BIOINFORMATICS ORIGINAL PAPER
Vol. 26 no. 12 2010, pages 1506–1512
doi:10.1093/bioinformatics/btq207

Gene expression

Advance Access publication April 21, 2010

Inference of combinatorial Boolean rules of synergistic gene sets from cancer microarray datasets

Inho Park, Kwang H. Lee* and Doheon Lee*
Department of Bio and Brain Engineering, KAIST, 373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Republic of Korea

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Gene set analysis has become an important tool for the functional interpretation of high-throughput gene expression datasets. Moreover, pattern analyses based on inferred gene set activities of individual samples have shown the ability to identify more robust disease signatures than individual gene-based pattern analyses. Although a number of approaches have been proposed for gene set-based pattern analysis, the combinatorial influence of deregulated gene sets on disease phenotype classification has not been studied sufficiently.

Results: We propose a new approach for inferring combinatorial Boolean rules of gene sets for a better understanding of cancer transcriptome and cancer classification. To reduce the search space of the possible Boolean rules, we identify small groups of gene sets that synergistically contribute to the classification of samples into their corresponding phenotypic groups (such as normal and cancer). We then measure the significance of the candidate Boolean rules derived from each group of gene sets; the level of significance is based on the class entropy of the samples selected in accordance with the rules. By applying the present approach to publicly available prostate cancer datasets, we identified 72 significant Boolean rules. Finally, we discuss several identified Boolean rules, such as the rule of *glutathione metabolism* (*down*) *and prostaglandin synthesis regulation* (*down*), which are consistent with known prostate cancer biology.

Availability: Scripts written in Python and R are available at http://biosoft.kaist.ac.kr/~ihpark/. The refined gene sets and the full list of the identified Boolean rules are provided in the Supplementary Material.

Contact: khlee@biosoft.kaist.ac.kr; dhlee@biosoft.kaist.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 14, 2009; revised on March 19, 2010; accepted on April 16, 2010

1 INTRODUCTION

With the advent of microarray technologies, genome-wide gene expression profiling has become a crucial tool in biomedical research during the past few decades. In cancer research, genome-wide gene expression profiling has been used to discover new cancer subtypes (Golub *et al*., 1999; Lapointe *et al*., 2004; Sorlie *et al*., 2003); to develop gene expression signatures for cancer diagnosis, prognosis or prediction of drug responsiveness (Nevins and Potti, 2007; Potti *et al*., 2006; van't Veer *et al*., 2002); and to identify cancer associated signaling pathways or cellular processes (Bild *et al*., 2006; Heiser *et al*., 2009).

In a typical analysis of cancer gene expression profiles, individual genes are ranked by the statistical significance of the differential expression between two different experimental conditions; several tens of top-ranked genes are then selected for further analysis, such as cancer classification (Golub *et al*., 1999; Lapointe *et al*., 2004; Nevins and Potti, 2007; Potti *et al*., 2006; Sorlie *et al*., 2003; van't Veer *et al*., 2002) and functional enrichment analysis (Al-Shahrour *et al*., 2007; Dennis *et al*., 2003; Huang *et al*., 2009). To investigate the combinatorial influences of deregulated genes on disease phenotype classification, researchers have proposed a number of multivariate approaches (Bo and Jonassen, 2002; Mukherjee *et al*., 2009; Varadan and Anastassiou, 2006). However, all of these individual gene-based approaches tend to produce unstable gene lists or combinations of genes due to the small sample size used in individual studies (Ambroise *et al*., 2002; Chuang *et al*., 2007).

A recently developed type of gene set analysis aims to directly evaluate the statistical significance of coordinated expression changes of genes belonging to specific pathways or functional categories without selecting an arbitrary number of highly ranked genes in advance, and the gene sets are usually derived from the Gene Ontology (Ashburner *et al*., 2000), Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa *et al*., 2004), Reactome (Vastrik *et al*., 2007), and other knowledge bases. These approaches have the ability to detect gene sets whose constituent genes show 'subtle but coordinated expression changes,' which might not be detected by conventional functional enrichment methods (Dinu *et al*., 2009; Mootha *et al*., 2003; Nam and Kim, 2008; Subramanian *et al*., 2005; Tian *et al*., 2005). Moreover, several cancer classification approaches based on the gene set activities inferred from individual samples perform better than other cancer classification approaches based on the expression levels of individual genes (Edelman *et al*., 2006; Lee *et al*., 2008a; Levine *et al*., 2006; Pang *et al*., 2006).

Besides the gene set based cancer classification approaches, various other approaches to the study of interdependencies among gene modules have been proposed. Segal *et al*. (2004, 2003) proposed a module network for learning about transcriptionally coexpressed modules and their dependency structure so that gene expression patterns in different types of cancer could be characterized. That model differs from other gene set dependency models in that it learns about the gene modules and their dependency structure simultaneously from data rather than from

*To whom correspondence should be addressed.

predefined gene sets. Tomlins *et al.* (2006) defined more than 14 000 molecular concepts as gene sets of functionally related genes or differentially expressed genes detected from cancer gene expression datasets; they also constructed a molecular concept map that connects two molecular concepts provided the number of overlapping genes between them is statistically significant. Along with the advantage of the unprecedented large collection of gene sets, the approach of Tomlins *et al.* (2006) can relate a list of the differentially expressed genes of an experiment to other previously conducted experiments and to the curated concepts. Tomlins *et al.* (2006) used the molecular concept map to analyze prostate cancer expression profiles, and they could detect enriched subnetworks related to prostate cancer progression models (Tomlins *et al*., 2006). However, with their approach there is still a need to select a number of differentially expressed genes between the two experimental conditions, so that enriched subnetworks can be detected; and that requirement makes it difficult to deal with heterogeneous patterns of gene set activities among cancer samples. Pang *et al.* proposed the use of a random forest algorithm for pathway clustering. They use the genes in each pathway to build a random forest classifier and then use the class votes of each classifier to construct feature vectors of each pathway. The feature vectors are then applied to a tight clustering algorithm for identification of the pathway clusters (Pang *et al*., 2008).

Inferring gene set dependency models based on gene set activities inferred for each sample has been proposed recently. Edelman *et al.* proposed a method of detecting pathway dependencies in prostate cancer progression models with 639 canonical pathways available at the Molecular Signature Database (MSigDB). In their approach, they used ASSESS to measure enrichment scores of pathways in individual samples (Edelman *et al*., 2006). They also used the covariance structure of the enrichment scores to infer a pathway dependency structure relevant to the progression models (Edelman *et al*., 2008). Although their approach was successful in identifying pathways relevant to the cancer progression models, they have not explicitly used the pathway dependency models to construct predictive models.

In this work, we propose a new approach to the task of inferring combinatorial Boolean rules of gene sets for a better understanding of cancer gene expression profiles as well as cancer classification. To this end, we first identify coherently expressed submodules of each gene set belonging to the 639 canonical pathways available at the MSigDB (Subramanian *et al*., 2005) and use the submodules as background gene sets for further analysis. Second, we infer gene set activities in individual samples by using gene expression profiles and then binarize them. Third, we construct a gene set synergy network (Watkinson *et al*., 2008); we use that network to search for small groups of synergistic gene sets that provide rich information on the disease status of samples. Finally, we extract significant Boolean rules of the gene sets within each identified group and validate the rules by using independent test datasets. For comparison with other approaches, we use a random forest classifier for classification analysis of the identified Boolean rules (Breiman, 2001).

2 METHODS

2.1 Datasets and preprocessing

We analyzed publicly available prostate cancer gene expression datasets with our proposed approach. The datasets of Lapointe *et al*. (2004) and Singh *et al*. (2002) were used for learning the Boolean rules of gene sets, and the datasets of Tomlins *et al*. (2006) and Yu *et al*. (2004) were used for testing the learned Boolean rules. Table 1 summarizes the properties of the datasets.

Table 1. Prostate cancer microarray datasets

Dataset	Platform	n_n	n_t
Traing set			
Singh *et al.* (2002)	HG-U95A	52	50
Lapointe *et al.* (2004)	Spotted cDNA	41	62
Test set			
Yu *et al.* (2004)	HG-U95A	18	65
Tomlins *et al.* (2006)	Spotted cDNA	27	32

The parameters n_n and n_t are the number of normal samples and cancer samples in a dataset.

Where <20% of the values of the probe set in a dataset were missing, we used the LLSimpute method (Kim *et al*., 2005) to impute the missing values. Whenever a higher portion of values was missing, we excluded the probe set from the subsequent analysis. We performed quantile normalization among the arrays (Bolstad *et al*., 2003), and we finally normalized the expression profile of each probe set to approximate a standard normal distribution across the samples. The expression profiles of the probe sets for each gene were summarized with the mean of their expression values.

2.2 Refinement of gene sets

As for gene sets, we downloaded the 639 canonical pathways from the MSigDB (Subramanian *et al*., 2005). Because the inclusion of genes that are not coherent with the other genes in a gene set can hinder the performance of our analysis, we redefined the 639 canonical pathways by identifying tightly coregulated submodules in the training datasets.

The coherence of the expression values of a gene set can be measured in a number of ways. One most widely used measurement approaches is to use the fraction of significantly correlated pairs of genes out of all the possible pairs of genes in a gene set (Pilpel *et al*., 2001). To identify tightly coregulated submodules for each gene set, we first calculated the pairwise rank correlation coefficients of all the pairs of genes in the gene set. We then constructed a correlation gene network that connects genes that have a significant positive rank correlation coefficient ($P < 0.05$) and applied a hierarchal graph clustering method called the fitHRG algorithm (Clauset *et al*., 2008) to the gene correlation network so that we could identify densely connected subnetworks. For each gene set, we selected maximal subnetworks with a connection ratio >0.8. Subnetworks composed of less than three genes were excluded from further analysis. As a result, we obtained 809 gene sets from the analysis of the gene correlation networks derived from the 639 canonical pathways. The redefined gene sets are provided in the Supplementary Materials.

2.3 Inference of gene set activity

After assuming that genes in a refined gene set are coherent, we used the normalized mean of expression values as the activity of a gene set in an individual sample. Let z_{kj} be a normalized expression value of a gene, g_k , in a sample, j , and let P_i be a set of genes. The activity of gene set P_i in a sample, j, can then be measured with the following equation (Jiang and Gentleman, 2007; Tian *et al*., 2005):

$$a_{ij} = \frac{\sum_{g_k \in P_i} z_{kj}}{\sqrt{|P_i|}}. \quad (1)$$

We binarized the activity values of each gene set by using a median binarization method for mining Boolean rules of gene sets.

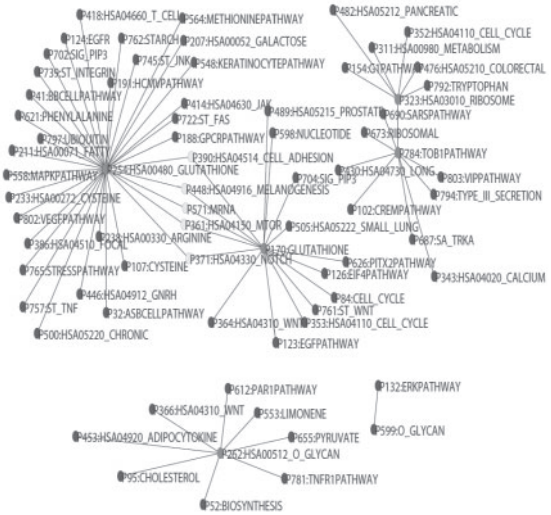

Fig. 1. A gene set synergy network extracted from the training datasets. In a simplification of the network, this figure shows only pairs of gene sets that have high mutual information with sample phenotypes ($P<0.01$) among the synergistic gene set pairs ($P<0.05$). The network is visualized with the Pajek software (Batagelj and Mrvar, 2002).

2.4 Gene set synergy network

To search for interdependency structures among the refined 809 gene sets, we used the recently developed information theoretic measure of synergy to measure the level of cooperativity between all the pair of gene sets. The synergy between two predictive random variables, X and Y, with respect to a dependent random variable, C, which denotes the disease status of a sample in our framework, is defined as follows (Anastassiou, 2007; Hanczar *et al*., 2007; Watkinson *et al*., 2008):

$$\mathrm{Syn}(X,Y;C)=I(X,Y;C)-[I(X;C)+I(Y;C], \quad (2)$$

where the $I(X;C)$ is used to denote the mutual information between the two random variables X and C. Thus, the synergy, $\mathrm{Syn}(X,Y;C)$, can be interpreted as the amount of additional information about C when the two predictive variables X and Y are considered simultaneously rather than individually. To assess the statistical significance of synergy, we use a permutation test that randomly shuffles gene set activities within each class of disease phenotype. As a result, we identified 421 pairs of gene sets that have significant synergy ($P<0.05$) as well as significant mutual information with the sample phenotypes ($P<0.05$) from the training datasets. Figure 1 shows the synergy network constructed with the pairs of synergistic gene sets. The synergy network shows that prostate cancer related gene sets, namely the glutathione metabolism and *O*-glycan biosynthesis gene sets, have many synergistic partner gene sets for prostate cancer classification.

2.5 Synergistic hierarchical clustering

We searched synergistic clusters of gene sets that have rich information about the sample phenotypes within the synergy network. For this purpose, we constructed dendrograms with the following procedure:

(1) Each gene set is considered an isolated gene set cluster.

(2) The pairwise synergies among the gene set clusters are computed with the Equation (3).

(3) Two clusters that have highest synergy are merged.

(4) Steps 2 and 3 are repeated until only k clusters remain.

Equation (3) is expressed as follows:

$$\mathrm{Syn}_\alpha(PS_i,PS_j;C)=I(PS_i\cup PS_j;C)^\alpha-[I(PS_i;C)^\alpha+I(PS_j;C)^\alpha]. \quad (3)$$

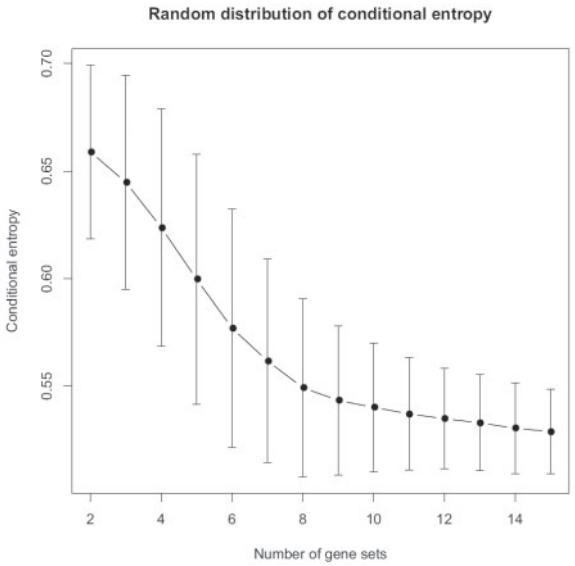

Fig. 2. A random distribution of the entropy of the sample phenotypes conditioned on a group of gene sets. We randomly sample 10 000 groups of gene sets for groups of each size.

where $PS_i=\{P_{i1},\cdots,P_{in}\}$ and $PS_j=\{P_{j1},\cdots,P_{jm}\}$ are sets of gene sets and $I(PS_i;C)$ is an abbreviation of $I(P_{i1},\cdots,P_{in};C)$. We introduce the parameter α to adjust a bias of synergy towards a negative value when $I(PS_i;C)$ and $I(PS_j;C)$ are large (as shown in the Supplementary Fig. S1). A larger α tends to greedily merge two gene set clusters to maximize joint mutual information between sample phenotypes and gene sets in a merged cluster, whereas a smaller α tends to merge two gene set clusters that have a large synergy value between the clusters. We use $\alpha=1,2,\cdots,5$ to produce gene set dendrograms and select subclusters of gene set dendrograms if the conditional entropy of sample phenotypes is <0.4. (See Fig. 2 for the random distribution of conditional entropy.) We might consider each gene set clusters produced with this approach as co-regulated functional modules in a specific

phenotype condition but not in all the conditions. The detailed procedure for constructing gene set synergy dendrograms using a hierarchical clustering algorithm is described in the Algorithm 1.

Data: $A=(a_{ij})$ is a binarized gene set activity profiles; $C=(c_j)$ denotes the disease status of a sample j.
Input: $G=(V;E)$ is a gene set synergy network described with a set of vertices (gene sets), V, and edges, E, and α is a control parameter of merge process
Output: Gene set synergy dendrograms

Dendrograms = $\{D_i=(null,null,\{P_i\})|P_i\in V\}$;
//A dendrogram is defined as (left child, right child, a set of elements)
$\Delta=\emptyset$; //Δ contains synergy values of pairs of gene set clusters

foreach *pair of* (D_i,D_j) *in Dendrograms* **do**
  synergy $=syn_\alpha(\text{ElementsOf}(D_i),\text{ElementsOf}(D_j);C)$;
  add $(D_i,D_j,\text{synergy})$ to Δ ;
end

while $|Dendrograms|==k$ **do**
  //k is the number of connected components in the graph G
  $D_{ml},D_{mr}=argmax_{D_i,D_j}\texttt{SynergyOf}(D_i,D_j;\Delta)$;
  $\Delta\setminus\{(D_i,D_j,synergy)|\{D_{ml},D_{mr}\}\cap\{D_i,D_j\}\neq\emptyset\}$;
  $D_{merged}=Merge(D_{ml},D_{mr})$;
  Dendrograms $\setminus\{D_{ml},D_{mr}\}$
  foreach D_k *in Dendrograms* **do**
    if *Connected*$(D_{merged}, D_k; G)$ **then**
      PS_k = ElementsOf(D_k) ;
      PS_{merged} = ElementsOf(D_{merged}) ;
      synergy $=syn_\alpha(PS_k,PS_{merged};C)$;
      add $(D_k,D_{merged},\text{synergy})$ to Δ ;
    end
  end
  add D_{merged} to Dendrograms
end

//Each element in V corresponds to a random variable that represent row vectors of $A=(a_{ij})$
//ElementsOf(D_i) returns a set of elements included in D_i
//Connected$(D_i,D_j;G)$ is TRUE if there is a path between a set of nodes in D_i and D_j in the graph G
//Merge(D_i,D_j) returns a dendrogram
$D_m=(D_i,D_j,$ the union of elements in D_i and $D_j)$

Algorithm 1: Construction of gene set synergy dendrograms

2.6 Entropy estimation

For a small sample size, the maximum likelihood estimator, $\hat{H}_{MLE}$, tends to underestimate the entropy of observations. When the conditional entropy of sample phenotypes is calculated, the number of samples that fall in each Boolean state of the gene set activities decreases as the number of predictive gene sets increases. To avoid underestimation of the maximum likelihood entropy estimator for such a small sample size, we use the Miller–Madow estimator, $\hat{H}_{MM}$, which is expressed as follows (Paninski *et al*., 2003):

$$\hat{p}_i=\frac{n_i}{N},$$
$$\hat{H}_{MLE}=\sum_{i=1}^{m}\hat{p}_i\log\hat{p}_i, \quad (4)$$
$$\hat{H}_{MM}=\hat{H}_{MLE}+\frac{m-1}{2N}.$$

Table 2. Significanct Boolean rules

ID	Rule	Training $H(C\|R_i)$	Testing $p_{tomlins}$	(rank-sum test) p_{yu}
$*R_1$	$\neg P_{170}\neg P_{640}$	0.0078	0.0002	<0.0001
$*R_2$	$\neg P_{352}P_{457}$	0.0093	0.0002	0.0002
R_3	$\neg P_{311}P_{457}$	0.0094	0.0001	<0.0001
R_4	$P_{457}\neg P_{476}$	0.0098	0.0025	0.0008
R_5	$P_{224}\neg P_{352}$	0.0100	0.0003	0.0001
$*R_6$	$P_{224}\neg P_{588}$	0.0100	0.0003	0.0001
R_7	$\neg P_{311}P_{673}$	0.0100	<0.0001	<0.0001
R_8	$\neg P_{352}P_{579}$	0.0106	<0.0001	0.0002
$*R_9$	$P_{114}\neg P_{154}$	0.0109	0.0004	0.0005
R_{10}	$P_{114}\neg P_{476}$	0.0109	0.0071	0.0002
R_{11}	$P_{224}\neg P_{311}$	0.0111	0.0002	<0.0001
R_{12}	$\neg P_{311}P_{321}$	0.0111	0.0001	<0.0001
R_{13}	$\neg P_{476}P_{579}$	0.0111	<0.0001	<0.0001
R_{14}	$\neg P_{132}P_{323}\neg P_{784}$	0.0119	<0.0001	<0.0001
R_{15}	$\neg P_{311}P_{692}$	0.0119	0.0002	0.0002
R_{16}	$\neg P_{352}P_{457}\neg P_{476}$	0.0125	0.0003	<0.0001
R_{17}	$\neg P_{170}P_{505}$	0.0128	0.0003	0.0006
$*R_{18}$	$\neg P_{405}\neg P_{431}P_{775}$	0.0128	0.0025	<0.0001
R_{19}	$P_{132}\neg P_{323}P_{784}$	0.0128	<0.0001	<0.0001
R_{20}	$P_{342}\neg P_{436}$	0.0132	<0.0001	<0.0001

The term $H(C|R_i)$ denotes the class entropy of samples selected by rule R_i. The selected rules (*) are investigated to find known or putative synergistic relations among the gene sets. Table 3 shows the gene sets that appear in the rules.

With Equation (4), the bias of the maximum likelihood estimator, $\hat{H}_{MLE}$, can be corrected by adding the estimated bias term $\frac{m-1}{2N}$, where is the number of possible observations (that is, normal and cancer) and m is the number of samples within each Boolean condition of the gene set activities. We use n_i and $\hat{p}_i$ to denote the number of samples and estimated probability in each phenotypic state under a Boolean condition of gene set activities.

3 RESULTS AND DISCUSSION

3.1 Boolean rules of synergistic gene sets

Synergistic hierarchical clustering methods identified 31 groups of synergistic gene sets with a low threshold of conditional entropy; i.e. $H(C|P_{i1},\cdots,P_{ik})<0.4$. From each identified small group of gene sets, we were able to extract significant Boolean rules by evaluating the significance of all the possible conjunctive Boolean functions.

3.2 Validation with independent datasets

We validated the identified Boolean rules by using the datasets of Tomlins *et al*. (2006) and Yu *et al*. (2004). To quantify the level of agreement between the Boolean rules and the gene expression profiles of each test dataset, we devised a following scheme from the log-rank sum approach (Breitling *et al*., 2004).

With the Boolean rule $R=(P_{p1}\wedge\cdots\wedge P_{pk})\wedge(\neg P_{n1}\wedge\cdots\wedge\neg P_{nm})$, we divided the literals into $PL=\{P_{p1},\ldots,P_{pk}\}$ and $NL=\{P_{n1},\ldots,P_{nm}\}$ to, respectively, represent the sets of gene sets that positively and negatively participate in the rule. We then constructed rank profiles for each element of PL and NL in ascending and descending order, respectively. Finally, we calculated the sum of the log-ranks from the rank profiles of the literals for each sample. To test the significance of each rule, we performed a rank-sum test for the sum of the log-ranks of the literals of the normal and cancer samples.

Table 3. Gene sets in the identified Boolean rules

ID	Name	Genes
P_{114}	DNA replication	MCM5, NACA, ORC3L, POLD2, POLD4, RPA1, RPS27A
P_{132}	ERK pathway	EGFR, MKNK1, PDGFRA, SOS1
P_{154}	G1 pathway	CDKN1A, SMAD3, TGFB3
P_{170}	Glutathione meta.	GSTM3-5, GPX2-3, GSTA2
P_{224}	Purine meta.	NME1, NME4, POLR2F
P_{311}	Meta. xenobiotics	ADH5, GSTP1, MGST3
P_{321}	ABC transporters	ABCA2, ABCA5, ABCA8, ABCC4, ABCD3, TAP1
P_{323}	Ribosome	FAU, RPL proteins, RPS proteins
P_{342}	Calcium signaling	ITPR3, P2RX4, SLC25A6
P_{352}	Cell cycle	CDKN1A, SMAD3, TGFB3
P_{405}	Complement and coagulation cascades	C1R, CD59, CFD, CFH, CFI, SERPINA5, SERPING1
P_{431}	Olfactory transduction	CALML3, CAMK2A, CAMK2G, CLCA2, GNAL, GUCA1A, PDE1C, PRKG1, PRKX
P_{436}	Regulation of actin cytoskeleton	ACTN1, PPP1R12B, ROCK2
P_{457}	Type 1 Diabetes mellitus	CPE, ICA1, PTPRN2
P_{476}	Colorectal cancer	FZD1, FZD7, PDGFRA, SMAD3, SOS1, TCFL2, TGFB3
P_{505}	Small lung cancer	AKT2, BCL2L1, CCNE1, CDK6, CDKN2B, COL4A4, E2F1, E2F2, FHIT, IKBKB, IKBKG, ITGA2B, ITGAV, NFKB2, NOS1, PIAS2, PIAS4, PIK3CB, PIK3CG, PIK3R2-3, RXRG, TP53, TRAF1, TRAF3
P_{579}	Ndkdynamin path.	EPS15, NME1, PICALM, SYNJ2
P_{588}	Nicotinate meta.	AOX1, CD38, NNT, NT5E
P_{640}	Prostaglandin synthesis regulation	ANXA1, ANXA4, EDNRA, EDNRB, HSD11B1, PTGER2, PTGIS, PTGS2
P_{673}	Ribosomal proteins	FAU, MRPL19, RPL proteins, RPS proteins
P_{692}	SET pathway	APEX1, CREBBP, NME1, SET
P_{775}	TERT pathway	HDAC1, MYC, MZF1, SP3
P_{784}	Tob1 pathway	SMAD, TGFB3, TGFBR3

A full list of the refined gene sets is provided in the Supplementary Material.

Table 2 shows top 20 Boolean rules validated from the test datasets; Table 3 shows the gene sets that appear in the selected rules. Several selected rules are discussed below. Figure 3 shows scatter plots of samples; these plots are based on the activities of the gene sets included in the most significant rules, R_1 and R_2.

- R_1: *glutathione metabolism* (*down*) *and prostaglandin synthesis regulation* (*down*) The glutathione s-transferases (GSTs) are known to be involved in the metabolism of carcinogens and in the defense against reactive oxygen species. Moskaluk *et al*. (1997) confirmed the down-regulation of π-class GSTs in adenocarcinoma of the prostate. In many types of cancer, COX-2 is known to be overexpressed, and the overexpression of COX-2 leads overproduction of prostaglandin E2 (Eruslanov *et al*., 2009). Note also that the highly selective inhibitor of COX-2, celecoxib, has been researched for use in cancer treatments as a supplement to pain treatments. Our analysis shows no significant change in the COX-2 expression level for the prostate cancer samples, though the other member genes PTGER2, PTGIS, EDNRA, and EDNRB are significantly down-regulated. (Jakobsson *et al*., 1999) identified a glutathione-dependent prostaglandin E synthase, which might be a cause of the down-regulation of prostaglandin synthesis regulation in the primary prostate cancer. Recently, GSTP1 and PTGDS have been proposed as key molecules in prostate cancer classification (Varadan and Anastassiou, 2006; Watkinson *et al*., 2008).

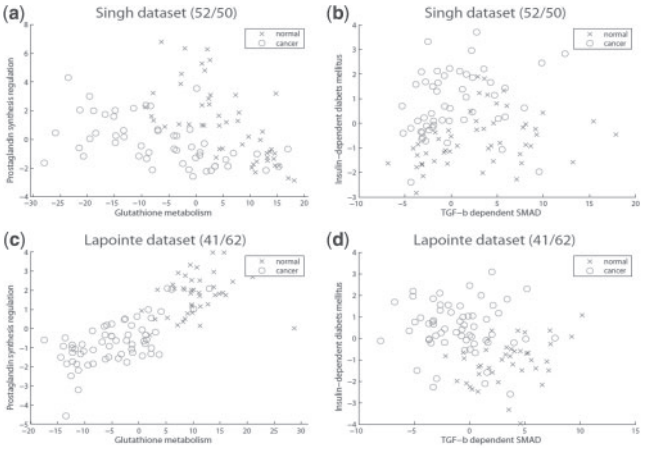

Fig. 3. The scatter plots of samples using activities of gene sets in the Boolean rules R_1 (**a** and **c**) and R_2 (**b** and **d**).

- R_2: *insulin-dependent diabetes mellitus* (*up*) *and transforming growth factor* (*TGF*)-β*-dependent SMAD signaling* (*down*) The association of prostate cancer risk and insulin-dependent diabetes has been reported in several population-based studies (Hsing *et al*., 2001; Pierce *et al*., 2008). In our analysis, the submodule of insulin-dependent diabetes mellitus that contains ICA1, PTPRN2, and CPE shows a higher level of activation in primary prostate cancer. On the other hand, an androgen receptor (AR) has been known to negatively modulate TGF-β-dependent SMAD signalling in AR-dependent prostate cancer (van der Poel *et al*., 2005); thus TGF-β-dependent SMAD signalling is down-regulated in AR-dependent prostate cancer. Interestingly, Lin *et al*. (2009) reported that the TGF-β signalling effector SMAD3 represses insulin gene transcription in pancreatic islet cells.
- R_6: *purine metabolism* (*up*) *and nicotine and nicotinate metabolism* (*down*) Alteration of the purine metabolism either facilitates production or retards degradation of adenosine, which is the main element of ATP (Linden *et al*., 2006). Obajimi *et al*. (2009) showed that inhibition of *de novo* purine synthesis in LNCaP cells results in ATP depletion. On the other hand, the nicotin and nicotinate metabolism, which includes AOX1 and NTE5, is related to antioxidation of reactive oxygen species (Dong *et al*., 2008), and those genes show significantly decreased expression patterns in the primary prostate cancer.
- R_9: *DNA replication* (*up*) *and G1 pathway* (*down*) AR has been suggested as a licensing factor for DNA replication in androgen-sensitive prostate cancer cells

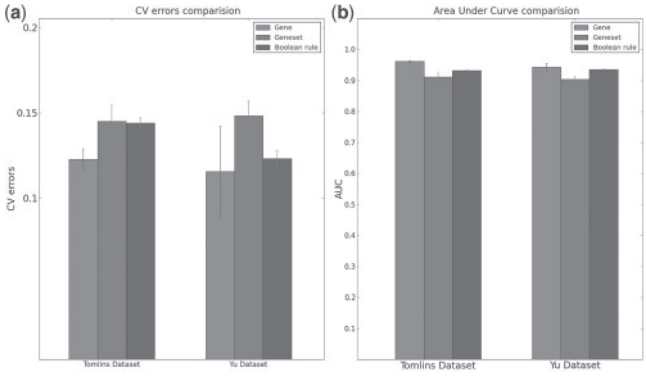

Fig. 4. Comparison of classification performance: a tenfold cross-validation. The figure shows (**a**) cross-validation errors and (**b**) the area under t he curve of random forest classifiers, which are based on Boolean rules for genes, gene sets, and the sum of log-rank. The most significant 50 features are selected from the training datasets of each classifier.

(Litvinov *et al*., 2006); in addition, the DNA replication regulation protein MCM7 has been suggested for use as a proliferation marker in prostate cancer (Padmanabhan *et al*., 2004). On the other hand, the G1 pathway is an important checkpoint of cell proliferation. Thus, the tumor cells undergo uncontrolled proliferation by destroying the G1 pathway while activating the DNA replication pathway.

- R_{18}: *complement and coagulation* (*down*)*, calmodulin* (*down*) *and TERT pathway* (*up*) The genes SERPINA5 and SERPING1, which are involved in the complement and coagulation pathway, are important for preventing cancer cell growth and metastasis in breast and prostate cancer (Sieben *et al*., 2005). Calmodulin is known as a negative modulator of ARs, which play an important role in AR breakage (Cifuentes *et al*., 2004). Therefore, the loss of expression of those genes promotes the development of prostate cancer. In the TERT pathway, the genes HDAC1, MYC, MZF1 and SP3 show higher expression levels in prostate cancer; they contribute to the survival and proliferation of cancer cells (Lee *et al*., 2008b).

3.3 Classification analysis

We used the discovered Boolean rules of the gene sets to perform classification analysis. We obtained feature vectors in individual samples for the Boolean rules with the log-rank sum of literals in the rule described in the previous section, and we constructed random forest classifiers (Breiman, 2001) for the test datasets by using the features selected from the training datasets. Using the cross-validation errors and the area under the curve (as shown in Fig. 4), we compared the overall classification performances with the classification based on the genes and gene sets. For independent datasets, our results are comparable the classification based on genes and gene sets.

4 CONCLUSION

We have presented a method of inferring Boolean rules of gene sets from cancer microarray datasets. We first identified small subsets of gene sets that are synergistic; we then enumerated all the possible conjunctive Boolean functions of the gene sets within each group of gene sets; and, finally, we evaluated the predictive powers of the Boolean functions. Although several approaches have been proposed to reveal the Boolean logic of individual molecules (Ruczinski *et al*., 2003; Mukherjee *et al*., 2009; Varadan and Anastassiou, 2006), our method is different with those methods in that we try to identify relevant variables by using a synergy network before we search for the Boolean rules. We have identified and discussed several significant Boolean rules of gene sets that are frequently observed in prostate cancer. Our results show the possibility of using Boolean rules of gene sets for cancer classification, and the identified Boolean rules provide insights into the combinatorial influences of deregulated gene sets on cancer developments. The coverage of gene set analysis increases as the information on pathway interactions is accumulated. Thus, further research directions should include the pathway interaction information for the dissection of coherent submodules in a gene set.

Funding: Samsung Biomedical Research Institute (grant C-A7-101-2); Korean National Research Laboratory (grant 2005-01450 to D.L.) of the Ministry of Science and Technology; Information Technology Research Center (grant IITA-2008-C1090-0801-0001 to D.L.).

Conflict of Interest: none declared.

REFERENCES

Al-Shahrour,F. *et al*. (2007) FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Res.*, **35**, W91–W96.

Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA*, **99**, 6562–6566.

Anastassiou,D. (2007) Computational analysis of the synergy among multiple interacting genes. *Mol. Syst. Biol.*, **3**, 83.

Ashburner,M. *et al*. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Batagelj,V. and Mrvar,A. (2002) Pajek - analysis and visualization of large networks. *Graph Drawing*, **2265**, 477–478.

Bild,A.H. *et al*. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**, 353–357.

Bo,T. and Jonassen, I. (2002) New feature subset selection procedures for classification of expression profiles. *Genome Biol.*, **3**, RESEARCH0017.

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Breitling,R. *et al*. (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*, **573**, 83–92.

Chuang,H.Y. *et al*. (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140.

Cifuentes,E. *et al*. (2004) Physical and functional interaction of androgen receptor with calmodulin in prostate cancer cells. *Proc. Natl Acad. Sci. USA*, **101**, 464–469.

Clauset,A. *et al*. (2008) Hierarchical structure and the prediction of missing links in networks. *Nature*, **453**, 98–101.

Dennis,G. *et al*. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, P3.

Dinu,I. *et al*. (2009) Gene-set analysis and reduction. *Brief. Bioinform.*, **10**, 24–34.

Dong,X.Y. *et al*. (2008) SnoRNA U50 is a candidate tumor-suppressor gene at 6q14.3 with a mutation associated with clinically significant prostate cancer. *Hum. Mol. Genet.*, **17**, 1031–1042.

Edelman,E.J. *et al*. (2006) Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics*, **22**, e108–116.

Edelman,E.J. *et al*. (2008) Modeling cancer progression via pathway dependencies. *PLoS Comput. Biol.*, **4**, e28.

Eruslanov *et al*. (2009) Altered expression of 15-hydroxyprostaglandin dehydrogenase in tumor-infiltrated CD11b myeloid cells: a mechanism for immune evasion in cancer. *J. Immunol*., **182**, 7548–7557.
Golub,T.R. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
Hanczar,B. *et al*. (2007) Feature construction from synergic pairs to improve mi-croarray-based classification. *Bioinformatics*, **23**, 2866–2872.
Heiser,L.M. *et al*. (2009) Integrated analysis of breast cancer cell lines reveals unique signaling pathways. *Genome Biol*., **10**, R31.
Hsing,A.W. *et al*. (2001) Prostate cancer risk and serum levels of insulin and leptin: a population-based study. *J. Natl Cancer. Inst.*, **93**, 783–789.
Huang da,W. *et al*. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
Jiang,Z. and Gentleman,R. (2007) Extensions to gene set enrichment. *Bioinformatics*, **23**, 306–313.
Ruczinski,I. *et al*. (2003) Logic regression. *J. Comput. Graph. Stat.*, **12**, 475–511.
Jakobsson,P.J. *et al*. (1999) Identification of human prostaglandin E synthase: a microsomal, glutathione-dependent, inducible enzyme, constituting a potential novel drug target. *Proc. Natl Acad. Sci. USA*, **96**, 7220–7225.
Kanehisa,M. *et al*. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–280.
Kim,H. *et al*. (2005) Missing value estimation for DNA microarray gene expres-sion data: local least squares imputation. *Bioinformatics*, **21**, 187–198.
Lapointe,J. *et al*. (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl Acad. USA*, **101**, 811–816.
Lee,E. *et al*. (2008) Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.*, **4**, e1000217.
Lee,J. *et al*. (2008) TERT promotes cellular and organismal survival independently of telomerase activity. *Oncogene*, **27**, 3754–3760.
Levine,D.M. *et al*. (2006) Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways. *Genome Biol.*, **7**, R93.
Lin,H.M. *et al*. (2009) Transforming growth factor-beta/SMAD3 signaling regulates insulin gene transcription and pancreatic islet beta-cell function. *J. Biol. Chem*., **284**, 12246–12257.
Linden,J. (2006) Adenosine metabolism and cancer. Focus on 'Adenosine downre-gulates DPPIV on HT-29 colon cancer cells by stimulating protein tyrosine phosphatases and reducing ERK1/2 activity via a novel pathway'. *Am. J. Physiol. Cell. Physiol.*, **291**, C405–406.
Litvinov,I.V. *et al*. (2006) Androgen receptor as a licensing factor for DNA replication in androgen-sensitive prostate cancer cells. *Proc. Natl Acad. Sci. USA*, **103**, 15085–15090.
Mootha,V.K. *et al*. (2003) PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.
Moskaluk,C.A. *et al*. (1997) Immunohistochemical expression of pi-class gluta-thione S-transferase is down-regulated in adenocarcinoma of the prostate. *Cancer*, **79**, 1595–1599.
Mukherjee,S. *et al*. (2009) Sparse combinatorial inference with an application in cancer biology. *Bioinformatics*, **25**, 265–271.
Nam,D. and Kim,S.Y. (2008) Gene-set approach for expression pattern analysis. *Brief. Bioinform.*, **9**, 189–197.
Nevins,J.R. and Potti,A. (2007) Mining gene expression profiles: expression signa-tures as cancer phenotypes. *Nat. Rev. Genet.*, **8**, 601–609.
Obajimi,O. *et al*. (2009) Inhibition of de novo purine synthesis in human prostate cells results in ATP depletion, AMPK activation and induces senescence. *Prostate*, **69**, 1206–1221.
Padmanabhan,V. *et al*. (2004) DNA replication regulation protein Mcm7 as a marker of proliferation in prostate cancer. *J. Clin. Pathol.*, **57**, 1057–1062.
Pang,H. *et al*. (2006) Pathway analysis using random forests classification and regression. *Bioinformatics*, **22**, 2028–2036.
Pang,H. and Zhao,H. (2008) Building pathway clusters from random forests classification using class votes. *BMC Bioinformatics*, **9**, 87.
Paninski,L. (2003) Estimation of entropy and mutual information. *Neural Comput.*, **15**, 1191–1253.
Pierce,B.L. *et al*. (2008) Diabetes mellitus and prostate cancer risk. *Prostate*, **68**, 1126–1132.
Pilpel,Y. *et al*. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet*., **29**, 153–159.
Potti,A. *et al*. (2006) Genomic signatures to guide the use of chemotherapeutics. *Nat. Med*., **12**, 1294–1300.
Segal,E. *et al*. (2004) A module map showing conditional activity of expression modules in cancer. *Nat. Genet*., **36**, 1090–1098.
Segal,E. *et al*. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet*., **34**, 166–176.
Sieben,N.L. *et al*. (2005) Differential gene expression in ovarian tumors reveals Dusp 4 and Serpina 5 as key regulators for benign behavior of serous borderline tumors. *J. Clin. Oncol*., **23**, 7257–7264.
Singh,D. *et al*. (2002) Gene expression correlates of clinical prostate cancer beha-vior. *Cancer Cell*, **1**, 203–209.
Sorlie,T. *et al*. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. USA*, **100**, 8418–8423.
Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
Tian,L. *et al*. (2005) Discovering statistically significant pathways in expression profiling studies. *Proc. Natl Acad. Sci. USA*, **102**, 13544–13549.
Tomlins,S.A. *et al*. (2006) Integrative molecular concept modeling of prostate cancer progression. *Nat. Genet.*, **39**, 41–51.
van der Poel,H.G. (2005) Androgen receptor and TGFbeta1/Smad signaling are mutually inhibitory in prostate cancer. *Eur. Urol.*, **48**, 1051–1058.
van 't Veer,L.J. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
Varadan,V. and Anastassiou,D. (2006) Inference of disease-related molecular logic from systems-based microarray analysis. *PLoS Comput. Biol.*, **2**, e68.
Vastrik,I. *et al*. (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.
Watkinson,J. *et al*. (2008) Identification of gene interactions associated with disease from gene expression data using synergy networks. *BMC Syst. Biol.*, **2**, 10.
Yu,Y.P. *et al*. (2004) Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. *J. Clin. Oncol.*, **22**, 2790–2799.