Characterization and Clustering of GO Terms by Feature Importance Vectors Obtained from Microarray Data

JOVAN DAVID REBOLLEDO-MENDEZ†, MASANORI HIGASHIHARA*, YOICHI YAMADA†, KENJI SATOU†

† Graduate School of Natural Science and Technology
Kanazawa University
Kakuma-machi, Kanazawa 920-1192
JAPAN
jovan@leo.ec.t.kanazawa-u.ac.jp http://leo.ec.t.kanazawa-u.ac.jp/~jovan/

* Graduate School of Knowledge Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292
JAPAN

Abstract: - In this paper it is explained a new approach for clustering Gene Ontology (GO) terms by examining microarray data related to them. By segmenting the entire ontology in a single specific level, and applying techniques as discrimination and ranking of features to those GO terms that are contained in that level, it is produced a characterization of the contained terms, as feature importance vectors related to the gene expression patterns that are included in the microarray dataset. By utilizing data mining techniques to cluster the vectors, it is concluded that this new approach may help to obtain relations that are normally hidden among GO terms, not only the ones in the same contained ontology, but also getting a trans-ontological relationship of them.

Key-Words: - Gene Ontology, Microarray data, Random forest, Feature ranking, Hierarchical clustering, Data mining, Machine learning, Ontology clustering

1 Introduction

There have been different technologies that have helped to understand genetics, like the inventions of Polymerase Chain Reaction (PCR), where an enzyme of DNA is used to create copies of a DNA piece (among several genetic manipulations) [1]; and the whole genome shotgun sequencing, for sequencing the entire human genome [2]. Among other techniques, DNA microarray started a new era of high-throughput measurement of multiple gene expression levels at a time. Starting from cDNA microarray developed at Stanford University [3], improved methods of this technique were proposed, invented, and some of these were commercialized. In consequence, now it is well-liked to measure all the different gene expression levels in a cell sample taken from a species under some pre-established conditions, e.g. samples of various tissues, diseased samples including tumor cell, etc., giving to each gene expression a value (which usually could be visualized at different colors in the heat maps) in the result of its data analysis.

These data are usually clustered in groups for getting a better representation of the results. There are different methods that permit clustering these data [4] however, the comparison of two or more microarray data (each one typically having between 10 and 100 samples), obtained from samples under different conditions, may reveal biologically meaningful relationships among genes and samples of that same microarray, as explained in [5], and [6]. As a result of gene expression analysis, a subset of all the genes in a certain species is frequently described in terms of expression levels, which is potentially meaningful in certain context. Consequently, it is needed to interpret the meaning of the subset (gene set), for instance in order to find correlations among genes and several illnesses. The use of ontology information in bioinformatics has been treated, as [7]; and among others techniques [6], [8], and [9], the potential use of Gene Ontology (GO) [10], which is the most comprehensive and

authorized hierarchy of biological concepts (controlled vocabulary), could benefit in the search of finding new meaningful relationships. In that research line, GO TermFinder [11] is a popular software to perform that specific task. It computes statistical significance between an input gene set and the terms in Gene Ontology. There have been also other methods to identify correlations, like [5] and [12], in which it is tried to correlate the GO terms in order to characterize the gene set. However, this kind of task requires a specific and limited gene set as an input, and, as a consequence, the method lacks of comprehensiveness in the analysis. What it means is that most of the expression data were used up in the gene expression analysis method step, and they were discarded before referring to Gene Ontology.

The work in this paper is a proposal of a new method in which all the expression data of a microarray is used, and by making a boundary in the level of the abstracted tree in Gene Ontology, and the resulted subtree of a specific GO term is used for a characterization with a consecutive classification to find new ways to correlate the GO terms among them. The remaining of this paper is organized as described next: in section 2 there is a description of the data and software that were used in the experiments; in section 3, the experimental results are shown with some analysis and interpretation. Finally, in section 4, it is presented a summary and conclusion of the results from this paper, including possibilities and future directions of this research.

2 Materials and Methods

2.1 Gene Ontology

A couple of examples about the use of ontology in bioinformatics have been researched before in different ways as [13] and [14]. Among different ontologies, Gene Ontology could be defined as a set up of three different ontologies: Biological Process Ontology, Cellular Component Ontology, and Molecular Function Ontology (see Fig.1); and combined, they contain currently more than 26,000 terms. The amount of terms changes constantly with their addition or correction. Each single ontology is formed by a hierarchical parent-child tree-like directed acyclic graph (DAG) data structure, including information about its terms and the relationships with other terms of the same ontology. Every single GO term is related with at least one different GO term, and always coming in pairs of terms. Such relationships between a pair of GO terms are "*is_a*", and "*part_of*", among others. A term can just have one type of relationship with another term. Also, a term is not limited to have just one parent, but it could have two or more; the same applies to the number of children. Each single term has a nomenclature which makes it unique among the others. The database of Gene Ontology can be obtained at the GO Consortium (http://www.geneontology.org).

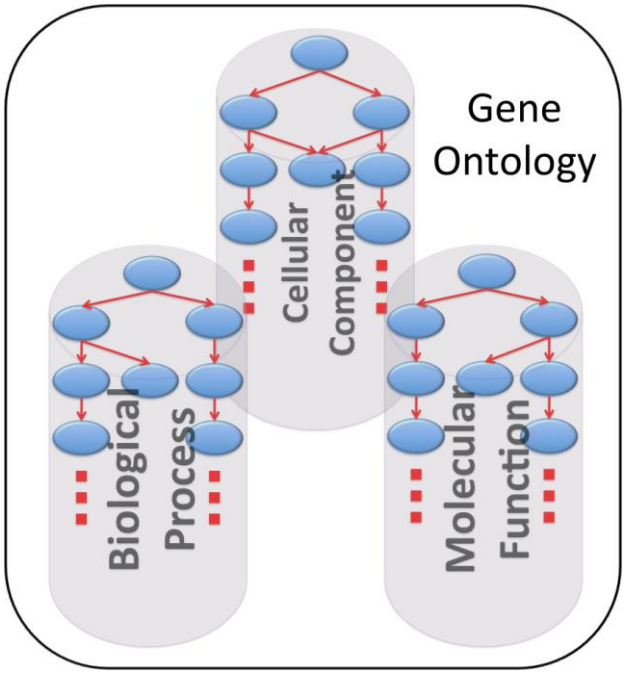

Fig.1. Gene Ontology representation.

By giving a GO term *x*, in which a set of GO terms *descendant(x)* in the subtree rooted at *x* can be considered, that might share some conceptual characteristics represented by the GO term *x*. Additionally, since links from genes in microarray to GO terms are provided, it could be considered a set of genes linked to *descendant(x)*.

Therefore, if the microarray data is considered appropriate to discriminate the *descendant(x)* from other GO terms with respect to gene expression pattern, a machine learning algorithm could classify the gene expression patterns linked to *descendant(x)* with high precision (Fig.2 illustrates the process). Furthermore, by using a technique called feature ranking, it is possible to evaluate the importance of each feature (i.e. a sample of the gene expression) in this discrimination process. It means that a GO term *x* can be represented as a vector of feature importance.

In case of microarray data on tissues (i.e. a sample corresponds to a tissue), *x* may be well discriminated by specific tissues (e.g. colon and small intestine). By conducting this computational

method on many GO terms that are related with that specific tissue and belonging to a certain constant distance from the root GO term of the ontology, it is possible to find hidden similarities among them (i.e. similarity of feature importance vectors) in terms of discrimination by expression pattern. Algorithms similar to our method are studied as "hierarchical classification" mainly in the field of text categorization [15], [16], where an annotation reference of genes is used in order to classify them. However, in this work a method of discrimination and feature ranking at many GO terms (limited by common boundaries of deepness in the tree of the ontology) is conducted for getting a characterization of each of them (not only for prediction of specific category).

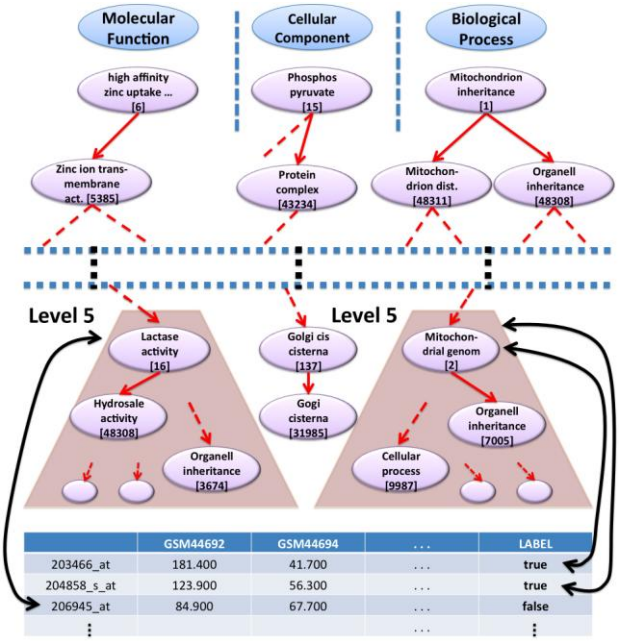

Fig.2. Conceptual figure of the proposed method.

2.2 Microarray data

The expression of most of genes is given in their transcription and translation, having protein biosynthesis as a result, and being this indispensable for life. Much of the information contained in thousands of the expressions of genes can be obtained by microarray techniques that have become so known in a wide range of fields. There are different techniques that permit the extraction of the information, like gene expression profiling, comparative genomic hybridization, and SNP detection, among others [17].

There are also different databases [18] that contain the expression of gene data, like ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae), the European Bioinformatics Institute (EBI), CIBEX and the Gene Expression Omnibus (GEO; http://ncbi.nlm.nih.gov/projects/geo) of the National Center for Biotechnology Information (NCBI) at the National Institutes of Health, each one having a repository of microarray data, usually of samples with different properties and from different species. The microarray data consists of a series of affixed DNA segments, known as probes or reporters which measure the different genes that are put on chips, giving a different value to each gene and probe. In this paper, it was used the microarray data GSE2361 taken from GEO, and which contains a matrix with 36 samples of human tissue as columns and 22,281 human genes as rows. Information that has been linked from genes to GO terms was extracted from the annotation file for the platform GPL96, Affymetrix GeneChip Human Genome U133 Array Set HG-U133A, which was employed to measure the expression data in GSE2361. And for the comparison study with the yeast, the dataset GSE3635 was used.

2.3 Setting of the positive and negative examples for discrimination and feature ranking

Although it is possible to conduct discrimination and feature ranking at all nodes in GO, it was used only the GO terms which are found at certain boundary or level as the number of links from root node of each ontology, the distance of links from the root node of Biological Process, Molecular Function, or Cellular Component. In addition to avoiding the problem of scale (i.e. around 26,000 terms are too many to compute them all), it is considered that this limitation is useful for making a fair comparison (i.e. only the GO terms at the same or similar abstraction level are compared).

In order to prepare better input for a classifier, the microarray data GSE2361 was normalized by using Distribution Free Weighted (DFW) algorithm, by utilizing variability estimates to "identify and down-weight probes that may be especially affected by non-specific and cross-hybridization" [19]. By using the DFW summarization method in R, it could be obtained data that became statistically more suitable to be processed for later classification. For yeast data GSE3635, normalization was not needed since the data has been normalized.

After previous normalization method, the data preparation was made separately for each ontology. Defining a set of GO terms $T_{BP} = \{t_1,...,t_n\}$ that is taken from a certain level of Biological Process (in this study, level 5 was adopted), first, examples corresponding to each GO term t_k in T_{BP} were prepared. The examples are gene expression data for

the genes $G(\{t_k\})$, where $G(\{t_k\})$ denotes all the genes linked to *descendant*(t_k). Therefore, positive and negative examples were prepared for each t_k according to this rule: by attaching class labels "true" or "false" to all of the examples in $G(\{t_k\})$ and $G(T_{BP}\text{-}\{t_k\})$, by which positive and negative examples were obtained, respectively. For a schematic visualization, see Fig.2.

2.4 Feature ranking by random forest

In order to characterize a GO term in terms of the features included in the specified microarray data, positive and negative examples corresponding to the GO term are input to the random forest algorithm. Random forest [20] is a kind of ensemble learning algorithm developed by L. Breiman. Besides its ability of classification, in this research it was used for feature ranking: to obtain importance of feature, i.e. contribution to discriminate positive and negative examples [21].

As an implementation of random forest algorithm, randomForest package for R was adopted. From given examples, it performs training on the data and, as a by-product, it outputs a value called mean decrease Gini for each feature, which can be used as an importance of the feature. In this study, each GO term in level 5 (from the root term of a single ontology) is characterized as a vector of feature importance of that tree. If a GO term t_k (more precisely, a set of genes in $G(\{t_k\})$) is well discriminated from others by some features (e.g. Brain and Hippocampus), it can be said that t_k potentially has a close relationship to the features, in terms of the microarray data.

2.5 Clustering GO terms

In order to interpret the results of the GO term characterization based on the microarray data GSE2361, a cluster analysis was conducted on the feature importance vectors obtained previously. From among various methods of cluster analysis, hierarchical clustering was adopted for this computation.

As a distance measure of hierarchical clustering, a distance based on Spearman's rank correlation was used. About cluster linkage method, UPGMA (also known as average linkage) was used. Computation of hierarchical clustering was performed by using Cluster 3.0 software, and among different graph drawing tools [22], the result was visualized by Java TreeView software.

Fig.3. Clustering result on (B) Biological Process, (M) Molecular Function, and (C) Cellular Component.

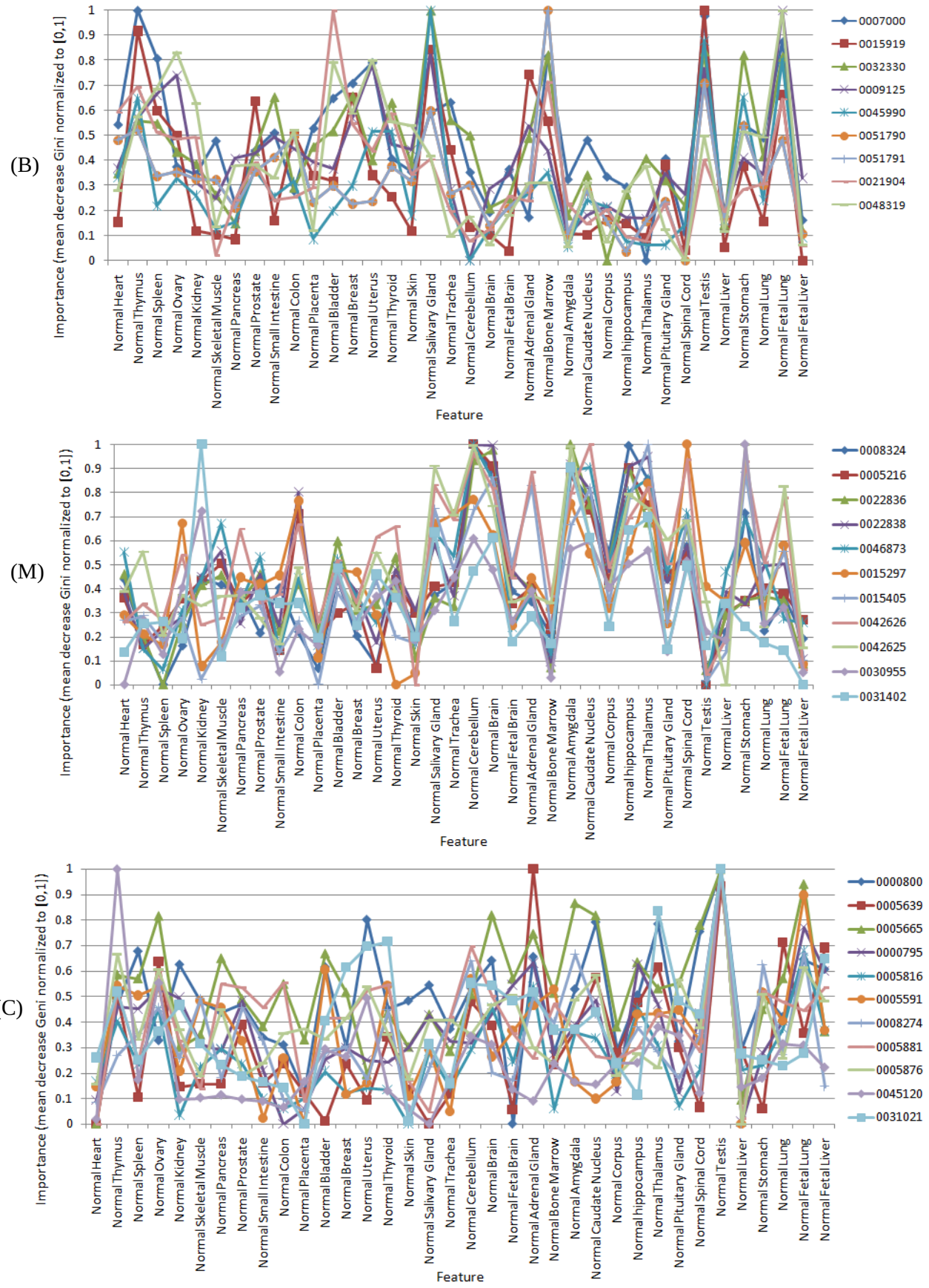

Fig.4. Feature importance vectors of GO terms in Fig.3, where (B) Biological Process, (M) Molecular Function, and (C) Cellular Component.

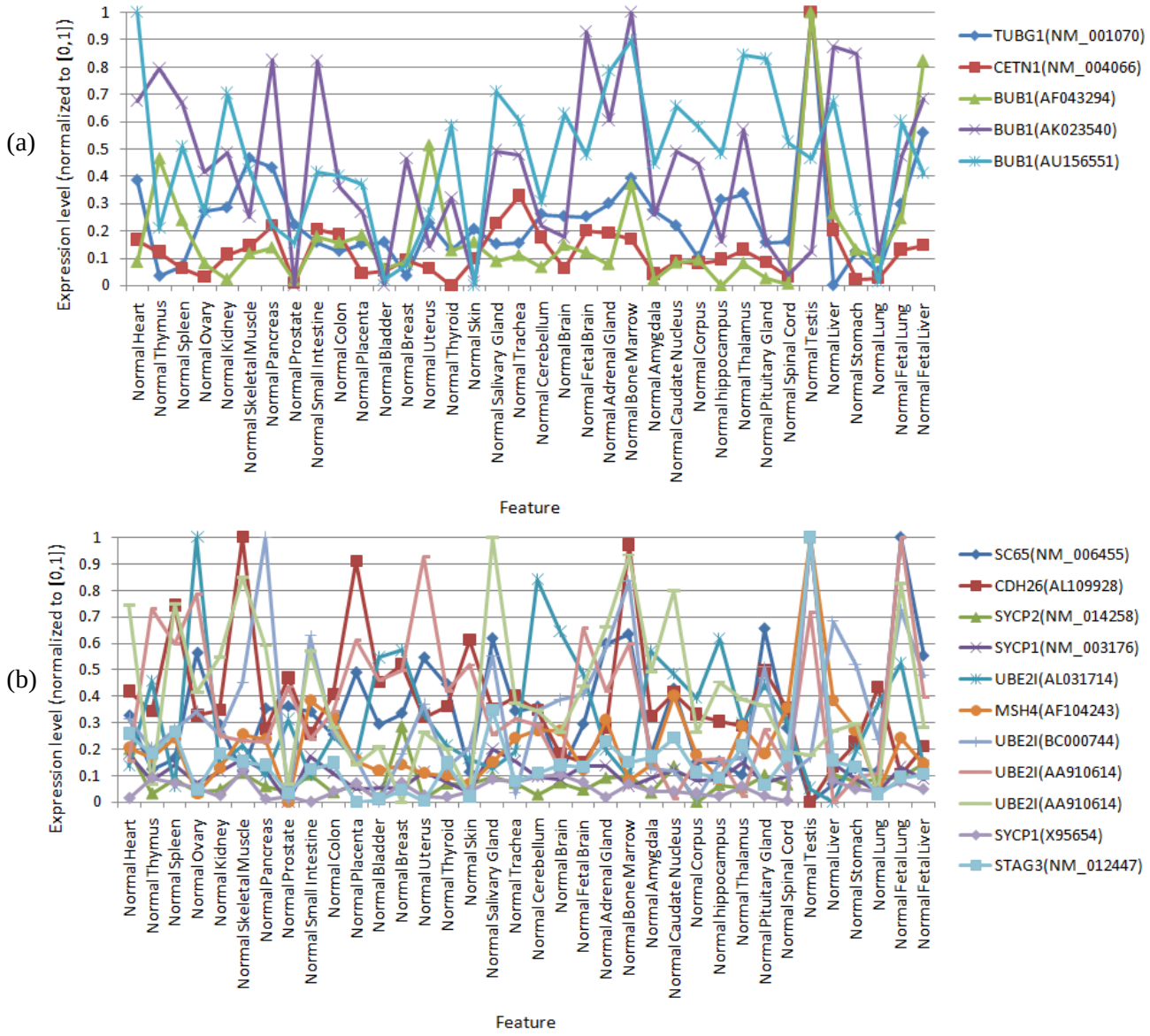

Fig.5. Expression patterns of the genes linked to the subtrees rooted at GO terms (a) 0005816 and (b) 0000796.

3 Experimental Results

3.1 Clustering in each of three ontologies

In Fig.3, it is partially shown the results of clustering the GO terms by feature importance vectors separately in each ontology (Biological Process, Molecular Function, and Cellular Component).

Feature importance vectors used in the clustering are shown in Fig.4. In Fig.4(B), it can be clearly seen that some features including "Normal Salivary Gland", "Normal Bone Marrow", and "Normal Testis" have significantly high importance in discrimination. It means that the GO terms in the figure are well characterized by the features. On the other hand, the GO terms in Fig.4(M), mainly related to some ion transport, are well-characterized by brain-related tissues ("Normal Cerebellum", "Normal Brain", "Normal Amyglada", "Normal Caudate Nucleus", "Normal hippocampus", and "Normal Thalamus"). These features have relatively low importance in Fig.4(B) and two sets of GO terms in Fig.4(B) and Fig.4(M) have contrasting patterns in feature importance. In Fig.4(C), though correlation among feature importance vectors are not so high as Fig.4(B) and Fig.4(M), we can see some important features including "Normal Adrenal Gland" and "Normal Testis". The GO terms in Fig.4(C) are mainly related to DNA recombination and cell division.

One of the advantages in this method is that it can find hidden relationships among dissimilar

expression patterns. For instance, the expression patterns of two sets of genes linked to the subtrees rooted at GO terms 0005816 and 0000795 are shown in Fig.5. Though the expression patterns in Fig.5(a) and Fig.5(b) are completely different, this proposed method detects their similarity from the viewpoint of feature importance vectors in discrimination from other genes. In addition, even if two GO terms are distantly located in the original ontology (e.g. GO terms 0005816 and 0000795 in Cellular Component), this new method can detect the similarity between them. It can be partially seen in Fig.3; the GO term IDs attached to the end of each line indicate the parents of the GO term in the line, and basically different in each line. It means that the GO terms in a subtree in Fig.3 do not have common parents and placed at different location of the ontology.

3.2 Clustering united data

More interestingly, it is possible to conduct cluster analysis on the united set of GO terms from the three different ontologies since all terms are represented in the same type of feature importance vectors. As Fig.6 illustrates, related to the human microarray data, the result of clustering contains subtrees with mixture of GO terms from three ontologies. In the figure, (B), (M), and (C) stand for Biological Process, Molecular Function, and Celullar Component to which the particular GO term belongs. In this figure, it can be seen that nearly half of the GO terms contain some keywords clearly related to muscle ("muscle", "myofibril", "actin", "striated", and "stress fiber").

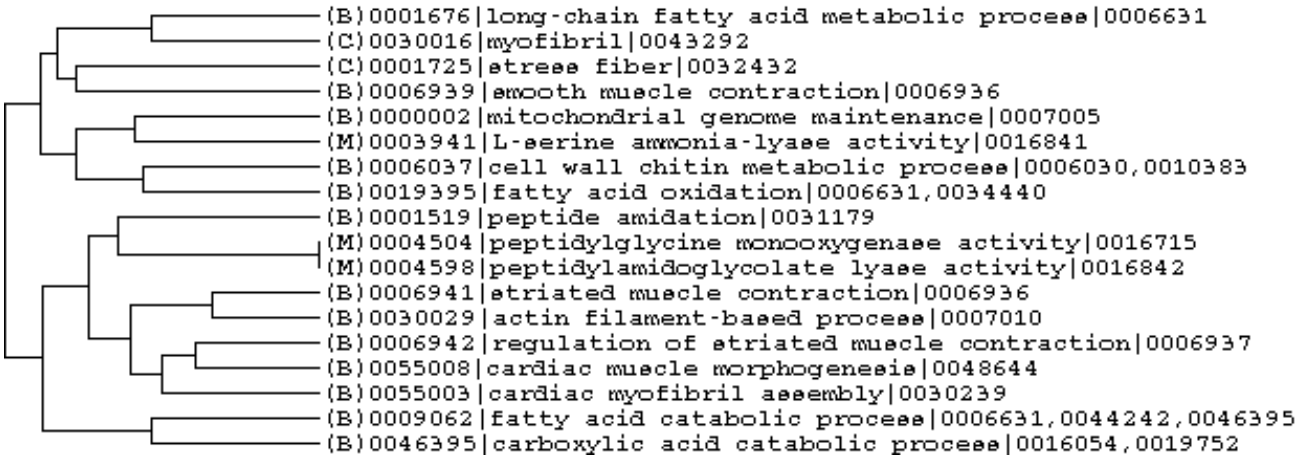

Fig.6. Clustering result on the united data.

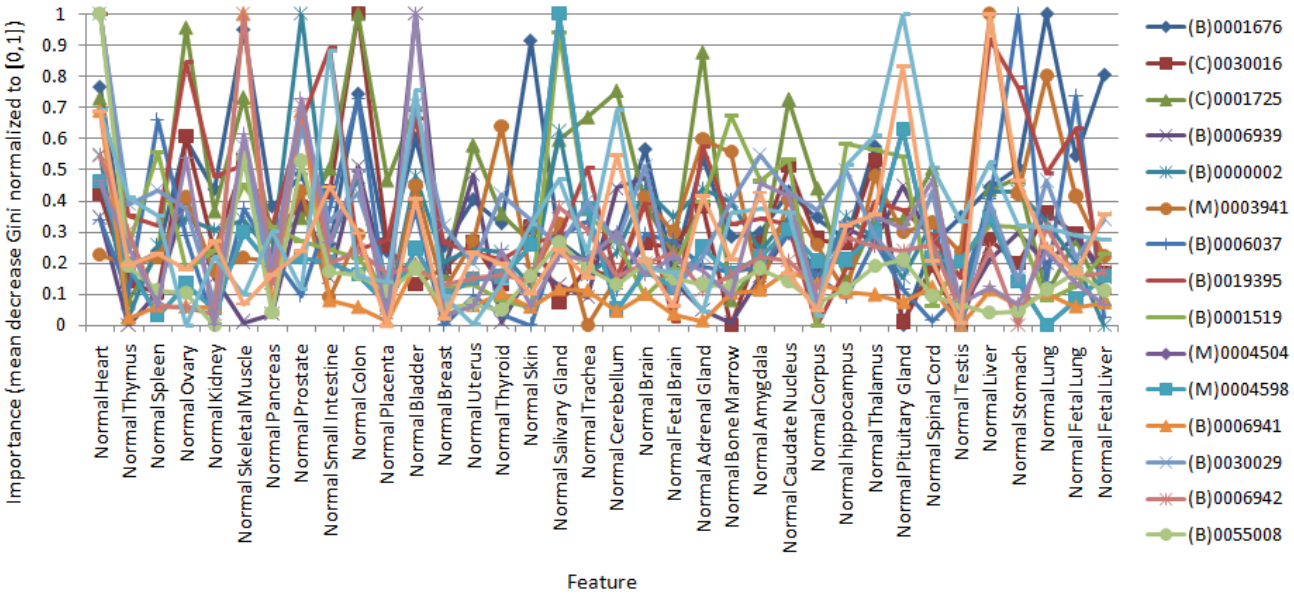

Fig.7. Feature importance vectors of GO terms in Fig.6.

In addition, since two of them contain the keyword "cardiac", this subtree might represent something about heart muscle. If so, it is expected that these GO terms have feature importance vectors with high importance in the tissues related to heart and muscle. This hypothesis can be confirmed by observing the plot of feature importance (scaled between 0 and 1) in Fig.7. In most of the GO terms, the features "Normal Skeletal Muscle" and "Normal Heart" have relatively higher importance in each vector. However, it can be seen that some other features also have high values (e.g. "Normal Colon" and "Normal Bladder"). Further inspection might be needed to know more detailed meanings of correlated GO terms in a certain subtree.

3.3 Clustering yeast time-course data

To show the applicability of this new method, it was conducted the same analysis for a microarray data obtained from a different species (i.e. *Saccharomyces Cerevisiae*) with different kind of dataset (i.e. time-course data). The dataset, GSE3635, consists of 13 samples taken at every 10 minutes after synchronization of yeast cell with alpha factor. 120 minutes correspond to 2 cell cycles. Result of clustering is partially shown in Fig.8.

In Fig.8, most of the GO terms are related to cell division. It is consistent to the well known fact that in time-course data, expression levels of some genes related to cell cycle have periodical pattern of increase and decrease. In addition, it is surprising that feature importance vectors are also periodic. Fig.9 shows that discrimination of genes linked to the subtrees rooted at these GO terms, samples that are near to M phase of the cell cycle are significantly important. In contrast, samples near to S phase have low importance. This result demonstrates that feature importance vectors and correlation among them can be a clue to find hidden relationship among biological concepts in Gene Ontology and microarray data.

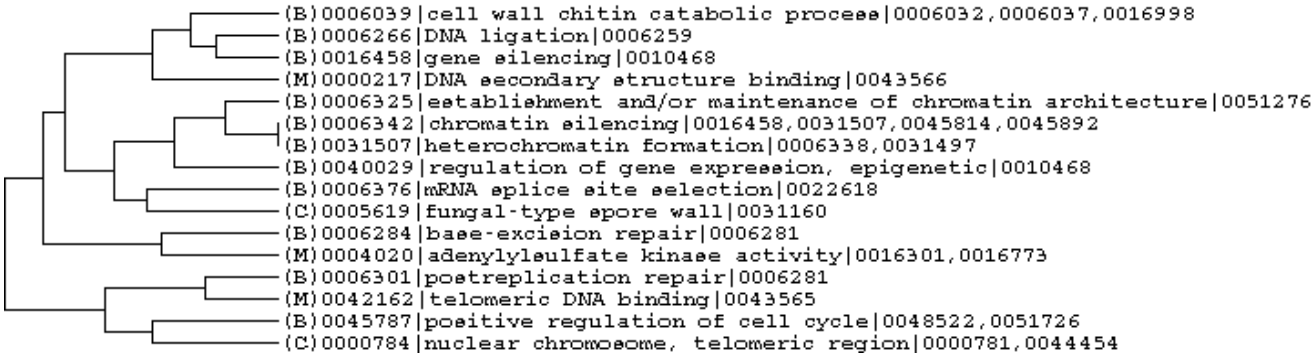

Fig.8. Clustering result on the united data (yeast time-course).

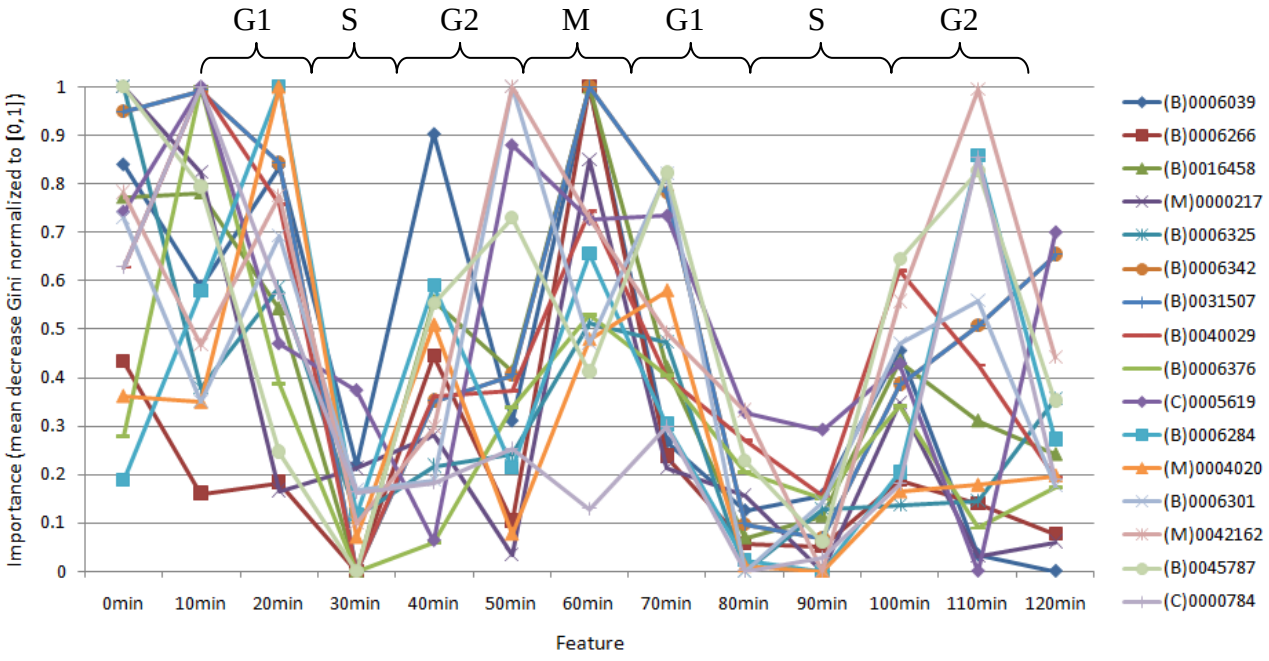

Fig.9. Feature importance vectors of GO terms in Fig.8. M, G1, S, and G2 indicate phases of cell cycle.

4 Conclusion

In this study it was proposed a novel method to analyze gene expression data with GO. By mapping microarray data to GO and performing feature ranking at each GO term, GO terms could be characterized as feature importance vectors with respect to the microarray data. It was also demonstrated that through hierarchical clustering on feature importance vectors, hidden relationships among GO terms can be discovered even if two GO terms are distantly located in the hierarchy of GO. More interestingly, this method could also discover the relationships among GO terms belonging to different ontologies. In future work, this research line will try to integrate this new method based on feature importance vector and various microarray analysis methods based on correlation in expression pattern [23] [24].

Acknowledgements

Part of the research was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), of the Government of Japan.

References:

[1] K.B. Mullis, "The Unusual Origin of the Polymerase Chain Reaction", *Scientific American*, Vol.262, No.4, 1990, pp.56–61, 64–5.

[2] J.L. Weber and E.W. Myers, "Human Whole-Genome Shotgun Sequencing", *Genome Research*, Vol.7, No.5, 1997, pp.401-409.

[3] J. DeRisi, L. Penland, P.O. Brown, M.L. Bittner, et al., "Use of a cDNA microarray to analyse gene expression patterns in human cancer", *Nature Genetics*, Vol.14, No.4, 1996, pp.457-460.

[4] D. Jiang, C. Tang, and A. Zhang, "Cluster Analysis for Gene Expression Data: A Survey", *Transactions on Knowledge and Data Engineering*, Vol.16 No.11, 2004, pp.1370-1386.

[5] Y. Su, T.M. Murali, V. Pavlovic, M. Schaffer, and S. Kasif, "RankGene: Identification of Diagnostic Genes Based on Expression Data", *Bioinformatics*, Vol.19, No.12, 2003, pp.1578 – 1579.

[6] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression", *Bioinformatics*, Vol.20, No.15, 2004, pp.2429 – 2437.

[7] G. Alterovitz, M. Xiang, and M.F. Ramoni, "An Information Theoretic Framework for Ontology-based Bioinformatics," *Information Theory and Applications Workshop*, 2007, pp.16-19.

[8] J.A. Lee, R.S. Sinkovits, D. Mock, et al., "Components of the antigen processing and presentation pathway revealed by gene expression microarray analysis following B cell antigen receptor (BCR) stimulation", *BMC Bioinformatics*, Vol.7, 2006, pp.237.

[9] A. Statnikov, C.F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis", *Bioinformatics*, Vol.21, No.5, 2005, pp.631-643.

[10] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dollinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Mattese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology", *Nature Genetics*, Vol.25, No.1, 2000, pp. 25-29.

[11] E.I. Boyle, S. Weng, J. Gollub, H. Jin, D. Botstein, J.M. Cherry, and G. Sherlock, "GO::TermFinder open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes", *Bioinformatics*, Vol.20, No.18, 2004, pp.3710-3715.

[12] H. Wang, F. Azuaje, O. Bodenreider, and J. Dopazo, "Gene Expression Correlation and Gene Ontology-Based Similarity: An Assessment of Quantitative Relationships", *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology 2004*, Vol.7, No.8, 2004, pp.25-31.

[13] R. Stevens, C. Goble, I. Horrocks, and S. Bechhofer, "Building a bioinformatics ontology using OIL," *IEEE Transactions on Information Technology in Biomedicine*, Vol.6, No.2, 2002, pp.135-141.

[14] H. Menager, Z. Lacroix, and P. Tuffery, "Bioinformatics Services Discovery Using Ontology Classification," *Services, 2007 IEEE Congress on*, Vol.9, No.13, 2007, pp.106-113.

[15] J. Rousu, C. Sauders, S. Szedmak, and J. Shawe-Taylor, "Learning Hierarchical Multi-

Category Text Classification Models", *Journal of Machine Learning Research*, Vol.7, 2006, pp.1601 – 1626.

[16] S. Kiritchenko, S. Matwin, and F. Famili, "Functional Annotation of Genes Using Hierarchical Text Categorization", *B In Proceedings of BioLINK SIG: Linking Literature, Information and Knowledge for Biology*, 2005.

[17] M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C. Walsh, T.S. Furey, M. Ares, and D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines", *Proceedings of the National Academy of Sciences*, Vol.97, No.1, 2000, pp.262-267.

[18] C.J. Stoeckert Jr., H.C. Causton, and C.A. Ball, "Microarray databases: standards and ontologies", *Nature Genetics*, Vol.32, Supplement - Chipping Forecast II, 2002, pp.469-473.

[19] Z. Chen, M. McGee, Q. Liu, and R.H. Scheuermann, "A distribution free summarization method for Affymetrix GeneChip arrays", *Bioinformatics*, Vol.23, No.3, 2007, pp.321-327.

[20] L. Breiman, "Random Forests", *Machine Learning*, Vol.45, No.1, 2001, pp.5-32.

[21] M. Higashihara, J.D. Rebolledo-Mendez, Y. Yamada, and K. Satou, "Application of a Feature Selection Method to Nucleosome Data: Accuracy Improvement and Comparison with Other Methods", *WSEAS Transactions on Biology and Biomedicine*, Vol.5, No.5, 2008, pp.95-104.

[22] K.C. Wiese and C. Eicher, "Graph Drawing Tools for Bioinformatics Research: An Overview," *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, 2006, pp.653-658.

[23] Y. Yamada, Y. Miyata, M. Higashihara, and K. Satou, "Comparison of Cluster Identification Methods for Selection of GO Terms related to Gene Clusters", *WSEAS Transactions on Biology and Biomedicine*, Vol.5, No.3, 2008, pp.54-63.

[24] B. Sathiyabhama and N.P. Gopalan, "Enhanced Correlation Search, Technique For Clustering Cancer Gene Expression Data", *WSEAS, Transactions on Biology and Biomedicine*, Vol.3, No.12, 2006, pp.2477-2484.