://dx.doi.org/doi:10.1016/j.procs.2015.03.178>.
- [42] Yu, B. Zhang, Y. & Zhao, L. (2015). "Cancer Classification by a Hybrid Method Using Microarray Gene Expression Data", *Journal of Computational and Theoretical Nanoscience*, Vol. 12, No. 10, pp. 3194-3200.
DOI: <https://doi.org/10.1166/jctn.2015.4101>.
- [43] Netto, O. P. Nozawa, S. R. Mitrowsky, R. A. R. Macedo, A. A. & Baranauskas, J. A. (2010). "Applying decision trees to gene expression data from dna microarrays: A leukemia case study". In XXX Congress of the Brazilian Computer Society, X Workshop on Medical Informatics, pp.1-10.
- [44] Chen, K. H. Wang, K. J. Tsai, M. L. Wang, K. M. Adrian, A. M. Cheng, W. & al., (2014). "Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm", *BMC bioinformatics*, Vol. 15, No. 49, pp. 1-10.
DOI: <http://dx.doi.org/10.1186/1471-2105-15-49>.
- [45] Agier, M. Petit, J-M. Chabaud, V. Pradeyrol, C. Bignon, I-J. & Vidal, V. (2004). "Different types of rules for expression of genes database Application to database of mammaire tumor", In XXIIème Congrès INFORSID, pp. 351-367. Biarritz : France. [original reference in French].
- [46] Agier, M. "Different types of rules for the reconstruction of networks of genes from expression data". (2007). *Revue I3 Information Interaction-Intelligence*, numéro hors série, pp. 161-81, Cépaduès Editions. [original reference in French].
- [47] Chen, S. C. Tsai, T. H. Chung, C. H. & Li, W. H. (2015). "Dynamic association rules for gene expression data analysis", *BMC genomics*, Vol. 16, No. 1, pp. 786.
DOI: <http://dx.doi.org/10.1186/s12864-015-1970-x>.
- [48] Selvaraj, S. & Natarajan, J. (2011). "Microarray Data Analysis and Mining Tools", *Biomedical Informatics*, Vol. 6, No. 3, pp. 95-99.
DOI: <http://dx.doi.org/10.6026/97320630006095>.
- [49] Berardini, T.Z. Li, D. Huala, E. Bridges, S. Burgess, S. McCarthy, F. & al., (2010). "The Gene Ontology in 2010: extensions and refinements", *Nucleic Acids Res*, Vol. 38, (Database issue): D331-D335. (cf: <http://www.geneontology.org>).
- [50] Huntley, R. P. Sawford, T. Mutowo-Meullenet, P. Shypitsyna, A. Bonilla, C. Martin, M. J. & O'Donovan, C. (2015). "The GOA database: gene ontology annotation updates for 2015. *Nucleic acids research*", Vol. 43(D1), pp. D1057-D1063. (cf: <http://www.ebi.ac.uk/GOA>).
- [51] Carbon, S. Ireland, A. Mungall, C. J. Shu, S. Marshall, B. Lewis, S. & Web Presence Working Group. (2009). "AmiGO: online access to ontology and annotation data", *Bioinformatics*, Vol. 25, No. 2. pp. 288-289.
DOI: <http://dx.doi.org/10.1093/bioinformatics/btn615>.
- [52] Ghazvinian, A. Noy, N. F. & Musen, M. A. (2011). "How orthogonal are the OBO Foundry ontologies?", *Journal of biomedical semantics*, Vol. 2, No. 2, pp. S2.
DOI: <http://dx.doi.org/10.1186/2041-1480-2-S2-S2>.

- [53] Whetzel, P. L. Noy, N. F. Shah, N. H. Alexander, P. R. Nyulas, C. Tudorache, T. & Musen, M. A. (2011). "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications", *Nucleic Acids Research*, Vol. 39, No. suppl 2 (Web Server issue), pp. W541-W545.
DOI: <https://doi.org/10.1093/nar/gkr469>.
- [54] Eilbeck, K. Lewis, S.E. Mungall, C.J. Yandell, M. Stein, L. Durbin, R. & Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology*, Vol. 6, No. 5, pp. r44.
DOI: <http://dx.doi.org/10.1186/gb-2005-6-5-r44>.
- [55] Parkinson, H. Kapushesky, M. Shojatalab, M. Abeygunawardena, N. Coulson, R. Farne, A. (2007). "ArrayExpress—a public database of microarray experiments and gene expression profiles". *Nucleic acids research*, Vol. 35, No. (suppl 1), pp. D747-D750.
DOI: <http://dx.doi.org/10.1093/nar/gkl995>.
- [56] arrett, T. Troup, D. B. Wilhite, S. E. Ledoux, P. Rudnev, D. Evangelista, C. & Edgar, R. (2007). "NCBI GEO: mining tens of millions of expression profiles—database and tools update", *Nucleic Acids Research*, Vol. 35, No. (suppl 1), pp. D760- D765.
DOI: <http://dx.doi.org/10.1093/nar/gkl887>.
- [57] Guérin, E. Marquet, G. Burgun, A. Loréal, O. Berti-Equille, L. Leser, U. & Moussouni, F. (July 2005). "Integrating and warehousing liver gene expression data and related biomedical resources in GEDAW". In *International Workshop on Data Integration in the Life Sciences*, pp. 158-174. Springer Berlin Heidelberg.
- [58] Griffin, J. L. & Steinbeck, C. (2010). "So what have data standards ever done for us? The view from metabolomics". *Genome Medicine*, Vol. 2, No. 6. pp. 38.
DOI: <http://dx.doi.org/10.1186/gm159>.
- [59] Pesquita, C. Faria, D. Falcao, A. O. Lord, P. & Couto, F. M. (2009). "Semantic similarity in biomedical ontologies". *PLoS Comput Biol*, Vol. 5, No. 7, pp. e1000443.
DOI=<http://dx.doi.org/10.1371/journal.pcbi.1000443>.
- [60] Resnik, P. (1999). "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language". *Journal of Artificial Intelligence* 11, pp. 95–130.
- [61] Lin, D. (July 1998), "An Information-Theoretic Definition of similarity", In : *Proceedings of The Fifteenth International Conference on Machine Learning (ICML'98)*, pp. 296-304.
- [62] Rada, R. Mili, H. Bicknell, E. & Blettner, M. (1989). "Development and application of ametric on semantic nets", *IEEE Transaction on Systems, Man, and Cybernetics*, Vol 19, No. 1, pp.17-30.
- [63] Wu, Z. & Palmer, M. (June 1994). "Verb semantics and lexical selection". In: *Proceedings of The 32nd Annual Meeting of the Associations for Computational Linguistics*, pp. 133–138.
- [64] Hirst, G. & Budanitsky, A. (2005). "Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, Vol. 1, No. 1, pp 87-111.
- [65] Jiang, J. J. & Conrath, D. W. (1997). "Semantic similarity based on corpus statistics and lexical taxonomy", In : *Proceedings of The International Conference on Research in Computational Linguistics*, arXiv preprint [cmp-lg/9709008](https://arxiv.org/abs/cmp-lg/9709008). Taiwan.
- [66] Leacock, C. & Chodorow, M. (1998). "Combining Local Context and WordNet Similarity for Word Sense Identification. In *WordNet: An Electronic Lexical Database*", Vol.49, No. 2, pp. 265-283.
- [67] Chagoyen, M. Carazo, J. & Pascual-Montano, A. (2008). "Pairwise similarity scores using functional annotations: review and comparison". In *8th Spanish Symposium on Bioinformatics and Computational Biology*: 2008.
- [68] Teng, Z. Guo, M. Liu, X. Dai, Q. Wang, C. & Xuan, P. (2013). "Measuring gene functional similarity based on group-wise comparison of GO terms". *Bioinformatics*, pp. 1-9.
DOI: <http://dx.doi.org/10.1093/bioinformatics/btt160>.
- [69] Lord, P. W. Stevens, R. D. Brass, A. & Goble, C. A. (October 2003). "Semantic similarity measures as tools for exploring the gene ontology". In *Pacific Symposium on Biocomputing*, Vol. 8, No. 4, pp. 601-612.
- [70] Lord, P. W. Stevens, R. D. Brass, A. & Goble, C. A. (2003). "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation". *Bioinformatics*, Vol. 19, No. 10, pp.1275-1283.
DOI: <http://dx.doi.org/10.1093/bioinformatics/btg153>.

- [71] Al-Mubaid, H. & H.A. Nguyen. (August 2006). "A cluster-based approach for semantic similarity in the biomedical domain", In Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE, pp. 2713-2717.
- [72] Couto, F. M. Silva, M. J. & Coutinho, P. M. (October 2005). "Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors", In : Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 343-344.
DOI: <http://dx.doi.org/10.1145/1099554.1099658>.
- [73] Othman, R. M. Deris, S. & Illias, R. M. 2008. "A genetic similarity algorithm for searching the Gene Ontology terms and annotating anonymous protein sequences", Journal of biomedical informatics, Vol. 41, No.1, pp. 65-81.
DOI: <http://dx.doi.org/10.1016/j.jbi.2007.05.010>.
- [74] Bettembourg, C. Diot, C. & Dameron, O. (2014). "Semantic particularity measure for functional characterization of gene sets using gene ontology". PloS one, Vol. 9, No.1, pp. e86525.
DOI: <http://dx.doi.org/10.1371/journal.pone.0086525>.
- [75] Chabalier, J. Mosser, J. & Burgun, A. (2007). "A transversal approach to predict gene product networks from ontology-based similarity". BMC bioinformatics, Vol. 8, No. 235, pp 1-12.
DOI: <http://dx.doi.org/10.1186/1471-2105-8-235>.
- [76] Devignes, M. D. Benabderrahmane, S. Smaïl-Tabbone, M. Napoli, A. & Poch, O. (2012). "Functional classification of genes using semantic distance and fuzzy clustering approach: evaluation with reference sets and overlap analysis". International journal of computational biology and drug design, Vol. 5, No. 3-4, pp. 245-260.
DOI: <http://dx.doi.org/10.1504/IJCBDD.2012.049207>.
- [77] Benabderrahmane, S. "Formal Concept Analysis and Knowledge Integration for Highlighting Statistically Enriched Functions from Microarrays Data". (2014). International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO 2014, Granada, Spain, Granada.
- [78] Wang, X. Chen, M. Khodursky, A. B. and Xiao, G. (2012). "Bayesian joint analysis of gene expression data and gene functional annotations". Statistics in Biosciences, Vol. 4, No. 2, pp. 300-318.
DOI: <http://dx.doi.org/10.1007/s12561-012-9065-6>.
- [79] Meng, J. Li, R. and Luan, Y. (2015). "Classification by integrating plant stress response gene expression data with biological knowledge". Mathematical Biosciences, Vol. 266, pp. 65-72.
DOI: [10.1016/j.mbs.2015.06.005](https://doi.org/10.1016/j.mbs.2015.06.005).