Finding Hidden Relationship among Biological Concepts in Gene Ontology

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

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

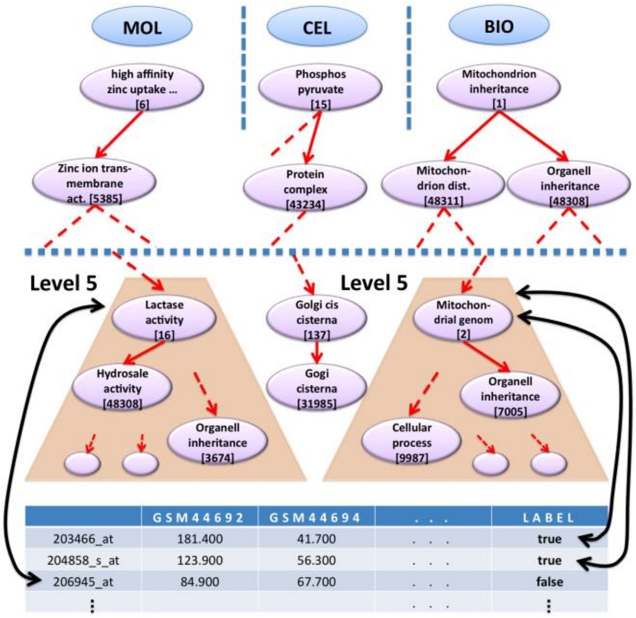
Abstract: - A novel method of microarray data analysis using Gene Ontology (GO) is proposed in this paper. Through the discrimination and feature ranking at each GO term, it was characterized as a feature importance vector with respect to the gene expression pattern contained in a microarray data. In combination with the use of hierarchical clustering on the vectors, it is demonstrated that the method could help to discover hidden relationship among GO terms not only the ones in the same category but also inter-category relationships.

Key-Words: - Gene Ontology, Microarray data, Random forest, Feature ranking, Hierarchical clustering

1 Introduction

Similar to the inventions of PCR [1] and whole genome shotgun sequencing [2] in genome sequencing, microarray opened a new era of high-throughput measurement of multiple gene expression levels at a time. Starting from cDNA microarray developed at Stanford [3], improved methods were proposed and some of them were commercialized. Accordingly, now it is popular to measure all the gene expression levels in the cell sample taken from a species under some conditions, e.g. samples of various tissues, diseased samples including tumor cell, etc. Comparison of two or more (typically 10~100) microarray data obtained from samples under different conditions may reveal biologically meaningful relationship among genes and samples. As a result of gene expression analysis, we frequently find a subset of all the genes in a species, which is potentially meaningful in some context. Then, we need to interpret the meaning of the subset (gene set). GO TermFinder [4] is popular software to do the task. It computes statistical significance between input gene set and terms in Gene Ontology (GO) [5], which is the most comprehensive and authorized hierarchy of biological concepts (controlled vocabulary). Then it outputs the GO terms which characterize the gene set. However, these kinds of tasks require a gene set as an input, and lack comprehensiveness in analysis. In other words, most of the expression data were consumed in the phase of gene expression analysis are discarded before referring GO.

In this paper, we propose a novel method to combine GO and gene expression data obtained from microarray experiment. Given a GO term *x*, we can consider a set of GO terms *descendant(x)* in the subtree rooted at *x*, which might share some conceptual characteristics represented by the GO term *x*. In addition, since links from genes in microarray to GO terms are provided, we can consider a set of genes linked to *descendant(x)*. So, if the microarray data is appropriate to discriminate *descendant(x)* from other GO terms with respect to gene expression pattern, a machine learning algorithm can classify the gene expression patterns linked to *descendant(x)* with high precision (Fig.1). Furthermore, by using a technique called feature ranking, we can evaluate the importance of each feature (i.e. sample) in this discrimination. It means that we can represent a GO term *x* 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 computation on many GO terms, we can detect hidden similarities among GO terms (i.e. similarity of feature importance vectors) in terms of discrimination by expression pattern. Algorithms similar to our method are studied as "hierarchical classification" in mainly in the field of text categorization [6], however we conduct discrimination and feature ranking at many GO terms for characterization of them (not only for prediction of category). The rest of this paper is organized as follows. In section 2, data and software used in the experiment are described. In section 3, experimental results are shown with some analysis and interpretation. Finally, section 4 concludes this paper.


MOL
CEL
BIO
high affinity zinc uptake ... [6]
Phosphos pyruvate [15]
Mitochondrion inheritance [1]
Zinc ion trans-membrane act. [5385]
Protein complex [43234]
Mitochon-drion dist. [48311]
Organell inheritance [48308]
Level 5
Level 5
Lactase activity [16]
Golgi cis cisterna [137]
Mitochon-drial genom [2]
Hydrosale activity [48308]
Gogi cisterna [31985]
Organell inheritance [7005]
Organell inheritance [3674]
Cellular process [9987]

	GSM44692	GSM44694	. . .	LABEL
203466_at	181.400	41.700	. . .	true
204858_s_at	123.900	56.300	. . .	true
206945_at	84.900	67.700	. . .	false
⋮			. . .	⋮

Fig.1. Conceptual figure of the proposed method

2 Materials and Methods

2.1 Gene Ontology (GO)

In [5], it is explained the definition of Gene Ontology (GO), which it is intended to provide a database that relates either biological processes, cellular components, or molecular functions, the three ontologies that form Gene Ontology, into a hierarchical parent-child tree-like directed acyclic graph (DAG) data structure containing only information about its terms (in each ontology) and their relationships to other terms. Currently there are more than 26,000 terms in the three ontologies of all the species that are being included in them. This number changes constantly when terms are added or corrected.

Each term can be related to another term with the relation "is_a", and "part_of" (though, there are different relationships: "consider", and "relationship" among others). The result is a DAG where its elements are interrelated in a hierarchical structure.

At the top level of the hierarchical structure of Gene Ontology, there are three GO terms, each one corresponding to the top level of its respective ontology. The database of GO can be obtained in the GO Consortium site (http://www.geneontology.org.)

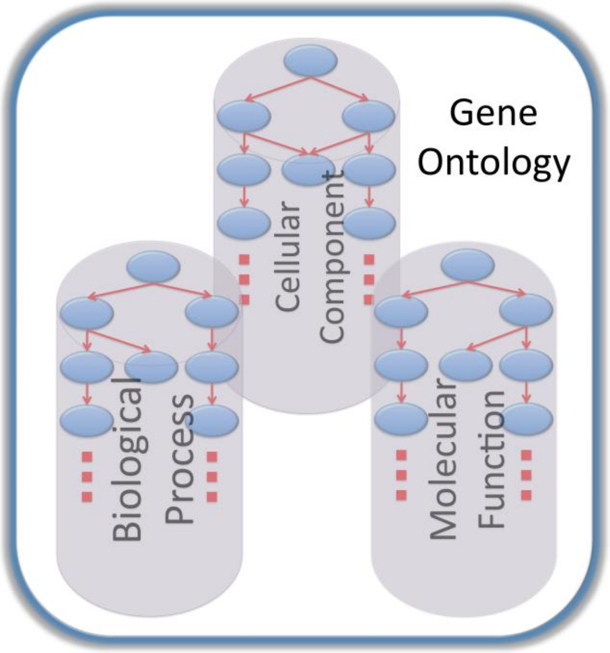


Fig.2. Gene Ontology

2.2 Microarray data

The expression of most of genes is given in the transcription and translation, having protein biosynthesis as a result, and is indispensable for life. Much of the information contained in thousands of the expressions of genes can be obtained by microarray techniques. There are different techniques that permit the extraction of the information, like gene expression profiling, comparative genomic hybridization, and SNP detection, among others [7].

There are also different databases [8] 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. The micro array data consists of a series of affixed DNA segments, known as probes or reporters which measure the different genes that are put on the tablets or chips, giving to each one a different value to each gene and probe. In this paper, we used a microarray data GSE2361 which contains a matrix with 36 samples of human tissue as columns and 22,281 human genes as rows. Link information from genes to GO terms was extracted from the annotation file for the platform GPL96, Affymetrix GeneChip Human Genome U133 Array Set HG-U133A, used to measure the expression data in GSE2361.

2.3 Preparation of positive and negative examples for discrimination and feature ranking

Though it is possible to conduct discrimination and feature ranking at all nodes in GO, in this study we used only the GO terms at some level (the number of links from 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), this limitation is useful for 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 is normalized. The normalization method used the DFW algorithm [9], in R, obtaining data that became statically better suitable to be processed.

After the normalization, data preparation was separately performed on Biological Process, Molecular Function, and Cellular Component. Suppose that a set of GO terms $T_{BP} = \{t_1,...,t_n\}$ is taken from a certain level of Biological Process (in this study, level 5 was adopted). First, we prepare examples corresponding to each GO term t_k in T_{BP}. The examples are gene expression data for the genes $G(\{t_k\})$, where $G(\{t_k\})$ denotes all the genes linked to $descendant(t_k)$. Then, for each t_k we prepare positive and negative examples as follows: by attaching class labels “true” and “false” to all the examples in $G(\{t_k\})$ and $G(T_{BP}\text{-}\{t_k\})$, we obtain positive and negative examples for feature ranking, respectively.

2.4 Feature ranking by random forest

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 random forest algorithm. Random forest [10] is a kind of ensemble learning algorithm developed by L. Breiman. Besides its ability of classification, we used it for feature ranking: to obtain importance of feature, i.e. contribution to discriminate positive and negative examples. As an implementation of random forest, randomForest package for R was adopted. From given examples, it performs training and as a by-product, it outputs a value called mean decrease Gini for each feature, which can be used as an importance of a feature. In this study, each GO term in level five is characterized as a vector of feature importance. 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), we can say that t_k potentially has close relationship to the features.

2.5 Clustering GO terms

To interpret the result of GO term characterization with respect to the microarray data GSE2361, we conducted cluster analysis on the feature importance vectors. Among various method of cluster analysis exists, hierarchical clustering was adopted. 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 Cluster 3.0 software, and the result was visualized by Java TreeView software.

3 Experimental Results

3.1 Clustering in each of three ontologies

Fig.3~5 partially show the result of clustering GO terms (feature importance vectors) separately in each of Biological Process, Molecular Function, and Cellular Component, respectively. Feature vectors of the GO terms in a subtree are highly correlated. Wide variation of the GO term number(s) at the end of each line, parent(s) of a GO term, indicates that the GO terms in a tree came from different places in GO.

3.1 Clustering united data

More interestingly, we can conduct cluster analysis on the united set of GO terms from three different ontologies since all terms are represented in the same type of feature importance vectors. As Fig.6 illustrates, here we can see the subtrees with mixture of GO terms from three categories. In the figure, (B), (M), and (C) stand for Biological Process, Molecular Function, and Celullar Component to which a 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"). In addition, since two of them contain the keyword "cardiac", this subtree might represents some knowledge about heart muscle. If so, it is expected that these GO terms have feature importance vectors with high importance in the tissues related heart and muscle. This hypothesis can be confirmed by the plot of feature importance (normalized to 0~1) in Fig.7. In most of the GO terms, the features "Normal Skeltal 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 the detailed meaning of correlated GO terms in a subtree.

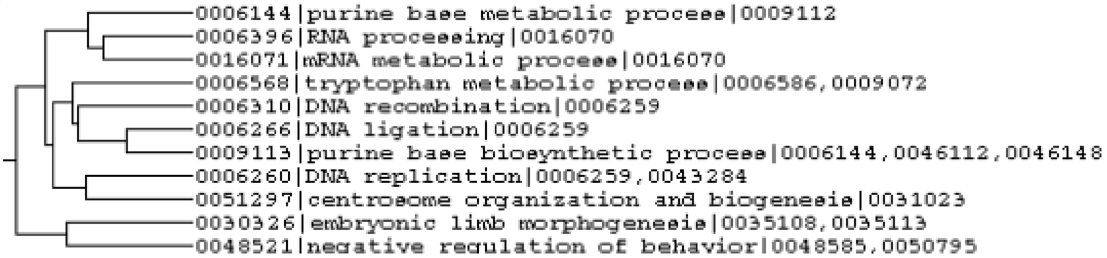

0006144|purine base metabolic process|0009112
0006396|RNA processing|0016070
0016071|mRNA metabolic process|0016070
0006568|tryptophan metabolic process|0006586,0009072
0006310|DNA recombination|0006259
0006266|DNA ligation|0006259
0009113|purine base biosynthetic process|0006144,0046112,0046148
0006260|DNA replication|0006259,0043284
0051297|centrosome organization and biogenesis|0031023
0030326|embryonic limb morphogenesis|0035108,0035113
0048521|negative regulation of behavior|0048585,0050795

Fig.3. A subtree of GO terms from the result of clustering in Biological Process

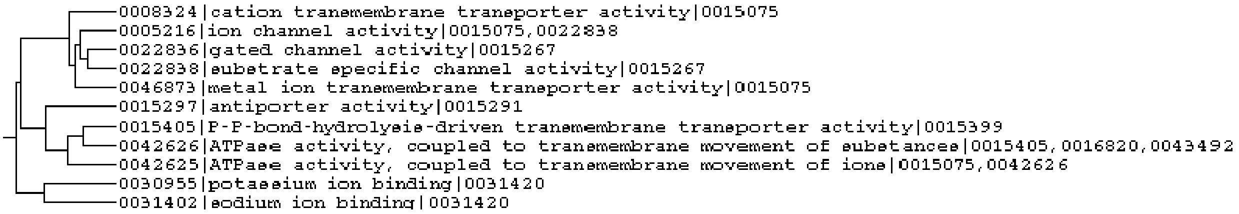

0008324|cation transmembrane transporter activity|0015075
0005216|ion channel activity|0015075,0022838
0022836|gated channel activity|0015267
0022838|substrate specific channel activity|0015267
0046873|metal ion transmembrane transporter activity|0015075
0015297|antiporter activity|0015291
0015405|P-P-bond-hydrolysis-driven transmembrane transporter activity|0015399
0042626|ATPase activity, coupled to transmembrane movement of substances|0015405,0016820,0043492
0042625|ATPase activity, coupled to transmembrane movement of ions|0015075,0042626
0030955|potassium ion binding|0031420
0031402|sodium ion binding|0031420

Fig.4. A subtree of GO terms from the result of clustering in Molecular Function

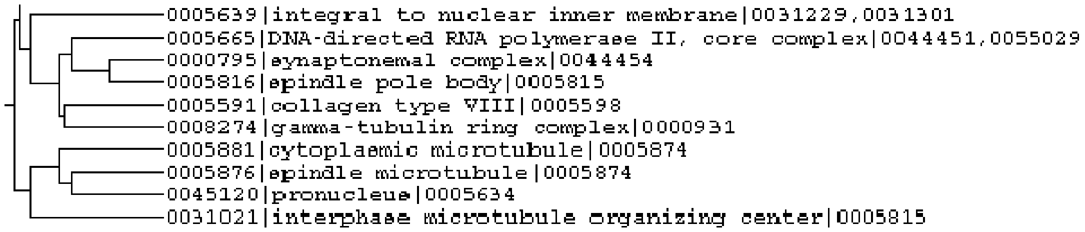

0005639|integral to nuclear inner membrane|0031229,0031301
0005665|DNA-directed RNA polymerase II, core complex|0044451,0055029
0000795|synaptonemal complex|0044454
0005816|spindle pole body|0005815
0005591|collagen type VIII|0005598
0008274|gamma-tubulin ring complex|0000931
0005881|cytoplasmic microtubule|0005874
0005876|spindle microtubule|0005874
0045120|pronucleus|0005634
0031021|interphase microtubule organizing center|0005815

Fig.5. A subtree of GO terms from the result of clustering in Cellular Component

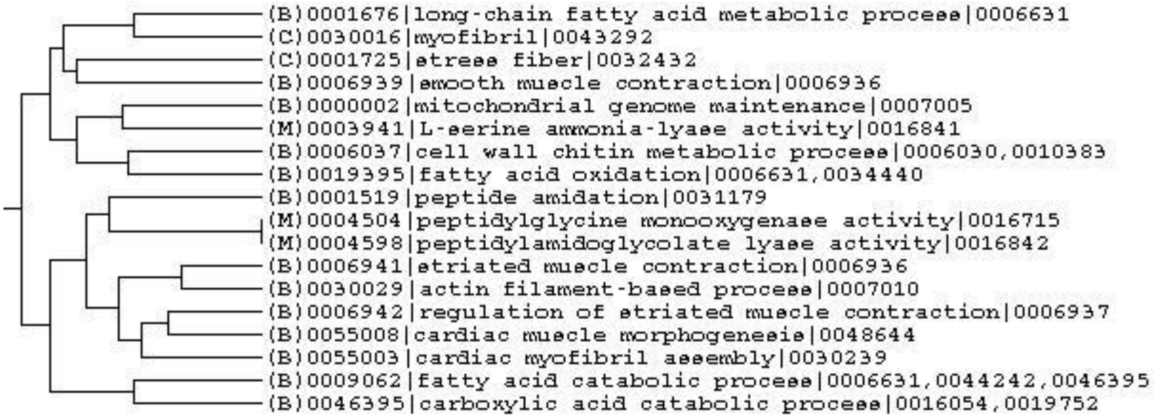

(B)0001676|long-chain fatty acid metabolic process|0006631
(C)0030016|myofibril|0043292
(C)0001725|stress fiber|0032432
(B)0006939|smooth muscle contraction|0006936
(B)0000002|mitochondrial genome maintenance|0007005
(M)0003941|L-serine ammonia-lyase activity|0016841
(B)0006037|cell wall chitin metabolic process|0006030,0010383
(B)0019395|fatty acid oxidation|0006631,0034440
(B)0001519|peptide amidation|0031179
(M)0004504|peptidylglycine monooxygenase activity|0016715
(M)0004598|peptidylamidoglycolate lyase activity|0016842
(B)0006941|striated muscle contraction|0006936
(B)0030029|actin filament-based process|0007010
(B)0006942|regulation of striated muscle contraction|0006937
(B)0055008|cardiac muscle morphogenesis|0048644
(B)0055003|cardiac myofibril assembly|0030239
(B)0009062|fatty acid catabolic process|0006631,0044242,0046395
(B)0046395|carboxylic acid catabolic process|0016054,0019752

Fig.6. A subtree of muscle-related GO terms from the result of clustering on the united data

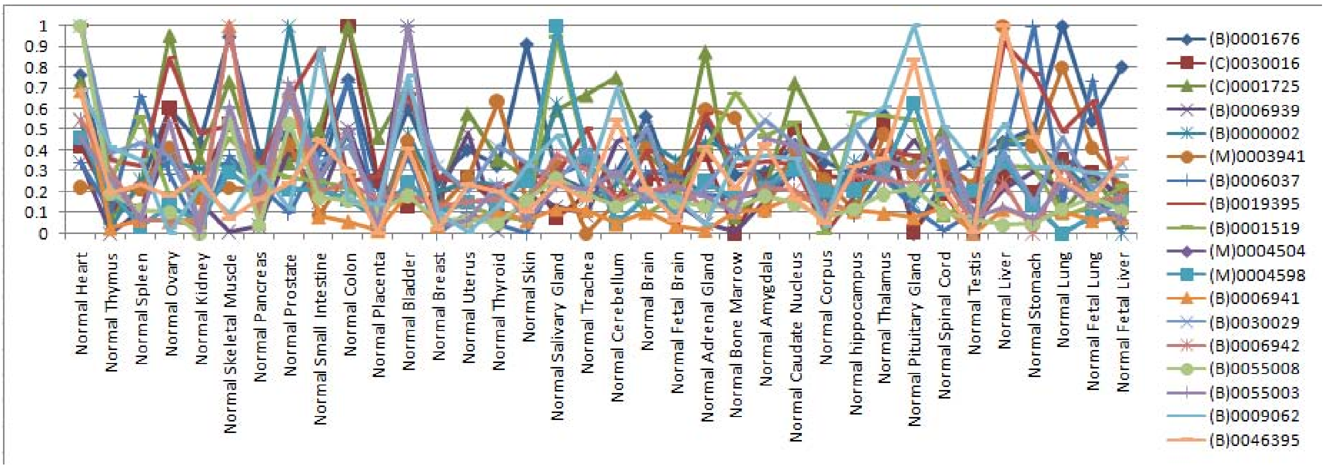


Fig.7. Similarity of feature importance vectors

4 Conclusion

In this study, we 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, we could characterize GO terms as feature importance vectors with respect to the microarray data. It was also demonstrated that through hierarchical clustering on feature importance vectors, we might discover hidden relationship among GO terms even if two GO terms are distantly placed in the hierarchy of GO. More interestingly, this method could also discover the relationship among GO terms belonging to different categories.

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*, 1990.

[2] J.L Weber, and E.W. Myers, Human Whole-Genome Shotgun Sequencing, *Genome Research*, 1997, 7/5 401 - 409.

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

[4] Elizabeth I. Boyle, Shuai Weng, J. G. H. J. D. B. J. M. C. Sherlock, G., GO::TermFinderopen source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes, *BioInformatics*, 2004, 20 3710 - 3715.

[5] 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*, 2000, 25, pp. 25–29.

[6] Kiritchenko, S., M. S. Famili, F., Functional Annotation of Genes Using Hierarchical Text Categorization, *BioLINK SIG Meeting on Text Data Mining at ISMB'05*, 2005.

[7] M.P.S. Brown, W. Noble 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*, 45 1, 2001, pp 5 - 32.

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

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

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