BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 13 2011, pages 1846–1853
doi:10.1093/bioinformatics/btr283

Systems biology

Advance Access publication May 6, 2011

Integrative gene network construction for predicting a set of complementary prostate cancer genes

Jaegyoon Ahn[1], Youngmi Yoon[2], Chihyun Park[1], Eunji Shin[1] and Sanghyun Park[1,*]

[1]Department of Computer Science, Yonsei University, Seoul and [2]Division of Information Engineering, Gachon University of Medicine and Science, Inchon, South Korea

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Diagnosis and prognosis of cancer and understanding oncogenesis within the context of biological pathways is one of the most important research areas in bioinformatics. Recently, there have been several attempts to integrate interactome and transcriptome data to identify subnetworks that provide limited interpretations of known and candidate cancer genes, as well as increase classification accuracy. However, these studies provide little information about the detailed roles of identified cancer genes.

Results: To provide more information to the network, we constructed the network by incorporating genetic interactions and manually curated gene regulations to the protein interaction network. To make our newly constructed network cancer specific, we identified edges where two genes show different expression patterns between cancer and normal phenotypes. We showed that the integration of various datasets increased classification accuracy, which suggests that our network is more complete than a network based solely on protein interactions. We also showed that our network contains significantly more known cancer-related genes than other feature selection algorithms. Through observations of some examples of cancer-specific subnetworks, we were able to predict more detailed and interpretable roles of oncogenes and other cancer candidate genes in the prostate cancer cells.

Availability: http://embio.yonsei.ac.kr/~Ahn/tc.php.

Contact: sanghyun@cs.yonsei.ac.kr

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

Received on December 20, 2010; revised on April 27, 2011; accepted on April 28, 2011

1 INTRODUCTION

Diagnosis and prognosis of cancer is one of the most important research areas in bioinformatics. Enormous efforts have been made to identify biomarkers of cancer, and have predominantly focused on the analysis of transcriptome data. Recently, diagnostic and prognostic predictive performances have been improved by incorporating interactome data (Chuang *et al*., 2007; Taylor *et al*., 2009). Incorporating interactome data has the additional benefit that it can suggest detailed roles of potential cancer-related genes.

Chuang *et al*. showed that with metastasis of breast cancer, integration of interactome and transcriptome data can be useful to extract coexpressed functional subnetworks, as well as to obtain higher classification accuracy. These subnetworks contain many known breast cancer genes that could not be detected in previous studies which analyzed only transcriptome data.

Taylor *et al*. suggest that the organization of interactome data is changed by altered gene expression in breast cancer, which affects disease outcome. To accommodate for this, they searched for changes in global modularity in protein interaction networks. They calculated the average Pearson Correlation Coefficient (PCC) of a hub protein and its interacting partners, and revealed many interactions that displayed altered PCCs as a function of disease outcome. Their analysis is based on the concept of date and party hub (Han *et al*., 2004). A date (intermodular) hub shows low correlation of coexpression with its interacting partner and acts as a global connector in the protein interaction network, while a party (intramodular) hub shows high correlation of coexpression and acts as a local connector. They found that intermodular hub proteins tend to have more partners which form altered interactions than intramodular hub proteins.

Analyses of coexpressed subnetworks or hub proteins have been helpful for the understanding of the metastasis of cancer at the molecular level. However, providing interpretability to a gene or protein network would also be valuable. For example, if the network could provide information on what makes networks change and what the effects of the changes are, then candidates to be validated or methods to validate would be more specific, and more effective targeting for therapies or drugs would be possible. To achieve this goal, we integrated genetic interactions and gene regulatory pathways, based on previous studies using protein–protein interactions and gene expression profiles.

Functional dependencies revealed by genetic interactions are known to provide abundant information regarding biological pathways (Battle *et al*., 2010; Beltrao *et al*., 2010). Recently, Lin *et al*. constructed genome-wide maps of human genetic interactions using radiation hybrid genotyping data (Lin *et al*., 2010). They suggested that their genetic interaction network approached saturation, suggesting the network did not show a scale-free distribution of connectivity, but was Gaussian-like. Thus, we can expect that the addition of genetic interactions would help make the network more global, provide functional dependencies between genes at a genome-wide level and give more accurate and abundant explanation to each individual pathway.

The other dataset we used for our network is gene regulation information. Gene regulatory pathways give more detailed explanations of cancer genes and their oncogenesis. Combinatorial

*To whom correspondence should be addressed.

interactions among transcription factors are critical to directing tissue-specific gene expression (Ravasi *et al*., 2010). As well as the transcription of the gene, we expect that the integration of gene regulations, physical interactions and genetic interactions can explain the combinatorial alterations of complexes which function as activators or inhibitors of translocation, and protein complex modification processes that are changed in the tumor cells.

An integrated network using various kinds of data including gene expression profiles, protein interactions, genetic interactions and gene regulatory pathways could provide more accurate diagnosis than the networks constructed using a single dataset. This explains the advantage of integration of multiple types of interactions. We also confirmed that the classification accuracy outperformed previously studied feature selection and classification algorithms. We validated our network with Fisher's exact test using the cancer-related gene list provided by the Cancer Genome Project, and confirmed that our network was enriched by cancer-related genes better than gene sets from other feature selection algorithms.

We could see that many cancer-related genes are involved in the cancer-specific gene network, and that many of those were hubs of protein–protein interactions or genetic interactions. Many cancer-related genes could be detected using only one type of interaction set, which also suggests that the network is more complete with the integration of multiple interaction types. We also found evidences that combinations of interactions including protein–protein interactions (PPIs), genetic interactions (GIs) and inferred protein–protein interactions (IPPIs) might influence the modification of a complex and the translocation of a protein, as well as the transcription of genes. It was observed that most cancer-related genes or cancer candidate genes play a role as a member of complex, which influences the transcription, modification or translocation processes, and also as an entity that is influenced by these processes.

2 METHODS

We integrated the PPI dataset, GI dataset and gene regulatory networks to construct the initial network. We then identified the subnetworks of which interactions showed different behavior between tumor and normal samples. We assumed that more accurate subnetworks result in better classification accuracy, when those subnetworks are used as a classifier. Therefore, we obtained the optimal parameter through several cross-validations, and used it for constructing the cancer-specific network.

2.1 Data description

To identify cancer-specific interactions, we analyzed the DNA microarray measurements of the expression of human mRNAs. This microarray dataset is composed of 12 600 measurements of 52 prostate cancer samples and 50 normal samples (Singh *et al*., 2002). We used two other microarray datasets for independent tests (LaTulippe *et al*., 2002; Welsh *et al*., 2001). Both datasets used 12 600 probes which is the same as Singh *et al*. The dataset of Welsh *et al.* is composed of 24 prostate cancer samples and 9 normal samples, and dataset of LaTulippe *et al*. is composed of 23 prostate cancer samples and 3 normal samples. We converted 12 600 probes into 8828 gene symbols by averaging the expression values of the probes that are mapped into the same gene. Then we normalized each expression profile by z-scoring transformation.

We used three different types of interactions. First, we downloaded 194 988 human PPIs from the I2D database on October 2010 (Brown and Jurisica, 2007), which includes known, experimental and predicted PPIs for human, as well as five other organisms. The proteins in those PPIs were mapped into gene symbols using UniPROT. After removing duplicated PPIs and PPIs that contain proteins that are not mapped into a gene symbol, we obtained 108 544 PPIs.

Table 1. Summary of collected data

Name	Description	Quantity	Reference
PPI	Protein–protein interaction	108 544	I2D database
GI	Genetic interaction	337 235	Lin *et al*., 2010
GR	Directed interaction which activates regulation processes	14 015 (modification) 1822 (transcription) 320 (translocation)	
	Directed interaction which inhibits regulation processes	618 (modification) 493 (transcription) 24 (translocation)	Pathway Interaction Database
IPPI	Inferred protein–protein interaction from protein complexes	17 305	
Reference Gene Set	Cancer-related genes	16 (prostate cancer) 411 (all types of cancer)	Cancer Genome Project

Second, we downloaded 7 248 479 GIs inferred from radiation hybrid genotypes (Lin *et al*., 2010). Among those, we used only 377 235 GIs of which gene symbols can be mapped into the gene symbol list in the Entrez Gene database.

Third, we downloaded human pathways from the Pathway Interaction Database (Schaefer *et al*., 2009) on October 2010. These include 114 pathways curated by NCI-Nature, and 332 pathways imported from BioCarta and Reactome database. Each pathway is mainly composed of protein complex modification, transcriptions of genes or translocation of a protein or protein complex. Modification, transcription or translocation processes can be activated or inhibited by proteins or protein complexes. We refer these interactions as GRs (Gene Regulations).

We assume that all the proteins in the protein complex take part in the activation or inhibition. For example, if a protein complex with four proteins activates transcription of a gene, then we obtain four binary activations of transcription. If the protein complex with three proteins inhibits the modification of the protein complex with four proteins, we can get 12 binary inhibitions of modification. Additionally, we inferred the interactions among proteins which constitute known protein complexes in the human pathways of Pathway Interaction Database. We refer these inferred interactions as IPPIs. We assumed that the protein complex would form a complete graph. For example, we obtain six IPPIs from the protein complex with four proteins. We extracted 14 015 activations and 681 inhibitions of modification, 1822 activations and 493 inhibitions of transcription, 320 activations and 24 inhibitions of translocation and 17 305 inferred interactions.

We also downloaded the cancer-related gene list from the Cancer Genome Project as a reference gene set on March 2010. This included 411 cancer genes and 16 prostate cancer genes. All data described above are summarized in Table 1.

2.2 Construction of cancer-specific gene network

PPIs, GIs, GRs and IPPs are commonly binary interactions of genes. Among these interactions, GRs are directional and others are non-directional.

Identifying interactions that should be included in a cancer-specific network from each type of interaction dataset is the key obstacle in constructing the network. Therefore, we propose a novel scoring measure for this purpose. The optimal threshold for the measure maximizes the classification accuracy of the network. Accordingly, we also propose a new prediction method which makes use of the network as a classifier.

A network with such interactions can form a wide map of gene interactions that comprises 7261 genes. To identify a cancer-specific gene network using this network, we took a similar approach as Taylor *et al.*, which exploited the difference of strengths of the interactions.

The strengths of some interactions can be different between normal and tumor cells. The changes of interaction levels from normal to tumor state can be causes or effects of tumorigenesis. Suppose that a protein complex is modified during tumorigenesis, and this protein complex is responsible for changing the strength of regulation of some proteins. Then we can say that changes in the interaction and regulation are the causes of tumorigenesis. The modified protein complex and altered regulations can affect various interactions and regulations, and these changes can be regarded as a result of tumorigenesis.

The changes in interactions can be represented as changes in degree of dependencies between two interactors, in this case, genes. As the dependency of two genes increases, the correlation of their mRNA expressions would also increase. To measure the dependency between two genes, we calculate the PCC between them. A large difference of PCC values means that there were significant changes of correlation of two genes between two groups of samples.

Based on the rationale mentioned previously, each interaction of the whole network is tested to determine if the PCC of mRNA expression values of two genes is different between normal samples and tumor samples. For a given interaction of which two interactors are a and b, we can say that they show significantly different correlations if the interaction satisfies the following equation:

$$\text{Score of an interaction} = |\ \text{PCC}(v_{at}, v_{bt}) - \text{PCC}(v_{an}, v_{bn})| > \text{threshold},$$

where v_{at} and v_{an} are vectors of mRNA expression values of gene a on tumor and normal samples, respectively, and v_{bt} and v_{bn} are vectors of mRNA expression values of gene b on tumor and normal samples, respectively.

An interaction that satisfies the equation is included in the cancer-specific network. Parameter threshold can be interpreted as minimal significance of difference between two groups of samples. As the threshold increases, the significance level becomes more stringent. Therefore, a smaller number of more definite interactions and regulations are selected. In other words, false positive and false negative rates of the cancer-specific network would decrease and increase, respectively, as the threshold increases. We should find the optimal threshold that yields the best result in the trade-offs between false negative and false positive rates. To identify the optimal threshold value, we performed leave one out cross-validation (LOOCV) while lowering the threshold by 0.01.

When performing LOOCV, the cancer-specific gene network functions as a classifier. Each edge of the cancer-specific network satisfies the *score of an interaction* equation. For a given sample s of which the class label is unknown, we can predict its class label using the edges of this network by the following procedure.

(i) Assign two scores, $score_t$ and $score_n$ to unknown sample s and initialize to zero.

(ii) For each edge $e = (a, b)$ in the cancer-specific gene network:

(a) Calculate $\text{PCC}(v'_{at}, v'_{bt})$ and $\text{PCC}(v'_{an}, v'_{bn})$, where $v'_{at} = v_{at} + x_a$, $v'_{bt} = v_{bt} + x_b$, $v'_{an} = v_{an} + x_a$, $v'_{bn} = v_{bn} + x_b$, and x_a and x_b be s's two mRNA expression values of a and b, respectively.

(b) Calculate n and t where,
$t = |\ \text{PCC}(v'_{at}, v'_{bt}) - \text{PCC}(v_{at}, v_{bt})|$
$n = |\ \text{PCC}(v'_{an}, v'_{bn}) - \text{PCC}(v_{an}, v_{bn})|$.

(c) If $t \geq n$, $\text{score}_t = \text{score}_t + 1$, else $\text{score}_n = \text{score}_n + 1$.

(iii) If $\text{score}_t \geq \text{score}_n$, s is labeled as tumor and otherwise, s is labeled as normal.

In the procedure above, t and n represent the changed values of PCC when the given sample s is added to the tumor and normal sample set, respectively. If s is normal sample, n is likely to be bigger than t. Therefore, score_n would be also increased if proper interactions are selected as a classifier, and s is labeled as normal if $\text{score}_t < \text{score}_n$.

3 RESULTS

First, we performed experiments to obtain the optimal threshold for each type of interactions. Then we constructed a tumor-specific gene network using optimal thresholds, and performed independent and comparison tests. Lastly, we analyzed our network with the known cancer-related gene list.

3.1 Obtaining optimal threshold

As stated above, we assumed that the more accurate the cancer-specific network is, the higher classification accuracy it has. Thus, for a more accurate cancer-specific network, we needed to obtain an optimal *threshold*. We measured the accuracy, sensitivity and specificity by LOOCV varying a threshold for each set of PPI, GI, GR and IPPI networks (Fig. 1).

In Figure 1, we can see that sensitivity generally decreases while specificity generally increases as the threshold increases. In other words, a low threshold can result in high false positive rates, and a high threshold can result in high false negative rates. We selected the threshold value that results in the best accuracy. When numbers of test and control cases are different, accuracy is not a good measure for classification quality. However, in our experiment, the numbers of tumor (=52) and normal samples (=50) are nearly identical, thus there is no reason not to use accuracy measure.

Once we obtained the optimal threshold for each type of interaction, we performed LOOCV again with integrated interactions which satisfy the threshold of each type. For example, PPIs of which score > threshold of PPI, or GIs of which score > threshold of GI can be used for the LOOCV test. Table 2 summarizes the results of these tests. Note that the number of interactions is an averaged value because training sample sets are changed for each run in LOOCV. As we can see in Table 2, we obtained better LOOCV results when using integrated interactions than when using individual set of interactions. The only test that shows similar classification results is IPPIs, which are a minority among all interaction types. Comparison with Taylor *et al.* and Chuang *et al.* is also exhibited in Table 2. A cancer network constructed from Chuang *et al.* is not a classifier. They use it as a feature selection. We applied SMO (Platt, 1999), Naïve Bayesian (John and Langley, 1995), k-NN (k-Nearest Neighbor, Aha and Kibler, 1991) and Random Forest (Breiman, 2001) for a classification method. SMO is a sequential minimal optimization algorithm for training a Support Vector Machine (SVM). Chuang *et al.* show a large variance depending on the classification algorithm.

For a more thorough test, we performed independent tests using two datasets from Welsh *et al.* and LaTulippe *et al.* Because these datasets were normalized by z-scoring transformation, they can be integrated into one dataset composed of 47 tumor samples and 12 normal samples. First, we performed four independent tests

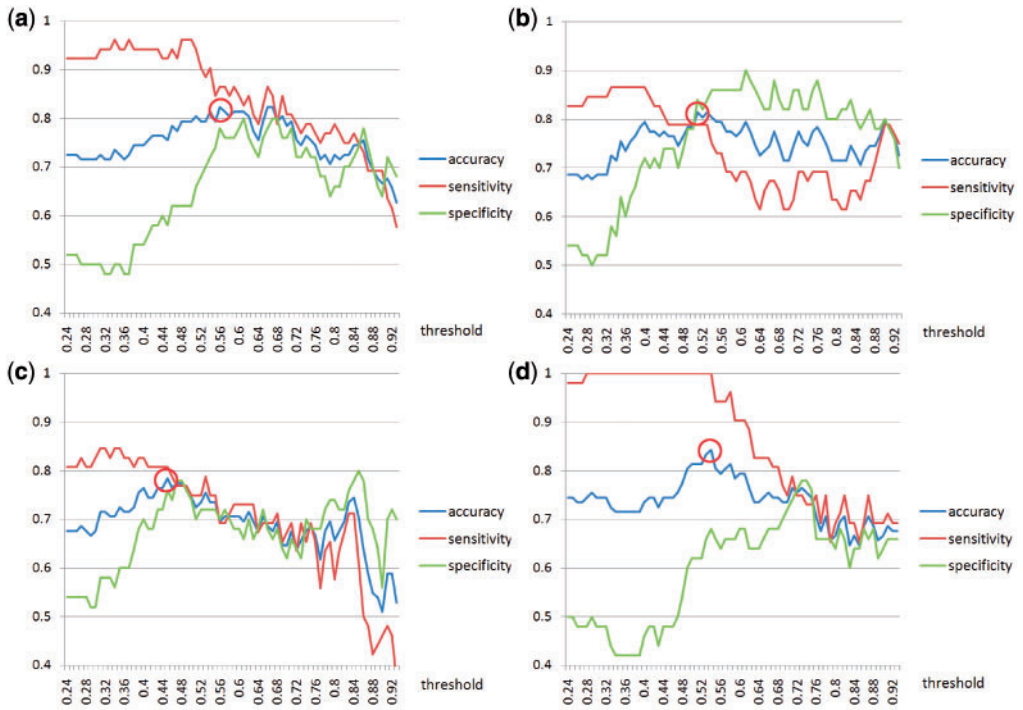

Fig. 1. Determination of the optimal *threshold*. Accuracy (blue line), sensitivity (red line) and specificity (green line) measured for (**a**) PPI network, (**b**) GI network, (**c**) GR network and (**d**) IPPI network. Points that yield the best accuracy (optimal *threshold*) are marked with red circles.

Table 2. Comparison of LOOCV classification accuracy for ours, Taylor *et al.* and *Chaung et al.*

Algorithm		Interaction type	Optimal *threshold* value	Number of interactions	Number of genes	Accuracy (%)	Sensitivity (%)	Specificity (%)
Ours		PPI	0.56	1869.29	2548.53	82.35 (84/102)	86.54 (45/52)	78.00 (39/50)
		GI	0.51	1788.37	1027.59	81.37 (83/102)	78.85 (41/52)	84.00 (42/50)
		GR	0.45	793.45	935.29	78.43 (80/102)	80.77 (42/52)	76.00 (38/50)
		IPPI	0.54	527.42	657.38	84.31 (86/102)	100.00 (52/52)	68.00 (34/50)
		Integrated	–	5033.45	2798.84	84.31 (86/102)	90.38 (47/52)	78.00 (39/50)
Taylor *et al.*		PPI	–	34106.10	5192.30	68.63 (70/102)	75.00 (39/52)	62.00 (31/50)
Chuang *et al.*	Naïve Bayesian	PPI	–	4944.31	594.12	87.25 (89/102)	86.54 (45/52)	88.00 (44/50)
	SMO					87.25 (89/102)	90.38 (47/52)	84.00 (42/50)
	k-NN					81.37 (83/102)	88.46 (46/52)	74.00 (37/50)
	Random Forest					82.35 (84/102)	82.69 (43/52)	82.00 (41/50)

with each of the four sets of cancer-specific interactions. Then, we performed an independent test again using integrated cancer-specific interactions. The results, shown in Table 3, suggest that using integrated interactions improves the classification performance, which is the same result shown in Table 2. These observations are indirect evidence that each interaction type is complementary to each other and they altogether result in a richer cancer-specific gene network while minimizing the false positives. Comparison with Taylor *et al.* and Chuang *et al.* for independent test is also exhibited in Table 3. Our method outperforms these two methods.

Next, we performed tests to compare the classification performance of our network with previously published classification algorithms. When we use the cancer-specific gene network as a classifier, it can be considered as preprocessed data by feature selection. Therefore, we applied feature selection algorithms including relief-F (Kira and Rendell, 1992), correlation-based filter (Yu and Liu, 2003), information gain (Mitchell, 1997), gain ratio (Pano, 1961) and chi-squared (Liu and Setiono, 1995) to the training dataset and then performed an independent test using classification algorithms including Naïve Bayesian, SMO, k-NN and Random Forest. Those feature selection and classification algorithms were implemented in Weka v3.5 (Witten and Frank, 2005), a publicly available open-source software package. We selected the top ranked 2772 genes, or the same number of genes as in our classifier,

Table 3. Comparison of classification accuracy with independent data for ours, Taylor *et al.* and *Chaung et al.*

Algorithm		Interaction type	Optimal *threshold* value	Number of interactions	Number of genes	Accuracy (%)	Sensitivity (%)	Specificity (%)
Ours		PPI	0.56	1835	1671	93.22 (55/59)	91.49 (43/47)	100.00 (12/12)
		GI	0.51	1776	854	88.14 (52/59)	87.23 (41/47)	91.67 (11/12)
		GR	0.45	821	690	88.14 (52/59)	93.62 (44/47)	66.67 (8/12)
		IPPI	0.54	524	460	89.83 (53/59)	89.36 (42/47)	91.67 (11/12)
		Integrated	–	4956	2772	96.61 (57/59)	97.87 (46/47)	91.67 (11/12)
Taylor *et al.*		PPI	–	35333	5316	91.53 (54/59)	95.74 (45/47)	75.00 (9/12)
Chuang *et al.*	Naïve Bayesian	PPI	–	3967	606	94.92 (56/59)	93.62 (44/47)	100.00 (12/12)
	SMO					94.92 (56/59)	95.74 (45/47)	100.00 (11/12)
	k-NN					93.22 (55/59)	91.49 (43/47)	100.00 (12/12)
	Random Forest					94.92 (56/59)	100.00 (47/47)	75.00 (9/12)

Table 4. Comparison of classification accuracy with independent data for feature selection algorithms

	No feature selection	Chi-squared	Gain ratio	Information gain	Relief-F	Correlation-based filter
Naïve Bayesian	96.61 (57/59)[a]	96.61 (57/59)	96.61 (57/59)	96.61 (57/59)	98.31 (58/59)	96.61 (57/59)
	97.87 (46/47)	97.87 (46/47)	97.87 (46/47)	97.87 (46/47)	97.87 (46/47)	97.87 (46/47)
	91.67 (11/12)	91.67 (11/12)	91.67 (11/12)	91.67 (11/12)	100.00 (12/12)	91.67 (11/12)
SMO	96.61 (57/59)	94.92 (56/59)	94.92 (56/59)	96.61 (57/59)	96.61 (57/59)	96.61 (57/59)
	95.57 (45/47)	93.62 (44/47)	93.62 (44/47)	95.57 (45/47)	95.57 (45/47)	95.57 (45/47)
	100.00 (12/12)	100.00 (12/12)	100.00 (12/12)	100.00 (12/12)	100.00 (12/12)	100.00 (12/12)
k-NN	76.27 (45/59)	88.14 (52/59)	88.14 (52/59)	88.14 (52/59)	89.83 (53/59)	88.14 (52/59)
	70.21 (33/47)	85.11 (40/47)	85.11 (40/47)	85.11 (40/47)	89.36 (42/47)	85.11 (40/47)
	100.00 (12/12)	100.00 (12/12)	100.00 (12/12)	100.00 (12/12)	91.67 (11/12)	100.00 (12/12)
Random Forest	83.05 (49/59)	84.75 (50/59)	91.53 (54/59)	91.53 (54/59)	93.22 (55/59)	94.92 (56/59)
	100.00 (47/47)	100.00 (47/47)	100.00 (47/47)	95.74 (45/47)	97.87 (46/47)	97.87 (46/47)
	16.67 (2/12)	25.00 (3/12)	58.33 (7/12)	75.00 (9/12)	75.00 (9/12)	83.33 (10/12)

[a]Values are given in % (n/N). First, second and third rows of each cell represent accuracy, sensitivity and specificity, respectively.

after each gene is ranked by each feature selection algorithm. The results are shown in Table 4. Naïve Bayesian and SMO show similar classification accuracy with our method. However, these two methods do not seem to gain from feature selection algorithms. Random Forest and k-NN definitely gain from feature selection algorithms, but they generally show a lower accuracy than our solution.

3.2 Analysis of cancer-specific gene network

Table 5 shows the summary information for the cancer-specific gene network, which is used for analysis throughout the rest of the article. To determine if genes in the pathway were previously published in the cancer-related studies, we searched GeneRIF database by keyword. The keywords used were as following: tumo(u)r, cancer, onco(gene), carcino(genesis), neopla(sm), adenocarcino(ma), leukem(ogenesis), astrocytoma, glio(ma), meningioma, thymoma, lymphoma, myeloma, ameloblastoma and hamartoma, in a case-insensitive manner.

We found 118 among 283 cancer-related genes in the protein interaction network, and found 156 among 286 cancer-related genes in the integrated network. This again shows that each type of interactions is complementary, as we have seen in Tables 2 and 3.

Table 5. Summary of the cancer-specific gene network

Number of genes	2772 total genes (Supplementary Table S1)
	1653 genes without cancerrelated references in GeneRIF
	963 genes with cancerrelated references in GeneRIF
	156 cancer-related genes
	3 prostate cancer-related genes
Number of interactions	4956 total interactions
	1835 PPIs
	1776 GIs
	821GRs
	524 IPPIs

We can see that the *P*-value of the integrated network is very low. In contrast, GI and IPPI networks show very high *P*-values. This is likely due to the fact that the cancer genes in those networks are relatively unknown cancer-related genes. Because we have confirmed that other networks also have discriminative power (Tables 2 and 3), it is less likely that a protein interaction

Table 6. Significance level of selection power of our cancer-specific gene network

Algorithm	Interaction type	Number of genes in whole network[a]	Number of cancer-related genes in whole network	Number of genes in cancer-specific network[b]	Number of cancer-related genes in cancer-specific network	*P*-value
Ours	PPI	6703	283	1671	118	2.811e-06
	GI	1246	21	854	11	0.1139
	GR	1969	157	690	73	0.01016
	IPPI	1751	138	460	40	0.06499
	Integrated	7126	286	2772	156	0.0002012
Chi-squared	–				125	0.02500
Gain ratio	–				123	0.02913
Information gain	–				119	0.03704
Relief-F	–				116	0.04180
Correlation-based filter	–				123	0.02913

[a]Number of total genes in a network made with whole interactions of each type.
[b]Number of total genes in a network made with only cancer-specific interactions of each type.

Table 7. Over and under-expressed genes using two sample *t*-test ($\alpha = 0.01$)

Overexpressed genes	Underexpressed genes
FAM107A, NELL2, CALM1, ANXA2, DPYSL2, PTGDS, RCAN2, CDC42BPA, MAL, JAK1, KANK1, DIP2C, CETN2, CCND2	XBP1, RPS2, P4HB, RPLP0, CLDN3, TSPAN1, NME2, FASN, RPL13, EEF2, RPL13A, GUCY1A3, RPL12, RPS10, WWC1, EEF1G, RPL29, RPS18, RPSA, RPL14, TNFSF10, RPS8, RPS17, RPS28, RPL8, RPS4Y1

network only is more significant in describing the cancer-specific gene network, and we can expect that genes that were selected from other networks are also significant.

Table 6 also shows the *P*-values using other feature selection algorithms. For each selection algorithm, we selected 2772 top ranked genes and counted the known cancer-related genes among 2772 genes. We can confirm that the selective power of our interaction based method is definitely higher than other gene-based feature selection algorithms.

Among 2772 genes (156 known cancer-related genes) in our cancer-specific network, we selected 517 genes of which degree is $\geqslant 5$ as hub genes. Among 517 hub genes, there were 39 cancer-related genes (7 prostate cancer-related genes), thus 25% (=39/156) of cancer-related genes were hub genes. There were 2616 genes which are not known as cancer-related genes based on Caner Genome Project, and 478 genes among them are hubs, thus 18.27% (=478/2616) of them were hub genes. Thus, cancer-related genes had higher rate to be a hub gene than non-cancer-related genes.

Table 7 lists 14 overexpressed genes and 26 underexpressed genes in tumor samples from two sample *t*-test. We used $\alpha = 0.01$ as the genome-wide significance level, and applied the Bonferroni adjustment to deal with multiple comparisons. The detailed information for each informative gene is available in the Supplementary Table S4. We used MeV software (Saeed *et al.*, 2006) to detect over- and underexpressed genes.

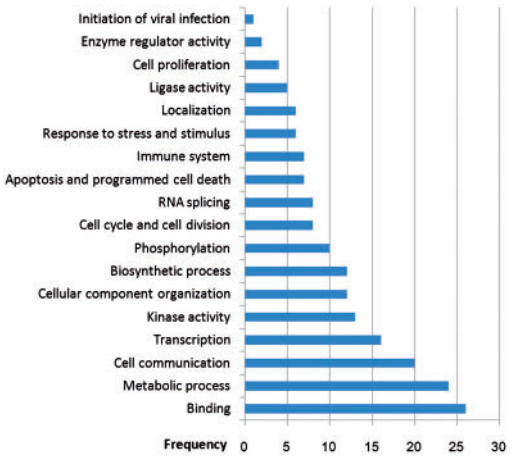

Fig. 2. Functional analysis on prostate cancer gene subnetworks. The length of the bar denotes the frequency of clusters indicating major biological processes and molecular functions. More frequent functions comprise larger part of our cancer-specific network.

4 DISCUSSIONS

To analyze the cancer-specific gene network, we firstly clustered it into modules using network clustering algorithm (Ahn *et al.*, 2010), which hierarchically detect the modules in the network while allowing overlapping nodes (genes) between modules. Then we enriched these modules with Gene Ontology (GO) database using FuncAssociate (Berriz *et al.*, 2003). We have found 312 clusters which has more than 3 genes. Among 312 clusters, 86 clusters were enriched with $P < 0.01$. Major biological processes and molecular functions of those 86 clusters are summarized in Figure 2. Figure 2 includes well-known processes which have been implicated in oncogenesis, and seems to cover global map of biological processes and cellular functions. Detailed enrichment results and genes of each cluster are provided in Supplementary Table S2.

To see more details of some cancer-specific subnetworks, we visualized the cluster 76, 128, 167, 200 and 430 of Supplementary Table S2, in Figure 3. Visualization was done using the Cytoscape (Shannon *et al.*, 2003).

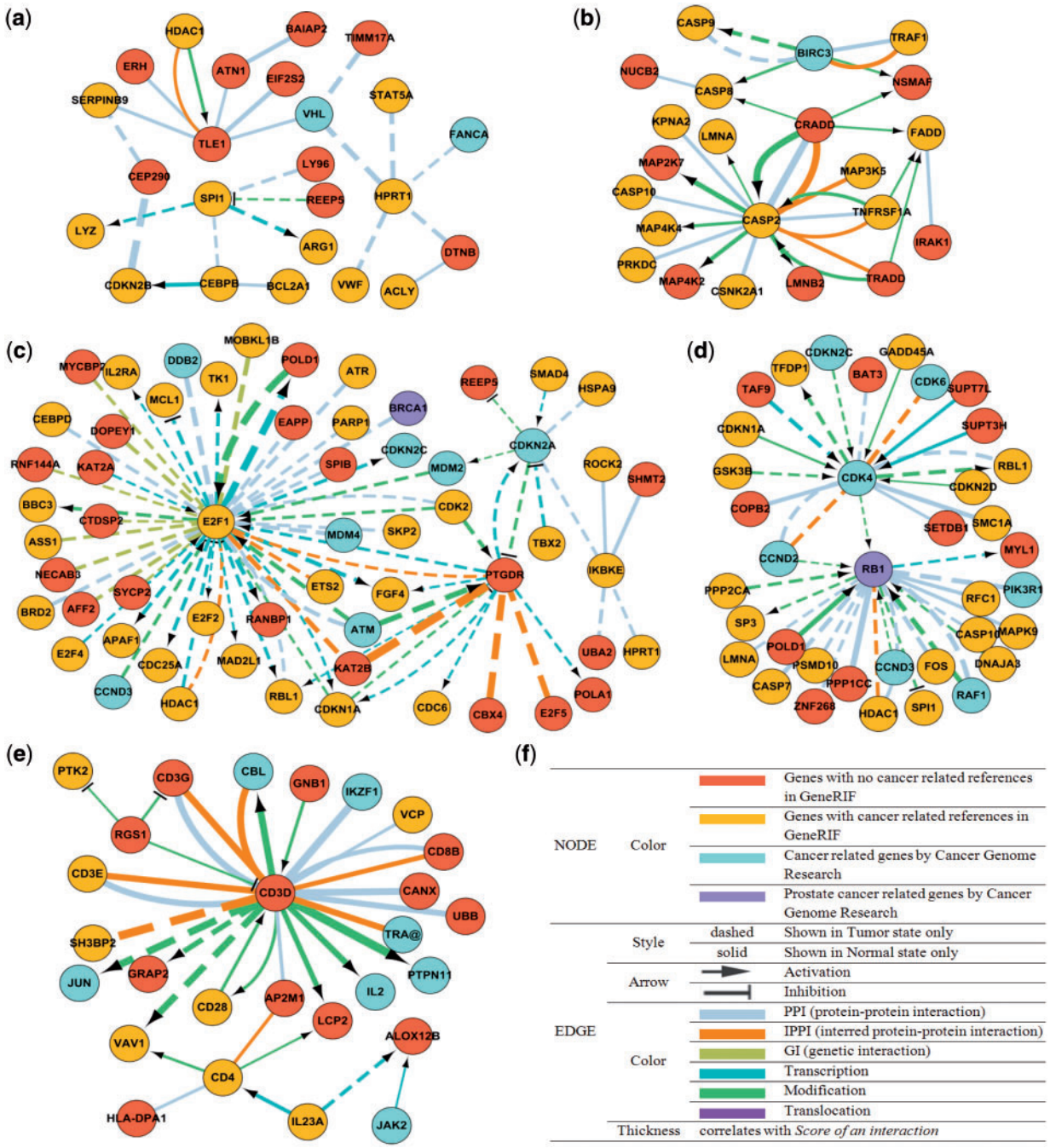

Fig. 3. Visualization of prostate cancer-specific gene subnetworks of Supplementary Table S2. (**a**) Negative regulation of apoptosis (cluster 128); (**b**) positive regulation of apoptosis (cluster 167); (**c**) cell cycle (cluster 430); (**d**) cell cycle (cluster 200); (**e**) immune system process (cluster 76); (**f**) legend for subnetworks.

We can confirm that activation or deactivation of interactions in tumor state explains well-known processes which are related to oncogenesis. Many edges of Figure 3a (negative regulation of apoptosis) are observed in tumor state only, which means apoptosis are negatively regulated in tumor state. On the contrary to this, almost all the edges of Figure 3b (positive regulation of apoptosis) are observed in normal state only. This means apoptosis process does not function normally in tumor state. Similarly, most edges of Figure 3e (immune system process) can be observed in normal state only. This explains there can be defects in immune system in tumor cells. Many edges of subnetworks for cell cycle (Fig. 3c and d) are activated in tumor state only, which shows changes of cell cycle process in network level.

Using these subnetworks, we can predict more detailed roles of each gene. For example, hub gene CD3D in Figure 3e, which is known to be involved in T-cell development, seems to form a

complex with many genes including cancer-related genes (IKZF1, TRA@) and activate modification of complexes in which cancer-related genes (IL2, PTPN11 and CBL) are involved. E2F1 in Figure 3c can be an evidence to support the fact that combinatorial interactions among proteins direct the gene expression level (Ravasi *et al*., 2010). E2F1 has many interactions including PPI, GI and IPPIs, and the combinations of those interactions may influence the transcription of POLD1, TK1, MCL1, etc. It is also interesting that many cancer-related genes are observed to be involved in multiple cancer-specific subnetworks, and many play important roles. For example, RB1 in Figure 3d is a hub of PPIs.

We identified a large number of genes that are not mentioned in any cancer-related publications, but are shown to play a crucial role. Those include TLE1 in Figure 3a, PTGDR in Figure 3c and CD3D in Fig ure 3e. These could be strong candidate genes for further biological investigation. In addition to the genes listed above, there are many more such candidates in our cancer-specific network with their predicted roles.

Throughout the cancer-specific subnetworks, GIs, PPIs, GRs and IPPIs are shown to be slightly overlapped. Moreover, we observed a large number of cancer-related genes which can be detected using only one type of those interactions, and those genes must be complementary to each other. Therefore, we conclude that a network with integrated interaction is better at finding cancer-related genes than a network with only one type of interaction.

Beside the subnetworks discussed above, we provide subnetwork files in SIF format for visualization in Cytoscape at http://embio.yonsei.ac.kr/∼Ahn/tc.php.

5 CONCLUSIONS

In this study, we proposed a novel cancer-specific network construction method and showed that the cancer-specific gene network contains many cancer-related genes. The selective power of our method was better than previously studied feature selection algorithms. We also proposed that the classification method can make use of our network. When constructing the network, integration of PPI, GI, GR and IPPI increased the classification accuracy. This means more complete and complementary network can be achieved by various datasets, rather than using only PPIs.

By analyzing the cancer-specific gene network, we confirmed that cancer-related genes played an important role in the network, and could suggest more detailed and interpretable roles of cancer-related genes and cancer candidate genes in the prostate cancer cells. These roles include activator/inhibitor (or activated/inhibited genes) of transcription, translocation and protein complex modification processes, as well as the membership of protein complexes that are in charge of or result in cancer-specific transcriptional regulation. We expect that more significant and interesting observations are possible.

ACKNOWLEDGEMENTS

The data were obtained from the Wellcome Trust Sanger Institute Cancer Genome Project web site http://www.sanger.ac.uk/genetics/CGP.

Funding: This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0005154).

Conflict of Interest: none declared.

REFERENCES

Aha,D. and Kibler,D. (1991) Instance-based learning algorithms. *Mach. Learn.*, **6**, 37–66.

Ahn,Y. *et al.* (2010) Link communities reveal multiscale complexity in networks. *Nature*, **466**, 761–765.

Battle,A. *et al.* (2010) Automated identification of pathways from quantitative genetic interaction data. *Mol. Syst. Biol.*, **6**, 379.

Beltrao,P. *et al.* (2010) Quantitative genetic interactions reveal biological modularity. *Cell*, **141**, 739–745.

Berriz,G.F. *et al.* (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics*, **19**, 2502–2504.

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

Brown,K.R. and Jurisica,I. (2007) Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.*, **8**, R95.

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

Han,J.D. *et al.* (2004) Evidence for dynamically organized modularity in the yeast proteinprotein interaction network. *Nature,* **430**, 88–93.

John,G.H. and Langley,P. (1995) Estimating continuous distributions in Bayesian classifiers. In *Proceedings of 11th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 338–345.

Kira,K. and Rendell,L.L. (1992) A practical approach to feature selection. In *Proceedings of 9th International Workshop on Machine learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 249–256.

LaTulippe,E. *et al.* (2002) Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease. *Cancer Res.*, **61**, 4499–4506.

Lin,A. *et al.* (2010) A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes. *Genome Res.*, **20**, 1122–1132.

Liu,H. and Setiono,R. (1995) Chi2: feature selection and discretization of numeric attributes. In *Proceedings of IEEE 7th International Conference on Tools with Artificial Intelligence*, IEEE Computer Society, Los Alamitos, CA, USA, pp. 338–391.

Mitchell,T. (1997) *Machine Learning*. McGraw-Hill, New York.

Pano,R. (1961) *Transmission of Information*. MIT Press, Cambridge, MA.

Platt,J.C. (1999) Fast training of support vector machines using sequential minimal optimization. In Schölkopf,B. *et al*. (eds) *Advances in Kernel Methods: Support Vector Learning*. MIT Press, Cambridge, MA.

Ravasi,T. *et al.* (2010) An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, **140**, 744–752.

Saeed,A.I. *et al.* (2006) TM4 microarray software suite. *Methods Enzymol.*, **411**, 134–193.

Schaefer,C.F. *et al.* (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674–D679.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Singh,D. *et al.* (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203–209.

Taylor,I.W. *et al*. (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.*, **27**, 199–204.

The International Cancer Genome Consortium. (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.

Welsh,J.B. *et al.* (2001) Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res.*, **61**, 5974–5978.

Witten, I.H. and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn. Morgan Kaufmann, San Francisco.

Yu,L. and Liu,H. (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. In Proc*eedings of 12th International Conference on Machine Learning*, Springer, Berlin, Heidelberg, pp. 856–863.