FUZZY *C*-MEANS CLUSTERING BY INCORPORATING BIOLOGICAL KNOWLEDGE AND MULTI-STAGE FILTERING TO IMPROVE GENE FUNCTION PREDICTION

SHAHREEN KASIM

UNIVERSITI TEKNOLOGI MALAYSIA

FUZZY *C*-MEANS CLUSTERING BY INCORPORATING BIOLOGICAL KNOWLEDGE AND MULTI-STAGE FILTERING TO IMPROVE GENE FUNCTION PREDICTION

SHAHREEN KASIM

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

NOVEMBER 2011

Bismillahirrahmanirrahim. Dengan nama Allah Yang Maha Pemurah Lagi. Maha Mengasihani.

Sekalung penghargaan buat:

Yang paling dirindui Allahyarham ayahanda, **Haji Kasim Bin Haji Abdullah**: Terima kasih di atas segala didikanmu, semoga Allah menempatkan ayahanda di tempat orang-orang yang soleh.

Khas buat ibunda, **Hajah Zainon Binti Haji Mohd. Said**: Terima kasih di atas doamu ibu, doa yang sangat bernilai bagiku. Restu dan kata-kata semangat ibunda memberi laluan lurus dalam hidup kami.

Yang terutama, suami tercinta, sahabat karib, kekasih awal dan akhir **Muhammad Edzuan Bin Zainodin**: Terima kasih kerana menjadi suami yang penyayang, penyabar, pendorong, peneman setia, bersama-sama susah dan senang mengharungi perjalanan pengajian ini.

Penyambung warisan, anakanda **Muhammad Eiskandar**, **Sara Arissa**, dan **Muhammad Eirshad**: Kamulah penghibur dan penguat semangat ibu.

Yang diingati keluarga yang sentiasa memberi sokongan: Bonda dan ayahanda mertua, along, ngah, uda, Allahyarham Sabariah, dan ipar-duai.

ACKNOWLEDGEMENTS

This thesis would not have been possible without the help and support of many people. Firstly I would like to thank my supervisors, Prof. Dr. Safaai Deris and Dr. Muhamad Razib Othman, for their excellent supervision, knowledge, belief, patience, and interest in the work, encouragement, and motivation throughout these three years. I am also very grateful to Prof. Dr. Richard Spears for his hours of patient and detailed proofreading. My thanks also go to everyone who has provided support or advice in one way or another, including people in Artificial Intelligence and Bioinformatics Research Group (AIBIG) and Laboratory of Computational Intelligence and Biotechnology (LCIB). My greatest thanks should also credit to all staff at Faculty of Computer Science and Information Technology (FSKTM), Universiti Tun Hussien Onn Malaysia (UTHM) for their understanding and support.

Most importantly, none of this would have been possible without the love and patience from my family and has been a constant source of love, concern, support and strength all these years: my husband, my sons, and my daughter for their understanding and sacrifices. I thank my mother for her belief and sacrifice in making me what I am today.

ABSTRACT

Gene expression is a process by which information from a gene is used in the synthesis of a functional gene product. Comprehensive studies of gene expression are useful for predicting gene functions, which includes predicting annotations for unknown gene functions. However, there are several issues that need to be addressed in gene function prediction, namely: solving multiple fuzzy clusters using biological knowledge and biological annotations in some existing databases. This includes, handling the high level expression and low level expression values. Therefore, this research was aimed at clustering gene expressions by incorporating biological knowledge in order to handle these issues. The basic Fuzzy *c*-Means (FCM) algorithm was introduced to address multiple fuzzy clusters in gene expression. Clustering Functional Annotation (CluFA) was developed to deal with insufficient knowledge via incorporating Gene Ontology (GO) datasets and multiple functional annotation databases. The GO datasets were used to determine number of clusters as well as clusters for genes. Meanwhile, the evidence codes in functional annotation databases were used to compute the strength of the association between data element and a particular cluster. The multi stage filtering-CluFA (msf-CluFA) was implemented by conducting filtering stages and applying an enhanced *apriori* algorithm in order to handle the high level expression and low level expression values. The performance of the proposed method was evaluated in terms of compactness and separation, consistency, and accuracy, using Eisen and Gasch datasets. Biological validation was also used to validate the gene function prediction, by cross checking them with the most recent annotation database. The results show that the proposed computational method achieved better results compared with other methods such as GOFuzzy, FuzzyK, and FuzzySOM in predicting unknown gene function.

ABSTRAK

Ekspresi gen merupakan satu proses di mana maklumat mengenai gen digunakan untuk mensintesis fungsi sesuatu produk gen. Kajian menyeluruh terhadap ekspresi gen adalah penting untuk meramalkan anotasi bagi fungsi gen yang belum dikenalpasti. Walau bagaimanapun, terdapat beberapa isu dalam peramalan fungsi gen yang perlu ditangani, antaranya ialah: menyelesaikan pelbagai kelompok kabur menggunakan pengetahuan dan anotasi biologi di dalam pangkalan data sedia ada. Ini juga termasuk mengatasi nilai ekspresi gen yang rendah dan tinggi. Oleh itu, kajian ini telah dilaksanakan bertujuan untuk mengelompokkan ekspresi gen dengan menggabungkan pengetahuan biologi di dalam menangani isu-isu tersebut. Algoritma *fuzzy c-means* (FCM) asas telah diperkenalkan untuk menyelesaikan pelbagai kelompok kabur dalam ekspresi gen. Seterusnya, Anotasi Kefungsian Pengelompokan (CluFA) pula telah dibangunkan bagi mengatasi isu ketidakcukupan pengetahuan biologi melalui penggunaan Ontologi Gen (GO) dan beberapa pangkalan data berkaitan kefungsian anotasi. Data GO telah digunakan untuk mengenalpasti bilangan kelompok dan menentukan kelompok bagi gen-gen. Sementara itu, kod bukti di dalam pangkalan data telah digunakan untuk mengira kekuatan pertalian di antara elemen data dengan kelompok tersebut. Tapisan Pelbagai Peringkat dengan Anotasi Kefungsian Pengelompokan (msf-CluFA) telah dilaksanakan melalui pelaksanaan beberapa penyaringan menggunakan algoritma *apriori* yang dikembangkan bagi mengatasi nilai ekspresi gen yang rendah dan tinggi. Prestasi kajian telah dinilai menggunakan beberapa ukuran seperti kepadatan dan pemisahan, konsisten serta ketepatan terhadap dua data pengujian iaitu data Eisen dan Gasch. Pengesahan biologi juga telah dijalankan untuk menentusahkan ramalan fungsi gen melalui semakan anotasi data yang terkini. Hasil menunjukkan kajian yang dijalankan mencapai keputusan yang lebih baik berbanding kaedah-kaedah lain seperti *GOFuzzy*, *FuzzyK* dan *FuzzySOM* di dalam meramalkan fungsi gen yang masih belum dikenalpasti.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIDS	-	Acquired Immune Deficiency Syndrome
BGED	-	Brain Gene Expression Database
BicAT	-	Biclustering Analysis Toolbox
BioGRID	-	Biological General Repository
C	-	Cellular Component
CLIFF	-	Clustering via Iterative Feature Filtering
CluFA	-	Clustering Functional Annotation
CPU	-	Central Processing Unit
CS	-	Compactness and Separation
CT	-	Consistency
CTWC	-	Couple Two Way Clustering
DAG	-	Directed Acyclic Graph
DHC	-	Density-based Hierarchical Clustering
DIP	-	Database of Interacting Proteins
DNA	-	Deoxyribonucleic Acid
EST	-	Expressed Sequence Tag
F	-	Molecular Function
FCM	-	Fuzzy *c*-means
FLOC	-	Flexible Overlapped Clustering
GA	-	Genetic Algorithm
GO	-	Gene Ontology
GXD	-	Gene Expression Database
HD	-	Hypergeometric Distribution
HPRD	-	Human Protein Reference Database

IC	-	Inferred by Curator
IDA	-	Inferred from Direct Assay
IEA	-	Inferred from Electronic Annotation
IEP	-	Inferred from Expression Pattern
IGI	-	Inferred from Genetic Interaction
IMP	-	Inferred from Mutant Phenotype
IPI	-	Inferred from Physical Interaction
ISS	-	Inferred from Sequence or Structural Similarity
KEGG	-	Kyoto Encyclopedia of Genes and Genomes
MaizeGDB	-	Maize Genetics and Genomics Database
MeSH	-	Medical Subject Headings
MGD	-	Mouse Genome Database
MGI	-	Mouse Genome Informatics
MINT	-	Molecular Interaction Database
MIPS	-	The Yeast Database at Munich Information Centre for Protein Sequences
mRNA	-	Messenger of Ribonucleic Acid
msf-CluFA	-	Multi Stage Filtering-Clustering Functional Annotation
NAS	-	Non-traceable Author Statement
NCBI	-	National Center for Biotechnology Information
ND	-	No Biological Data Available
OBO	-	Open Biomedical Ontologies
P	-	Biological Process
PCA	-	Principal Component Analysis
PCD	-	Programmed Cell Death
RAM	-	Random Access Memory
RCA	-	Inferred from Reviewed Computational Analysis
RED	-	Rice Expression Database
RNA	-	Ribonucleic Acid
SGD	-	*Saccharomyces* Genome Database
SGMD	-	Soybean Genomics and Microarray Database
SOM	-	Self-Organizing Map
SRBCT	-	Small Round Blue Cell Tumors

TAS	-	Traceable Author Statement
TRIPLES	-	Transposon-Insertion Phenotypes, Localization and Expression in *S. cerevisiae*
YPD	-	Yeast Protein Database

CHAPTER 1

INTRODUCTION

1.1 Background

The evolution of Deoxyribonucleic Acid (DNA) microarray has lead to the study of variations in genes on a genome-wide scale. The relative abundance of the messenger of Ribonucleic Acid (mRNA) of a gene under a specific experimental condition is called expression level of a gene. The expression level of a large number of genes of an organism under various experimental conditions can be arranged in a data matrix, also known as gene expression data matrix, where rows represent genes and columns represent conditions. Currently, the gene expression dataset consists of thousand of genes that have encouraged numerous experiments such as those related to gene function prediction. The various methods of gene function prediction have quite naturally led to varying results. The benefit of implementing the gene function prediction has been widely known in the area of health (Hu *et al.*, 2007; Wassenaar *et al.*, 2007) and biotechnology (Pandit *et al.*, 2010; Arakaki *et al.*, 2009).

Due to the large size of genes and the complexity of biological networks, a computational method is needed to analyze the gene expression dataset. One of the objectives of gene expression dataset analysis is to group genes according to their expression under a variety of conditions. It is known that genes in the same group are similar while genes in different groups are dissimilar. This grouping method is essential in the process of gene function prediction. The grouping method can utilize biological knowledge in order to guide the process and thus provide a limited form of supervision. However, only a few studies have applied biological knowledge, for example Gene Ontology (GO: as established by the Gene Ontology Consortium, 2000), for guidance. The GO has been used in many gene expression analyses and defines the results in GO terms with GO annotations. The GO term is a standard terminology to describe features of gene product. Meanwhile, the GO annotation reflects connection between GO terms and biological types that are represented in the GO using GO evidence. The GO terms are organized as Directed Acyclic Graph (DAG). The nodes represent GO terms and arcs represent relationships between the GO terms. In GO, the terms have been categorized into cellular component, biological process, and molecular function. GO annotations are derived from various functional annotation databases derived from various species; for example *S. cerevisiae* (SGD: *Saccharomyces* Genome Database; http://www.yeastgenome.org/), *D. melanogaster* (FlyBase; http://flybase.org/), and *M. musculus* (MGI: Mouse Genome Informatics; http://www.informatics.jax.org/). Incorporating GO information of gene expression dataset in the clustering process increase the potential of identifying similarities in biological expression. This identification has the capacity to describe the function of an unknown gene function by predicting the GO annotation for the gene.

The high dimensionality in the gene expression dataset makes clustering a challenging task. This is due to the existence of extraneous attributes that inhibit the determination of the existence of clusters. The traditional clustering algorithms use all the attributes in the data to calculate the distances between two genes. Among other drawbacks in the traditional clustering algorithms are their difficulty in determining the number of clusters, random initialization of genes, and conflicts in gene function domination. In the traditional clustering, the number of clusters require several trial and error attempts on the part of the user. In addition, the random initialization of genes produces unstable results and requires excessive time due to the multiple runs needed for the experiments. Furthermore, conflicts of gene function domination will arise due to multiple functions being assigned to the genes. Therefore, semi-supervised clustering methods are introduced to provide solutions to the above problems. A semi-supervised clustering is usually performed with some information provided to guide the clustering process effectively. There are still loopholes in the semi-supervised clustering method, since the implementations of non-comprehensive biological knowledge with the best identification of multiple functional annotations still remain the main problem to be addressed. Furthermore, with the multiple functions being assigned to a gene, the most dominant functions and the degree to which they may be confidently predicted remain ambiguous.

The following sections in this chapter discuss the challenges involved in producing an accurate gene function prediction. This is followed by a brief discussion of the current efforts among those designing computational approaches to predict gene function. The aims of this research and a summary of its objectives were explained in the next section. Also, the scope and significance of this study is outlined before presenting the overview of the organization of this thesis.

1.2 Challenges for Gene Function Prediction

Application of the gene expression clustering algorithm provides a powerful tool for studying the functional relationships of genes in the biological process. Identifying the optimum cluster in the gene expression dataset represents the basic challenge to gene function prediction. The first challenge in this study is to handle changes of the mRNA during the experiment conducted on the gene expression dataset. The results from the experiments show that positive numbers represent an increase in expression, while negative numbers represent decreases in expression. This situation produces intensive data which leads to the computational challenge. Furthermore, the nature of gene itself allows one gene can belong to one or more (multiple) functions. This introduces to the challenge in characterizing the gene function. This situation happens because biological functions involve the integrated activities of many genes. The same gene may have different functions depending on context, which is in turn be defined partly by the presence of other gene products.

The second challenge is relative to the lack of similar expression profiles in gene expression datasets, thus bring challenges to data quality issues. These circumstances will affect the genes with similar functions as they may not be in the same group. To find similarity among expression profiles in a gene expression dataset, researchers have to utilize information from both aspects: biological and expressional, in order to achieve biological meaning. While most researchers concentrate in similarity of the gene expression dataset, the expression profile which is associated to biological function is also important in predicting the gene function.

The third challenge is related to high level expression and low level expression value in gene expression datasets. When utilizing a high level expression during the gene expression analysis, these values will also reflect a high membership value in multiple groups. These circumstances will degrade the performance thus unable to identify the dominant function. Furthermore, the handling of low level expression during gene expression analysis brings challenge to the researcher. This is due to the belonging of many genes in the same group but having lower membership values. This situation generates a lower degree of confidence to those particular genes with low membership values. Meanwhile, some of the genes are grouped with higher confidence due to their high membership values.

1.3 Current Methods for Gene Function Prediction

Basically, current methods for gene function prediction for gene expression dataset can be divided into two approaches; experimental and computational analyses (the details are presented in Chapter 2):

(i) Experimental gene expression is a method that predicts genes from a physical characterization of a gene or gene product during *in vivo* or *in vitro* analysis. The gene function prediction in the experimental approach is based on direct assay (Perry *et al.,* 2009; Tripathi *et al.,* 2009), genetic interaction (Bugnicourt *et al.,* 2008; Motley *et al.,* 2008), phenotype (Perry *et al.,* 2009; Wanat *et al.,* 2008), physical interaction (Arifuzzaman *et al.,* 2006; Taverna *et al.,* 2006), and expression pattern (Deng *et al.,* 2005; Basrai *et al.,* 1999).

(ii) Computational gene expression analysis is a method that predicts genes from an *in silico* analysis of the gene sequence and/or other data. The gene function prediction in computational analysis approach is based on co-expression (Cai *et al.,* 2010; Oti *et al.,* 2008), sequence (Punta and Ofran, 2008; Deng *et al.,* 2005), phylogenetic profile (Jiang, 2008; Taşan *et al.,* 2008), interaction (Zare *et al.,* 2006; Bhardwaj and Lu, 2005), and gene neighbourhood (Ruan, 2010; Pandey *et al.,* 2009).

1.4 Problem Statement

The solution of the gene function prediction problem was briefly described as follows:

> Given a gene expression dataset, the challenge is to cluster the intensive dataset while characterizing the gene function. In addition, quality of the data needs to be tackled, without degrading the performance analysis, and also handling with the uncertainty degree. At the same time, the computational method must be capable of producing highly compacted clusters with furthest separation (CS), high consistency (CT), and accuracy (precision, recall, and F-measures).

In light of the above statement, a gene expression which incorporates GO knowledge and multiple functional annotation databases extracted from high similarity gene expression will be able to produce optimum cluster resulting in the prediction of gene function. However, in order to realize this, three factors need to be considered. The first factor relates to the ambiguity of gene expression datasets and comes from *ab initio* process which results in inaccurate data. The aim is to group genes that exhibit more than one function due to the nature of genes.

The second factor relates to insufficient knowledge related to the similarity of gene expression profiles. The insufficient biological knowledge has been determined to be a key factor affecting the data quality issues. This biological knowledge is useful in determining the functional relationship between genes in order to characterize their function and to predict unknown gene function. The aim is to produce a systematic and automatic method for predicting gene function by applying GO as underlying biological knowledge together with multiple functional annotation databases which supports multiple annotation databases formats.

The third factor that needs to be considered is the inaccuracy of the results obtained from the gene expression analysis. The inaccuracy is the result of multiple high membership values where in certain situations a gene could belong to more than one function. These high membership values bring imprecise results in identifying their dominant function. The precision is also affected when some other genes of the same function but having a lower membership value produce genes that are not assigned to that function with high confidence.

1.5 Objective of the Study

The goal of this study is to develop computational method to cluster the gene expression with incorporation of biological knowledge in order to predict gene function. Therefore, this study has the following objectives:

(i) To study and evaluate the current methods of gene function prediction in order to understand the domains, data, and processes involved.

(ii) To develop a fuzzy *c*-means algorithm that can handle data ambiguity in order to solve their intensity and redundancy.

(iii) To incorporate GO and multiple functional annotation databases in the fuzzy *c*-means algorithm so that it might handle insufficient knowledge in the solution the data quality issues.

(iv) To improve the algorithm by implementing multi-stage filtering that is able to handle the inaccuracies and consequently solve the degrading performance and the uncertainty degree.

1.6 Scope and Significance of the Study

In this study, we only cover three data types; the GO datasets, functional annotation databases, and testing datasets. The GO datasets are used to form and assign genes into their clusters. This data is obtained from the GO Consortium website: http://www.geneontology.org/GO.downloads.database.shtml. At the same time, the functional annotation databases which are SGD, the Yeast Database at Munich Information Centre for Protein Sequences (MIPS), and Entrez are used to extract the annotation evidence code for the particular genes in order to calculate their membership values. The testing datasets were downloaded from http://titan.biotec.uiuc.edu/clustering/ and http://genome.www.stanford.edu/clustering/ and served as input to test and evaluate the proposed computational method. This research scopes also involves a novel computational method named msf-CluFA, which has been developed to show the capabilities of the proposed semi-supervised clustering of gene expression datasets. The msf-CluFA consists of only four components: fuzzy *c*-means clustering (msf-CluFA-0), achieving dominant cluster (msf-CluFA-1), improving confidence level (msf-CluFA-2), and combination (msf-CluFA-3). In the component of msf-CluFA-0, there are two stages: (i) the preparation of GO datasets, functional annotation databases, and testing datasets and (ii) a fuzzy *c*-means clustering to find the optimal clusters. In combination with the three GO term categories (biological process, molecular function, and cellular component) and the functional annotation databases, the msf-CluFA is able to determine the number of clusters and reduce random initialization. By employing double filtering in msf-CluFA-1 and enhanced the *apriori* algorithm in msf-CluFA-2, our new computational method will be capable determining the dominant clusters and improving the gene's confidence level for lower membership values. This new computational method is also able to predict unknown gene functions. The evaluation measurements of msf-CluFA only cover computational evaluation (compactness and separation, consistency, precision, recall, and F-measure, hypergeometric distribution, *z*-score, and cluster profile) and biological validation (cross check with the latest annotation database).

Gene function prediction is the focus in this research with the goal being to develop a better computational method to get the optimum cluster from which to select informative genes. The significances of this research can be seen in its impact in the areas of cancer informatics and pharmacogenomic. An optimal cluster output can be used in cancer research (McKibbin *et al.,* 2008; Sausville and Holbeck, 2004) with the latest advances in the application of bioinformatics and computational biological toward the discovery of new knowledge in oncology and cancer biology, and toward the clinical translation of that knowledge to increase the efficacy of practicing oncologists, radiologists, and pathologists. The study of clustering gene expression dataset can also lead to the field of pharmacogenomics. Pharmacogenomics represents the union of genomic information to the clinical practice of medicine and thus extends the pharmacogenomic paradigm to drug discovery, for example the research done by Zhou *et al.* (2008) and Young and Winzeler (2005). Further research has also been done by van Baarsen *et al.* (2008), Li *et al.* (2007), and Zhao *et al.* (2007) who also used gene expression datasets as input for their prediction of gene functions relevant in multiple sclerosis, bacteriophages, and pancreatic cancer, all of which now use clustering as the algorithm. Currently, the research in gene function prediction is in high demand where about 40% of the proteins encoded in eukaryotic genomes still have unknown functions (Horan *et al.*, 2008). Therefore, with the immense progress of algorithms, more researches has been conducted using diverse species of gene expression datasets and applying clustering as their algorithm, as in the research of Bradford *et al.* (2010), Zeng *et al.* (2010), Zhang *et al.* (2008), and Brown *et al.* (2006) for gene function prediction. These efforts are believed to be beneficial to the health, human beings, and nature.

1.7 Organization of the Thesis

This thesis is organized into seven chapters. A brief description of the contents for each chapter is given as follows:

(i) Chapter 1 describes the challenges, problems, current methods, objectives, scopes, and significance of the study.

(ii) Chapter 2 reviews the main subjects used in the study; which include gene function prediction, the approach of computational analysis for gene expression, co-expression methods, biological knowledge, and functional annotation.

(iii) Chapter 3 describes the design of the computational method adopted to achieve the objectives of the study. This includes analysis, instrumentation, and data sources.

(iv) Chapter 4 describes the development of fuzzy *c*-means to handle data ambiguity in the gene expression clustering. The algorithm consists of four components: cluster initialization, fuzzy membership initialization, centroid calculation, and membership update. This algorithm has been validated by using a membership function that can express the variable strength of the association by allowing genes to have membership in multiple clusters.

(v) Chapter 5 describes CluFA as a new computational method that is able to deal with insufficient knowledge by incorporating GO and multiple functional annotation databases in a fuzzy *c*-means algorithm. The incorporation of GO slim was used to automatically define the number of clusters in the cluster initialization where three GO term categories were used to cover all terms in the GO. The multiple functional annotation databases were then used to reduce random initialization. The CluFA method has shown that the results are improved in terms of compactness and separation, and therefore produce more consistent and accurate clusters.

(vi) Chapter 6 describes the enhancement of CluFA that can handle the inaccuracies resulting from conducting the filtering stages and applying the enhanced *apriori* algorithm. Filtering the genes membership values and calculation of the genes *specificity* was found to have achieved the dominant clusters (msf-CluFA-1). Concurrrently, the enhanced *apriori* algorithm was used in order to increase the confidence level for genes with low membership values (msf-CluFA-2). The msf-CluFA method has shown the capability of finding the dominant cluster and being able to increase the confidence level for genes with low membership value, while maintaining the best value for *HD* and *z*-scores. Our msf-CluFA has also shown promising results in predicting an unknown gene function.

(vii) Chapter 7 draws the general conclusions of the achieved results and presents the contributions together with a discussion of suggested topics for future studies.

REFERENCES

Abba, M. C., Drake, J. A., Hawkins, K. A., Hu, Y., Sun, H., Notcovich, C., Gaddis, S., Sahin, A., Baggerly, K., Aldaz, C. M. (2004). Transcriptomic Changes in Human Breast Cancer Progression as Determined by Serial Analysis of Gene Expression. *Breast Cancer Research*. 6(5): R499-R513.

Adler, P., Kolde, R., Kull, M., Tkachenko, A., Peterson, H., Reimand, J., Vilo, J. (2009). Mining for Coexpression Across Hundreds of Datasets Using Novel Rank Aggregation and Visualization Methods. *Genome Biology*. 10(12): R139.

Agrawal, R., Imielinski, T., Swami, A. (1993). Mining Association Rules Between Sets of Items in Large Databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data*. May 26-28. Washington, USA: ACM, 207-216.

Agrawal, R., Srikant, R. (1994). Fast Algorithms for Mining Association Rules in Large Databases. *Proceedings of the 20th International Conference on Very Large Databases*. September 12-15. Santiago, Chile: ACM, 487-499.

Al-Shahrour, F., Díaz-Uriarte, R., Dopazo, J. (2004). Fatigo: A Web Tool for Finding Significant Associations of Gene Ontology Terms with Groups of Genes. *Bioinformatics*. 20(4): 578-580.

Alter O., Brown P. O., Bostein D. (2000). Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling. *The National Academy of Sciences*. 97(18): 10101-10106.

Andreopoulos, B., An, A., Wang, X. (2007). Hierarchical Density-Based Clustering of Categorical Data and a Simplification. In Zhou, Z. H., Li, H., Yang, Q. (Eds.) Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science. 4275. (pp. 11-22). Berlin, Germany: Springer-Verlag.

Angiulli, F., Cesario, E., Pizzuti, C. (2008). Random Walk Biclustering for Microarray Data. *Information Sciences*. 178(1): 1479-1497.

Arakaki, A. K., Huang, Y., Skolnick, J. (2009). Eficaz^2: Enzyme Function Inference by a Combined Approach Enhanced by Machine Learning. *BMC Bioinformatics*. 10(1): 107.

Arifuzzaman, M., Maeda, M., Itoh, A., Nishikata, K., Takita, C., Saito, R., Ara, T., Nakahigashi, K., Huang, H. C., Hirai, A., Tsuzuki, K., Nakamura, S., Altaf-Ul-Amin, M., Oshima, T., Baba, T., Yamamoto, N., Kawamura, T., Ioka-Nakamichi, T., Kitagawa, M., Tomita, M., Kanaya, S., Wada, C., Mori, H. (2006). Large-Scale Identification of Protein-Protein Interaction of Escherichia Coli K-12. *Genome Research*. 16(5): 686-691.

Arima, C., Hanai, T., Okamoto, M. (2003). Gene Expression Analysis using Fuzzy K-Means Clustering. *Genome Informatics*. 14(1): 334-335.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J. E., Ringwald, M., Rubin, G. M., Sherlock, G. (2000). Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*. 25(1): 25-29.

Balasubramaniyan, R., Hüllermeier, E., Weskamp, N., Kämper, J. (2005). Clustering of Gene Expression Data Using a Local Shape-Based Similarity Measure. *Bioinformatics*. 21(7): 1069-1077.

Bandyopadhyay, S., Mukhopadhyay, A., Maulik, U. (2007). An Improved Algorithm for Clustering Gene Expression Data. *Bioinformatics*. 23(21): 2859-2865.

Barkow, S., Bleuler, S., Prelic, A., Zimmermann, P., Zitzler, E. (2006). Bicat: A Biclustering Analysis Toolbox. *Bioinformatics*. 22(10): 1282-1283.

Basrai, M. A., Velculescu, V. E., Kinzler, K. W., Hieter, P. (1999). NORF5/HUG1 is a Component of the MEC1-Mediated Checkpoint Response to DNA Damage and Replication Arrest in Saccharomyces Cerevisiae. *Molecular and Cellular Biology*. 19(10): 7041-7049.

Bays, N., Margolis, J. (2004). Yeast as a Budding Technology in Target Validation. *Drug Discovery Today: Technologies*. 1(2): 157-162.

Bereta, M., Burczyński, T. (2009). Immune *K*-Means and Negative Selection Algorithms for Data Analysis. *Information Sciences*. 179(10): 1407-1425.

Berget, I., Mevik, B. H., Næs, T. (2008). New Modifications and Applications of Fuzzy C-Means Methodology. *Computational Statistics and Data Analysis*. 52(5): 2403-2418.

Bezdek, J. C. (1973). *Fuzzy Mathematics in Pattern Classification*. PhD Thesis, Cornell University, USA.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, USA: Kluwer Academic Publishers.

Bhardwaj, N., Lu, H. (2005). Correlation between Gene Expression Profiles and Protein-Protein Interactions within and Across Genomes. *Bioinformatics*. 21(11): 2730-2738.

Bolshakova, N., Azuaje, F. (2003). Cluster Validation Techniques for Genome Expression Data. *Signal Process*. 83(1): 825-833.

Botet, J., Mateos, L., Revuelta, J. L., Santos, M. A. (2007). A Chemogenomic Screening of Sulfanilamide-Hypersensitive Saccharomyces Cerevisiae Mutants Uncovers ABZ2, the Gene Encoding a Fungal Aminodeoxychorismate Lyase. *Eukaryotic Cell*. 6(11): 2102-2111.

Bourqui, R., Auber, D., Lacroix, V., Jourdan, F. (2006). Metabolic Network Visualization Using Constraint Planar Graph Drawing Algorithm. *Proceedings of the 10th International Conference on Information Visualisation*. 5-7 July, London, UK: IEEE, 489-496.

Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., Sherlock, G. (2004). GO::Termfinder--Open Source Software for Accessing Gene Ontology Information and Finding Significantly Enriched Gene Ontology Terms Associated with a List of Genes. *Bioinformatics*. 20(18): 3710-3715.

Bradford, J. R., Needham, C. J., Tedder, P., Care, M. A., Bulpitt, A. J., Westhead, D. R. (2010). GO-At: In Silico Prediction of Gene Function in Arabidopsis Thaliana by Combining Heterogeneous Data. *The Plant Journal*. 61(4): 713-721.

Brameier, M., Wiuf, C. (2007). Co-Clustering and Visualization of Gene Expression Data and Gene Ontology Terms for Saccharomyces Cerevisiae Using Self-Organizing Maps. *Journal of Biomedical Informatics*. 40(2): 160-173.

Brecheisen, S., Kriegel, H. P., Pfeifle, M. (2006). Multi-Step Density-Based Clustering. *Knowledge and Information Systems*. 9(3): 284-308.

Broët, P., Lewin, A., Richardson, S., Dalmasso, C., Magdelenat, H. (2004). A Mixture Model-Based Strategy for Selecting Sets of Genes in Multiclass Response Microarray Experiments. *Bioinformatics*. 20(16): 2562-2571.

Brown, J. A., Sherlock, G., Myers, C. L., Burrows, N. M., Deng, C., Wu, H. I., Mccann, K. E., Troyanskaya, O. G., Brown, J. M. (2006). Global Analysis of Gene Function in Yeast by Quantitative Phenotypic Profiling. *Molecular System Biology*. 2(1): 2006.0001.

Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guenoche, A., Jacq, B. (2003). Functional Classification of Proteins for the Prediction of Cellular Function from a Protein-Protein Interaction Network. *Genome Biology*. 5(1): R6.

Bugnicourt, A., Mari, M., Reggiori, F., Haguenauer-Tsapis, R., Galan, J. M. (2008). Irs4p And Tax4p: Two Redundant EH Domain Proteins Involved in Autophagy. *Traffic*. 9(5): 755-769.

Burington, B., Barlogie, B., Zhan, F., Crowley, J., Shaughnessy, Jr. J. D. (2008). Tumor Cell Gene Expression Changes Following Short-Term in Vivo Exposure to Single Agent Chemotherapeutics are Related to Survival in Multiple Myeloma. *Clinical Cancer Research*. 14(1): 4821- 4829.

Cai, J., Xie, D., Fan, Z., Chipperfield, H., Marden, J., Wong, W. H., Zhong, S. (2010). Modeling Co-Expression Across Species for Complex Traits: Insights to the Difference of Human and Mouse Embryonic Stem Cells. *PloS Computational Biology*. 6(3): E1000707.

Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A. (2007). GENECODIS: A Web-Based Tool for Finding Significant Concurrent Annotations in Gene Lists. *Genome Biology*. 8(1): R3.

Carter, G. W., Prinz, S., Neou, C., Shelby, J. P., Marzolf, B., Thorsson, V., Galitski, T. (2007). Prediction of Phenotype and Gene Expression for Combinations of Mutations. *Molecular System Biology*. 3(1): 96.

Cesa-Bianchi, N., Valentini, G. (2009). Genome-Wide Hierarchical Classification of Gene Function. *The 3rd International Workshop on Machine Learning in Systems Biology*. 5-6 September. Ljubljana, Slovenia, 14-29.

Chan, Z. S. H., Collins, L., Kasabov, N. (2006). An Efficient Greedy K-Means Algorithm for Global Gene Trajectory Clustering. *Expert Systems with Applications*. 30(1): 137-141.

Chen, X., Wang, L., Smith, J. D., Zhang, B. (2008). Supervised Principal Component Analysis for Gene Set Enrichment of Microarray Data with Continuous or Survival Outcomes. *Bioinformatics*. 24(21): 2474-2481.

Cheng Y., Church G. M. (2000). Biclustering of Expression Data. *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*. 19-23 August. California, USA: AAAI Press, 93-103.

Cheng, Y., Miura, R. M., Tian, B. (2006). Prediction of mRNA Polyadenylation Sites by Support Vector Machine. *Bioinformatics*. 22(19): 2320-2325.

Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., Botstein, D. (1998). SGD: Saccharomyces Genome Database. *Nucleic Acids Research*. 26(1): 73-80.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., Davis, R. W. (1998). A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle. *Molecular Cell*. 2(1): 65-73.

Choi, E., Dial, J. M., Jeong, D. E., Hall, M. C. (2008). Unique D Box and KEN Box Sequences Limit Ubiquitination of Acm1 and Promote Pseudosubstrate Inhibition of the Anaphase-Promoting Complex. *The Journal of Biological Chemistry*. 283(35): 23701-23710.

Choi, Y., Kendziorski, C. (2009). Statistical Methods for Gene Set Co-Expression Analysis. *Bioinformatics*. 25(21): 2780-2786.

Chua, H. N., Sung, W. K., Wong, L. (2006). Exploiting Indirect Neighbours and Topological Weight to Predict Protein Function from Protein-Protein Interactions. *Bioinformatics*. 22(13): 1623-1630.

Clare, A., King, R. D. (2003). Predicting Gene Function in Saccharomyces Cerevisiae. *Bioinformatics*. 19(Suppl 2): 42-49.

Collins, S. R., Kemmeren, P., Zhao, X. C., Greenblatt, J. F., Spencer, F., Holstege, F. C., Weissman, J. S., Krogan, N. J. (2007). Toward a Comprehensive Atlas of the Physical Interactome of Saccharomyces Cerevisiae. *Molecular and Cellular Proteomics*. 6(3): 439-450.

Costa, I. G., Krause, R., Optiz, L., Schliep, A. (2007). Semi-Supervised Learning for the Identification of Syn-Expressed Genes from Fused Microarray and In Situ Image Data. *BMC Bioinformatics*. 8(Suppl 10): S3.

Costanzo, M. C., Hogan, J. D., Cusick, M. E., Davis, B. P., Fancher, A. M., Hodges, P. E., Kondu, P., Lengieza, C., Lew-Smith, J. E., Lingner, C., Roberg-Perez, K. J., Tillberg, M., Brooks, J. E., Garrels, J. I. (2000). The Yeast Proteome Database (YPD) and Caenorhabditis Elegans Proteome Database (WormPD): Comprehensive Resources for the Organization and Comparison of Model Organism Protein Information. *Nucleic Acids Research*. 28(1): 73-76.

Costello, J. C., Dalkilic, M. M., Beason, S. M., Gehlhausen, J. R., Patwardhan, R., Middha, S., Eads, B. D., Andrews, J. R. (2009). Gene Networks in Drosophila Melanogaster: Integrating Experimental Data to Predict Gene Function. *Genome Biology*. 10(9): R97.

Cui, G., Cao, X., Wang, Y., Cao, L., Huang, B., Yang, C. (2006). Wavelet Packet Decomposition-Based Fuzzy Clustering Algorithm for Gene Expression Data. *Proceedings of the Asia Pacific Conference on Circuits and Systems*. 30 November - 3 December. Macao, China: IEEE, 1027-1030.

Damjanović, V., Behrendt, W., Plössnig, M., Holzapfel, M., Franconi, E., Kifer, M. (2007). Developing Ontologies for Collaborative Engineering in Mechatronics. In Franconi, E., Kifer, M., May, M. (Eds.), The Semantic Web: Research and Applications. Lecture Notes in Computer Science. (pp. 336-350). Berlin/German: Springer-Verlag.

Del Bimbo, A., Bertini, M., Torniai, C. (2007). Multimedia Ontologies for Video Digital Libraries. *International Journal of Parallel, Emergent and Distributed Systems*. 22(6): 407-416.

Dembélé, D., Kastner, P. (2003). Fuzzy-C-Means Method for Clustering Microarray Data. *Bioinformatics*. 19(8): 973-980.

Deng, Y., He, T., Wu, Y., Vanka, P., Yang, G., Huang, Y., Yao, H., Brown, S. J. (2005). Computationally Analyzing the Possible Biological Function of YJL103C--An ORF Potentially Involved in the Regulation of Energy Process in Yeast. *International Journal of Molecular Medicine*. 15(1): 123.

DiMaggio Jr, P. A., Mcallister, S. R., Floudas, C. A., Feng, X. J., Rabinowitz, J. D., Rabitz, H. A. (2008). Biclustering via Optimal Re-Ordering of Data Matrices in Systems Biology: Rigorous Methods and Comparative Studies. *BMC Bioinformatics*. 9(1): 458.

Ding**,** C. H. Q. (2002). Analysis of Gene Expression Profiles: Class Discovery and Leaf Ordering. *Proceedings of the 6th International Conference Research in*

Computational Molecular Biology. 18-21 April. Washington, USA: ACM, 127-136.

Ding, Y., Han, H., Liu, F. (2010). Intelligent Integrated Data Processing Model for Oceanic Warning System. *Knowledge-Based Systems*. 23(1): 61-69

Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T., Muller, T. (2008). Identifying Functional Modules in Protein-Protein Interaction Networks: An Integrated Exact Approach. *Bioinformatics*. 24(13): I223-I231.

Dong, A., Li, H. (2006). Multi-Ontology Based Multimedia Annotation for Domain-Specific Information Retrieval. *Proceedings of the IEEE International Conference on Sensor Networks, Ubiquitous, and Trustworthy Computing*. 5-7 June. Taichung, Taiwan: IEEE, 158-165.

Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C., Conklin, B. R. (2003). Mappfinder: Using Gene Ontology and Genmapp to Create a Global Gene-Expression Profile from Microarray Data. *Genome Biology*. 4(1): 7-18.

Dortet-Bernadet, J. L. (2007). Model-based Clustering on the Unit Sphere with an Illustration Using Gene Expression Profiles. *Biostatistics*. 9(1): 66-68.

Duda, R. O., Hart, P. E., Stork, D. G. (2001). *Pattern classification*. (2nd ed.). New York, USA: John Wiley and Sons.

Dunn, J. C. (1973). A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*. 3(3): 32-57.

Dwight, S. S., Harris, M. A., Dolinski, K., Ball, C. A., Binkley, G., Christie, K. R., Fisk, D. G., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D., Cherry, J. M. (2002). Saccharomyces Genome Database (SGD) Provides Secondary Gene Annotation Using the Gene Ontology (GO). *Nucleic Acids Research*. 30(1): 69-72.

Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D. (1998). Cluster Analysis and Display of Genome-wide Expression Patterns. *The National Academy of Sciences*. 95(25): 14863-14868.

Elferink, M. G. L., Olinga, P., Draaisma, A. L., Merema, M. T., Bauerschmidt, S., Polman, J., Schoonen, W. G., Groothuis, G. M. M. (2008). Microarray Analysis in Rat Liver Slices Correctly Predicts In Vivo Hepatotoxicity. *Toxicology and Applied Pharmacology*. 229(3): 300-309.

Enquist-Newman, M., Sullivan, M., Morgan, D. O. (2008). Modulation of the Mitotic Regulatory Network by APC-Dependent Destruction of the Cdh1 Inhibitor Acm1. *Molecular Cell*. 30(4): 437-446.

Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., Mcbroom Cerajewski, L., Robinson, M. D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y. V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J. P., Duewel, H.S., Stewart, I. I., Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S. L., Moran, M. F., Morin, G. B., Topaloglou, T., Figeys, D. (2007). Large-Scale Mapping of Human Protein-Protein Interactions by Mass Spectrometry. *Molecular Systems Biology*. 3(1): 89.

Fang, Z., Li, Y., Luo, Q., Liu, L. (2006). Knowledge Guided Analysis of Microarray Data. *Journal of Biomedical Informatics*. 39(4): 401-411.

Fernandez, E. A., Balzarini, M. (2007). Improving Cluster Visualization in Self-Organizing Maps: Application in Gene Expression Data Analysis. *Computers in Biology and Medicine*. 37(12): 1677-1689.

Fielden, M. R., Brennan, R., Gollub, J. (2007). A Gene Expression Biomarker Provides Early Prediction and Mechanistic Assessment of Hepatic Tumor Induction by Nongenotoxic Chemicals. *Toxicological Sciences*. 99(1): 90-100.

Fielden, M. R., Nie, A., Mcmillian, M., Elangbam, C. S., Trela, B. A., Yang, Y., Dunn, R. T. 2nd, Dragan, Y., Fransson-Stehen, R., Bogdanffy, M., Adams, S. P., Foster, W. R., Chen, S. J., Rossi, P., Kasper, P., Jacobson-Kram, D., Tatsuoka, K. S., Wier, P. J., Gollub, J., Halbert, D. N., Roter, A., Young, J. K., Sina, J. F., Marlowe, J., Martus, H. J., Aubrecht, J., Olaharski, A. J., Roome, N., Nioi, P., Pardo, I., Snyder, R., Perry, R., Lord, P., Mattes, W., Car, B. D., Predictive Safety Testing Consortium, Carcinogenicity Working Group. (2008). Interlaboratory Evaluation of Genomic Signatures for Predicting Carcinogenicity Rat. *Toxicological Sciences*. 103(1): 28-34.

Fraley, C., Raftery, A. E. (1998). How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis. *The Computer Journal*. 41(8): 578-588.

Frank, D., Kuhn, C., Brors, B., Hanselmann, C., Lüdde, M., Katus, H. A., Frey, N. (2008). Gene Expression Pattern in Biomechanically Stretched

Cardiomyocytes: Evidence for a Stretch-Specific Gene Program. *Hypertension*. 51(2): 309-318.

Fu, L., Medico, E. (2007). FLAME, a Novel Fuzzy Clustering Method for the Analysis of DNA Microarray Data. *BMC Bioinformatics*. 8(1): 3.

Fung, B. Y. M., Ng, V. T. Y. (2003). Classification of Heterogeneous Gene Expression Data. Special Issue on Microarray Data Mining. *SIGKDD Explorations*. 5(2): 69-78.

Gamberoni, G., Storari, S., Volinia, S. (2006). Finding Biological Process Modifications in Cancer Tissues by Mining Gene Expression Correlations. *BMC Bioinformatics*. 7(1): 6.

Gasch, A. P., Eisen, M. B. (2002). Exploring the Conditional Coregulation of Yeast Gene Expression Through Fuzzy K-Means Clustering. *Genome Biology*. 3(11): 1-22.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., Brown, P. O. (2000). Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Molecular Biology of the Cell*. 11(12): 4241-4257.

Gertz, E. M., Sengupta, K., Difilippantonio, M. J., Ried, T., Schäffer, A. A. (2009). Evaluating Annotations of an Agilent Expression Chip Suggests that Many Features Cannot be Interpreted. *BMC Genomics*. 10(1): 566.

Getz, G., Levine, E., Domany, E. (2000). Coupled Two-Way Clustering Analysis of Gene Microarray Data. *The National Academy of Sciences*. 97(22): 12079-12084.

Ghosh, D., Chinnaiyan, A. M. (2002). Mixture Modelling of Gene Expression Data from Microarray Experimens. *Bioinformatics*. 18(1): 275-286.

Ghouila, A., Yahia, S.B., Malouche, D., Jmel, H., Laouini, D., Guerfali, F. Z., Abdelhak, S. (2009). Application of Multi-SOM Clustering Approach to Macrophage Gene Expression Analysis. *Infection, Genetics and Evolution*. 9(3): 328-336.

Gibbons, F., Roth, F. (2002). Judging the Quality of Gene Expression-Based Clustering Methods Using Gene Annotation. *Genome Research*. 12(10): 1574-1581.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H.

W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S. G. (1996). Life with 6000 Genes. *Science*. 274(5287): 563-567.

Gonzales, R., Gaudreau, C., Deest, G., Parrott, L., Cardille, J. (2008). An Interactive Cartography and Ecoinformatics Tool to Facilitate the Exchange and Visualization of Canada-Wide Forestry Data. *The 6th International Conference on Ecological Informatics*. 2-5 December. Cancun, Mexico.

Gruca, A., Kozielski, M., Sikora M. (2009). Fuzzy Clustering and Gene Ontology based Decision Rules for Identification and Description of Gene Groups. In: Cyran, K.A., Kozielski, S., Peters, J. F., Stánczyk, U., Wakulicz-Deja, A. (Eds.) Man-Machine Interactions. Lecture Notes in Computer Science. 59. (pp. 141-149). Berlin, Germany: Springer-Verlag.

Gu, J., Liu, J., S. (2008). Bayesian Biclustering of Gene Expression Data. *BMC Genomics*. 9(1): 4-13.

Guo, J., Zhu, P., Wu, C., Yu, L., Zhao, S., Gu, X. (2003). In Silico Analysis Indicates a Similar Gene Expression Pattern Between Human Brain and Testis. *Cytogenetic and Genome Research*. 103(1-2): 58-62.

Guan, J., Gan, Y., Wang, H. (2009). Discovering Pattern-Based Subspace Clusters by Pattern Tree. *Knowledge-Based Systems*. 22(8): 569-579.

Halkidi, M., Batistakis, Y., Vazirgiannis, M. (2001). On Clustering Validation Techniques. *Journal of Intelligenet Information System*. 17(2-3): 107-145.

Hanisch, D., Zien, A., Zimmer, R., Lengauer, T. (2002). Co-Clustering of Biological Networks and Gene Expression Data. *Bioinformatics*. 18(Suppl 1): 145-154.

Hastie, A. R., Pruitt, S. C. (2007). Yeast Two-Hybrid Interaction Partner Screening Through In Vivo Cre-Mediated Binary Interaction Tag Generation. *Nucleic Acids Research*. 35(21): E141.

Herbert, A., Gerry, N. P., Mcqueen, M. B, Heid, I. M., Pfeufer, A., Illig, T., Wichmann, H. E., Meitinger, T., Hunter, D., Hu, F. B., Colditz, G., Hinney, A., Hebebrand, J., Koberwitz, K., Zhu, X., Cooper, R., Ardlie, K., Lyon, H., Hirschhorn, J. N., Laird, N. M., Lenburg, M. E., Lange, C., Christman, M. F. (2006). A Common Genetic Variant is Associated with Adult and Childhood Obesity. *Science*. 312(5771): 279-283.

Herrero, J., Valencia, A., Dopazo, J. (2001). A Hierarchical Unsupervised Growing Neural Network for Custering Gene Expression Patterns. *Bioinformatics*. 17(2): 126-136.

Herszberg, B., Mata, X., Giulotto, E., Decaunes, P., Piras, F. M., Chowdhary, B. P., Chaffaux, S., Guérin, G. (2007). Characterization of the Equine Glycogen Debranching Enzyme Gene (AGL): Genomic and cDNA Structure, Localization, Polymorphism and Expression. *Gene*. 404(1-2): 1-9.

Hestilow, T. J., Huang, Y. (2009). Clustering of Gene Expression Data Based on Shape Similarity. *Journal on Bioinformatics and Systems Biology*. 2009(1): 195712-195724.

Hlynialuk, C., Schierholtz, R., Vernooy, A., Van Der Merwe, G. (2008). Nsf1/Ypl230w Participates in Transcriptional Activation During Non-Fermentative Growth and in Response to Salt Stress in Saccharomyces Cerevisiae. *Microbiology*. 154(8): 2482-2491.

Holleman, A., Cheok, M. H., Den Boer, M. L., Yang, W., Veerman, A. J., Kazemier, K. M., Pei, D., Cheng, C., Pui, C. H., Relling, M. V., Janka-Schaub, G. E., Pieters, R., Evans, W. E. (2004). Gene-Expression Patterns in Drug-Resistant Acute Lymphoblastic Leukemia Cells and Response to Treatment. *The New England Journal of Medicine*. 351(6): 533-542.

Hong, E. L., Balakrishnan, R., Dong, Q., Christie, K. R., Park, J., Binkley, G., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Krieger, C. J., Livstone, M. S., Miyasato, S. R., Nash, R. S., Oughtred, R., Skrzypek, M. S., Weng, S., Wong, E. D., Zhu, K. K., Dolinski, K., Botstein, D., Cherry, J. M. (2008). Gene Ontology Annotations at SGD: New Data Sources and Annotation Methods. *Nucleic Acids Research*. 36(Database Issue): D577-D581.

Horan, K., Jang, C., Bailey-Serres, J., Mittler, R., Shelton, C., Harper, J. F., Zhu, J. K., Cushman, J. C., Gollery, M., Girke, T. (2008). Annotating Genes of Known and Unknown Function by Large-Scale Coexpression Analysis. *Plant Physiology*. 147(1): 41-57.

Hu, P., Bader, G., Wigle, D. A., Emili, A. (2007). Computational Prediction of Cancer-Gene Function. *Nature Reviews Cancer*. 7(1): 23-34.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., Barkai, N. (2002). Revealing Modular Organization in the Yeast Transcriptional Network. *Nature Genet*ics. 31(4): 370-377.

Jiang, D., Pei, J., Zhang, A. (2003). DHC: A Density-Based Hierarchical Clustering Method for Timeseries Gene Expression Data. *Proceeding of the 3rd IEEE*

International Symposium on Bioinformatics and Bioengineering. 10-12 March. Bethesda, USA: IEEE, 393-400.

Jiang, Z. (2008). Protein Function Predictions based on the Phylogenetic Profile Method. *Critical Reviews in Biotechnology*. 28(4): 233-238.

Jourdan, F., Melancon, G. (2003). Tool for Metabolic and Regulatory Pathways Visual Analysis. *Proceedings of the Visualization and Data Analysis*. 19-24 October. Washington, USA: IEEE, 46-55.

Kaba, B., Pinet, N., Lelandais, G., Sigayret, A., Berry, A. (2007). Clustering Gene Expression Data Using Graph Separators. *Journal In Silico Biology*. 7(4): 433-452.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C., Silverman, R., Wu, A. Y. (2002). An Efficient K-Means Clustering Algorithm: Analysis and Implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 24(7): 881-892.

Kanungo, T., Mount, D. M., Netanyahu, N., Piatko, C., Silverman, R., Wu, A. Y. (2004). A Local Search Approximation Algorithm for K-Means Clustering. *Computational Geometry: Theory and Applications*. 28(1): 89-112.

Kelley, R., Ideker, T. (2005). Systematic Interpretation of Genetic Interactions Using Protein Networks. *Nature Biotechnology*. 23(5): 561-566.

Kelso, R. J., Buszczak, M., Quiñones, A. T., Castiblanco, C., Mazzalupo, S., Cooley, L. (2004). Flytrap, a Database Documenting a GFP Protein-Trap Insertion Screen in Drosophila Melanogaster. *Nucleic Acids Research*. 1(32): D418-D420.

Kensche, P. R., Van Noort, V., Dutilh, B. E., Huynen, M. A. (2008). Practical and Theoretical Advances in Predicting the Function of a Protein by its Phylogenetic Distribution. *Journal of the Royal Society Interface*. 5(19): 151-170.

Kiechle, M., Manivasakam, P., Eckardt-Schupp, F., Schiestl, R. H., Friedl, A. A. (2002). Promoter-Trapping in Saccharomyces Cerevisiae by Radiation-Assisted Fragment Insertion. *Nucleic Acids Research*. 30(24): E136.

Kim, D.W., Kang, B.Y. (2006). Iterative Clustering Analysis for Grouping Missing Data in Gene Expression Profiles. In Ng, W. K., Kitsuregawa, M., Li, J. (Eds.) Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science. 3918. (pp. 129-138). Berlin, Germany: Springer-Verlag.

Kim, S. A. (2007). A Graph-Theoretic Classification of Gene Expression Microarray Data of Cancer. *Proceedings of the IEEE Frontiers in the Convergence of Bioscience and Information Technologies*. 11-13 October. Jeju Island, Korea: IEEE, 179-182.

Kivioja, T., Tiirikka, T., Siermala, M., Vihinen, M. (2008). Dynamic Covariation between Gene Expression and Genome Characteristics. *Gene*. 410(1): 53-66.

Köhler, J., Philippi, S., Specht, M., Rüegg, A. (2006). Ontology Based Text Indexing and Querying for the Semantic Web. *Knowledge-Based Systems*. 19(8): 744-754.

Kotsiantis, S., Kanellopoulos, D. (2006). Association Rules Mining: A Recent Overview. *Global Engineering, Science, and Technology Society International Transactions on Computer Science and Engineering*. 32(1): 71-82.

Kourmpetis, Y. A. I., Van Dijk, A. D. J., Bink, M. C. A. M., Van Ham, R. C. H. J., Ter Braak, C. J. F. (2010). Bayesian Markov Random Field Analysis for Protein Function Prediction Based on Network Data. *Plos ONE*. 5(2): E9293.

Krause, S. A., Xu, H., Gray, J. V. (2008). The Synthetic Genetic Network Around PKC1 Identifies Novel Modulators and Components of Protein Kinase C Signaling in Saccharomyces Cerevisiae. *Eukaryotic Cell*. 7(11): 1880-1887.

Kumar, A., Cheung, K. H., Tosches, N., Masiar, P., Liu, Y., Miller, P., Snyder, M. (2002). The TRIPLES Database: A Community Resource for Yeast Molecular Biology. *Nucleic Acids Research*. 30(1): 73-75.

Lamas-Maceiras, M., Núñez, L., Rodríguez-Belmonte, E., González-Siso, M. I., Cerdán, M. E. (2007). Functional Characterization of Klhap1: A Model to Foresee Different Mechanisms of Transcriptional Regulation by Hap1p in Yeasts. *Gene*. 405(1-2): 96-107.

Lawrence, C. J., Harper, L. C., Schaeffer, M. L., Sen, T. Z., Seigfried, T. E., Campbell, D. A. (2008). MaizeGDB: The Maize Model Organism Database for Basic, Translational, and Applied Research. *International Journal Plant Genomics*. 2008(496957): 1-10.

Lazzeroni, L., Owen A. (2002). Plaid Models for Gene Expression Data. *Statistica Sinica*. 12(1): 61-86.

Lee, H., Kong, S. W., Park, P. J. (2008). Integrative Analysis Reveals the Direct and Indirect Interactions Between DNA Copy Number Aberrations and Gene Expression Changes. *Bioinformatics*. 24(7): 889-896.

Lerat, E., Sémon, M. (2007). Influence of the Transposable Element Neighborhood on Human Gene Expression in Normal and Tumor Tissues. *Gene*. 396(2): 303-311.

Li, A., Horvath, S. (2007). Network Neighborhood Analysis with the Multi-Node Topological Overlap Measure. *Bioinformatics*. 23(2): 222-231.

Li, A., Tuck, D. (2009). An Effective Tri-Clustering Algorithm Combining Expression Data with Gene Regulation Information. *Gene Regulation and Systems Biology*. 3(1): 49-64.

Li, G., Wang, Z. (2008). Incorporating Protein-Protein Interactions Knowledge in Clustering Gene Expression Data. *Proceedings of the International Conference on Bioinformatics and Biomedical Engineering*. 8-10 October. Athens, Greece: IEEE, 207-210.

Li, J., Halgamuge, S. K., Kells, C. I., Tang, S-L. (2007). Gene Function Prediction based on Genomic Context Clustering and Discriminative Learning: An Application to Bacteriophages. *BMC Bioinformatics*. 8(Suppl 4): S6.

Li, X. L., Tan, Y. C., Ng, S. K. (2006). Systematic Gene Function Prediction from Gene Expression Data by Using a Fuzzy Nearest-Cluster Method. *BMC Bioinformatics*. 7(Suppl 4): S23.

Lin, H. K., Harding, J. A. (2007). A Manufacturing System Engineering Ontology Model on the Semantic Web for Inter-Enterprise Collaboration. *Computers in Industry*. 58(5): 428-437.

Lin, K. K., Chudova, D., Hatfield, G. W., Smyth, P., Andersen, B. (2004). Identification of Hair Cycle-Associated Genes from Time-Course Gene Expression Profile Data by Using Replicate Variance. *The National Academy of Sciences*. 101(45): 15955-15960.

Liu, J., Xu, M. (2008). Kernelized Fuzzy Attribute C-Means Clustering Algorithm. *Fuzzy Sets and Systems*. 159(18): 2428-2445.

Lobenhofer, E. K., Auman, J. T., Blackshear, P. E., Boorman, G. A., Bushel, P. R., Cunningham, M. L., Fostel, J. M., Gerrish, K., Heinloth, A. N., Irwin, R. D., Malarkey, D. E., Merrick, B. A., Sieber, S. O., Tucker, C. J., Ward, S. M., Wilson, R. E., Hurban, P., Tennant, R. W., Paules, R. S. (2008). Gene

Expression Response in Target Organ and Whole Blood Varies as A Function of Target Organ Injury Phenotype. *Genome Biology*. 9(6): R100.

Lobo, I. (2008). Pleiotropy: One Gene Can Affect Multiple Traits. *Nature Education*. 1(1): 1.

Lu, Z., Hunter, L. (2005). GO Molecular Function Terms are Predictive of Subcellular Localization. *Proceedings of Pacific Symposium on Biocomputing*. 4-8 January. Maui, USA: World Scientific, 151-161.

Luo, F., Khan, L., Bastani, F., Yen, I. L., Zhou, J. (2004). A Dynamically Growing Self-Organizing Tree (DGSOT) for Hierarchical Clustering Gene Expression Profiles. *Bioinformatics*. 20(16): 2605-2617.

Luo, Z., van Vuuren, H. J. (2009). Functional Analyses of PAU Genes in Saccharomyces Cerevisiae. *Microbiology*. 155(Part 12): 4036-4049.

Luukka, P. (2009). Similarity Classifier Using Similarities Based on Modified Probabilistic Equivalence Relations. *Knowledge-Based Systems*. 22(1): 57-62.

Maere, S., Van Dijck, P., Kuiper, M. (2008). Extracting Expression Modules from Perturbational Gene Expression Compendia. *BMC Systems Biology*. 2(1): 33

Mao, J., Habib, T., Shenwu, M., Kang, B., Allen, W., Robertson, L., Yang, J. Y., Deng, Y. (2008). Transcriptome Profiling of Saccharomyces Cerevisiae Mutants Lacking C2H2 Zinc Finger Proteins. *BMC Genomics*. 9(Suppl 1): S14.

Martin-Granados, C., Riechers, S. P., Stahl, U., Lang, C. (2008). Absence of See1p, a Widely Conserved Saccharomyces Cerevisiae Protein, Confers Both Deficient Heterologous Protein Production and Endocytosis. *Yeast*. 25(12): 871-877.

Maulik, U., Mukhopadhyay, A. (2009). Simulated Annealing based Automatic Fuzzy Clustering Combined with ANN Classification for Analyzing Microarray Data. *Computers and Operations Research*. 37(8): 1369-1380.

McGarry, K., Sarfraz, M., Macintyre, J. (2007). Integrating Gene Expression Data from Microarrays Using the Self-Organising Map and the Gene Ontology. In Rajapakse, J. C., Schmidt, B., Volkert, G. (Eds.) Pattern Recognition in Bioinformatics. Lecture Notes in Computer Science. 4774. (pp. 206-217). Berlin, Germany: Springer-Verlag.

McKibbin, T., Frei, C. R., Greene, R. E., Kwan, P., Simon, J., Koeller, J. M. (2008). Disparities in the Use of Chemotherapy and Monoclonal Antibody Therapy for Elderly Advanced Colorectal Cancer Patients in the Community Oncology Setting. *Oncologist*. 13(8): 876-885.

Mewes, H. W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S., Frishman, D. (2002). MIPS: A Database for Genomes and Protein Sequences. *Nucleic Acids Research*. 30(1): 31-34.

Miller, J. A., Oldham, M. C., Geschwind, D. H. (2008). A Systems Level Analysis of Transcriptional Changes in Alzheimer's Disease and Normal Aging. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*. 28(6): 1410-1420.

Mollinedo, F., López-Pérez, R., Gajate, C. (2008). Differential Gene Expression Patterns Coupled to Commitment and Acquisition of Phenotypic Hallmarks During Neutrophil Differentiation of Human Leukaemia HL-60 Cells. *Gene*. 419(1-2): 16-26.

Morbach, J., Yang, A., Marquardt, W. (2007). Ontocape: A Large-Scale Ontology for Chemical Process Engineering. *Engineering Applications Artificial Intelligence*. 20(2): 147-161.

Morgan, A. A., Lu, Z., Wang, X., Cohen, A. M., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., Sun, C., Liu, H. H., Torres, R., Krauthammer, M., Lau, W. W., Liu, H., Hsu, C. N., Schuemie, M., Cohen, K. B., Hirschman, L. (2008). Overview of Biocreative II Gene Normalization. *Genome Biology*. 9(Suppl 2): S3.

Mosley, A. L., Florens, L., Wen, Z., Washburn, M. P. (2009). A Label Free Quantitative Proteomic Analysis of the Saccharomyces Cerevisiae Nucleus. *Journal of Proteomics*. 72(1): 110-120.

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., Morris, Q. (2008). GeneMANIA: A Real-Time Multiple Association Network Integration Algorithm for Predicting Gene Function. *Genome Biology*. 9(Suppl 1): S4.

Motley, A. M., Ward, G. P., Hettema, E. H. (2008). Dnm1p-Dependent Peroxisome Fission Requires Caf4p, Mdv1p and Fis1p. *Journal of Cell Science*. 121(Part 10): 1633-1640.

Nacu, S., Critchley-Thorne, R., Lee, P., Holmes, S. (2007). Gene Expression Network Analysis and Applications to Immunology. *Bioinformatics*. 23(7): 850-858.

Naphade, M., Smith, J. R., Tesic, J., Chang, S., Hsu, W., Kennedy, L., Hauptmann, A., Curtis, J. (2006). Large-Scale Concept Ontology for Multimedia. *IEEE Multimedia*. 13(3): 86-91.

Nariai, N., Kolaczyk, E. D., Kasif, S. (2007). Probabilistic Protein Function Prediction from Heterogeneous Genome-Wide Data. *PloS ONE*. 2(3): E337.

Natarajan, J., Ganapathy, J. (2007). Functional Gene Clustering via Gene Annotation Sentences, Mesh and GO Keywords from Biomedical Literature. *Bioinformatics*. 2(5): 185-193.

Nayak, R. R., Kearns, M., Spielman, R. S., Cheung, V. G. (2009). Coexpression Network Based on Natural Variation in Human Gene Expression Reveals Gene Interactions and Functions. *Genome Research*. 19(11): 1953-1962.

Nikolsky, Y., Nikolskaya, T., Bugrim, A. (2005). Biological Networks and Analysis of Experimental Data in Drug Discovery. *Drug Discovery Today*. 10(9): 653-662.

Norden-Krichmar, T. M., Holtz, J., Pasquinelli, A. E., Gaasterland, T. (2007). Computational Prediction and Experimental Validation of Ciona Intestinalis Microrna Genes. *BMC Genomics*. 8(1): 445.

Okada, Y., Sahara, T., Mitsubayashi, H., Ohgiya, S., Nagashima, T. (2005). Knowledge-assisted Recognition of Cluster Boundaries in Gene Expression Data. *Artificial Intelligence in Medicine*. 35(1-2): 171-183.

Ostapenko, D., Burton, J. L., Wang, R., Solomon, M. J. (2008). Pseudosubstrate Inhibition of the Anaphase-Promoting Complex by Acm1: Regulation by Proteolysis and Cdc28 Phosphorylation. *Molecular and Cellular Biology*. 28(15): 4653-4664.

Oti, M., Van Reeuwijk, J., Huynen, M. A., Brunner, H. G. (2008) Conserved Co-Expression for Candidate Disease Gene Prioritization. *BMC Bioinformatics*. 9(1):208.

Ovaska, K., Laakso, M., Hautaniemi, S. (2008). Fast Gene Ontology Based Clustering for Microarray Experiments. *Biodata Mining*. 1(1): 11.

Pan, W., Lin, J., Le, C. T. (2002). Model-Based Cluster Analysis of Microarray Gene Expression Data. *Genome Biology*. 3(2): 9-16.

Pandey, G., Myers, C. L., Kumar, V. (2009). Incorporating Functional Inter-Relationships into Protein Function Prediction Algorithms. *BMC Bioinformatic*s. 10(1): 142.

Pandit, S. B., Brylinski, M., Zhou, H., Gao, M., Arakaki, A. K., Skolnick, J. (2010). PSiFR: An Integrated Resource for Prediction of Protein Structure and Function. *Bioinformatics*. 26(5): 687-688.

Pang, H., Zhao, H. (2008). Building Pathway Clusters from Random Forests Classification Using Class Votes. *BMC Bioinformatics*. 9(1): 87.

Park, H. S., Cho, S. B. (2007). Evolutionary Fuzzy Cluster Analysis with Bayesian Validation of Gene Expression Profiles. *Journal of Intelligent and Fuzzy Systems: Applications in Engineering and Technology*. 18(6): 543-559.

Pascual-Marqui, R. D., Pascual-Montano, A. D., Kochi, K., Carazo, J. M. (2001). Smoothly Distributed Fuzzy C-Means: A New Self-Organizing Map. *Pattern Recognition*. 34(12): 2395-2402.

Perry, R. J., Mast, F. D., Rachubinski, R. A. (2009). Endoplasmic Reticulum-Associated Secretory Proteins Sec20p, Sec39p, and Dsl1p are Involved in Peroxisome Biogenesis. *Eukaryotic Cell*. 8(6): 830-843.

Pireddu, L., Poulin, B., Szafron, D., Lu, P., Wishart, D. S. (2005). Pathway Analyst-Automated Metabolic Pathway Prediction. *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. 14-15 November. California, USA: IEEE, 1-8.

Pireddu, L., Szafron, D., Lu, P., Greiner, R. (2006). The Path-A Metabolic Pathway Prediction Web Server. *Nucleic Acids Research*. 34(Web Server Issue): W714-W719.

Popescu, M., Keller, J. M., Mitchell, J. A. (2006). Fuzzy Measures on the Gene Ontology for Gene Product Similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 3(3): 263-274.

Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Buhlmann, P., Gruissem, W., Hennig, L., Thiele, L., Zitzler, E. (2006). A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data. *Bioinformatics*. 22(9): 1122-1129.

Premsler, T., Zahedi, R. P., Lewandrowski, U., Sickmann, A. (2009). Recent Advances in Yeast Organelle and Membrane Proteomics. *Proteomics*. 9(20): 4731-4743.

Prokisch, H., Andreoli, C., Ahting, U., Heiss, K., Ruepp, A., Scharfe, C., Meitinger, T. (2006). Mitop2: The Mitochondrial Proteome Database-Now Including Mouse Data. *Nucleic Acids Research*. 1(34): D705-D711.

Punitha, A., Santhanam, T. (2007). Feature Space Optimization in Breast Cancer Diagnosis Using Linear Vector Quantization. *Information Technology Journal*. 6(8): 1258-1263.

Punta, M., Ofran, Y. (2008). The Rough Guide to In Silico Function Prediction, or How to Use Sequence and Structure Information to Predict Protein Function. *PLoS Computational Biology*. 4: E1000160.

Qin, J., Lewis, D. P., Noble, W. S. (2003). Kernel Hierarchical Gene Clustering from Microarray Expression Data. *Bioinformatics*. 19(16): 2097-2104.

Ray, S. S., Bandyopadhyay, S., Pal, S. K. (2007). Gene Ordering in Partitive Clustering Using Microarray Expressions. *Journal Bioscience*. 32(5): 1019-1025.

Ray, S. S., Bandyopadhyay, S., Pal, S. K. (2009). Combining Multi-Source Information Through Functional Annotation Based Weighting: Gene Function Prediction in Yeast. *IEEE Transactions on Biomedical Engineering*. 56(2): 229-236.

Re, M., Valentini, G. (2010). Integration of Heterogeneous Data Sources for Gene Function Prediction Using Decision Templates and Ensembles of Learning Machines. *Neurocomputing*. 73(7-9): 1533-1537.

Rhee, S. Y., Wood, V., Dolinski, K., Draghici, S. (2008). Use and Misuse of the Gene Ontology Annotations. *Nature Reviews, Genetics*. 9(7): 509-515.

Robbins, K. R., Zhang, W., Bertrand, J. K. (2007). The Ant Colony Algorithm for Feature Selection in High-Dimension Gene Expression Data for Disease Classification. *Mathematical Medicine and Biology*. 24(4): 413-426.

Ruan, J. (2010). A Top-Performing Algorithm for the DREAM3 Gene Expression Prediction Challenge. *PloS Computational Biology*. 5(2): E8944.

Sachs, J., Parr, C., Parafiynyk, A., Pan, R., Han, L., Ding, L., Finin, T., Hollender, A., Wang, T. (2006). Using the Semantic Web to Support Ecoinformatics. *Proceedings of the Symposium on Semantic Web for Collaborative Knowledge Acquisition*. 12-15 October. Arlington, USA: AAAI Press, 56-61.

Sato, E., Yamaguchi, T., Harashima, F. (2004). Networked Intelligence by Using Ontology. *Proceedings of the IEEE International Conference on Fuzzy Systems*. 22-25 May. Nevada, USA: IEEE, 311-316.

Sausville, E. A., Holbeck, S. L. (2004). Transcription Profiling of Gene Expression in Drug Discovery and Development: The NCI Experience. *European Journal of Cancer*. 40(17): 2544-2549.

Scherens, B., Goffeau, A. (2004). The Uses of Genome-Wide Yeast Mutant Collections. *Genome Biology*. 5(7): 229.

Schietgat, L., Vens, C., Struyf, J., Blockeel, H., Kocev, D., Dzeroski, S. (2010). Predicting Gene Function using Hierarchical Multi-Label Decision Tree Ensembles. *Bioinformatics*. 11(1): 2.

Schlenoff, C., Ivester, R., Knutilla, A. (1998). A Robust Process Ontology for Manufacturing Systems Integration. *Proceedings of the 2nd International Conference on Engineering Design and Automation*. 9-12 August. Maui, USA: IEEE, 7-14.

Schuler, G. D., Epstein, J. A., Ohkawa, H., Kans, J. A. (1996). Entrez: Molecular Biology Database and Retrieval System. *Methods Enzymol*. 266(1): 141-162.

Seo, D., Wang, T., Dressman, H., Herderick, E., Iversen, E.S., Dong, C., Vata, K., Milano, C. A., Rigat, F., Pittman, J., Nevins, J. R., West, M., Goldschmidt-Clermont, P. J. (2004). Gene Expression Phenotypes of Atherosclerosis. *Arteriosclerosis, Thrombosis, and Vascular Biology*. 24(10): 1922-1927.

Shan, J., Yuan, L., Budman, D. R., Xu, H. P. (2002). WTH3, a New Member of the Rab6 Gene Family, and Multidrug Resistance. *Biochimica Et Biophysica Acta (BBA) - Molecular Cell Research*. 1589(2): 112-123.

Sharan, R., Maron-Katz, A., Shamir, R. (2003). Click and Expander: A System for Clustering and Visualizing Gene Expression Data. *Bioinformatics*. 19(14): 1787-1799.

Sharma, A., Podolsky, R., Zhao, J., Mcindoe, R. A. (2009). A Modified Hyperplane Clustering Algorithm Allows for Efficient and Accurate Clustering of Extremely Large Datasets. *Bioinformatics*. 25(9): 1152-1157.

Shmueli, G., Patel, N. R., Bruce, P. C. (2006). *Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel With XLMiner*. New Jersey, USA: Wiley.

Shoemaker, B. A., Panchenko, A. R. (2007). Deciphering Protein-Protein Interactions. Part I. Experimental Techniques and Databases. *PLoS Computational Biology*. 3(3): E42.

Simperl, E. P. B., Tempich, C. (2006). Ontology Engineering: A Reality Check. In Tari, Z., Meersman, R., Herrero, P. (Eds.), on the Move to Meaningful Internet Systems 2006: CoopIS, DOA, GADA, and ODBASE. Lecture Notes in Computer Science. 4275. (pp. 836-854). London, UK: Springer-Heidelberg.

Soong, T. T., Wrzeszczynski, K. O., Rost, B. (2008). Physical Protein-Protein Interactions Predicted from Microarrays. *Bioinformatics*. 24(22): 2608-2614.

Srivastava, S., Zhang, L., Jin, R., Chan, C. (2008). A Novel Method Incorporating Gene Ontology Information for Unsupervised Clustering and Feature Selection. *PloS ONE*. 3(12): E3860.

Stenson, P. D., Mort, M., Ball, E. V., Howells, K., Phillips, A. D., Thomas, N. S., Cooper, D. N. (2009). The Human Gene Mutation Database: 2008 Update. *Genome Medicine*. 1(1): 13.

Stuart, J. M., Segal, E., Koller, D., Kim, S. K. (2003). A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science*. 302(5643): 249-255.

Szeto, L. K., Liew, A. W. C., Yan, H., Tang, S. (2003). Gene Expression Data Clustering and Visualization Based on a Binary Hierarchical Clustering Framework. *Proceedings of the 1st Asia-Pacific Bioinformatics Conference on Bioinformatics*. 4-7 February. Adelaide, Australia: ACM, 145-152.

Tagkopoulos, I., Slavov, N., Kung, S. Y. (2005). Multi-Class Biclustering and Classification Based on Modeling of Gene Regulatory Networks. *Proceedings of the 5th IEEE International Symposium on Bioinformatics and Bioengineering*. 19-21 October. Minneapolis, USA: IEEE, 89-96.

Takahara, Y., Kobayashi, T., Takemoto, K., Adachi, T., Osaki, K., Kawahara, K., Tsujimoto, G. (2008). Pharmacogenomics of Cardiovascular Pharmacology: Development of an Informatics System for Analysis of DNA Microarray Data with a Focus on Lipid Metabolism. *Journal of Pharmacological Sciences*. 107(1): 1-7.

Tamayo, P., Solni, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., Golub, T. R. (1999). Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation. *The National Academy of Sciences*. 96(6): 2907-2912.

Tan, M. P., Smith, E. N., Broach, J. R., Floudas, C. A. (2008). Microarray Data Mining: A Novel Optimization-Based Approach to Uncover Biologically Coherent Structures. *BMC Bioinformatics*. 9(1): 268.

Tang C., Zhang L., Zhang A., Ramanathan M. (2001). Interrelated Two-Way Clustering: An Unsupervised Approach for Gene Expression Data Analysis. *Proceeding of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering*. Washington, USA: IEEE, 41-48.

Tang, C., Zhang, A. (2002). An Iterative Strategy for Pattern Discovery in High-Dimensional Data Sets. *Proceeding of the 11th International Conference on Information and Knowledge Management*. 4-9 November. Virginia, USA: ACM, 10-17.

Tanya, C.P., Steven, G.C. (2009). Computational Methods to Identify Novel Methyltransferases. *BMC Bioinformatics*. 10(Suppl 13): P7.

Tari, L., Baral, C., Kim, S. (2009). Fuzzy C-Means Clustering with Prior Biological Knowledge. *Journal Biomedical Informatics*. 42(1): 74-81.

Taşan, M., Tian, W., Hill, D. P., Gibbons, F. D., Blake, J. A., Roth, F. P. (2008). An En Masse Phenotype and Function Prediction System for Mus Musculus. *Genome Biology*. 9(Suppl 1): S8.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., Church, G. M. (1999). Systematic Determination of Genetic Network Architecture. *Nature Genetics*. 22(3): 281-285.

Taverna, S. D., Ilin, S., Rogers, R. S., Tanny, J. C., Lavender, H., Li, H., Baker, L., Boyle, J., Blair, L. P., Chait, B. T., Patel, D. J., Aitchison, J. D., Tackett, A. J., Allis, C. D. (2006). Yng1 PHD Finger Binding to H3 Trimethylated at K4 Promotes Nua3 HAT Activity at K14 Of H3 and Transcription at a Subset of Targeted ORFs. *Molecular Cell*. 24(5): 785-796.

Ternes, P., Sperling, P., Albrecht, S., Franke, S., Cregg, J. M., Warnecke, D., Heinz, E. (2006). Identification of Fungal Sphingolipid C9-Methyltransferases. Phylogenetic Profiling. *The Journal of Biological Chemistry*. 281(9): 5582-5592.

The Flybase Consortium (2008). Flybase: Integration and Improvements to Query Tools. *Nucleic Acids Research*. 36(1): D588-D593.

The Gene Ontology Consortium (2000). Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*. 25(1): 25-29.

The Mouse Genome Database Group (2008). The Mouse Genome Database (MGD): Mouse Biology and Model Systems. *Nucleic Acids Research*. 36(1): D724-D728.

The Reference Genome Group of the Gene Ontology Consortium. (2009). The Gene Ontology's Reference Genome Project: A Unified Framework for Functional Annotation Across Species. *PloS Computational Biology*. 5(7): E1000431.

Tjaden, B. (2006). An Approach for Clustering Gene Expression Data with Error Information. *BMC Bioinformatics*. 7(1): 17-31.

Tripathi, A., Ren, Y., Jeffrey, P. D., Hughson, F. M. (2009). Structural Characterization of Tip20p and Dsl1p, Subunits of the Dsl1p Vesicle Tethering Complex. *Nature Structural Molecular and Biology*. 16(2): 114-123.

Tseng, G. C. (2007). Penalized and Weighted K-Means for Clustering with Scattered Objects and Prior Information in High-Throughput Biological Data. *Bioinformatics*. 23(17): 2247-2255.

Tu, K., Yu, H., Li, X. Y. (2006). Combining Gene Expression Profiles and Protein-Protein Interaction Data to Infer Gene Functions. *Journal of Biotechnology*. 124(3): 475-485.

Ulitsky, I., Shamir, R. (2007). Pathway Redundancy and Protein Essentiality Revealed in the Saccharomyces Cerevisiae Interaction Networks. *Molecular System Biology*. 3(1): 104.

Valarmathie, P., Srinath, M. V., Ravichandran, T., Dinakaran, K. (2009). Hybrid Fuzzy C-Means Clustering Technique for Gene Expression Data. *International Journal of Research and Reviews in Applied Sciences*. 1(1): 33-37.

van Baarsen, L. G. M., Vosslamber, S., Tijssen, M., Baggen, J. M. C., Van Der Voort, L. F., Killestein, J., Van Der Pouw Kraan, T. C. T. M., Polman, C. H., Verweij, C. L. (2008). Pharmacogenomics of Interferon-Beta Therapy in Multiple Sclerosis: Baseline IFN Signature Determines Pharmacological Differences between Patients. *PloS Computational Biology*. 3(4): E1927-E1935.

van Berlo, R. J. P., Wessels, L. F. A., Martes, S. D. C., Reinders, M. J. T. (2005). Predicting Gene Function by Combining Expression and

Interaction Data. *Proceedings of the IEEE Computational Systems Bioinformatics Conference*. 8-11 August. California, USA: IEEE, 166-167.

Vashist, A., Kulikowski, C., Muchnik, I. (2005). Automatic Protein Function Annotation Through Candidate Ortholog Clusters from Incomplete Genomes. *Proceedings of the IEEE Computational Systems Bioinformatics Conference*. 16-19 August. California, USA: IEEE, 73-74.

Voy, B. H., Scharff, J. A., Perkins, A. D., Saxton, A. M., Borate, B., Chesler, E. J., Branstetter. L. K., Langston, M. A. (2006). Extracting Gene Networks for Low-Dose Radiation using Graph Theoretical Algorithms. *PLoS Computational Biology*. 2(7): E89.

Wanat, J. J., Kim, K. P., Koszul, R., Zanders, S., Weiner, B., Kleckner, N., Alani, E. (2008). Csm4, in Collaboration with Ndj1, Mediates Telomere-Led Chromosome Dynamics and Recombination during Yeast Meiosis. *PloS Genetics*. 4(9): E1000188.

Wang, W., Zhang, Y. (2007). On Fuzzy Cluster Validity Indices. *Fuzzy Sets and Systems*. 158(19): 2095-2117.

Wang, Y., Fang, Y., Wang, S. (2007). Clustering and Principal-Components Approach Based on Heritability for Mapping Multiple Gene Expressions. *BMC Proceedings*. 1(1): S121-S126.

Wassenaar, T., M., Gamieldien, J., Shatkin, J., Luber, P., Moyer, N., Carpenter, T., David, W. (2007). The Importance of Virulence Prediction and Gene Networks in Microbial Risk Assessment. *Ussery Human and Ecological Risk Assessment: An International Journal*. 13(2): 254-268.

Wiederhold, E., Veenhoff, L. M., Poolman, B., Slotboom, D. J. (2010). Proteomics of Saccharomyces Cerevisiae Organelles. *Molecular and Cellular Proteomics*. 9(3): 431-445.

Williams, R. J., Martinez, N. D., Golbeck, J. (2006). Ontologies for Ecoinformatics. *Web Semantics*: *Science, Services and Agents on the World Wide Web*. 4(4): 237-242.

Wilson, W. A., Wang, Z., Roach, P. J. (2002). Systematic Identification of the Genes Affecting Glycogen Storage in the Yeast Saccharomyces Cerevisiae: Implication of the Vacuole as a Determinant of Glycogen Level. *Molecular and Cellular Proteomics*. 1(3): 232-242.

Winden, K. D., Oldham, M. C., Mirnics, K., Ebert, P. J., Swan, C. H., Levitt, P., Rubenstein, J. L., Horvath, S., Geschwind, D. H. (2009). The Organization of the Transcriptional Network in Specific Neuronal Classes. *Molecular Systems Biology*. 5(1): 291.

Wingender, E., Hogan, J., Schacherer, F., Potapov, A. P., Kel-Margoulis, O. (2007). Integrating Pathway Data for Systems Pathology. *In Silico Biology*. 7(2 Suppl): S17-S25.

Witten, I. H., Frank, E. (2005). *Data mining: Practical Machine Learning Tools and Techniques*. San Francisco, USA: Morgan Kaufmann Publishers.

Wu, F. X., Zhang, W. J., Kusalik, A. J. (2003). A Genetic K-Means Clustering Algorithm Applied to Gene Expression Data. In Xiang, Y., Chaib-draa, B. (Eds.) Advances in Artificial Intelligence. Lecture Notes in Computer Science. 2671. (pp. 520-526). Berlin, Germany: Springer-Verlag.

Xie, X. L., Beni, G. (1991). A Validity Measure for Fuzzy Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 13(8): 841-847.

Xing, E. P., Karp, R. M. (2001). Cliff: Clustering of High-Dimensional Microarray Data via Iterative Feature Filtering using Normalized Cuts. *Bioinformatics*. 17(1): 306-315.

Xu, J., Zhang, Q., Shih, C. K. (2006). V-Cluster Algorithm: A New Algorithm for Clustering Molecules Based Upon Numeric Data. *Journal Molecular Diversity*. 10(3): 463-478.

Xutao, D., Huimin, G., Hesham, A. H. (2008). A Hidden Markov Model Approach to Predicting Yeast Gene Function from Sequential Gene Expression Data. *International Journal of Bioinformatics Research and Applications*. 4(3): 263-273.

Yang, D., Li, Y., Xiao, H., Liu, Q., Zhang, M., Zhu, J., Ma, W., Yao, C., Wang, J., Wang, D., Guo, Z., Yang, B. (2008). Gaining Confidence in Biological Interpretation of the Microarray Data: The Functional Consistence of the Significant GO Categories. *Bioinformatics*. 24(1): 265-271.

Yang, J., Wang, H., Wang, W., Yu, P. (2003). Enhanced Biclustering on Expression Data. *Proceedings of the 3rd IEEE Symposium on Bioinformatics and Bioengineering*. 10-12 March. Bethesda, USA: IEEE, 321-327.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., Ruzz, W. L. (2001). Model-Based Clustering and Data Transformations for Gene Expression Data. *Bioinformatics*. 17(10): 977-987.

Yoon, S., De Micheli, G. (2006). Computational Identification of MicroRNAs and Their Targets. *Computational Biology and Chemistry*. 78(2): 118-128.

Young, J. A., Winzeler, E. A. (2005). Using Expression Information to Discover New Drug and Vaccine Targets in the Malaria Parasite Plasmodium Falciparum. *Pharmacogenomics*. 6(1): 17-26.

Yuan, Y., Li, C. T., Wilson, R. (2008). Partial Mixture Model for Tight Clustering of Gene Expression Time-Course. *BMC Bioinformatics*. 9(287): 1471-2105.

Zare, H., Khodursky, A. B., Kaveh, M. (2006). Gene Clustering and Gene Function Prediction Using Multiple Sources of Data. *Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics*. May 28-30. Texas, USA: IEEE, 113-114.

Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S., Bussey, K. J., Riss, J., Barrett, J. C., Weinstein, J. N. (2003). GOMiner: A Resource for Biological Interpretation of Genomic and Proteomic Data. *Genome Biology*. 4(4): R28-R35.

Zeng, J., Zhu, S., Liew, A. W., Yan, H. (2010). Multiconstrained Gene Clustering Based on Generalized Projections. *BMC Bioinformatics*. 31(11): 164.

Zhang, J., Li, J., Deng, H. (2008). Class-Specific Correlations of Gene Expressions: Identification and Their Effects on Clustering Analyses. *The American Journal of Human Genetics*. 83(2): 269-277.

Zhang, M. L., Peña, J. M., Robles, V. (2009). Feature Selection for Multi-Label Naive Bayes Classification. *Information Sciences*. 179(19): 3218-3229.

Zhang, M., Therneau, T., Mckenzie, M. A., Li, P., Yang, P. (2008). A Fuzzy C-Means Algorithm Using a Correlation Metrics and Gene Ontology. *Proceedings of the International Conference on Pattern Recognition*. 15-17 October. Melbourne, Australia: IEEE, 1-4.

Zhang, Q., Zhang, Y. (2006). Hierarchical Clustering of Gene Expression Profiles with Graphics Hardware Acceleration. *Pattern Recognition Letters*. 27(6): 676-681.

Zhang, T., Li, L., Li, X., Wang, H. (2008). Unravelling the Hidden Relationship Between Subtype of Ion Channel and Channlopathy based on CTWC Approach. *Proceedings of the 2nd International Conference on Bioinformatics and Biomedical Engineering*. 16-18 May. Shanghai, China: IEEE, 676-679.

Zhao, T., Han, X. (2005). Auto-K Dynamic Clustering Algorithm. *Asian Journal of Information Technology*. 4(4): 467-471.

Zhao, X. M., Wang, Y., Chen, L., Aihara, K. (2008). Gene Function Prediction Using Labeled and Unlabeled Data. *BMC Bioinformatics*. 9:57.

Zhao, Y. P., Chen G., Feng B., Zhang T. P., Ma, E. L., Wu, Y. D. (2007). Microarray Analysis of Gene Expression Profile of Multidrug Resistance in Pancreatic Cancer. *Chinese Medical Journal*. 120(20): 1743-1752.

Zhong, S., Storch, F., Lipan, O., Kao, M. J., Weitz, C., Wong, W. H. (2004). Gosurfer: A Graphical Interactive Tool for Comparative Analysis of Large Gene Sets in Gene Ontology Space. *Applied Bioinformatics*. 3(4): 1-5.

Zhou, J., Dieng-Kuntz, R. (2004). Manufacturing Ontology Analysis and Design: Towards Excellent Manufacturing. *Proceedings of the International Conference on Industrial Informatics*. 24-26 June. Berlin, Germany: IEEE, 39-45.

Zhu, D. (2009). Semi-Supervised Gene Shaving Method for Predicting Low Variation Biological Pathways from Genome-Wide Data. *BMC Bioinformatics*. 10(Suppl 1): S54.

Zhu, J., Zhang, M. Q. (2000). Cluster, Function and Promoter: Analysis of Yeast Expression Array. *Proceedings of Pacific Symposium on Biocomputing*. 4-8 January. Maui, USA: IEEE, 467-486.

Zien, A., Küffner, R., Zimmer, R., Lengauer, T. (2000). Analysis of Gene Expression Data with Pathway Scores. *Proceedings of the International Conference on Intelligence System Molecular Biology*. 19-23 August. California, USA, 407-417.