LEAF: Leave-one-out Forward Selection Method for Informative Gene Discovery in DNA Microarray Data

Kentaro Fukuta and Yoshifumi Okada

***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 microarrays for candidate biomarkers. Here we discuss a gene selection method, which is called *LEAve-one-out Forward selection method* (LEAF), for discovering informative genes embedded in gene expression data, and propose an additional algorithm for extending LEAF's capabilities. LEAF is an iterative forward selection method incorporating the concept of leave-one-out cross validation (LOOCV) and provides a discrimination power score (DPS) for genes, which is a criterion for selecting the candidate of informative genes. We show that LEAF identifies genes that are practically used as biomarkers. Our method should be useful bioinformatics tool for biomedical, clinical, and pharmaceutical researchers.**

***Index Terms*—biomarkers, data mining, gene expression profiles, cancer classification.**

I. Introduction

Recent progress in bioinformatics technology has facilitated large-scale screening for candidate biomarkers [6]. 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 [5]. In our previous study, we have proposed a novel method called LEAve-one-out Forward selection method (LEAF) for analysis of gene expression data [8], [9]. 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 four public leukemia datasets (ALL/AML, ALL/MLL, and MLL/AML) [1], [7]. The results showed 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 conduct a biological functional analysis of the genes, using public gene function database.

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. (e-mail: fukuta@mail.svbl.muroran-it.ac.jp).

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. (e-mail: okada@epsilon2.csse.muroran-it.ac.jp).

II. Methodology

A. Datasets

We used three well-known leukemia datasets provided by Armstrong et al. and Golub et al., [1], [6]. These datasets are available at the Broad Institute [7]. 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 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_1 = 12,582 and TG_2 = 7,129.

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) [11]. To satisfy a maximal variance ratio (F-value) between two disease classes by using Mahalanobis distance, FSM cumulatively selects gene one-by-one and ultimately identifies a set of genes (a gene ranking) that is informative for disease classification.

The flow of the FSM algorithm is described as follows:

1) Calculation of the F-value (F_1) for all genes and selection of a gene having the maximum F_1 as the first gene.
2) For k (≥ 2)th gene, we pick up a k-th gene from the rest of genes, and add it into the set of $k-1$ genes.
3) Step 2 is repeated for all the genes in the rest set, and k-th genes is determined by selecting the gene with the maximum F_k.
4) Step 2 and Step 3 are repeated for $k \leq (Class1\_N + Class2\_N) - 2$ till the ranking of the genes is accomplished.

In fact, LEAF is an iterative FSM inspired by leave-one-out cross validation (LOOCV) [10]. Details of the algorithm have been published [8], [11]. 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

A) Preparation of Data "Leukemia dataset by Armstrong et al."

Class (#Genes : TG_1 = 12582)	Samples : N
ALL (Acute lymphocytic leukemia)	24
MLL (Mixed lineage leukemia)	20
AML (Acute myelogenous leukemia)	28

Dataset name	Class 1 (N)	Class 2 (N)
ALL1 vs. AML1	ALL (24)	AML (28)
ALL1 vs. MLL1	ALL (24)	MLL (20)
MLL1 vs. AML2	MLL (20)	AML (28)

Preparation of Data "Leukemia dataset by Golub et al."

Class (#Genes : TG_2 = 7129)	Samples : N
ALL (Acute lymphocytic leukemia)	47
AML (Acute myelogenous leukemia)	25

Dataset name	Class 1 (N)	Class 2 (N)
ALL2 vs. AML2	ALL (47)	AML (25)

B) Data matrix for FSM

	Sample ID	Sample Gene	Gene index number 1	2	⋯	k	⋯	TG
Class 1	1	S_1^1	g_1	g_2	⋯	g_k	⋯	$g_{12582 \text{ or } 7129}$
	2	S_2^1						
	⋮	⋮						
	Class1_N	$S_{Class1\_N}^1$						
Class 2	Class1_N + 1	$S_{Class2\_N}^2$						
	⋮	⋮						
	CN = Class1_N + Class2_N	S_{CN}^2						

Fig. 1. Preparation of dataset.

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.

The discrimination power of selected genes greatly depends on the genes that compose the ranking. In the method of this paper, the genes with low discrimination power are excluded by using threshold parameter $Thrs1$. In this research, we adopt $Thrs1 = 2.0$, and use the genes which satisfy $F_1 > 2.0$ for selecting the k (≥ 2)th gene.

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

Because previous work [8] 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. 3. The h-value is calculated by the following steps:

1) Descending sort of DPS (Fig. 3A).
2) Decision of h-value.
   a) Normalize the horizontal and vertical axes by dividing by their respective maximum values (Fig. 3B).
   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. Results and Discussion

A. Discrimination Power Score

Figure 4 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 are represented as peaks in the DPS graph.

B. Biological function analysis

TABLE I
h AND DPS VALUES OF LEUKEMIA DATASET

Dataset	$h1$	DPS	$h2$	DPS
ALL1 vs. AML1	129	0.0173	17	0.1149
ALL1 vs. MLL1	158	0.0187	14	0.0809
MLL1 vs. AML1	157	0.0180	21	0.0715
ALL2 vs. AML2	147	0.0242	18	0.1106

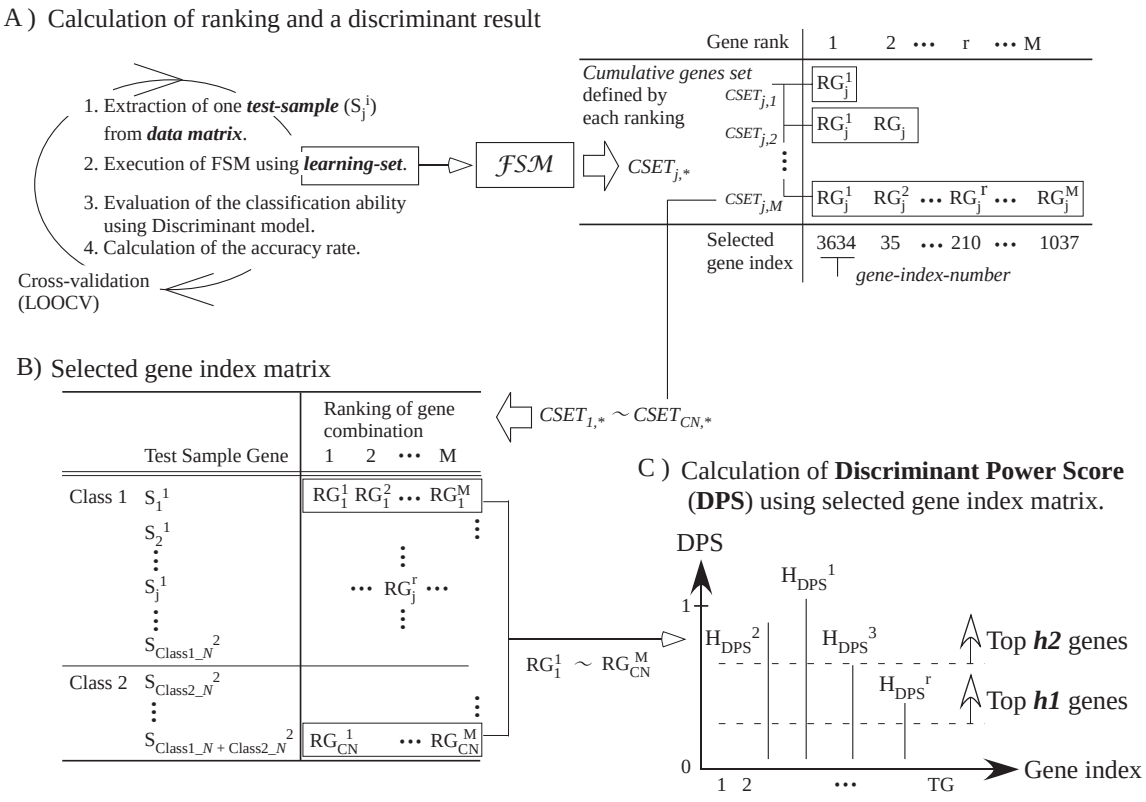

Fig. 2. Overview of LEAF's methodology.

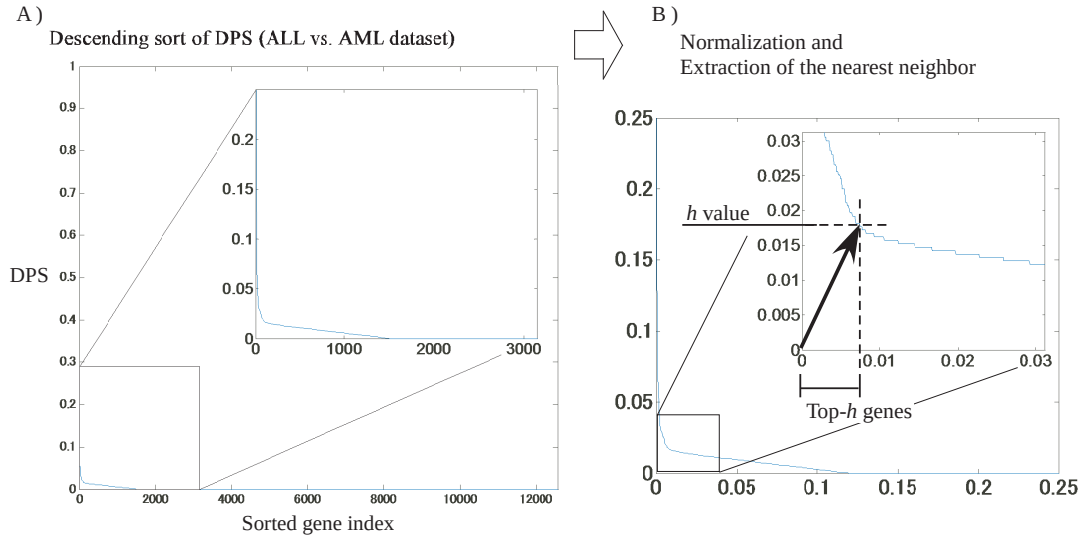

Fig. 3. Outline for defining h-value.

The h-values of each dataset are presented in Table I. 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 [2], [3] and the University of Washington's L2L microarray analysis tool [12]. Below we focus on the top-$h2$ genes' biological function.

Table. II 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, EEF1A1, IL8RB and IL18 [4], are selected from every dataset pair. For example, it is shown that TCL1A is a T-lymph cell of leukemia. In particular, it should be noted that ribosomal protein L38 (RPL38), interleukin 18 (IL18) and eukaryotic translation elongation factor (EEF1A1) are highly expressed in pancreatic cancer cell lines [13]–[15]. 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 III, IV, V and VI summarizes the L2L results. In the each datasets, we can observe that functions related to human cancer and tumor, such as colon carcinoma, glioma and breast cancer, exhibit statistical significance.

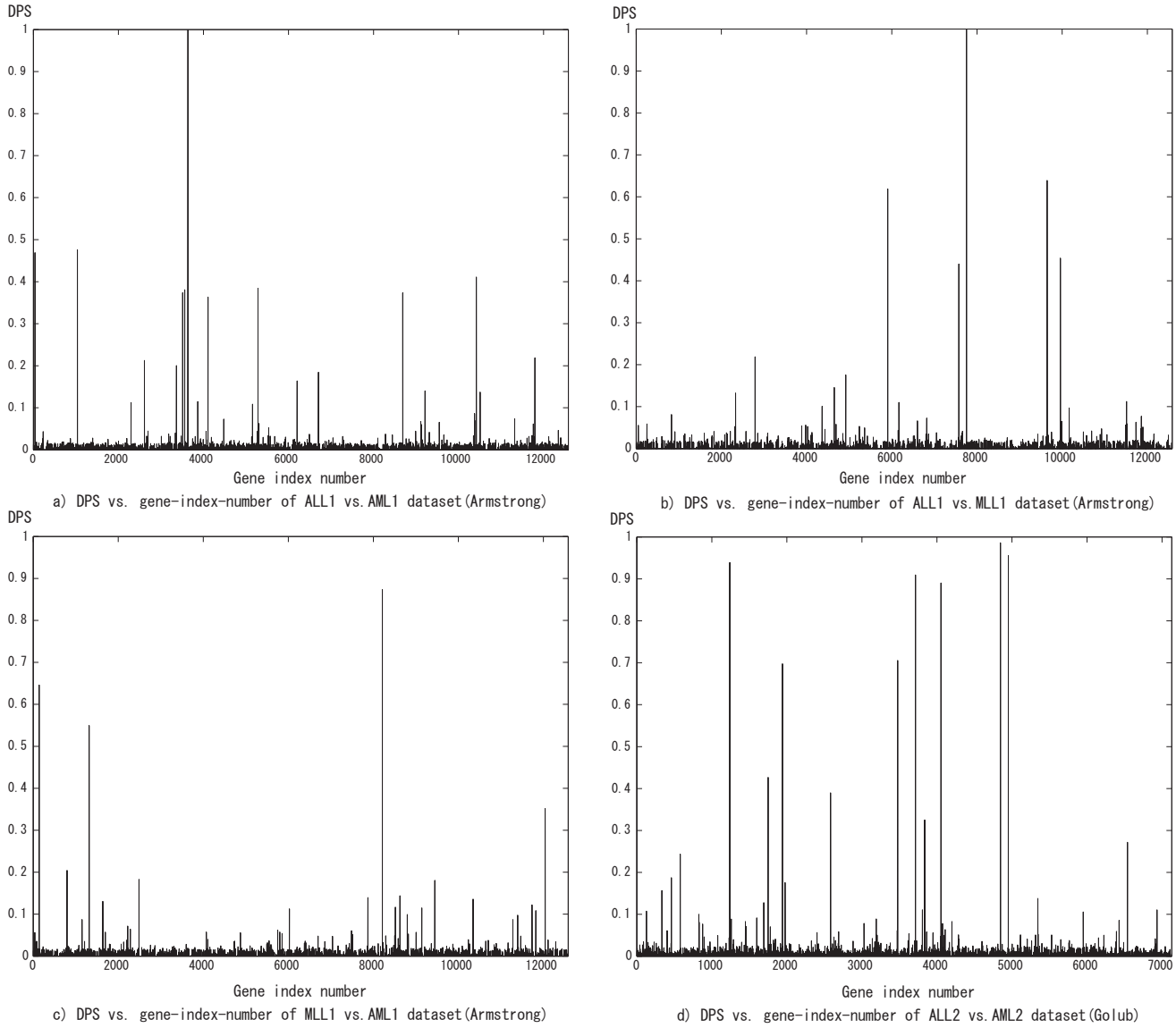

Fig. 4. DPS vs. gene-index-number of leukemia dataset.

C. Gene analysis method

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 method, which applies LEAF. An overview of the method (Fig. 5) illustrates the processes by which it operates.

1) Analysis of the dataset using LEAF, and display of DPS (Figs. 5A and B).
2) Calculation of h-values (Fig. 5C).
3) Extraction of the genes based on the h-value (Fig. 5D).
4) Analysis of top-$h2$ genes (Fig. 5E).
    a) Construction of a discriminant model.
    b) Output of a summary (*i.e.*, Table II).
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 5G 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 [16], [17]. We can apply probabilistic reasoning [18] and search for the biological process that supports discovery of a biomarker. Moreover, in this method, 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. Actually, GO does not have only a common vocabulary in biological science. In addition, it does provide just 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 [19], [20] that searches for a candidate biological process to built, BN. They change

TABLE II
SUMMARY OF THE TOP $h2$ GENES RANKED BY DPS

A) ALL1 vs. AML1 dataset (Armstrong)

DPS ranking	Input name / Gene name	Description
1	39318_at / **TCL1A**	**T-cell leukemia/lymphoma 1A**
2	34085_at / **RPL38**	**ribosomal protein L38**
3	AFFX-M27830_5_at / ---	---
4	32541_at / PPP3CC	protein phosphatase 3 (formerly 2B), catalytic subunit, gamma isoform
5	34717_s_at / FUSIP1	FUS interacting protein (serine/arginine-rich) 1
6	39243_s_at / PSIP1	PC4 and SFRS1 interacting protein 1
7	38955_at / AVPR1A	arginine vasopressin receptor 1A
8	36982_at / USP14	ubiquitin specific peptidase 14 (tRNA-guanine transglycosylase)
9	40749_at / MS4A1	membrane-spanning 4-domains, subfamily A, member 1
10	910_at / ---	---
11	36259_at / ---	---
12	38569_at / NRF1	nuclear respiratory factor 1
13	39373_at / FADS1	fatty acid desaturase 1
14	37913_at / DHFR	dihydrofolate reductase
15	38463_s_at / ---	---
16	33110_at / SOX2	SRY (sex determining region Y)-box 2
17	40030_at / ---	---

B) ALL1 vs. MLL1 dataset (Armstrong)

DPS ranking	Input name / Gene name	Description
1	33412_at / ---	---
2	39857_at / STX11	syntaxin 11
3	36897_at / ---	---
4	40887_g_at / **EEF1A1**	**eukaryotic translation elongation factor 1 alpha 1**
5	32755_at / ACTA2	actin, alpha 2, smooth muscle, aorta
6	36798_g_at / ---	---
7	33277_at / MTMR2	myotubularin related protein 2
8	32054_at / CCNT2	cyclin T2
9	35383_at / ---	---
10	1203_at / ---	---
11	37640_at / HPRT1	hypoxanthine phosphoribosyltransferase 1 (Lesch-Nyhan syndrome)
12	41605_at / ---	---
13	41332_at / POLR2E	polymerase (RNA) II (DNA directed) polypeptide E, 25kDa
14	33042_r_at / ---	---

C) MLL1 vs. AML1 dataset (Armstrong)

DPS ranking	Input name / Gene name	Description
1	35307_at / GDI2	GDP dissociation inhibitor 2
2	31397_at / ---	
3	35083_at / FTL	ferritin, light polypeptide
4	664_at / **IL8RB**	**interleukin 8 receptor, beta**
5	33008_at / ---	
6	35896_at / ---	
7	39175_at / PFKP	phosphofructokinase, platelet
8	36678_at / TAGLN2	transgelin 2
9	33889_s_at / ---	
10	41818_at / CARD10	caspase recruitment domain family, member 10
11	32294_g_at / LHCGR	luteinizing hormone/choriogonadotropin receptor
12	979_g_at / ---	
13	36571_at / TOP2B	topoisomerase (DNA) II beta 180kDa
14	38391_at / CAPG	capping protein (actin filament), gelsolin-like
15	37257_at / PRUNE	prune homolog (Drosophila)
16	892_at / TM4SF1	transmembrane 4 L six family member 1
17	37332_r_at / ALDH4A1	aldehyde dehydrogenase 4 family, member A1
18	1307_at / XPA	xeroderma pigmentosum, complementation group A
19	34556_at / DAPK2	death-associated protein kinase 2
20	1420_s_at / ---	
21	34912_at / ---	

D) ALL2 vs. AML2 dataset (Golub)

DPS ranking	Input name / Gene name	Description
1	X95735_at / ZYX	zyxin
2	Y07604_at / NME4	non-metastatic cells 4, protein expressed in
3	L07633_at / PSME1	proteasome (prosome, macropain) activator subunit 1 (PA28 alpha)
4	U77604_at / MGST2	microsomal glutathione S-transferase 2
5	X04143_at / ---	---
6	U60205_at / SC4MOL	sterol-C4-methyl oxidase-like
7	M31994_at / ---	---
8	M16424_at / ---	---
9	U01212_at / ---	---
10	U82313_at / ---	---
11	X85116_rna1_s_at / ---	---
12	D79994_at / ANKRD15	ankyrin repeat domain 15
13	D49950_at / **IL18**	**interleukin 18 (interferon-gamma-inducing factor)**
14	M34455_at / INDO	indoleamine-pyrrole 2,3 dioxygenase
15	D28532_at / SLC17A1	solute carrier family 17 (sodium phosphate), member 1
16	M58026_at / CALML3	calmodulin-like 3
17	M12759_at / ---	---
18	U80034_at / MIPEP	mitochondrial intermediate peptidase

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).

IV. 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 method 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 investigate the effect of threshold parameter ($Thrs1$) on the result of function analysis to 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.

TABLE III
FUNCTION ENRICHMENT ANALYSIS (L2L) FOR THE TOP-$h2$ GENES OF ALL1 VS AML1 DATASET (ARMSTRONG)

Function name	*p*-Value < 0.05	Enrichment	Description
elongina_ko_dn	1.53e-04	15.28	Downregulated in MES cells from elongin-A knockout mice
gamma_unique_fibro_dn	7.77e-04	15.28	Down-regulated at any timepoint by treatment of human fibroblasts with gamma radiation, but not by UV lght or 4-NQO
hdaci_colon_tsa48hrs_dn	8.81e-04	113.06	Downregulated by TSA at 48 hrs in SW260 **colon carcinoma cells**
senescence_rep-ind_dn	0.01	11.19	Down-regulated in models of both replicative (high-passge human foreskin fibroblast) and induced (repression of E7 in HeLa) cellular senescence.
aom-dss_colon_10wks_up	0.02	43.48	Up-regulated in mouse colonic mucosa after 10 weeks of treatment with the **colon carcinogens** azoxymethane (AOM) and 2% dextran sodium sulfate (DSS) vs. untreated controls.
senescence_hff_dn	0.03	7.20	Down-regulated in primary human foreskin fibroblasts at replicative senescence (passage 26) compared to active replication (passage 8).
bcnu_glioma_nomgmt_24hrs_up	0.03	37.69	Up-regulated in an MGMT-deficient **glioma cell** line (A172) at 24 hours following treatment with BCNU
oxstress_breastca_up	0.03	28.26	Upregulated by H2O2, Menadione and t-BH in **breast cancer cells**
tff2_ko_up	0.03	31.41	Up-regulated in pyloric atrium tissue from Trefoil Factor 2 (Tff2) knockout mice, compared to wild-type controls
hsc_hsc_adult	0.04	6.01	Up-regulated in mouse hematopoietic stem cells from adult bone marrow (HSC Shared + Adult)
idx_tsa_dn_cluster6	0.04	23.55	Strongly down-regulated at 2 hours during differentiation of 3T3-L1 fibroblasts into adipocytes with IDX (insulin, dexamethasone and isobutylxanthine), vs. fibroblasts treated with IDX + TSA to prevent differentiation (cluster 6)
hdaci_colon_tsa48hrs_up	0.04	24.58	Upregulated by TSA at 48 hrs in SW260 **colon carcinoma cells**

TABLE IV
FUNCTION ENRICHMENT ANALYSIS (L2L) FOR THE TOP-$h2$ GENES OF ALL1 VS MLL1 DATASET (ARMSTRONG)

Function name	*p*-Value < 0.05	Enrichment	Description
elongina_ko_dn	1.53e-04	15.28	Downregulated in MES cells from elongin-A knockout mice
gamma_unique_fibro_dn	7.77e-04	15.28	Down-regulated at any timepoint by treatment of human fibroblasts with gamma radiation, but not by UV lght or 4-NQO
hdaci_colon_tsa48hrs_dn	8.81e-04	113.06	Downregulated by TSA at 48 hrs in SW260 **colon carcinoma cells**
senescence_rep-ind_dn	0.01	11.19	Down-regulated in models of both replicative (high-passge human foreskin fibroblast) and induced (repression of E7 in HeLa) cellular senescence.
aom-dss_colon_10wks_up	0.02	43.48	Up-regulated in mouse colonic mucosa after 10 weeks of treatment with the **colon carcinogens** azoxymethane (AOM) and 2% dextran sodium sulfate (DSS) vs. untreated controls.
senescence_hff_dn	0.03	7.20	Down-regulated in primary human foreskin fibroblasts at replicative senescence (passage 26) compared to active replication (passage 8).
bcnu_glioma_nomgmt_24hrs_up	0.03	37.69	Up-regulated in an MGMT-deficient **glioma cell** line (A172) at 24 hours following treatment with BCNU
oxstress_breastca_up	0.03	28.26	Upregulated by H2O2, Menadione and t-BH in **breast cancer cells**
tff2_ko_up	0.03	31.41	Up-regulated in pyloric atrium tissue from Trefoil Factor 2 (Tff2) knockout mice, compared to wild-type controls
hsc_hsc_adult	0.04	6.01	Up-regulated in mouse hematopoietic stem cells from adult bone marrow (HSC Shared + Adult)
idx_tsa_dn_cluster6	0.04	23.55	Strongly down-regulated at 2 hours during differentiation of 3T3-L1 fibroblasts into adipocytes with IDX (insulin, dexamethasone and isobutylxanthine), vs. fibroblasts treated with IDX + TSA to prevent differentiation (cluster 6)
hdaci_colon_tsa48hrs_up	0.04	24.58	Upregulated by TSA at 48 hrs in SW260 **colon carcinoma cells**

TABLE V
FUNCTION ENRICHMENT ANALYSIS (L2L) FOR THE TOP-$h2$ GENES OF MLL1 VS AML1 DATASET (ARMSTRONG)

Function name	*p*-Value < 0.05	Enrichment	Description
hdaci_colon_but12hrs_dn	5.79e-04	18.81	Downregulated by butyrate at 12 hrs in SW260 **colon carcinoma cells**
hdaci_colon_but16hrs_dn	1.32e-03	14.15	Downregulated by butyrate at 16 hrs in SW260 **colon carcinoma cells**
hdaci_colon_cur2hrs_up	1.50e-03	35.20	Upregulated by curcumin at 2 hrs in SW260 **colon carcinoma cells**
breastca_three_classes	4.64e-03	19.60	Gene set that can be used to differentiate BRCA1-linked, BRCA2-linked, and sporadic primary **breast cancers**
hdaci_colon_but_dn	0.01	6.13	Downregulated by butyrate at any timepoint up to 48 hrs in SW260 **colon carcinoma cells**
sarcomas_leiomyosarcoma_up	0.02	50.85	Top 20 positive significant genes associated with calponin negative **leiomyosarcoma tumors**, versus **other soft-tissue tumors**.
mkk6ee_up	0.02	41.6	Upregulated by expression of constitutively active MKK6
hdaci_colon_but48hrs_dn	0.02	9.53	Downregulated by butyrate at 48 hrs in SW260 **colon carcinoma cells**
fsh_ovary_mcv152_dn	0.02	9.44	Down-regulated in ovarian epithelial cells (MCV152) 72 hours following FSH treatment, compared to untreated
hdaci_colon_cur_up	0.02	9.24	Upregulated by curcumin at any timepoint up to 48 hrs in SW260 **colon carcinoma cells**
hdaci_colon_but24hrs_dn	0.02	8.89	Downregulated by butyrate at 24 hrs in SW260 **colon carcinoma cells**
fsh_granulosa_up	0.02	8.80	Up-regulated in human granulosa cells stimulated with follicle stimulation hormone (FSH)
diab_neph_up	0.02	8.72	Upregulated in the glomeruli of cadaver kidneys from patients with diabetic nephropathy, compared to normal controls
lh_granulosa_up	0.02	8.55	Up-regulated in human granulosa cells stimulated with luteinizing hormone (LH)
hypoxia_fibro_up	0.03	38.13	Upregulated by hypoxia in normal fibroblasts from both young and old donors (Table 3)
parp_ko_dn	0.04	26.92	Downregulated in MEF cells from PARP knockout mice
bcnu_glioma_nomgmt_48hrs_up	0.04	24.09	Up-regulated in an MGMT-deficient **glioma cell** line (A172) at 48 hours following treatment with BCNU
brca_er_neg	0.04	2.74	Genes whose expression is consistently negatively correlated with estrogen receptor status in **breast cancer** - higher expression is associated with ER-negative tumors

TABLE VI
FUNCTION ENRICHMENT ANALYSIS (L2L) FOR THE TOP-$h2$ GENES OF ALL2 VS AML2 DATASET (GOLUB)

Function name	*p*-Value < 0.05	Enrichment	Description
gh_exogenous_middle_dn	3.71e-03	269.56	Down-regulated at middle time points (6-8 hours) following treatment of mammary **carcinoma cells** (MCF-7) with exogenous human growth hormone
bay_pbmc_30min_dn	7.41e-03	134.78	Down-regulated at 30 min following treatment of peripheral blood mononuclear cells (PBMC) with BAY 50-4798, an IL-2 receptor agonist.
human_tissue_kidney	0.02	53.91	Genes expressed specifically in human kidney tissue
mouse_tissue_kidney	0.02	44.93	Genes expressed specifically in mouse kidney tissue
bay_pbmc_6hr_up	0.02	5.12	Up-regulated at 6 hr following treatment of peripheral blood mononuclear cells (PBMC) with BAY 50-4798, an IL-2 receptor agonist.
scchn_hpvpos_dn	0.03	29.95	Down-regulated in HPV-positive **squamous cell carcinomas** of the head and neck (SCCHN) vs. normal oral epithelium.
cpr_low_liver_up	0.04	24.51	Up-regulated in mouse liver tissue from mice with reduced liver expression of NADPH-cytochrome P450 reductase (CPR), versus normal controls
cmv_hcmv_timecourse_8hrs_dn	0.04	22.46	Down-regulated in fibroblasts following infection with **human cytomegalovirus** (at least 3-fold, with Affymetrix change call, in at least two consectutive timepoints), with maximum change at 8 hours
idx_tsa_dn_cluster4	0.04	22.46	Strongly down-regulated at 8-48 hours during differentiation of 3T3-L1 fibroblasts into adipocytes with IDX (insulin, dexamethasone and isobutylxanthine), vs. fibroblasts treated with IDX + TSA to prevent differentiation (cluster 4)
tgfbeta_all_up	0.04	6.66	Upregulated by TGF-beta treatment of skin fibroblasts, at any timepoint

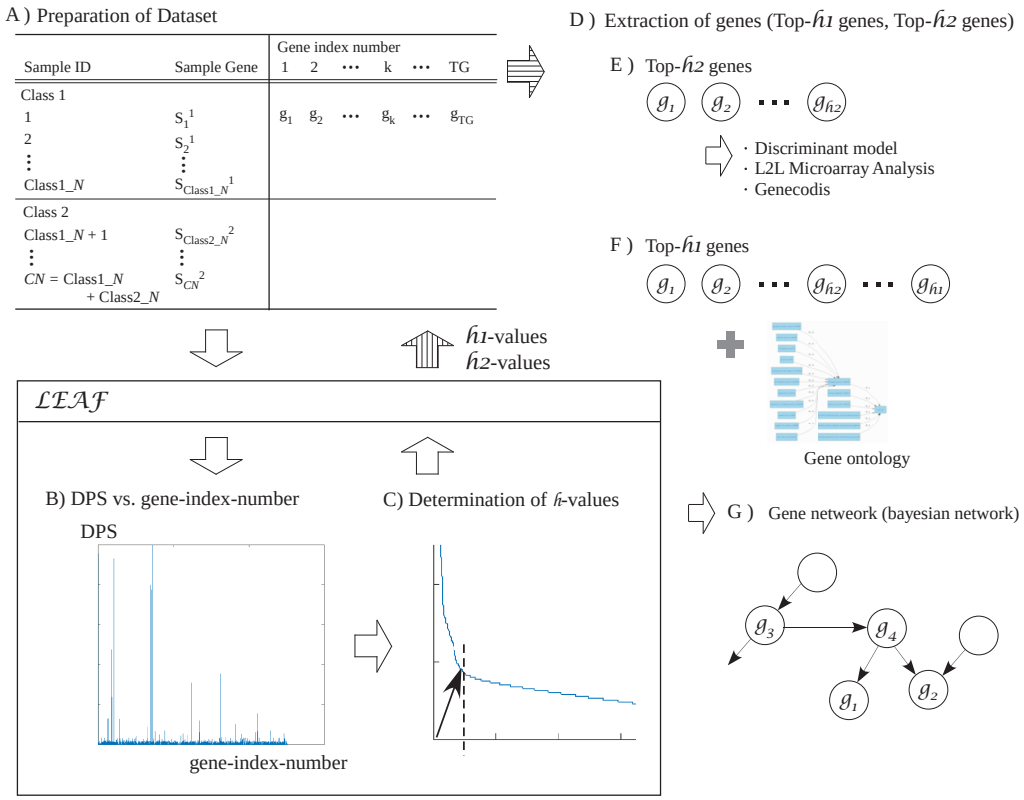

Fig. 5. Overview of the gene analysis method.

REFERENCES

[1] 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.

[2] Gene Ontology Consortium, "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, pp. 25-29, 2000.

[3] Gene Ontology Consortium, "The Gene Ontology," *http://www.geneontology.org/*, 2000.

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

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

[6] 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*, vol. 286, pp. 531-537, 1999.

[7] Broad Institute, "Broad Institute of MIT and Harvard," *http://www.broadinstitute.org/*, 2010.

[8] K. Fukuta, T. Nagashima, and Y. Okada, "Leaf: leave-one-out forward selection method for cancer classification using gene expression data," *Proceedings of The 9th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2010*, 18-20, August, 2010, pp. 31-36.

[9] K. Fukuta, T. Nagashima, and Y. Okada, "Leaf: leave-one-out forward selection method for cancer classification using gene expression data," *Lecture Notes in Engineering and ComputerScience: Proceedings of The International MultiConference of Engineers and Computer Scientists 2011, IMECS 2011, 16-18 March, 2011*, Hong Kong, pp. 175-180.

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

[11] 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.

[12] University of Washington, "L2L microarray analysis tool," *http://depts.washington.edu/l2l/*, 2007.

[13] 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.

[14] N. Srabovic, Z. Mujagic, JM. Mustedanagic, Z. Muminovic, and E. Cickusic, "Interleukin 18 expression in the primary breast cancer tumour tissue," *Med Glas Ljek komore Zenicko-doboj kantona*, vol. 8, no. 1, pp. 109-115, 2011.

[15] K. Lin and S. SOUCHELNYTSKYI, "Eukaryotic elongation factor eEF1A1 promotes and Ser300 mutants of eEF1A1 inhibit transition through the S and G2/M phases of the cell cycle," *Journal of Cell and Molecular Biology*, vol. 8, no. 2, pp. 125-130, 2010.

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

[17] E. N. Richard, *Learning Bayesian Networks*. Upper Saddle River, NJ: Prentice Hall, 2003.

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

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

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