Applying Gene Ontology to Microarray Gene Expression Data Analysis

Andy C. Yang, Hui-Huang Hsu*, Ming-Da Lu
Department of CSIE
Tamkang University, TKU
Taipei, Taiwan, R.O.C.
andyyung0215@gmail.com, *huihuanghsu@gmail.com

***Abstract*— Selecting informative genes from microarray gene expression data is the most important task while performing data analysis on the large amount of data. Mining genes having regulatory relations within thousands of genes is essential. To fit this need, a number of methods were proposed from various points of view. However, most existing methods solely focus on gene expression values themselves without using any external information of genes. Gene Ontology (GO) provides biological information of genes or proteins involved. It utilizes a hierarchical structure to give additional biological information of genes as the aid for data analysis. In this paper, we first give a brief description about the GO structure and give a review of existing literatures that take GO into account. Subsequently, we propose a novel method to identify regulatory gene pairs in a real microarray dataset based on dynamic time warping (DTW) algorithm and GO. Finally, we summarize this paper with a discussion on how GO can be used to facilitate the analysis of microarray gene expression data.**

***Keywords*— gene expression, microarray time series data, dynamic time warping, gene ontology**

I. INTRODUCTION

Microarray technology is widely used in this decade due to the high throughput data it can produce. These huge amounts of data are called microarray gene expression data and they provide informative meanings for biologists. Microarray gene expression data are time-series and matrix-liked that in the format of numeric values generated from specific computer tools. Each gene expression value in microarray time series data means different reaction degrees result from experiments. These quantitative values are in the format of logarithm which represents distinct intensity of expressions. These kinds of data provide a possible means for the inference of transcriptional regulatory relationships among the genes on the microarray gene chips. The discovery of specific gene pairs with highly-correlated relations could provide valuable information for biologists to predict important biological reactions [1].

Despite the informative meanings of these data, there remains a challenge for the analysis of them. On the one hand, biologists can retrieve significant information of genes from these data. On the other hand, these large amounts of gene expression data raise the need for effective approaches to deal with them. Typically, the aim of the analysis on microarray time-series data is to observe and find out whether there exists any pair of genes that have highly-correlated relations. Researches on this issue have been worked for these years, and a variety of approaches are proposed. Existing methods are generated from various aspects. Common proposed solutions include clustering analysis [2-5], spectral analysis [6, 7], similarity analysis [8-10], and Bayesian networks [11, 12]. Above approaches are applied for the inference and prediction of gene-gene relations in microarray time-series data. Although some of them may have a success for the analysis of microarray time series data, the effectiveness is very limited. This is because these methods only take gene expression values themselves into consideration and they lack for external or biological information of genes.

Gene ontology (GO) is a hierarchical structure of defined annotations for known genes or proteins. It consists of three independent domains: molecular function, biological process, and cellular component. Each known gene has its own annotations or terms that represent for the biological characteristic or similarity within the evolution process in terms of the three domains. Genes can hence be preprocessed or grouped based on GO before we actually deal with gene expression values to increase the efficiency. Moreover, analyses combining this external information with gene expression values can obtain more accurate results than merely performing similarity measurement on gene expression values [13].

In this paper, we first briefly mention about the GO structure. Subsequently, we give a review of existing methods that use gene ontology as additional information to improve the quality of microarray data analysis. We also propose a novel method for the prediction of gene regulation relationship based on dynamic time warping (DTW) that combines GO structure. Finally, we present a discussion on how GO can be applied to facilitate microarray gene expression data analysis.

The remaining of this paper is organized as follows. In Section II, GO structure is briefly described. A review of methods that utilize GO is discussed in Section III. The proposed method that combines DTW and GO is mentioned in Section IV. The discussion of experiment results and how GO can be used for the analysis of microarray gene expression data is given in Section V. The concluding remarks are made in Section VI.

II. GENE ONTOLOGY

The gene ontology (GO) is a definition and annotation for genes that are known and studied. Each known gene has a specific annotation (term) in GO structure within three independent domains: molecular function, biological process, and cellular component. Terms within three above domains consider different aspects respectively. Molecular function considers the biological or biochemical activity at the molecular level. Biological process can be said as the combination of molecular function. It denotes a biological objective which genes contribute to. Cellular component records the place in cells where a gene product is active. One gene may have more than one annotation in each domain. Moreover, one gene can have totally different annotations while considering different aspects of domain. A gene product might be associated with or located in one or more cellular components. It is active in one or more biological processes, during which it performs one or more molecular functions.

The structure of GO can be described as a directed acyclic graph (DAG), where each GO term is a node, and the relationships between the terms are arcs between the nodes. GO resembles a hierarchy, as child terms are more specialized and parent terms are less specialized. Although GO structure is hierarchical, each node in GO can have several parent nodes and several children nodes just in case that relations between each node do not form a cycle. The most commonly-used relations in GO are "is-a" relation and "part of" relation. For example, if the relation "term A is a term B" exists in GO, that means term A is a subtype of term B. By contrast, if the relation "term A is part of term B" stands, it means all children terms of term A with term A itself belong to term B. Each term in GO has one unique GO id for it, but the number of GO id does not represent the similarity between terms.

Fig. 1 illustrates an example of GO. For example, GO id 0015749 shown in Fig. 1 denotes a term "monosaccharide transport", which has the relation "is-a" with its parent term (GO id 0008643). Equally, the parent-children relation between terms at consequent levels starting from one specific node can be traced level by level to the root node. In Fig. 1, if we start from the term GO: 0015749, we can trace the path from the selected node to the root as "GO: 0015749->GO: 0008643->GO: 0006810->GO: 0051234->GO: 0008150". With this directed acyclic graph structure, we can easily query the GO annotation terms of each gene in microarray gene expression time series data to give a general view of the biological activities of the genes involved.

Since each gene may have different terms in the three independent domains, deciding which domain we are focusing on is hence an important issue. Besides, one gene may be annotated by more than one term even in the same domain. Moreover, each term can have more than "one-to-one" relation with its parent term or children term. Typically, a completed tracing path of GO annotation terms for one gene is somehow complicated. Therefore, the way how we can use gene ontology differs from data themselves and the algorithm we are applying. Sometimes it can also depend on the kind of analysis we are performing.

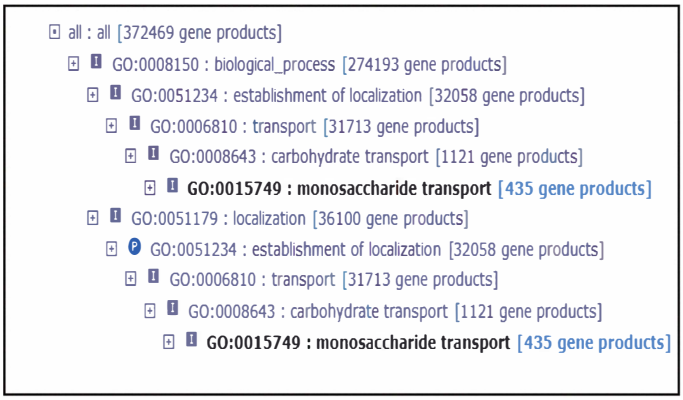

Figure 1. Example of Gene Ontology

With GO tern annotation, each gene can have a uniform representation across biological databases. For more information about gene ontology, please refer to the Gene Ontology website [14]. With GO annotations, we can hence acquire the relations for the genes involved in the experiment. The related closeness of two genes can be identified if we perform some quantitative assessments on the gene pair with their GO annotations. Following section mentions about our survey of existing methods that take gene ontology into consideration.

III. EXISTING METHODS USING GENE ONTOLOGY

The main task in microarray gene expression data analysis is to identify the gene pairs or groups that are highly co-expressed under individual experimental conditions. The most common procedure is performing similarity measurement or classification and clustering on gene expression values. Nevertheless, with the usage of gene ontology, this task can be done more efficiently and accurately.

To our knowledge, the first literature proposing methods that use gene ontology is [13]. In the study, the algorithm proposed by the authors first finds the sets of GO ids for the pair of genes that are being identified. A table recording the tracing path of all terms annotated for the both genes is then created. The algorithm calculates the probability of the occurrence of each term in the table, and then estimates all the parent-children relations of each term in the table to determine whether the two genes have common ancestors. If the two genes analyzed have shared parent nodes in the GO tracing path, they are marked as similar according to a probability threshold of occurrence. This study proposes a typical approach that combines GO with numeric values processing so that a better result is retrieved.

In [15], GO annotations are regarded as important parameters in a determining equation. The authors modify the FCM clustering algorithm with some adjustments. In the study, if we are going to determine the similarity of one gene with the other genes in a cluster, a dichotomy equation is given. This equation sets the parameter to 1 if two genes have the same annotation, otherwise the parameter will be set to 0 so that it influences on the original FCM algorithm. The usage of GO in this study is very simple, but it should not just be a dichotomy

determination for genes. Difference of levels at which the annotation terms locate should be considered.

Another example of using gene ontology is in [16]. In the paper, terms in GO are taken as labels that are capable of discriminative power of identifying whether a gene with the term is informative. This algorithm first sets a discriminative score equation for genes. For each GO term, if genes annotated with it are with higher discriminative scores, the term is defined as the informative term. Genes that can be annotated with these kinds of informative terms are so called informative genes. The algorithm uses the GO terms as a probe to determine which genes can be said as informative. This paper is based on the concept that co-expressed genes should have similar GO annotation terms.

Similar usage of GO is proposed in [17]. In the study, GO terms are used as the information content. Semantic closeness is defined if the most immediate parent node is shared by two annotation terms. The authors also merge various GO-based similarity measurement algorithms that consider intra and inter ontological relations by translating each relative term into a hierarchical relation within a smaller sub-ontology.

Among existing methods that apply gene ontology to discover relative genes or to pre-group genes with known similar biological functions, it seems GO does aid in the analysis of microarray gene expression data. However, there exist several kinds of approaches to the application of gene ontology. In the following section, we discuss the usage and argue the importance of applying GO correctly.

IV. GENE REGULATION PREDICTION BASED ON DTW AND GO

Here we propose a method that combines dynamic time warping and gene ontology. The proposed method takes both gene expression values and external information for the genes into consideration. In this section, we briefly introduce the dynamic time warping algorithm along with the procedures of the proposed method. We also mention about a real microarray gene expression time series dataset that we use for our experiments.

A. Dynamic time warping (DTW)

Dynamic time warping is a commonly-used algorithm which was first applied in voice and pattern recognition [18, 19]. Literatures have been shown that DTW performs well to find out the similarity for a pair of time series data [20, 21]. In this paper, we combine the DTW algorithm with GO as the similarity measurement to identify gene pairs with regulatory relations in microarray time-series data. In general, the dynamic time warping method is used to warp and match generic sequences of numbers that can be viewed as curves in a proper coordinate system. The aim of DTW is to obtain a precise matching along the temporal axis, and to maximize the number of point-wise matches between two time series. The alignment of temporal patterns by DTW has traditionally been used in the recognition of speech signals. This method is a widely-used algorithm for string comparison and for the alignment of time series data. If two series with time points are given as input, the DTW algorithm can select the best possible alignment between them by minimizing a local distance between the series points.

DTW is a recursive algorithm that matches each two-point pair from the first element to the last element on input sequences. After the table recording all local optimal paths and corresponding points is created, a multiple of its last computed value returns the DTW distance between the two sequences. With DTW mapping method, local similarity can be found as the best path within the two comparison sequences. As a result, if two genes with similar gene expression values at certain parts in microarray time series data are analyzed by DTW, it is more precise for similarity measurement because DTW can discover their similarity that cannot be identified with other similarity measurements such as Pearson correlation coefficient or Euclidean distance.

Therefore, we choose DTW as the similarity measurement to preliminarily find the similarity of gene pairs. Usually, the points that a DTW path goes through should not be totally diagonal in the matrix. Otherwise it would be the same with the Euclidean distance and hence makes DTW alignment meaningless. If the best warping path between the two input sequences is found, local similarity of the two sequences can thus be discovered. For more details about the DTW algorithm, please see [22].

B. Real microarray dataset

Spellman et al. and Cho et al. provided the yeast microarray dataset (http://genome-www.stanford.edu/cellcycle) [4, 23]. The data was obtained for genes of Yeast Saccharomyces cerevisiae cells that were collected with four synchronization methods: alpha-factor, cdc15, cdc28, and elutriation [24]. These four subsets of the Spellman's dataset contain totally 6178 gene ORF profiles with their expression values across individual amounts of time slots. For example, the alpha subset contains 18 time points with seven minutes as the time interval, while the cdc28 contains 17 time points with ten minutes as the time interval. These four kinds of subsets record the degree of gene expression reactions at various experimental time points during different phases in cell cycle.

Filkov et al. reviewed related literatures and collected all known gene regulations of alpha and cdc28 subsets in Spellman's yeast cell dataset [25]. A database for recording all these known gene regulations was also constructed. In our evaluation, the known gene regulations recorded in Filkov's database are taken as the validation datasets. In the database, the number of recorded gene activations and inhibitions for alpha subset is 343 and 96 respectively, while for cdc28 subset is 469 and 155 accordingly. All these regulations are in the format of A (+) B that denotes gene A is an activator that activates gene B. Similarly, C (-) D represents an inhibitor gene C inhibits gene D. For example, ABF1 (+) ACS1 is an activation regulation with gene ABF1 as the activator. However, among these regulations recorded in the database, there might be a widespread situation that one gene could be the activator or inhibitor for more than one genes. For instance, gene ABF1 stands for the activator for totally eight different genes in cdc28 subset. Therefore, the pre-processing of the raw data is necessary. First, we parse all regulations of alpha and

cdc28 subsets in Filkov's database and retrieve unrepeatable involved genes. The parsing result is shown in Table I.

TABLE I. PARSING RESULT FOR GENE REGULATIONS

Dataset	Content			
	No. of Genes	*No. of Activations*	*No. of Inhibitions*	*Total*
alpha	295	343	96	439
cdc28	357	466	155	621

After the involved genes are parsed out, the next step is to map these hundreds of genes to Spellman's datasets to match the corresponding gene expression values. Nevertheless, gene names in Filkov's database are denoted as the gene standard name, while the gene systematic names are used in Spellman's dataset. As a result, a mapping procedure between gene standard name and systematic name is required.

We perform the mapping task with the aid of the reference database called the Saccharomyces Genome Database (SGD, http://www.yeastgenome.org/) [26]. The SGD database acts as a platform for biologists to refer and query yeast gene information including the gene standard name and systematic name. During the process of gene name mapping, we found that some of the gene standard name in Filkov's database cannot be found in Spellman's dataset due to the different naming conventions. For example, the gene systematic name for the gene with standard name STA1 cannot be found in the SGD database. Consequently, regulations with gene STA1 are filtered that causes the decrease of gene activations in cdc28 subset from 469 to 466.

Also, we have purified the involved genes with their gene expression values and the corresponding gene standard name as the implementation dataset for the proposed method. Theoretically, the number of pairwise gene combinations for alpha subset is C_2^{295} which equals to 43365, and the number of pairwise gene combinations for cdc28 subset is C_2^{357} which equals to 63546. Eventually, some missing regulations are replenished and the final amount of pairwise gene combinations for alpha and cdc28 subsets is 43366 and 63548, respectively. Known regulations in Filkov's database are marked as the validation measurement to estimate the correctness of the proposed method. Afterward, we apply the proposed method on these gene pairwise combinations and count the number of potential regulatory gene pairs we find which are also listed in Filkov's database. Regulations of activations and inhibitions are summed up separately. The results are shown and discussed in Section V.

C. Method

Our gene regulation prediction method first calculates the DTW distance for all combinations of involved genes. The number of gene pairwise combinations of alpha and cdc28 subsets is 43365 and 63546 respectively. After DTW distances for all these combinations are calculated, we then compute the mean of all DTW distances. Assume the mean DTW distance of alpha subset is DTWmean_alpha, gene pairs with DTW distance smaller then DTWmean_alpha are retained and recorded. These recorded gene pairs are subsequently compared with the validation datasets from Filkov's database . Afterward, the number of mapping gene pairs between the validation datasets from Filkov's database and gene pairs which are identified based on DTW distance is gathered. Theoretically, since we suppose DTW distances reflect the better similarity measurement for the gene pairs, potential regulatory gene pairs should have smaller DTW distances compared with the others in all gene pair combinations. This is the concept of this procedure. The same operations are performed on cdc28 subset, and the number of mapping gene pairs of cdc28 subset is also gathered.

After the number of mapping gene pairs identified with DTW distance is gathered, we then add GO information into our method. As mentioned in Section II, GO structure gives genes or proteins standard annotation terms which are defined based on biological functions or attributes for the genes or proteins themselves. Therefore, we define a similar relationship for a gene pair if the two genes in the pair have GO annotation terms in common. However, from the GO website we can only find a relative file recording genes which are annotated with their corresponding GO terms [27]. To analyze each GO annotation term with genes annotated by it, we have to parse the annotation file from the GO website and construct a table recording the GO annotation terms and the genes annotated by these terms. After we construct the table recording annotations for the genes, gene products annotated by each GO annotation term are grouped. Here we simply make an intuitive assumption that genes annotated by the same GO annotation terms tend to have similar biological functions or attributes. As a result, we take this similarity measurement for the genes to supplement potential regulatory gene pairs which are not found with DTW distance. Gene pairs identified with GO similarity measurement are also compared with the validation database from Filkov's database. The number of mapping gene pairs between the validation datasets from Filkov's database and the gene pairs which are identified with the union of DTW distance and GO annotation terms is then gathered to assess the effectiveness of our method. The detail algorithm of the proposed method is described as follows.

Algorithm for the proposed method:

1) *For all gene pair combinations, find DTW distance of each gene pair, and then calculate the mean DTW distance DTWmean of all combinatios.*

2) *Record gene pairs with DTW distance smaller than DTWmean, assume S_{DTW}.*

3) *For all gene pair combinations, record gene pairs with more than one GO annotation terms in common, assume S_{GO}.*

4) *Find the union of S_{DTW} and S_{GO}, assume U.*

5) *Compare U with Filkov's datasets. Count the number of mapping gene pairs.*

V. EXPERIMENTAL RESULTS AND DISCUSSION

Table II shows the experimental results of our method and method from [7]. Activation regulations and inhibition regulations from Filkov's database are separated. The four subsets lying in the first column denote the known gene regulations from Filkov's database. The number of mapping gene pairs of the four methods, including Pearson Correlation Coefficient (PCC), the method from [7], only DTW similarity measurement, and DTW similarity measurement with GO information is listed in the corresponding grids of the table. Gene pairs are said to be similar if their PCC values are larger than 0.5 according to [7]. We can see that PCC can only find very few mapping known regulatory gene pairs, while the method from [7] gets better then PCC. However, in [7] only alpha subset is experimented. Therefore we mark the result of cdc28 subset of the method from [7] with N/A. The result of only DTW similarity measurement seems to be very similar to that of the method from [7]. Obviously, with our method we can find much more known regulatory gene pairs compared with other methods. In alpha activation regulations, we can even find almost 315/343 = 91% of known regulatory gene pairs and 401/469 = 85% of known regulatory gene pairs in cdc28 activation regulations. The results show that our method is effective.

TABLE II. NUMBER OF IDENTIFIED REGULATORY GENE PAIRS

Dataset/ # of Known gene pairs	**Method**			
	PCC	**[7]**	**DTW**	**DTW+GO**
alpha(+)/ 343	36	223	215	315
alpha(-)/ 96	5	55	56	77
cdc28(+)/ 469	66	N/A	287	401
cdc28(-)/ 155	14	N/A	87	121

Gene ontology is the well-structured annotation for genes concerning the three domains. It provides essential convenience for biologists to estimate the closeness among known genes. Due to the complicated annotation terms in GO, it is required to apply GO to microarray gene expression data properly. Deciding which GO-based approach to apply mainly depends on the data we are dealing with, and it is also related to which kind of analysis we are performing. In following paragraphs, we discuss about the issues in terms of two aspects while applying gene ontology.

A. How to use GO terms?

As mentioned in section II, GO is a directed acyclic graph that consists of many annotation terms for known genes. Each term in GO can have more than one parent node or children node. The tracing path starting from a given node to the root represents a biological activity if one gene is annotated by this node. As a result, the commonest way to utilize gene ontology is to judge whether two genes have similar annotation terms in GO. For example, if we are going to identify whether two genes are co-expressed, we can make a query on GO to retrieve GO annotation terms for the two genes. The comparison of annotation terms for the two genes is then performed. If the GO terms for the two genes are the same or similar (a threshold for similarity is needed), we can create a table recording similar gene lists for each gene in microarray gene expression data based on gene ontology annotation terms.

In some cases, GO terms are used just to mark whether two genes are similar or not. This dichotomy relies on a set of determining assessments. It may work well in some situations, but it is not a general method. Theoretically, genes should not be only defined as similar or not. Similarity degrees of each gene pair should be emphasized. For example, if there are three genes: Gene A, Gene B, and Gene C. Gene A and Gene B have the same GO terms at a specified level, while Gene B and Gene C have the same GO terms at a general level. From the definition of GO, we should regard the relation between Gene A and Gene B as closer than the relation between Gene B and Gene C. The various degrees of closeness are informative and should not be ignored.

Another way to use GO terms is tracing the path from a given node to the root first, and then finding the occurrence of shared ancestors of two terms. If two genes are annotated by two terms that have a common ancestor in their parent nodes or even at higher levels, these two genes are somehow similar corresponding to the threshold the user defines. Data mining techniques such as association rules can be applied for this need.

B. Which domain of GO terms should we focus on?

Gene ontology annotation terms are with in three independent domains. These three domains consist of totally different annotation terms. As for the number of terms in these three domains, there are 18480 terms in biological process, 2685 terms in cellular component, and 8687 terms in molecular function to date. Annotation terms in these three domains are considered as different aspects of gene activity. Molecular function focuses on the biological or biochemical activity at the molecular level. Biological process can be said as the combination of molecular function. It denotes a biological objective which genes contribute to. Cellular component records the place in cells where a gene product is active.

In this paper, we simply define a GO similarity relationship based on the same GO annotation terms. Actually, before applying gene ontology it is essential to decide which domain of GO terms we are going to focus on. Microarray gene expression data come from a number of conditional experiments, and the experimental results may differ from distinguish domains in which the genes are involved. What counts is which kind of aspects the data come from. For example, if we are analyzing gene expression data involved in biological process domain, choosing GO terms in molecular function or cellular component is meaningless. As a result, it would be of no use if we do not use GO terms that correspond to our microarray gene expression data involved. Only by choosing suitable methods and corresponding GO terms of domains can facilitate the analysis of microarray gene expression data.

VI. CONCLUSION

In this paper, we briefly describe the gene ontology structure and discuss about the important issues while applying gene ontology to microarray data analysis. We also propose a novel method combining DTW and GO to predict regulatory gene pairs in microarray time series data. Experimental results argue that gene ontology is the useful external information for genes within microarray time series data. We discuss the way how we can take gene ontology as a hint to help the analysis of microarray gene expression data. We believe that applying gene ontology in a proper manner facilitates the identification of informative genes in microarray data.

REFERENCES

[1] D.S.V. Wong, F.K. Wong, and G.R. Wood, "A multi-stage approach to clustering and imputation of gene expression profiles," Bioinformatics, Vol. 23, pp.998-1005, 2007.

[2] Y. Huang and P.S. Yu, "Adaptive query processing for time-series data," in: Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, pp.15-18, 1999.

[3] K. Kalpakis, D. Gada, and V. Puttagunta, "Distance measures for effective clustering of ARIMA time-series," in: Proceedings of the IEEE International Conference on Data Mining, pp.273-280, 2001.

[4] R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfsberg, A. Gabrielian, D. Landsman, and D. Lockhart, "A genome-wide transcriptional analysis of the mitotic cell cycle," Molecular Cell, Vol. 2, pp.65-73, 1998.

[5] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," in: Proceedings of the National Academy of Science, Vol. 95, pp.14863-14868, 1998.

[6] M.K. Choong, K.C. Lye, David Levy, and H. Yang, "Periodicity identification of microarray time series data based on spectral analysis," in: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, pp.1281-1285, 2006.

[7] L.K. Yeung, H. Yan, Alan W.C. Liew, L.K. Szeto, Michael Yang, and Richard Kong, "Measuring correlation between microarray time series data using dominant spectral component," in: Proceedings of The 2nd Asia-Pacific Bioinformatics Conference, Vol. 29, pp.309-314, 2004.

[8] M. Vlachos, G. Kollios, and G. Gunopulos, "Discovering similar multidimensional trajectories," in: Proceedings of the 18th International Conference on Data Engineering, pp.673-684, 2002.

[9] M.S. Lee, L.Y. Liu, and M.Y. Chen, "Similarity analysis of time series gene expression using dual-tree wavelet transform," in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.I-413-I-416, 2007.

[10] A.C. Yang, H.H. Hsu, M.D. Lu, "Outlier filtering for identification of gene regulations in microarray time-series data, " in: Proceedings of the 3rd International Conference on Complex, Intelligent and Software Intensive Systems, pp.854-859, 2009.

[11] S. Kim, S. Imoto, and S. Miyano, "Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data," Biosystems, Vol. 75, pp.57-65, 2004.

[12] N. Friedman, M. Linial, I. Nachman, and DanaPe'er, "Using Bayesian network to analyze expression data," in: Proceedings of the 4th Annual International Conference on Computational Molecular Biology, pp.601-620, 2000.

[13] J. Tuikkala, L. Elo, O.S. Nevalainen, and T. Aittokallio, "Improving missing value estimation in micorarray data with gene ontology," Bioinformatics, Vol. 22, pp.566-572, 2006.

[14] Gene Ontology website, URL: http://www.geneontology.org/, last accessed on March 22, 2010.

[15] A. Mohammadi and M.H. Saraee, "Estimating missing value in microarray data using fuzzy clustering and gene ontology," in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, pp.382-385, 2008.

[16] X. Xu and A. Zhang, "Selecting informative genes from microarray dataset by incorporating gene ontology," in: Proceedings of the 5th IEEE Symposium on Bioinformatics and Bioengineering, pp.241-245, 2005.

[17] A. Sanfilippo, B. Baddeley, N. Beagley, and B. Gopalan, "Enhancing automatic biological pathway generation with GO-based gene similarity," in: Proceedings of International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, pp.448-453, 2009.

[18] C. Furlanello, S. Merler, and G. Jurman, "Combining feature selection and DTW for time-varying functional genomics," IEEE Transactions on Signal Processing, Vol. 54, pp.2436-2443, 2006.

[19] H.M. Yu, W.H. Tsai, and H.M. Wang, "Query-by-singing system for retrieving karaoke music," IEEE Transactions on Multimedia, Vol. 10, pp.1626-1637, 2008.

[20] C. Myers, L. Rabiner, and A. Roseneberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-28, pp.623-635, 1980.

[21] L. Rabiner, A. Rosenberg, and S. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-26, pp.575-582, 1978.

[22] E. Keogh, "Exact indexing of dynamic time warping," in: Proceedings of the 28th international conference on Very Large Data Bases, pp.406-417, 2002.

[23] Paul T. Spellman, Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization." Mol. Biol. Cell, Vol.9, pp.3273-3297, 1998.

[24] The Yeast Cell Cycle website by Spellman et al. URL: http://genome-www.stanford.edu/cellcycle/, last accessed on March 22, 2010.

[25] V. Filkov, S. Skiena, and J. Zhi, "Analysis techniques for microarray time-series data." Proceedings of The Fifth Annual International Conference on Computational Molecular Biology, pp.124-131, 2001.

[26] The Saccharomyces Genome Database (SGD), URL: http://www.yeastgenome.org/,last accessed on March 22, 2010.

[27] The Gene Ontology Annotation relations, URL: http://www.geneontology.org/GO.current.annotations.shtml, last accessed on March 22, 2010.