International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 3 (2011) pp. 256-263

A Novel Gene Analysis Method for Biomarker Mining in DNA Microarray Data

Kentaro Fukuta[1], Tomomasa Nagashima[2], Takashi Uozumi[3] and Yoshifumi Okada[4]

[1]Satellite Venture Business Laboratory, Muroran Institute of Technology,
27-1, Mizumoto-cho, Muroran 050-8585, Japan
fukuta@mail.svbl.muroran-it.ac.jp

[2]College of Information and Systems, Muroran Institute of Technology,
27-1, Mizumoto-cho, Muroran 050-8585, Japan
nagasima@epsilon2.csse.muroran-it.ac.jp

[3]College of Information and Systems, Muroran Institute of Technology,
27-1, Mizumoto-cho, Muroran 050-8585, Japan
uozumi@epsilon2.csse.muroran-it.ac.jp

[4]College of Information and Systems, Muroran Institute of Technology,
27-1, Mizumoto-cho, Muroran 050-8585, Japan
okada@epsilon2.csse.muroran-it.ac.jp

Abstract: **Preventing, diagnosing, and treating disease is greatly facilitated by the availability of biomarkers. Recent improvements in bioinformatics technology have facilitated large-scale screening of DNA arrays for candidate biomarkers. Here we discuss a gene analysis method that we call the *LEAve-one-out Forward selection method* (LEAF) for discovering informative genes embedded in expression data, and propose an additional algorithm for extending LEAF's capabilities. An iterative forward selection method incorporating the concept of leave-one-out cross validation (LOOCV), LEAF provides a discrimination power score (DPS) for genes. We show that LEAF identifies genes that correspond to known biomarkers. Therefore, our method should provide a useful bioinformatics tool for biomedical, clinical, and pharmaceutical researchers.**

Keywords: biomarkers, data mining, gene expression profiles, cancer classification

I. Introduction

Recent progress in bioinformatics technology has facilitated large-scale screening for candidate biomarkers [1]. A biomarker, as the name implies, is a cell-derived substance such as a gene, protein or enzyme that can be used to elucidate physiological or pathological process [2]. In our previous study, we have proposed a novel method called LEAve-one-out Forward selection method (LEAF) for analysis of gene expression data [12]. This method enabled us to construct a ranking system of informative genes using a parameter reflecting the efficiency of the class discriminant designated the Discriminant Power Score (DPS).

We applied LEAF to three public leukemia datasets (ALL/AML, ALL/MLL, and MLL/AML) [3, 4]. The results show that our method yields a stable discriminant result with 100% accuracy using a three-gene set. Furthermore, some genes with high DPS values are cancer-related genes (top-h genes), as clarified by research in recent years.

Nevertheless, two problems remain to be resolved, namely: (1) We have not selected a criterion for defining the h-value. (2) The candidate list of associated genes is insufficient to assign a discrete biological function (correlation and causal relation between genes).

Here we briefly introduce LEAF and then propose a solution to address these problems. Thus, using public gene function database, we propose a simple and straightforward method for determining the top-h genes (h-value) and conduct a biological functional analysis of the genes. Subsequently, we designed a gene analysis framework for assembling the information supporting consideration of a biological process.

II. LEAF: LEAve-one-out Forward selection method

A. Materials: Cancer dataset

We used three well-known leukemia datasets provided by Armstrong et al., which includes acute lymphocytic leukemia (ALL), mixed lineage leukemia (MLL), and acute myelogenous leukemia (AML) [3]. These datasets are available at the Broad Institute [4]. Details of the datasets are summarized in Fig. 1A.

Fig. 1B presents two datasets are arranged in the form of a data matrix. The matrix size is $CN \times TG$, where CN denotes $Class1\_N + Class2\_N$. Furthermore, $Class1\_N$ and $Class2\_N$, respectively, represent the number of samples in

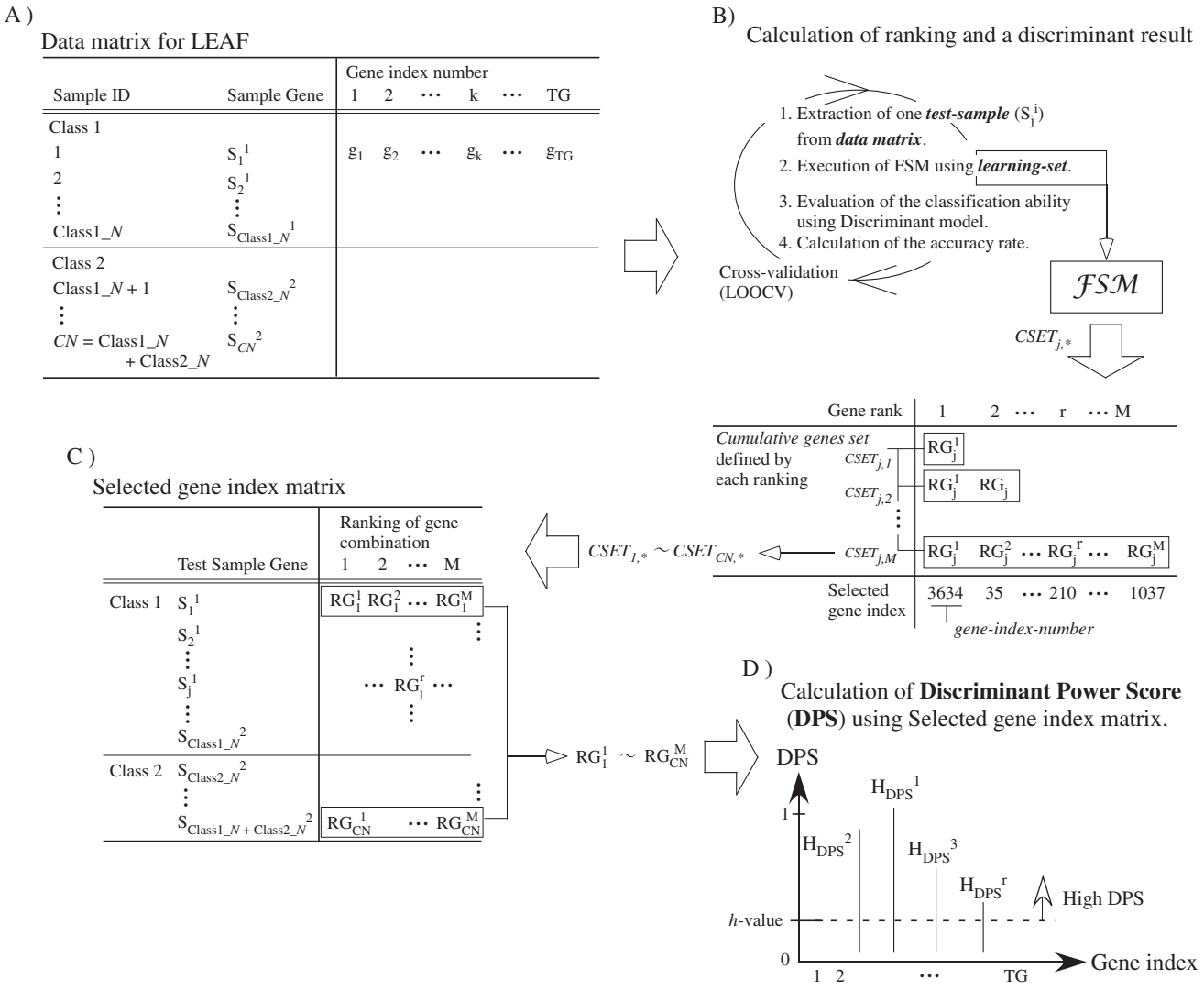

Figure. 2: Overview of LEAF's methodology

A)

Class (#Genes : 12,582)	Samples : *N*
ALL (Acute lymphocytic leukemia)	24
MLL (Mixed lineage leukemia)	20
AML (Acute myelogenous leukemia)	28

B)

Dataset name	Class 1 (*N*)	Class 2 (*N*)
ALL vs. AML	ALL (24)	AML (28)
ALL vs. MLL	ALL (24)	MLL (20)
MLL vs. AML	MLL (20)	AML (28)

Figure. 1: Cancer dataset "Leukemia dataset" [4]

Class 1 and Class 2, and g_k $(k = 1, 2, \ldots, TG)$ corresponds to a gene expression value, and TG signifies the total number of genes: TG = 12,582.

B. LEAF: LEAve-one-out Forward selection method

We have proposed a robust and accurate gene selection method based on forward selection called forward selection method (FSM) [10]. To satisfy a maximal variance ratio (F-value) between two disease classes, FSM cumulatively selects gene one-by-one and ultimately identifies a set of genes (a gene ranking) that is informative for disease classification. In fact, LEAF is an iterative FSM inspired by leave-one-out cross validation (LOOCV) [11]. Details of the algorithm have been published [12]. Figure 2 outlines the method. First, one test sample is taken from the dataset. Then the remaining samples are used as a learning set. Subsequently, we apply FSM to the learning set and obtain a gene ranking. These steps are repeated for every test sample.

Finally, we extract a highly robust set of genes in a classification based on discriminant power, called DPS. DPS is a parameter of the class discriminant ability defined for all genes. $DPS(k)$ $(1 \leqq k \leqq TG)$ represents the DPS value of the gene with the k-th gene-index-number.

Figure 7 displays the DPSs of genes calculated from the respective pairs of the leukemia datasets. The horizontal axis shows the gene index number, and the vertical axis indicates the DPS given for each gene. The DPS graph can help visualize genes' statistical importance. Genes with higher DPSs can be regarded as those contributing more significantly to discrimination between the classes. That is, significant genes

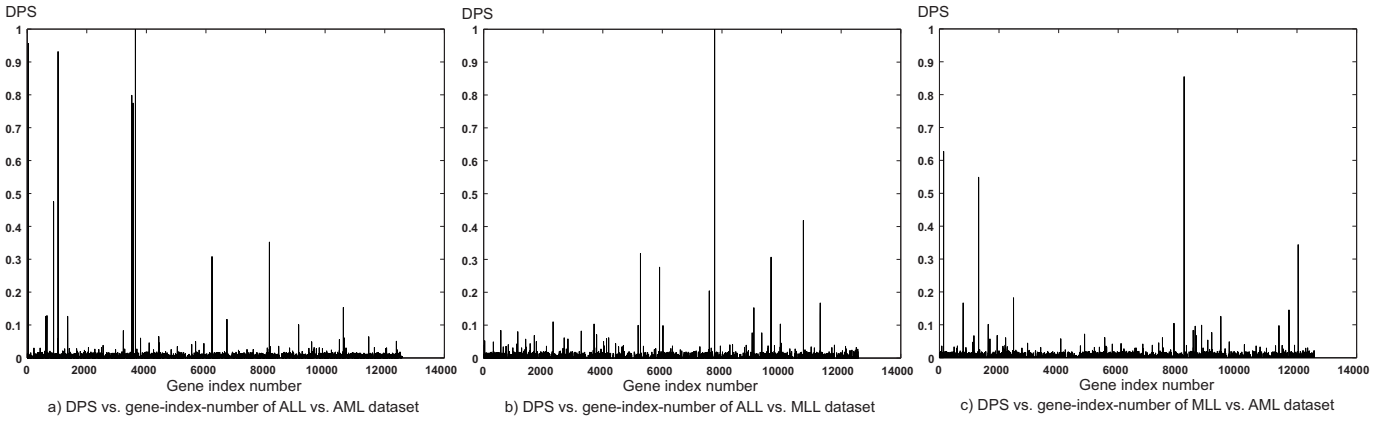

Figure. 3: DPS vs. gene-index-number of leukemia dataset.

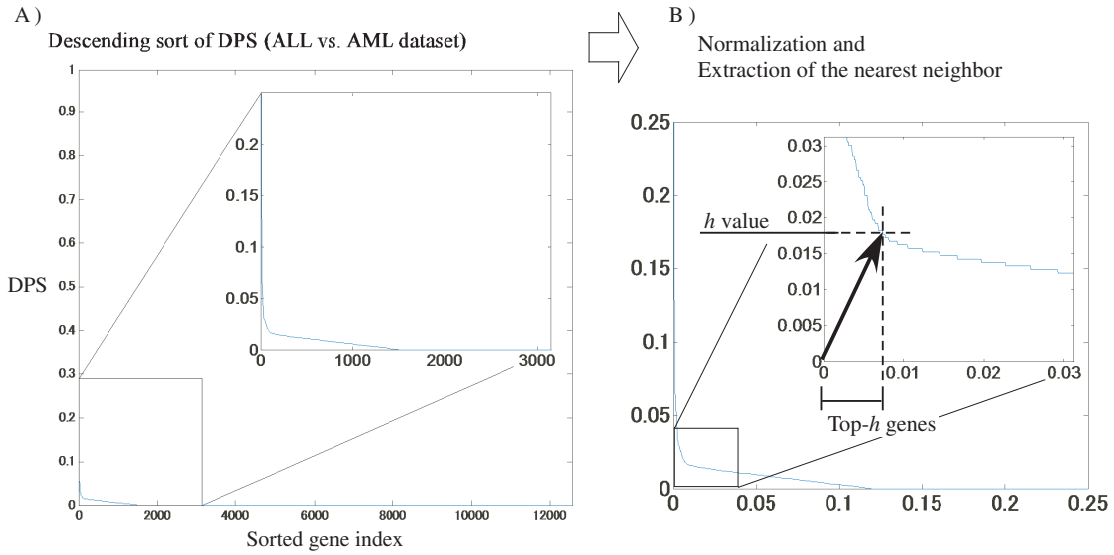

Figure. 4: Outline for defining h-value

are represented as peaks in the DPS graph.

C. Determination method of h-value (top-h genes)

Because previous work [12] has not provided any criterion (cut-off threshold) for obtaining a set of discriminative genes, here we introduce an interactive method for extracting the top-h genes that are used to generate a final discriminant function. The identification method of the h-value is illustrated in Fig. 4. The h-value is calculated by the following steps:

1. Descending sort of DPS (Fig. 4A).
2. Decision of h-value.
   (a) Normalize the horizontal and vertical axes by dividing by their respective maximum values (Fig. 4B).
   (b) Find the shortest Euclidean distance on the DPS graph to the origin. The abscissa value of the point is called the h-value.
   (c) Extract the set of genes having DPSs $\geqq$ h-value.
   (d) Recreate a DPS graph using only the gene set obtained in Step (c).
   (e) Repeat from Step (a) to Step (d) unless the number of points is 1 or all points take an identical distance.

Thus, we employ the nearest neighbor point (h-value) from the origin for detecting drastic curvature in the descending sorted-DPS graph. We can then extract genes having high DPSs, which are ranked higher than the h-value. This method narrows down top-h-genes by interactively iterating the above procedure. Obviously, many iterations drastically decrease gene numbers, potentially eliminating biologically meaningful genes. In this study, therefore, the number of iterations in the decision of h-value is set to two (the respective h-values are referred to as $h1$ and $h2$).

III. Evaluation and Verification

Figure 5 shows the evaluation procedure of our method, $HCSET_i(i = 1, 2, \ldots, h)$ represents a cumulative gene set of the top-i in the DPS ranking. For each $HCSET_i$, we construct discriminant models by a LOOCV and calculate a classification accuracy (Fig. 5A). Finally, a Discriminant result matrix as shown in Fig. 5B is generated, and the performance of our method is evaluated by using Accuracy rates derived

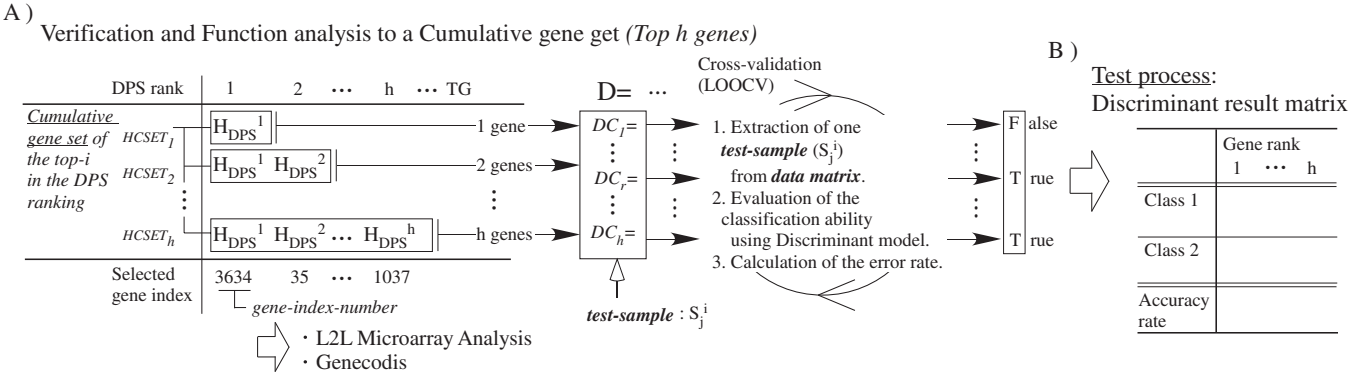

Figure. 5: Evaluation method

from each $HCSET_i$. In addition, we conduct a function enrichment analysis for $HCSET_h$ and discuss the biological significance of genes included in

A. *Classification accuracy and DPS*

We generate discriminant result matrices for every pair of the three datasets (ALL, MLL and AML). Figure 6 shows the classification accuracies obtained using the cumulative gene sets ($HCSET_1, HCSET_2, \ldots, HCSET_{20}$) for each pair of datasets. From these figures, we can see that perfect classifications are stably achieved by using three or more genes in all the pairs. Figure 7 exhibits the DPSs of genes calculated for the respective pairs of datasets. The genes with higher DPSs can be regarded as ones having greater contribution to discriminating between the classes.

IV. Biological function analysis

The h-values of each dataset are presented in Table 1. Ideally, it is preferred that the extracted genes provide biologically useful information in addition to imparting high discriminatory power to different classes. We conducted a biological function analysis of gene group in reference to the Gene ontology tool [5, 6] and the University of Washington's L2L microarray analysis tool [7]. Below we focus on the top-$h2$ genes' biological function.

In the L2L program, a p value for the significance of overlap between the given list and the function list of the databases is calculated by using the binomial distribution. Tables 3, 4 and 5 summarizes the L2L results. In the three datasets, we can observe that functions related to human cancer, such as colon carcinoma, gastric cancer, and breast cancer, exhibit statistical significance.

Table. 2 summarizes the primary functions of the top-$h2$ genes obtained using Gene ontology. As expected, genes related to leukemia in addition to leucocyte communication, such as TCL1A, RPL38, CALLA, and IL8RB [8], are selected from every dataset pair. In particular, it should be noted that ribosomal protein L38 (RPL38) is highly expressed in pancreatic cancer cell lines [9].

V. Gene analysis Framework

For basic biomedical and translational research purposes, it is not sufficient to list informative candidate genes without knowing the pathways in which their products participate. Our method for mining biomarkers is based upon differential gene expression analysis, thereby providing functional information. We propose this as a gene-analysis framework, which applies LEAF. An overview of the framework (Fig. 8) illustrates the processes by which it operates.

Table 1: h and DPS values of Leukemia Dataset

Dataset	$h1$	DPS	$h2$	DPS
ALL vs. AML	104	0.0168	10	0.1287
ALL vs. MLL	123	0.0192	11	0.1030
MLL vs. AML	139	0.0179	9	0.1042

1. Analysis of the dataset using LEAF, and display of DPS (Figs. 8A and B).
2. Calculation of h-values (Fig. 8C).
3. Extraction of the genes based on the h-value (Fig. 8D).
4. Analysis of top-$h2$ genes (Fig. 8E).
   (a) Construction of a discriminant model.
   (b) Output of a summary (*i.e.*, Table 2).
5. Gene-network analysis for top-$h1$ genes.
6. Output of the dependency rules based on probabilistic reasoning.

Interaction between genes can be inferred using the model of dependency structure (correlation and causal relationship). Figure 8G shows that gene-network analysis expresses a dependency using a graphical structure.

A graph node is a gene; an arrow represents the existence of dependency between nodes. One method of building gene networks uses a Bayesian network [13, 14]. We can apply probabilistic reasoning [15] and search for the biological process that supports discovery of a biomarker. Moreover, in this framework, we use biological ontology for the construction and interpretation of a Bayesian network.

Gene Ontology (GO) is a popular gene function database consisting of three independent ontologies: Biological process, molecular functions, and cellular components. Each node of the ontology corresponds to a certain biological function and includes one or more genes.

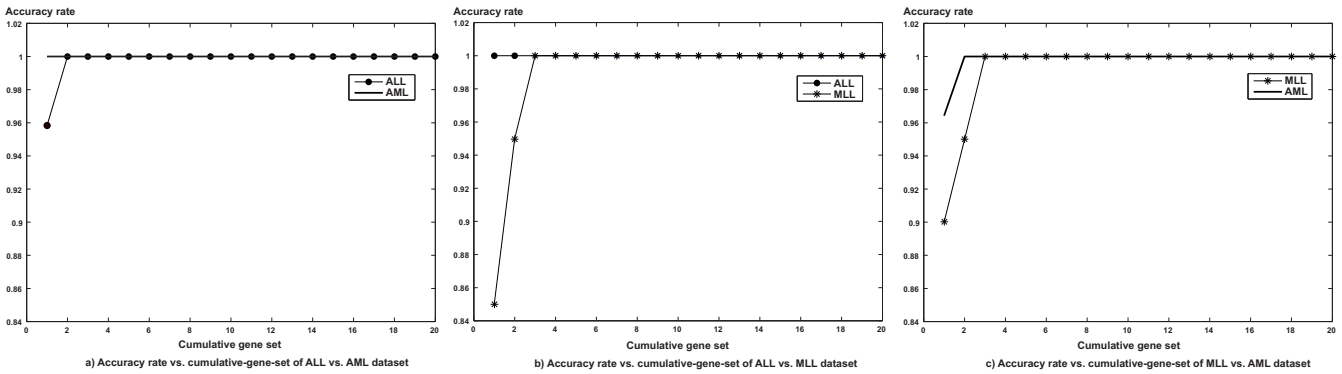

Figure. 6: Accuracy rate vs. cumulative gene set of leukemia dataset

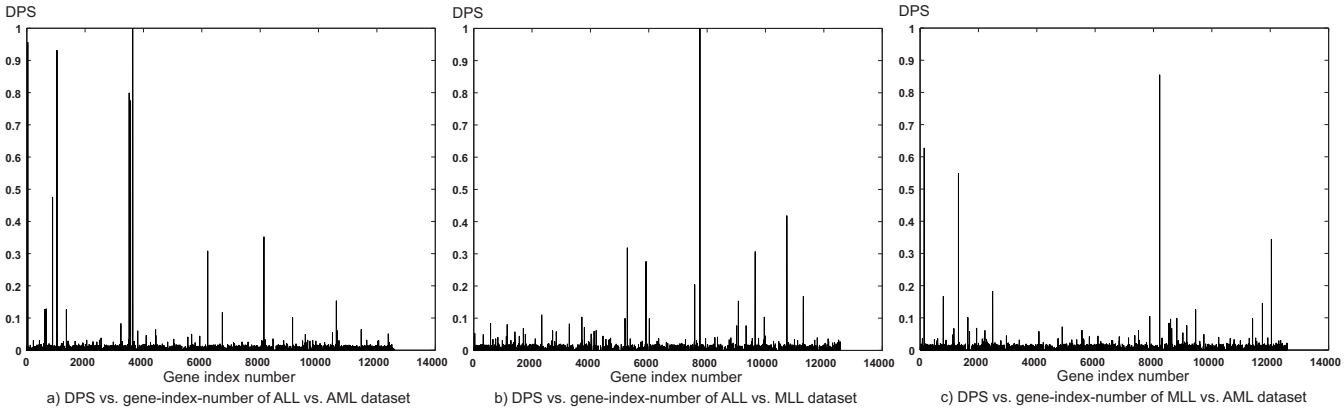

Figure. 7: DPS vs. gene-index-number of leukemia dataset

Actually, GO does not have only a common vocabulary in biological science. It does provide a classification tree of the concept of generalization and specialization (*i.e.*, the "part-of link" for which biological process A consists of a molecular interaction X and Y.).

We prepare software agents [16, 17] that searches for a candidate biological process to built, BN. They change the node value of a gene network variously, and perform probabilistic reasoning. We store the candidate of a biological process sought by the agent as a general knowledge format (OWL ontology).

VI. Conclusion

LEAF is an iterative FSM incorporating the concept of LOOCV; it also provides a DPS of genes. Moreover, we can determine the top-h according to the distribution of DPS value for each dataset using a simple algorithm for determining h-values. The h-values can be used as criteria for identifying candidate or informative genes. Our method shows that the biological functions of extracted genes correspond well with those reported in the literature. Finally, we propose a gene analysis framework for using LEAF for basic biomedical research and drug discovery. From these results, we expect that our method will provide a powerful tool to explore biomarker candidates and as a new method for disease diagnosis.

We plan to develop an automatic detection method of h-value based on information criterion such as AIC (Akaike Information Criterion) [18] and evaluate the usefulness of the method by applying it to other datasets.

Acknowledgments

A part of this work was supported by Promotion for Young Research Talent and Network from Northern Advancement Center for Science & Technology (NOASTEC Japan) and Grant-in-Aid for Young Scientists (B) No.21700233 from MEXT Japan.

References

[1] T.-R. Golub, D.-K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, M. J.-P., H. Coller, L. M.-L., J.-R. Downing, M.-A. Caligiuri, et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, 286(5439):531–537, 1999.

[2] Foundation for the National Institutes of Health, "The biomarkers consortium," `http://www.biomarkersconsortium.org/`, 2007.

[3] S. A. Armstrong, J. E. Staunton, L. B. Silverman, et al., "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *BioinNature Genetics*, vol. 30, no. 1, pp. 41–47, 2001.

[4] Broad Institute, "Broad institute of mit and harvard," `http://www.broadinstitute.org/`, 2010.

Table 2: Summary of the top $h2$ genes ranked by DPS

DPS ranking	A) ALL vs. AML dataset		B) ALL vs. MLL dataset		C) MLL vs. AML dataset	
h2 value	Input name / Gene name	Description	Input name / Gene name	Description	Input name / Gene name	Description
1	39318_at / TCL1A	**T-cell leukemia/lymphoma 1A**	33412_at / ---	---	35307_at / GDI2	GDP dissociation inhibitor 2
2	AFFX-M27830_5_at / ---	---	1984_s_at / ARHGDIB	Rho GDP dissociation inhibitor (GDI) beta	31397_at / ---	---
3	34085_at / RPL38	**ribosomal protein L38**	34699_at / CD2AP	CD2-associated protein	35083_at / FTL	ferritin, light polypeptide
4	38955_at / AVPR1A	arginine vasopressin receptor 1A	39857_at / STX11	syntaxin 11	664_at / IL8RB	**interleukin 8 receptor, beta**
5	39243_s_at / PSIP1	PC4 and SFRS1 interacting protein 1	36897_at / ---	---	35896_at / ---	---
6	33099_at / FUT5	fucosyltransferase 5 (alpha (1,3) fucosyltransferase)	32755_at / ACTA2	actin, alpha 2, smooth muscle, aorta	33008_at / ---	---
7	34863_s_at / SCCPDH	saccharopine dehydrogenase (putative)	1389_at / MME	membrane metallo-endopeptidase, **common ALL antigen (CALLA)**	979_g_at / ---	---
8	37913_at / DHFR	dihydrofolate reductase	38083_at / NOTCH2	Notch homolog 2 (Drosophila)	39175_at / PFKP	phosphofructokinase, platelet
9	33213_g_at / RRBP1	ribosome binding protein 1 homolog 180kDa (dog)	1388_g_at / VDR	vitamin D (1,25- dihydroxyvitamin D3) receptor	33889_s_at / ---	---
10	32397_r_at / ERCC2	excision repair cross-complementing rodent repair deficiency, complementation group 2	35383_at / ---	---		
11			39631_at / EMP2	epithelial membrane protein 2		

Table 3: Function enrichment analysis (L2L) for the top-$h2$ genes of ALL vs AML dataset

Function name	*p*-Value	Description
elongina_ko_dn	5.33e-04	Downregulated in MES cells from elongin-A knockout mice
uvb_nhek1_c1	4.08e-03	Upregulated by UV-B light in normal human epidermal keratinocytes, cluster 1
senescence_rep-ind_dn	5.11e-03	Down-regulated in models of both replicative (high-passge human foreskin fibroblast) and induced (repression of E7 in HeLa) cellular senescence.
hdaci_colon_tsa48hrs_dn	5.19e-03	Downregulated by TSA at 48 hrs in SW260 **colon carcinoma cells**

Table 4: Function enrichment analysis (L2L) for the top-$h2$ genes of ALL vs MLL dataset

Function name	*p*-Value	Description
hdaci_colon_cur12hrs_dn	4.57e-03	Downregulated by curcumin at 12 hrs in SW260 **colon carcinoma cells**
refractory_gastric_dn	9.12e-03	Downregulated in samples of **gastric cancer** refractory to 5-FU/cisplatin treatment, compared to chemosensitive controls

Table 5: Function enrichment analysis (L2L) for the top-$h2$ genes of MLL vs AML dataset

Function name	*p*-Value	Description
breastca_three_classes	8.83e-04	Gene set that can be used to differentiate BRCA1-linked, BRCA2-linked, and sporadic primary **breast cancers**
hypoxia_review	3.04e-03	Genes known to be induced by hypoxia
breastca_two_classes	9.99e-03	Gene set that can be used to differentiate BRCA1-linked and BRCA2-linked **breast cancers**

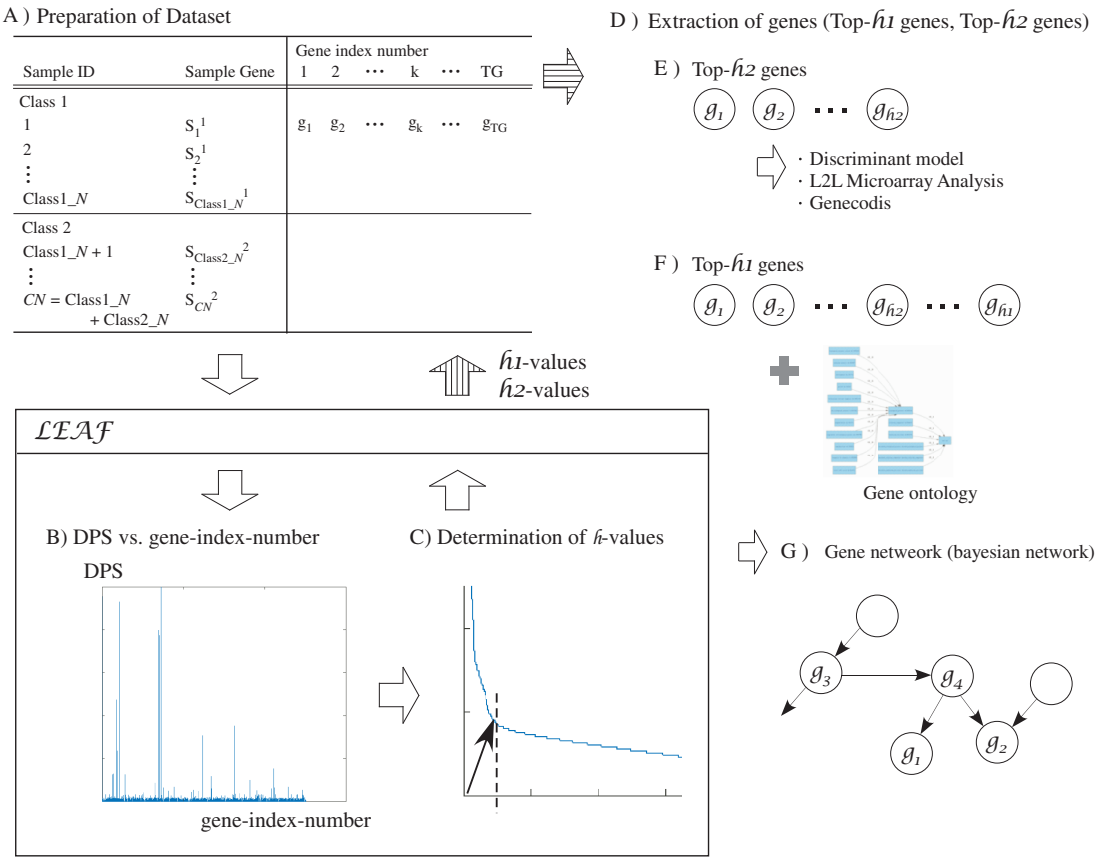

Figure. 8: Overview of the framework.

[5] Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[6] Gene Ontology Consortium. the gene ontology. *http://www.geneontology.org/*, 25:25–29, 2000.

[7] University of Washington, "L2l microarray analysis tool," `http://depts.washington.edu/l2l/`, 2007.

[8] T. Doan, R. Melvold, S. Viselli, Carl, and P. Waltenbaugh, *Lippincott's Illustrated Reviews: Immunology*. Lippincott Williams and Wilkins, 2007.

[9] F. Sahin, W. Qiu, R. E. Wilentz, C. A. Iacobuzio-Donahue, A. Grosmark, and G. H. Su, "RPL38, FOSL1, and UPP1 Are Predominantly Expressed in the Pancreatic Ductal Epithelium," *PANCREAS*, vol. 30, no. 2, pp. 158–167, 2005.

[10] H. Mitsubayashi, S. Aso, T. Nagashima, and Y. Okada, "Accurate and robust gene selection for disease classification using a simple statistic," *Bioinformation*, vol. 3, no. 2, pp. 68–71, 2008.

[11] P. A. Lachenbruch, *Discriminant Analysis*. Gendai-Sugakusha, Kyoto, 1979.

[12] F. Kentaro, N. Tomomasa, and O. Yoshifumi. Leaf: leave-one-out forward selection method for cancer classification using gene expression data. *9th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2010)*, 31–36, 2010.

[13] F. Jensen, *An introduction to Bayesian Networks*. University College London Press, 1996.

[14] E. N. Richard, *Learning Bayesian Networks*. Prentice Hall, 2003.

[15] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA, 1988.

[16] Wooldridge, M.J. and Jennings, N.R., Intelligent Agents: Theory and Practice. *Knowledge Engineering Review*, vol. 3, no. 2, pp. 115–152, 1995.

[17] Fabio Bellifemine, *Developing Multi-Agent Systems with JADE*. John Wiley and Sons, Ltd, 2004.

[18] Akaike, H. Petrov, B. N., and Csaki, F. (eds.), Information theory and an extention of the maximum likelihood principle. *2nd International Symposium on Information Theory, Akadimiai Kiado, Budapest*, 267–281, 1973.

Author Biographies

Kentaro Fukuta received his Ph.D. in engineering from Muroran Institute of Technology, Japan, in 2008. His current research interests are bioinformatics, Kansei engineering, and ontology engineering.

Tomomasa Nagashima received his Ph.D. in Physics from Waseda University, Japan, in 1978. He is currently Direc-

tor of Japanese Society of Kansei Engineering (JSKE) and Fellow of the Institute of Electronics, Information and Communication Engineers (IEICE), Japan. He is also the editor of IJBM. His research interests are Kansei engineering and bioinformatics.

Takashi Uozumi received his BE degree from Muroran Institute of Technology, Muroran, Japan, in 1973 and his DE degree from Hokkaido University, Sapporo, Japan, in 1980. Since 1988 he has been an Associate Professor in the Department of Computer Science and Systems Engineering, Muroran Institute of Technology. His research fields include evolutionary algorithms, image recognition, computer vision and machine learning.

Yoshifumi Okada received his Ph.D. in engineering from Muroran Institute of Technology, Japan, in 2002. His current research interests are bioinformatics, Kansei engineering, data mining, statistical pattern recognition, and signal processing.