Available online at www.sciencedirect.com
ScienceDirect
Progress in Natural Science 18 (2008) 1491–1500
Progress in Natural Science
www.elsevier.com/locate/pnsc

A feature ensemble technology to identify molecular mechanisms for distinction between multiple subtypes of lymphoma

Yueying Yang [a,1], Haiyun Wang [a,b,1], Xia Li [a,b,c,*], Xue Xiao [c], Su Fei [a], Chuanxing Li [a], Hongzhi Wang [d], Shaoqi Rao [a,c,e], Yadong Wang [d,*]

[a] College of Bioinformatics Science and Technology, Bio-pharmaceutical Key Laboratory of Heilongjiang Province and State, Harbin Medical University, Harbin 150081, China
[b] College of Life Science and Technology, Tongji University, Shanghai 200092, China
[c] The Biomedical Engineering Institute, Capital Medical University, Beijing 100054, China
[d] Department of Computer Science, Harbin Institute of Technology, Harbin 150080, China
[e] Department of Molecular Cardiology, Cleveland Clinic Foundation, Cleveland, OH 44195, USA

Received 26 February 2008; received in revised form 10 April 2008; accepted 15 April 2008

Abstract

Due to complexities and genetic heterogeneities of biological phenotypes, robust computational approaches are desirable to achieve high generalization performance with multiple classifiers, perturbations of the data structures, and biological interpretations. The purpose of this study is to extend our developed ensemble decision approach to distinguish multiple heterogeneous phenotypes and to elucidate the underlying molecular bridges that intertwine the subtypes. Our work identifies the significant molecular mechanisms (disease-relevant genes and functions) that underpin the complex molecular mechanisms for distinction between multiple phenotypes. Feature genes and hierarchical gene cores identified by our method have achieved high accuracy in the classification of multiple phenotypes. The results show that the proposed analysis strategy is feasible and powerful in the classification of biological subtypes and in the explanation of the molecular connections between clinical phenotypes. Biological interpretations with Gene Ontology revealed concerted genetic pathways for some lymphoma subtypes.

Keywords: Ensemble decision approach; Molecular mechanisms; Lymphoma; Microarray

1. Introduction

Complex diseases are the result of the collective actions of many genetic and non-genetic factors; therefore, the genetic dissection of complex diseases should be carried out by a new systematic technology. Gene expression microarray is an efficient and high throughout method for genetically profiling diseases and their associated treatments. Such a process involves a key step of biomarker identification, which is expected to be closely related to the disease. A most important task of these identified genes is that they can be used to construct a classifier which can effectively diagnose diseases and even recognize the disease subtypes. Binary classification, for example, for the diseased or healthy, in microarray data analysis has been successful; while multi-class classification, such as cancer subtyping, remains challenging. The patterns of up-regulation or down-regulation of gene activities can serve as secondary endpoints of biomarkers [1], and the method of finding such biomarkers is actually the application of traditional feature selection methods in the field of molec-

* Corresponding authors. Tel.: +86 451 8661 5912; fax: +86 451 8666 9617.
E-mail address: lixia@hrbmu.edu.cn (X. Li).
[1] These authors contributed equally to this work.

1002-0071/$ - see front matter © 2008 National Natural Science Foundation of China and Chinese Academy of Sciences. Published by Elsevier Limited and Science in China Press. All rights reserved.
doi:10.1016/j.pnsc.2008.04.012

ular biology and life sciences. Feature selection aims to pick out d features from D features ($D > d$) that can best discriminate between heterogeneous samples [2].

Traditional feature selection methods are devoted to producing a small set of biomarker genes which can be used in classifier construction such as linear discrimination analysis, nearest neighbor models, support vector machines, and logistic regression models [3–5]. These gene selection methods typically find the best feature subset; however, researchers cannot always get a good classifying performance and biological comprehension because such feature selection methods focus on finding the best feature subset which contains a few genes and which is also sensitive to many factors such as learning algorithms and training samples.

It is important to point out that most of these gene selection methods and the associated classifiers are usually worked on two-class datasets, though they can be theoretically extended to multi-class datasets through one-versus-all (OVA). The methods have been tested on multi-class dataset(s) [3,4].

Using ensembles of base classifiers to improve classifying performance has been a hot topic of machine learning [6]. While binary classification has been extensively explored, multi-class classification remains challenging in microarray data analysis. In this work, we focus on gene selection for multi-class classification and we demonstrate the power of the proposed method by applying it to the identification of a cancer subtype. We have extended our newly developed ensemble decision approach [7] to the analysis of multiple heterogeneous phenotypes and to elucidate the underlying molecular mechanisms that intertwine the phenotypes. Rather than simply maximizing prediction accuracy, we further identify the genes that are most relevant to a disease by retrieving 'redundant' genes which are excluded during the course of feature selection but actually are strongly relevant to the disease. Besides analyzing the role of key genes, we also unravel the molecular mechanisms of multiple phenotypes at the function level by mapping genes onto Gene Ontology [8–10].

2. Methods

2.1. Ensemble of feature selection

The whole process of ensemble of feature selection is simply depicted in Fig. 1 and described as follows.

The feature genes of the numerous subtypes of the lymphoma datasets are heterogeneous in feature space, which shows that individual genes are of unequal importance in different regions of the whole feature space. In our work, a multi-class problem is firstly transformed into a two-class (the positive and the negative) problem, more specifically, the positive samples are gathered from the samples of a certain lymphoma subtype, while the negative samples are collected randomly from the other lymphoma subtypes. The training sets and the test sets are achieved by boosting and randomly grouping the positive and the negative samples. Sets of original feature subsets are acquired by decision tree method, each feature subset G_j^c ($c = 1, 2, \ldots, C$.

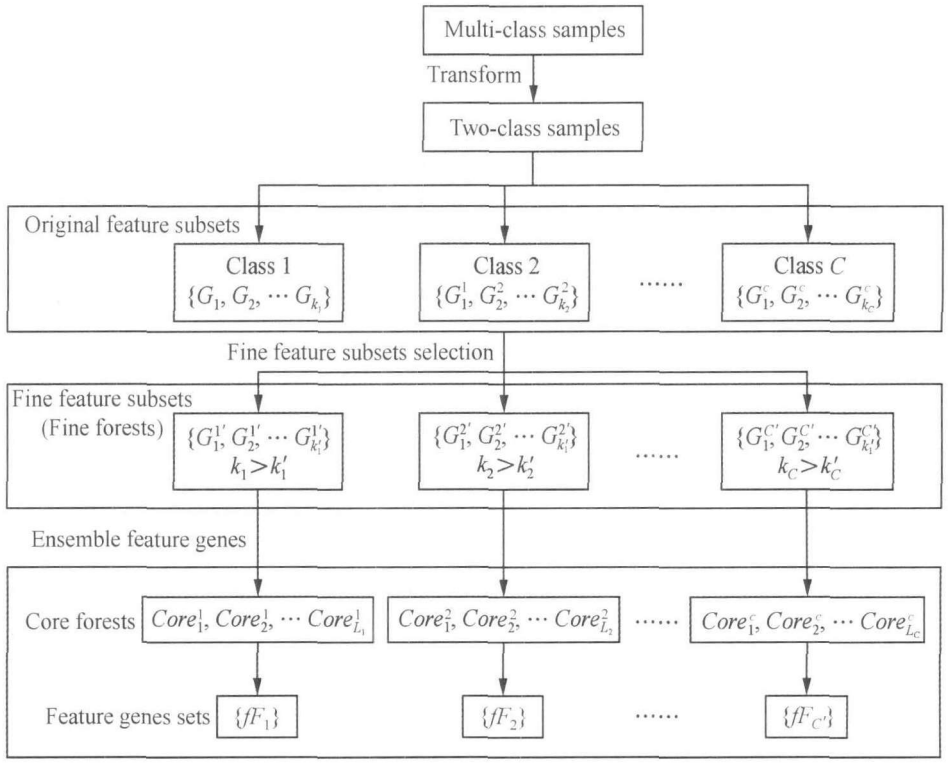

Fig. 1. Ensembles of feature selection of multiple phenotypes based on 4026 genes.

$j = 1, 2, \ldots, kc$) includes the split nodes and split rule in one decision tree, and can discriminate samples of the positive and the negative, where C denotes the number of subtypes, and j denotes the serial number of a feature subset. In this paper, we do not differentiate between feature subsets and decision trees clearly, that is, one decision tree corresponds to one feature subset. C original feature subsets constitute the original forest.

Then the feature subset with $\delta > 0.6$ is picked out as 'fine subset' G_j^c ($c = 1, 2, \ldots, C$, $j = 1, 2, \ldots, kc'$).

$$\delta = 2 \times p \times r/(p + r)$$

where δ is the classifying performance of each feature subset G_j^c under supervised learning, while $p = \mathrm{TP}/(\mathrm{FP} + \mathrm{TP})$, $r = \mathrm{TP}/(\mathrm{FN} + \mathrm{TP})$, TP is true positive, FP is false positive, TN is true negative, and FN is false negative. The δ is high if both false positive (FP) and false negative (FN) are low. Consequently, C fine feature subsets constitute the fine forest.

The next step is to ensemble feature genes. Genes with the high appearance frequency in the fine feature subsets are recognized as feature genes because they do not depend on the composition of training samples. The algorithm which we developed on the basis of algorithm Mining Core [11] is applied to determine gene cores in which high-frequency genes are grouped together. The inputs of this algorithm are the set of fine feature subsets, if there are genes shared by at least 2 feature subsets, these genes constitute one gene core. Moreover, we take gene cores found above as the input of the algorithm, thereby, the output is a group of hierarchical gene cores $Core_1^c, Core_2^c, \ldots, Core_{Lc}^c$.

Feature genes and gene cores are selected by the method of ensemble of the feature selections, but there are still some genes that are never selected because of their redundancy relative to the selected genes, namely their expression profiles are similar to those of selected genes. However, genes with similar expression profiles are inclined to be in the same metabolic pathway, same signal transduction pathway or be in the different components of a protein complex [12–14]. Without these redundant genes it would be difficult to understand the whole process of changes of a disease at the molecular level. Thus, we delete selected genes and perform the previous work repeatedly to retrieve more genes until the classifying performance no longer falls in five consecutive repeats.

Moreover, feature genes selected by a decision tree may not be up-regulated or down-regulated in the samples of a certain lymphoma subtype. So we further filter genes of normal versus lymphoma subtypes by Student's t-test [15].

2.2. Biological evaluation of feature genes

Based on gene ID and three other databases, GenBank [16], Unigene [17,18] and LocusLink [19], feature genes are mapped into Gene Ontology (GO). Fisher test [15] is applied to get function terms in GO with significance level of 0.05. The feature genes enriched in the function terms are partly selected for biological evaluation. The function terms belonging to different lymphoma subtypes can help us understand the relationship between the disease and function.

3. Results

3.1. Lymphoma dataset

The lymphoma expression profile dataset we have used [20,21] includes 86 samples of five classes and 4026 genes: 21 GCB-like-DLBCL samples (germinal centre B-like DLBCL) (class 1), 21 AB-like-DLBCL samples (activated B-like DLBCL) (class 2); 11 CLL samples (chronic lymphocytic leukaemia) (class 3); 9 FL samples (follicular lymphoma) (class 4) and 24 normal samples (class 5). Samples in different normal classes were merged into one class for their small sample numbers, and samples without

Table 1
The step-by-step algorithm of ensemble of the feature selection

Step 1. Randomly split all samples into two parts. One part (1/5) is the validation set which never takes part in feature selection process.
Step 2. The other part (4/5) is randomly divided into training set and testing set. Threefold cross-validation technology is used.
Step 3. Samples are transformed into those with the two-class label. Boosting builds a series of feature subsets belonging to each class while training set is fixed.
Step 4. Pick out fine feature subsets belonging to each class under the direction of testing set's classifying performance ($\delta > 0.6$).
Step 5. Repeat Step 2 to Step 4 twenty times to achieve a series of fine feature subsets belonging to each class $\{G_1^{C'}, G_2^{C'}, \ldots, G_{k_C'}^{C'}\}$ ($C = 1, 2, \ldots, 5$).
Step 6. Repeat Step 1 to Step 5 ten times.
Step 7. Achieve ten series of $\{G_1^{C''}, G_2^{C''}, \ldots, G_{k_C''}^{C''}\}$ after above steps from Step 1 to Step 6. Calculate frequency of each gene in each series of $\{G_1^{C'}, G_2^{C'}, \ldots, G_{k_C'}^{C'}\}$ and pick out genes with frequency $f \geqslant 2$ to enter feature genes set $\{F_c\}$. Then we calculate the accumulative frequency of genes which appeared in ten series of $\{F_c\}$ and pick out genes with a frequency higher than nine times to enter final feature genes set $\{fF_c\}$.
Step 8. Mine hierarchical gene cores.
Step 9. Classifying performance evaluation using the validation set.
Step 10. Delete selected genes and repeat the previous work again and again to retrieve more genes and add to $\{fF_c\}$.
Step 11. Biological evaluation of feature genes in the set $\{fF_c\}$ and gene cores.

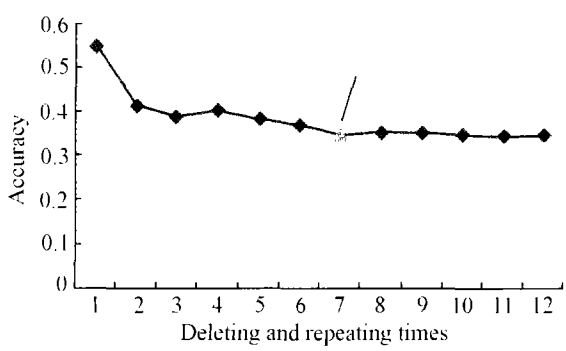

Fig. 2. Retrievial of the redundant genes while deleting selected genes and repeating the method of ensemble feature selection again and again.

Table 2
Gene cores of different phenotypes with high appearance frequency

Phenotypes	Gene cores	Appearance frequency	Gene ID	Gene name
GCB-like-DLBCL	Gene core 1	22	GENE1865X	Unknown UG Hs.124304 ESTs; Clone = 1358064
			GENE2758X	Unknown; Clone = 682995
	Gene core 2	19	GENE1836X	Unknown UG Hs.190487 ESTs; Clone = 1358277
			GENE2758X	Unknown; Clone = 682995
	Gene core 3	19	GENE3165X	Unknown; Clone = 1339226
			GENE2758X	Unknown; Clone = 682995
	Gene core 4	14	GENE1835X	Unknown; Clone = 1357915
			GENE1836X	Unknown UG Hs.190487 ESTs; Clone = 1358277
	Gene core 5	11	GENE1836X	Unknown UG Hs.190487 ESTs; Clone = 1358277
			GENE1933X	Unknown UG Hs.221606 ESTs; Clone = 1358190
AB-like-DLBCL	Gene core 1	24	GENE3939X	Unknown UG Hs.169081 ets variant gene 6 (TEL oncogene); Clone = 1355435
			GENE1639X	OSF-2os = osteoblast-specific factor = putative bone adhesion protein with homology with the insect
	Gene core 2	22	GENE1836X	Unknown UG Hs.190487 ESTs; Clone = 1358277
			GENE1639X	OSF-2os = osteoblast-specific factor = putative bone adhesion protein with homology with the insect
	Gene core 3	22	GENE1835X	Unknown; Clone = 1357915
			GENE1639X	OSF-2os = osteoblast-specific factor = putative bone adhesion protein with homology with the insect
	Gene core 4	12	GENE1835X	Unknown; Clone = 1357915
			GENE3939X	Unknown UG Hs.169081 ets variant gene 6 (TEL oncogene); Clone = 1355435
	Gene core 5	10	GENE3067X	Similar to dead box, Y isoform (DBY) = probable ATP-dependent RNA helicase; Clone = 1350869
			GENE1639X	OSF-2os = osteoblast-specific factor = putative bone adhesion protein with homology with the insect
CLL	Gene core 1	23	GENE1835X	Unknown; Clone = 1357915
			GENE2395X	Unknown UG Hs.59368 ESTs; Clone = 1353778
	Gene core 2	17	GENE2039X	Unknown UG Hs.29052 ESTs, Highly similar to (defline not available 4103857) [M. musculus]; Clone = 1316906
			GENE2395X	Unknown UG Hs.59368 ESTs; Clone = 1353778
	Gene core 3	7	GENE1835X	Unknown; Clone = 1357915
			GENE2039X	Unknown UG Hs.29052 ESTs, Highly similar to (defline not available 4103857) [M. musculus]; Clone = 1316906
			GENE2395X	Unknown UG Hs.59368 ESTs; Clone = 1353778
	Gene core 4	7	GENE1836X	Unknown UG Hs.190487 ESTs; Clone = 1358277
			GENE1141X	MAPKKK5 = ASK1 = mitogen-activated kinase kinase kinase 5; Clone = 504877
			GENE2319X	Unknown UG Hs.125719 ESTs; Clone = 1350862
	Gene core 5	5	GENE3072X	APC = adenomatous polyposis coli protein; Clone = 125294
			GENE2501X	Titin; Clone = 1251981
FL	Gene core 1	9	GENE1933X	Unknown UG Hs.221606 ESTs; Clone = 1358190
			GENE2310X	Unknown; Clone = 703659
	Gene core 2	7	GENE1835X	Unknown; Clone = 1357915
			GENE3068X	Unknown UG Hs.126784 ESTs; Clone = 826721
	Gene core 3	7	GENE1835X	Unknown; Clone = 1357915
			GENE3068X	Unknown UG Hs.126784 ESTs; Clone = 826721
			GENE3704X	CD45; Clone = 239287
	Gene core 4	7	GENE3068X	Unknown UG Hs.126784 ESTs; Clone = 826721
			GENE2415X	Unknown; Clone = 1289937
	Gene core 5	5	GENE1835X	Unknown; Clone = 1357915
			GENE2109X	IL-4 receptor alpha chain; Clone = 714453
Normal	Gene core 1	56	GENE1836X	Unknown UG Hs.190487 ESTs; Clone = 1358277
			GENE3795X	AIM2 = interferon-inducible protein = associated with chromosome 6-mediated suppression of melanoma
	Gene core 2	51	GENE3073X	SEC14-like; Clone = 685336
			GENE3795X	AIM2 = interferon-inducible protein = associated with chromosome 6-mediated suppression of melanoma
	Gene core 3	36	GENE1835X	Unknown; Clone = 1357915
			GENE3795X	AIM2 = interferon-inducible protein = associated with chromosome 6-mediated suppression of melanoma

Table 2 (*continued*)

Phenotypes	Gene cores	Appearance frequency	Gene ID	Gene name
	Gene core 4	28	GENE1931X	Unknown UG Hs.123650 ESTs; Clone = 1336983
			GENE3795X	AIM2 = interferon-inducible protein = associated with chromosome 6-mediated suppression of melanoma
	Gene core 5	28	GENE1835X	Unknown; Clone = 1357915
			GENE1836X	Unknown UG Hs.190487 ESTs; Clone = 1358277

a typical expression profile were deleted. Most of the cDNA clones on the microarray were chosen from a germinal centre B-cell library because of the suspected importance of the germinal centre B-cell to the genesis of non-Hodgkin's lymphomas.

3.2. Experiment

The whole work is summarized in Table 1. All algorithms in this paper are achieved using MATLAB 6.5 and JAVA 1.4. The retrieval process of the redundant genes is described in Fig. 2. In this figure, classifying performance falls when feature genes are deleted, and this retrieval process stops at the point where the accuracy has no apparent decline in the consecutive five repeating process. We retrieved the deleted genes before the point where these genes become important for classifying and biological interpretations.

3.3. Establishing molecular mechanisms of lymphoma subtypes

The target phenotypes are multiple subtypes of lymphoma. Thus, our work is conducted to identify the significant molecular mechanisms (disease-relevant genes and functions) that underpin the complex molecular mechanisms for the distinction of multiple phenotypes. Gene cores of multiple phenotypes are shown in Table 2. Although many of the genes lack the function annotation, these genes are worthy of further investigation.

Still some other core genes are intriguing: Osteoblast-specific factor (OSF-2os), a putative bone adhesion protein with homology in insects and named GENE1639X, is a key gene of five gene cores that are strongly related to AB-like-DLBCL samples and over-expressed in them. Gui et al. [22] developed a threshold gradient descent method for the Cox model to select genes that are relevant to DLBCL patients' survival. An osteoblast-specific factor is included in the genes that were identified to be related to the risk of death, which belongs to the lymph node signature defined by Rosenwald et al. [23] using clustering analysis of genes. Osteoblast-specific factor-2 has been described as a transcription factor of osteopontin (OPN), the protein associated with the progression and metastasis formation of various cancer types. Our work confirmed that OSF-2os is over-expressed in AB-like-DLBCL samples, and this finding might well be used for an understanding of the progression and metastasis of AB-like-DLBCL.

GENE2395X is a strongly relevant gene for the distinction of CLL samples from other samples and appears in three gene cores. Moreover, GENE2395X is mapped in MAPK signaling pathway in KEGG database. Another gene, GENE1141X, mitogen-activated kinase kinase kinase 5, is also relevant to CLL samples. This implies that the MAPK signaling pathway is affected in CLL development and progression, and both GENE2395X and GENE1141X play the key role in this pathway. More evidence described in the next text confirms that the MAPK signaling pathway is activated abnormally. Another core gene, Titin, was reported to be over-expressed in CLL samples [23].

BCL-2 gene in the set $\{fF_3\}$ associated with CLL samples is strongly over-expressed. Multiple lines of evidence from molecular biological studies imply that the over-expression of this gene occurs in many forms of leukaemia, and so contributes to the relentless accumulation of lymphocytes that fail to die and to their resistance to chemotherapy. But in our signature, over-expression of BCL-2 is only typical in CLL samples. High expression of the antiapoptotic protein BCL-2, a profound inhibitor of programmed cell death, has been reported in the vast majority of B-cell CLLs [24–26]. Fegan's research [27] has shown that BCL-2 protein is one of the several proteins that regulate cell death. BCL-2 protein inhibits programmed cell death and is consistently over-expressed in B-CLL patients. Over-expression of BCL-2 is present in over 90% of B-CLL patients.

For AB-like-DLBCL and GCB-like-DLBCL, it is hard to capture more information on the molecular links. The main causes are that some feature genes are lacking in function annotation and the significant function concepts of GO mapped by remnant feature genes are not specific enough to explain anything.

For FL and CLL, We further analyzed function concepts mapped by these relevant genes to understand the disease mechanism at the functional level.

The function concepts mapped by feature genes of CLL are listed in Table 3. There are some function concepts such as 'MAPKKK cascade', 'MAP kinase kinase kinase', 'protein tyrosine kinase', 'protein tyrosine phosphatase', 'activation of JUN kinase' and 'small GTPase-mediated signal transduction' to be selected. There are published papers reporting that JUN kinase activation has been implicated as a major player in the induction of apoptosis by a number of agents and has also recently been shown to result in p53 activation and subsequent p53-mediated apoptosis in sympathetic neurons [28,29]. These function

Table 3
Significant function concepts (empirical $P < 0.05$) mapped by feature genes of CLL samples

Gene name and note	Up/down-regulation	GO acc	Term name	Term type
JAK2 tyrosine kinase; Clone = 789379	1	GO:0000074	Regulation of cell cycle	Biological process
BCL-2; Clone = 342181	1	GO:0000074	Regulation of cell cycle	Biological process
E2F-4 = pRB-binding transcription factor; Clone = 51058	1	GO:0000074	Regulation of cell cycle	Biological process
Cdc25B = M-phase inducer phosphatase 2; Clone = 1354190	1	GO:0000074	Regulation of cell cycle	Biological process
Cdc25B = M-phase inducer phosphatase 2; Clone = 786067	−1	GO:0000074	Regulation of cell cycle	Biological process
Unknown; Clone = 1369098	1	GO:0000165	MAPKKK cascade	Biological process
MAPKKK5 = ASK1 = mitogen-activated kinase kinase kinase 5; Clone = 504877	1	GO:0000165	MAPKKK cascade	Biological process
ICSBP = Interferon concensus sequence binding protein; Clone = 290230	1	GO:0003702	RNA polymerase II transcription factor	Molecular function
BCL-2; Clone = 342181	1	GO:0003750	Cell cycle regulator	Molecular function
JAK2 tyrosine kinase; Clone = 789379	1	GO:0004672	Protein kinase	Molecular function
Unknown; Clone = 1369098	1	GO:0004709	MAP kinase kinase kinase	Molecular function
MAPKKK5 = ASK1 = mitogen-activated kinase kinase kinase 5; Clone = 504877	1	GO:0004709	MAP kinase kinase kinase	Molecular function
JAK2 tyrosine kinase; Clone = 789379	1	GO:0004713	Protein tyrosine kinase	Molecular function
SKAP55 = associates with the protein tyrosine kinase p59fyn in human T-lymphocytes; Clone = 1320051	1	GO:0004713	Protein tyrosine kinase	Molecular function
Lyn = tyrosine kinase; Clone = 1289379	1	GO:0004713	Protein tyrosine kinase	Molecular function
FGR tyrosine kinase; Clone = 347751	1	GO:0004713	Protein tyrosine kinase	Molecular function
Cdc25B = M-phase inducer phosphatase 2; Clone = 1354190	1	GO:0004725	Protein tyrosine phosphatase	Molecular function
Cdc25B = M-phase inducer phosphatase 2; Clone = 786067	1	GO:0004725	Protein tyrosine phosphatase	Molecular function
Unknown UG Hs.5103 ESTs; Clone = 1308105	1	GO:0004725	Protein tyrosine phosphatase	Molecular function
Unknown UG Hs.5103 ESTs; Clone = 1308810	1	GO:0004725	Protein tyrosine phosphatase	Molecular function
Unknown; Clone = 1370135	1	GO:0004725	Protein tyrosine phosphatase	Molecular function
Protein tyrosine phosphatase, non-receptor type 12; Clone = 289965	1	GO:0004726	Non-membrane spanning protein tyrosine phosphatase	Molecular function
Protein tyrosine phosphatase, non-receptor type 12; Clone = 289965	1	GO:0004726	Non-membrane spanning protein tyrosine phosphatase	Molecular function
Sphingomyelin phosphodiesterase 2,neutral membrane (neutral sphingomyelinase); Clone = 1319288	1	GO:0004767	Sphingomyelin phosphodiesterase	Molecular function
Neurotrophic tyrosine kinase, receptor, type 3 (TrkC); Clone = 35356	1	GO:0005016	Neurotrophin TRKC receptor	Molecular function
Zinc finger protein 42 MZF-1; Clone = 490387	1	GO:0006355	Regulation of transcription, DNA-dependent	Biological process
Protein phosphatase 2C gamma; Clone = 530950	1	GO:0006470	Protein amino acid dephosphorylation	Biological process
Protein tyrosine phosphatase, non-receptor type 12; Clone = 289965	1	GO:0006470	Protein amino acid dephosphorylation	Biological process
Protein tyrosine phosphatase, non-receptor type 12; Clone = 289965	1	GO:0006470	Protein amino acid dephosphorylation	Biological process
Unknown UG Hs.5103 ESTs; Clone = 1308105	1	GO:0006470	Protein amino acid dephosphorylation	Biological process
Unknown UG Hs.5103 ESTs; Clone = 1308810	1	GO:0006470	Protein amino acid dephosphorylation	Biological process
Unknown; Clone = 1370135	1	GO:0006470	Protein amino acid dephosphorylation	Biological process
Protein phosphatase 2C gamma; Clone = 1357352	−1	GO:0006470	Protein amino acid dephosphorylation	Biological process
Sphingomyelin phosphodiesterase 2,neutral membrane (neutral sphingomyelinase); Clone = 1319288	1	GO:0006684	Sphingomyelin metabolism	Biological process
MDA-7 = melanoma differentiation-associated 7 = anti-proliferative; Clone = 267158	1	GO:0006915	Apoptosis	Biological process
BCL-2; Clone = 342181	1	GO:0006916	Anti-apoptosis	Biological process
Titin; Clone = 1251981	1	GO:0006942	Regulation of striated muscle contraction	Biological process
Titin; Clone = 358640	1	GO:0006942	Regulation of striated muscle contraction	Biological process
BCL-2; Clone = 342181	1	GO:0006959	Humoral immune response	Biological process
CD1C; Clone = 428103	−1	GO:0006960	Antimicrobial humoral response (sensu Invertebrata)	Biological process
Unknown; Clone = 1369098	1	GO:0007257	Activation of JUN kinase	Biological process
MAPKKK5 = ASK1 = mitogen-activated kinase kinase kinase 5; Clone = 504877	1	GO:0007257	Activation of JUN kinase	Biological process

Table 3 (*continued*)

Gene name and note	Up/down-regulation	GO acc	Term name	Term type
Ack = p21cdc42Hs kinase; Clone = 1143183	1	GO:0007264	Small GTPase mediated signal transduction	Biological process
ABR = guanine nucleotide regulatory protein; Clone = 52408	1	GO:0007264	Small GTPase mediated signal transduction	Biological process
BCL-2; Clone = 342181	1	GO:0008189	Apoptosis inhibitor	Molecular function
Unknown UG Hs.117302 ESTs; Clone = 1234067	1	GO:0008283	Cell proliferation	Biological process
Vascular endothelial growth factor B; Clone = 1271813	1	GO:0008284	Positive regulation of cell proliferation	Biological process
Vascular endothelial growth factor B; Clone = 167296	1	GO:0008284	Positive regulation of cell proliferation	Biological process
Cdc25B = M-phase inducer phosphatase 2; Clone = 1354190	−1	GO:0008284	Positive regulation of cell proliferation	Biological process
Cdc25B = M-phase inducer phosphatase 2; Clone = 786067	−1	GO:0008284	Positive regulation of cell proliferation	Biological process
BCL-2; Clone = 342181	1	GO:0008285	Negative regulation of cell proliferation	Biological process
Titin; Clone = 1251981	1	GO:0008307	Structural constituent of muscle	Molecular function
Titin; Clone = 358640	1	GO:0008307	Structural constituent of muscle	Molecular function
FGR tyrosine kinase; Clone = 347751	1	GO:0009615	Response to viruses	Biological process
Protein phosphatase 2C gamma; Clone = 530950	1	GO:0015071	Protein phosphatase type 2C	Molecular function
Protein phosphatase 2C gamma; Clone = 1357352	−1	GO:0015071	Protein phosphatase type 2C	Molecular function

Table 4
Significant function concepts (empirical $P < 0.05$) mapped by feature genes of FL samples

Gene name and note	Up/down-regulation	GO acc	Term name	Term type
Unknown; Clone = 1300358	1	GO:0003700	Transcription factor	Molecular function
Unknown; Clone = 1300358	1	GO:0003713	Transcription co-activator	Molecular function
SMRT = silencing mediator of retinoid and thyroid hormone action = co-repressor; Clone = 235911	−1	GO:0003714	Transcription co-repressor	Molecular function
SMRT = silencing mediator of retinoid and thyroid hormone action = co-repressor; Clone = 723911	−1	GO:0003714	Transcription co-repressor	Molecular function
Adenosine triphosphatase, calcium; Clone = 1357222	1	GO:0004002	Adenosinetriphosphatase	Molecular function
Adenosine triphosphatase, calcium; Clone = 1335110	1	GO:0004002	Adenosinetriphosphatase	Molecular function
SIP-110 = signaling inositol polyphosphate 5 phosphatase; Clone = 1305138	1	GO:0004445	Inositol-1,4,5-trisphosphate	Molecular function
Unknown; Clone = 1241453	1	GO:0004725	Protein tyrosine phosphatase	Molecular function
Interferon gamma receptor beta chain; Clone = 1352434	1	GO:0004906	Interferon-gamma receptor	Molecular function
FGFR4 = Fibroblast growth factor receptor 4; Clone = 784224	−1	GO:0005007	Fibroblast growth factor receptor	Molecular function
CD151 = platelet-endothelial tetraspan antigen 3; Clone = 310348	−1	GO:0005194	Cell adhesion molecule	Molecular function
Adenosine triphosphatase, calcium; Clone = 1357222	1	GO:0005388	Calcium-transporting ATPase	Molecular function
Adenosine triphosphatase, calcium; Clone = 1335110	1	GO:0005388	Calcium-transporting ATPase	Molecular function
Unknown; Clone = 1300358	1	GO:0006366	Transcription from Pol II promoter	Biological process
Adenosine triphosphatase, calcium; Clone = 1357222	1	GO:0006832	Small molecule transport	Biological process
Adenosine triphosphatase, calcium; Clone = 1335110	1	GO:0006832	Small molecule transport	Biological process
CD151 = platelet-endothelial tetraspan antigen 3; Clone = 310348	−1	GO:0007155	Cell adhesion	Biological process
Unknown; Clone = 1300358	1	GO:0007517	Muscle development	Biological process
FGFR4 = Fibroblast growth factor receptor 4; Clone = 784224	−1	GO:0008543	FGF receptor signaling pathway	Biological process
Interferon gamma receptor beta chain; Clone = 1352434	1	GO:0009615	Response to viruses	Biological process
Interferon gamma receptor beta chain; Clone = 1352434	1	GO:0009619	Resistance to pathogenic bacteria	Biological process

concepts indicate that two pathways of cellular signal transduction, tyrosine protein kinase-mitogen activated protein kinase pathway (TPK-MAPK) and small GTPase-mediated signal transduction pathway, are both activated abnormally, which brings on an excessive proliferation of tumor cells. Genes mapped onto function concepts MAPKKK cascade, MAP kinase, protein tyrosine kinase and activation of JUN kinase are up-regulated entirely, which means that the TPK-MAPK signal transduction pathway is activated abnormally and persistently. From the information listed in Table 3, we surmise that one of the earlier processes is the activation of the protein tyrosine kinases (JAK2 tyrosine kinase, SKAP55, lyn, FGR tyrosine kinase), which results in the activation of MAPK cascade (MAPKKK, unknown clone = 1369098). Some researchers report that Protein kinase A (PKA) and mitogen-activated protein kinases (MAPKs) have been involved in the apoptosis of B-CLL cells [30,31]. However, all the genes associated with protein tyrosine phosphatase are also up-regulated. It is possible that the level of protein

tyrosine phosphatase is correspondingly up-regulated in order to keep a low level of tyrosine phosphorylation, while the expression level of tyrosine kinase is up-regulated. In addition, genes linked to another signal transduction pathway, a small GTPase-mediated signal transduction, are also over-expressed entirely. This means such a pathway is also activated, but now we fail to find any related research to confirm it. Significant function concepts we have found confirm such a theory further [32] that the occurrence of a tumor has a close relationship with signal transduction.

The function concepts mapped by feature genes of FL are listed in Table 4. Some function concepts such as 'transcription factor', 'transcription co-activator' and 'transcription co-repressor' are selected. Genes mapped onto 'transcription factor' and 'transcription co-activator' are over-expressed and 'transcription co-repressor' are down-expressed, which is consistent with tumor pathogenesis. 'Adenosinetriphosphatase', 'calcium-transporting ATPase', 'protein tyrosine phosphatase' and 'inositol-1,4,5-trisphosphate' function concepts are also selected, and genes mapped onto them are over-expressed; however, presently no evidence supports the relationship of these function concepts with FL.

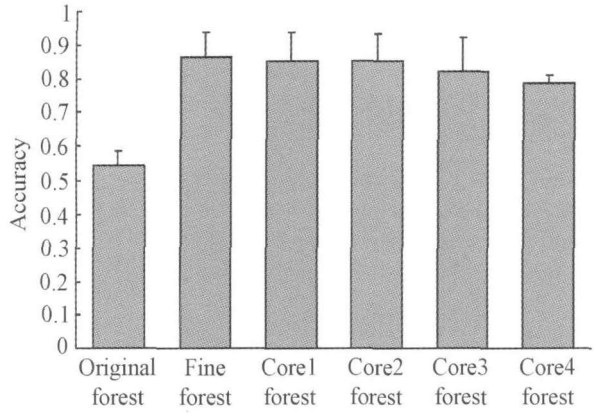

Fig. 3. Classifying performance evaluation of feature genes and hierarchical gene cores. Core N forest ($N = 1, 2, \ldots, 4$) is core forest at N level, which means that feature subsets for any class used to classify are fine feature subsets including gene cores at the Nth level.

To investigate whether the genes in the molecular mechanisms can classify these subtypes of lymphoma well, we use the Two-Level Integrating Evaluation Machine [7] to assess the classifying performance (accuracy) of the fine forest and hierarchical core forest and to compare it with the original forest when ten series of validation sets are fixed. Results in Fig. 3 show that the average accuracy of fine forest reaches 86.50%, which is remarkably higher than that of original forest (54.32%). Such a classifying performance is also high compared with that of other classifying algorithms for multi-class samples [6]. Moreover, accuracy of deeper-level core forest does not noticeably reduce along with the decrease of the number of genes in core forest.

In the feature selection approach, we performed external cross-validation [7], then with a validation set and separate classifiers we carried out feature gene subset selection. The classifiers considered are Fisher linear discrimination, Logit nonlinear discrimination, Mahal distance and K-nearest neighbor classifier. We applied a threefold cross-validation to assess the two-class discriminating ability of feature genes included in gene cores at the deepest level (core genes). At the same time, the same numbers of genes were sampled randomly from all genes to do the same work, and to assure randomness, the random sampling was repeated ten times. Results listed in Table 5 show that two-class discriminating ability of core genes is markedly higher than that of the randomly sampled genes, especially, when Fisher linear discrimination, Logit nonlinear discrimination and Mahal distance discrimination are used, the accuracy of core genes can reach around 90%.

4. Conclusion

Our approach reported in this paper extended our previously developed ensemble decision approach. Both approaches aim to mine disease-relevant genes for the classification of biological types. However, they are different. Firstly, our approaches analyze the multiple heterogeneous phenotypes with another discriminant index δ with which selected subsets would distinguish positive and negative

Table 5
Two-class classifying performance evaluation of core genes with Fisher linear discriminate, Logit nonlinear discriminate, Mahal distance and K-nearest neighbor classifier

Phenotypes		Classifier			
		Fisher linear discrimination (Mean ± SD)	Logit nonlinear discrimination (Mean ± SD)	K-nearest neighbor (Mean ± SD)	Mahal distance (Mean ± SD)
GCB-like-DLBC	Core genes	0.8536 ± 0.0996	0.8248 ± 0.1104	0.9114 ± 0.0348	0.6746 ± 0.1174
	Random genes	0.7449 ± 0.1020	0.7368 ± 0.0921	0.7658 ± 0.0740	0.5096 ± 0.1844
AB-like-DLBCL	Core genes	0.9116 ± 0.0631	0.9233 ± 0.0625	0.8678 ± 0.0839	0.7531 ± 0.0755
	Random genes	0.7574 ± 0.1050	0.6975 ± 0.1018	0.7863 ± 0.0823	0.4573 ± 0.2164
CLL	Core genes	1.0000 ± 0.0000	0.9976 ± 0.0118	0.9672 ± 0.0482	0.8156 ± 0.1070
	Random genes	0.8997 ± 0.0821	0.8929 ± 0.0636	0.9117 ± 0.0679	0.3384 ± 0.1647
FL	Core genes	0.9275 ± 0.0514	0.9299 ± 0.0379	0.8836 ± 0.0705	0.5531 ± 0.0792
	Random genes	0.8451 ± 0.0932	0.8883 ± 0.0662	0.9182 ± 0.0618	0.3493 ± 0.2209
Normal	Core genes	0.8869 ± 0.0925	0.8663 ± 0.0921	0.8389 ± 0.0601	0.8834 ± 0.0727
	Random genes	0.7453 ± 0.1071	0.8413 ± 0.0965	0.7990 ± 0.0742	0.4121 ± 0.1239

phenotypes in a highly unbalanced class distribution. Secondly, by the supervision of classifying the performance of some so-called 'redundant' genes we retrieved as many redundant genes as possible, which are very important in elucidating the complex genetic architecture of a complex disease. Finally, in addition to the single gene marker, significant function concepts found by mapping genes onto a gene function classifying frame Gene Ontology are analyzed, which avoids unnecessary loss of important genes during the microarray design, and improves the interpretability of the data mining results.

In our study, we proposed a method for the extraction of critical disease-relevant genes through multiple feature subsets, each being selected based on its classifying performance. By retrieving the 'redundant' genes, the majority of strongly relevant and partially relevant genes can be identified. Our findings support our speculation that retrieving feature genes is efficient for extracting 'redundant' genes. For example, gene SMRT, has two clones that have been both identified. By mapping these relevant genes onto GO, the molecular bridge of multiple phenotypes elucidating the disease mechanism is unraveled, which may provide valuable clues for the further research.

Feature genes and hierarchical gene cores identified by our method have achieved accuracy in the classification of multiple phenotypes. In the study, we performed external cross-validations. Using external classifiers we obtained an unbiased estimate of the classification performance of the deepest gene cores, and a high accuracy which shows gene cores definitely have discriminate ability of subtypes of lymphoma. In the future, genes in gene cores which lack function annotations need further investigations.

Acknowledgements

This work was supported in part by the National High Tech Development Project of China (Grant No. 2007AA02Z329), the National Natural Science Foundation of China (Grant Nos. 30571034, 30570424 and 20060213024), the grant for Outstanding Overseas Scientist, Department of Education, Heilongjiang Province (Grant No. 1055HG009), Natural Science Foundation of Heilongjiang Province (Grant Nos. ZJG0501, GB03C602-4 and F2004-02) and Health Department of Heilongjiang Province Key Project (2005-39).

References

[1] Otomo A, Hadano S, Okada T, et al. ALS2, a novel guanine nucleotide exchange factor for the small GTPase Rab5, is implicated in endosomal dynamics. Hum Mol Genet 2003;12(14):1671–87.
[2] Bian ZQ, Zhang X. Pattern recognition. Beijing: Tsinghua University Press; 2000.
[3] Diaz-Uriarte R, Alvarez de Andres S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006;7:3–16.
[4] Yang K, Cai Z, Li J, et al. A stable gene selection in microarray data analysis. BMC Bioinformatics 2006;7:228–44.
[5] Xiong M, Fang X, Zhao J. Biomarker identification by feature wrappers. Genome Res 2001;11(11):1878–87.
[6] Puuronen S, Tsymbal A. Local feature selection with dynamic integration of classifiers. Fundam Inform 2001;47:91–117.
[7] Li X, Rao S, Wang Y, et al. Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling. Nucleic Acids Res 2004;32(9):2685–94.
[8] Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25(1):25–9.
[9] Gruber TR. The role of common ontology in achieving sharable, reusable knowledge bases. In: Proceedings of the second international conference on principles of knowledge representation and reasoning; 1991. p. 601–2.
[10] Gruber TR. Toward principles for the design of ontologies used for knowledge sharing. Int J Hum-Comput Stud 1995;43:907–28.
[11] Kurra G, Niu W, Bhatnagar R. Mining microarray expression data for classifier gene-cores. In: Proceedings of the workshop on data mining in bioinformatics; 2001. p. 8–14.
[12] Eisen MB, Spellman PT, Brown PO, et al. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95(25):14863–8.
[13] Hughes TR, Marton MJ, Jones AR, et al. Functional discovery via a compendium of expression profiles. Cell 2000;102(1):109–26.
[14] Miki R, Kadota K, Bono H, et al. Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. Proc Natl Acad Sci USA 2001;98(5):2199–204.
[15] Jiang ZJ. Medical statistics. Beijing: Public Health Press; 1997. [in Chinese].
[16] Benson DA, Karsch-Mizrachi I, Lipman DJ, et al. GenBank: update. Nucleic Acids Res 2004;32(Database issue):D23–6.
[17] Boguski MS, Schuler GD. Establishing a human transcript map. Nat Genet 1995;10(4):369–71.
[18] Wheeler DL, Church DM, Federhen S, et al. Database resources of the National Center for Biotechnology. Nucleic Acids Res 2003;31(1):28–33.
[19] Pruitt KD, Maglott DR. RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Res 2001;29(1):137–40.
[20] Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403(6769):503–11.
[21] Doyen V, Fournier C, Bautin N, et al. Rheumatoid arthritis and cystic fibrosis. Rev Mal Respir 2005;22(4):667–71.
[22] Gui J, Li H. Threshold gradient descent method for censored data regression with applications in pharmacogenomics. Pac Symp Biocomput 2005:272–83.
[23] Rosenwald A, Alizadeh AA, Widhopf G, et al. Relation of gene expression phenotype to immunoglobulin mutation genotype in B cell chronic lymphocytic leukemia. J Exp Med 2001;194(11):1639–47.
[24] McConkey DJ, Chandra J, Wright S, et al. Apoptosis sensitivity in chronic lymphocytic leukemia is determined by endogenous endonuclease content and relative expression of BCL-2 and BAX. J Immunol 1996;156(7):2624–30.
[25] Schena M, Larsson LG, Gottardi D, et al. Growth- and differentiation-associated expression of bcl-2 in B-chronic lymphocytic leukemia cells. Blood 1992;79(11):2981–9.
[26] Thomas DC. Statistical methods in genetic epidemiology. Oxford: Oxford University Press; 2004.
[27] Fegan C. Molecular abnormalities in B-cell chronic lymphocytic leukaemia. Clin Lab Haematol 2001;23(3):139–48.
[28] Jarpe MB, Widmann C, Knall C, et al. Anti-apoptotic versus pro-apoptotic signal transduction: checkpoints and stop signs along the road to death. Oncogene 1998;17:1475–82.
[29] Morrison RS, Kinoshita Y. Development. p73 – guilt by association? Science 2000;289(5477):304–6.

[30] Kim DH, Lerner A. Type 4 cyclic denosine monophosphate phosphodiesterase as a therapeutic target in chronic lymphocytic leukemia. Blood 1996;92:2484–94.
[31] Pedersen IM, Buhl AM, Klausen P, et al. The chimeric anti-CD20 antibody rituximab induces apoptosis in B-cell chronic lymphocytic leukemia cells through a p38 mitogen activated protein-kinase-dependent mechanism. Blood 2002;99(4): 1314–9.
[32] Cheng YS. Tumor cell and molecular biology. Beijing: Public Surgeon Press; 2002, [in Chinese].