Cancer Informatics

ORIGINAL RESEARCH

Protein-protein Interaction Reveals Synergistic Discrimination of Cancer Phenotype

Jianghui Xiong[1,2], Juan Liu[1], Simon Rayner[3], Yinghui Li[2] and Shanguang Chen[2]

[1]School of Computer Science, Wuhan University, Wuhan, P.R. China. [2]State Key Lab of Space Medicine Fundamentals and Application, China Astronaut Research and Training Center, Beijing, P.R. China. [3]Bioinformatics Group, State Key Laboratory of Virology, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, P.R. China.
Email: laserxiong@gmail.com or liujuan@whu.edu.cn.

Abstract: Cancer is a disease associated with the deregulation of multiple gene networks. Microarray data has permitted researchers to identify gene panel markers for diagnosis or prognosis of cancer but these are not sufficient to make specific mechanistic assertions about phenotype switches. We propose a strategy to identify putative mechanisms of cancer phenotypes by protein-protein interactions (PPI). We first extracted the logic status of a PPI via the relative expression of the corresponding gene pair. The joint association of a gene pair on a cancer phenotype was calculated by entropy minimization and assessed using a support vector machine. A typical predictor is "*If Src high-expression, and Cav-1 low-expression, then cancer.*" We achieved 90% accuracy on test data with a majority of predictions associated with the MAPK pathway, focal adhesion, apoptosis and cell cycle. Our results can aid in the development of phenotype discrimination biomarkers and identification of putative therapeutic interference targets for drug development.

Keywords: cancer, biomarker, phenotype discrimination, protein-protein interaction

Cancer Informatics 2010:9 61–66

This article is available from http://www.la-press.com.

The evolution of systems biology out of molecular biology has redefined the concept of a biomarker from a traditional single parametric measure to that of a profile involving multiple genes.[1] This type of approach has identified several gene expression signatures of breast cancer for prognosis prediction,[2–6] although these signatures do not yet provide enough understanding of how these genes cooperatively predict the phenotypes. Recently, analysis of pathway-derived signatures achieved better prediction power in an independent cohort.[7] However, this method still lacks the capability to make specific mechanistic assertions about the phenotype discrimination. Thus, resolving pathway signatures into specific genes, or interaction of genes, can provide additional insight regarding the behavior of a system as a whole, and may assist in the identification of potential targets for future drug development.[8,9]

Here we propose a novel approach to identify synergistic protein-protein interactions associated with a cancer phenotype discrimination. The genome-wide protein interaction data provides unique prior knowledge as a physical basis of cellular signaling pathways. When coupled with gene expression profiling data, it becomes feasible to evaluate the role of protein interactions in a cancer phenotype discrimination. In this pilot study, gene pairs involved in protein-protein interactions were binarized into two states: "high-expression/low-expression" or "on/off". Thus there are four states (00, 01, 10, 11) for each gene pair and the uncertainty that a given state falls into a certain phenotype can be evaluated by a Shannon entropy calculation. The pairs which have minimum entropy for a given cancer phenotype were selected as the pairs most likely associated with that phenotype. The phenotype prediction performance of two-genes pairs were evaluated by a support vector machine (SVM) classifier. The output of the above analysis pipeline is a set of logic statements on phenotype discrimination in the form of: "*If gene A is high-expression and gene B low-expression, then this sample is cancerous (or normal)*".

Materials and Methods

Data set

Adjacent normal-tumor matched lung cancer samples were analysed by the Affymetrix GeneChip Human Genome U133 Array Set HG-U133A. A total of 66 samples were used for microarray analysis, including pair-wise samples from 27 patients.[10] The accession number in the Gene Expression Omnibus (GEO) is GSE7670. The protein-protein interaction data was downloaded from the Human Protein Reference Database (HPRD) (09/01/2007 release).

Entropy minimization

The joint association of gene pair expression states with phenotype was evaluated by calculation of the entropy. Here we adapted a simple formulation called Entropy Minimization and Boolean Parsimony (EMBP).[11] Firstly, the logarithmic transformed expression value of each gene was binarized into two states: 1 as "high-expression" and 0 as "low-expression" using the corresponding average value across samples as the threshold. Then there are four possible states for each gene pair (Table 1). For each state (S) we counted the number of times, N_0, that the state S appeared in normal samples and the number of times, N_1, that it appeared in cancerous samples. Then we defined $q = N_1/(N_0 + N_1)$ as our estimation of the probability that state S is encountered in the cancer phenotype.

The uncertainty of determining whether or not the state is encountered in a cancer phenotype was estimated from the entropy $H(q)$:

$$H(q) = -q\log_2(q) - (1-q)\log_2(1-q)$$

$H(q)$ approaches 0 for values of q that are close to either 0 or 1, and takes a maximum value of 1 for $q = 0.5$. If N_1 or N_0 is equal to 0, then q equal 0 or 1, thus $H(q)$ cannot be defined according to above formula. We set N_0 or N_1 equal to 1 in this situation. To find the most informative predictive gene pairs

Table 1. Example of entropy calculation from protein-protein interaction data for a gene pair.

Module state	Gene 1 state	Gene 2 state	N_0	N_1	H
[0 0]	0	0	28	27	0.9998
[1 0]	1	0	1	50	0.1396
[0 1]	0	1	30	16	0.9321
[1 1]	1	1	23	5	0.6769

N_0, number of times that the state S appeared in normal samples. N_1, the number of times it appeared in cancerous samples. H, Entropy.

associated with a phenotype, we therefore selected those with an $H(q)$ below a particular threshold value (<0.3) for further analysis.

We then performed a label-randomizing permutation test 10^5 times, to assess whether any selected predictive gene pairs differed significantly from those selected at random. The permutation P value were calculated by comparing the H with the randomly permutated H.

SVM classification

To evaluate the prediction performance of gene pairs as biomarkers, we use a support vector machine (SVM) as a classifier to test the classification power. For each gene pair, we used the expression values of the two genes as the inputs. The kernel was polynomial (degree 1), and the prediction accuracy on leave-one-out cross-validation was evaluated by the GeMS tool with default setting[12] (http://www.gems-system.org/).

Results and Discussion

The gene pairs most strongly associated with the phenotype for human lung cancer are listed in Table 2. Almost all of the 16 gene pairs listed showed a prediction accuracy $>90\%$ and all modules had a P-value $<10^{-5}$ in the permutation test. Three gene pairs, Pafah1b1-Ndel1, Cav1-Src and Nos3-Cav1, showed a clear distribution skewness in cancer samples ($N_1 = 26$ vs $N_0 = 1$).

To determine whether these identified gene pairs played a role in the mechanism of tumorigenesis, we further investigated the enriched gene functional categories and pathways, and the genetic association of these informative gene pairs with cancer using the National Cancer Institute (NCI) DAVID tool (http://david.abcc.ncifcrf.gov). A total of 354 genes involved in gene pairs which entropy <0.3 were selected for further analysis. We defined these 354 genes to be the "ensemble signatures", and the 187 genes that showed a "high-expression" status in cancer samples as the "cancer-specific signatures".

Of the 354 genes with ensemble signatures, there were 24 (7.1%) that had a genetic association with human cancer, and of the 187 genes with cancer-specific signatures, there were 18 (9.6%) with a similar genetic association (Table 3, gene-disease association is based on The Genetic Association Database, http://geneticassociationdb.nih.gov/). The most enriched gene function according to Gene Ontology function association in all signatures was "signal transduction" (39.6%, P-value 5.1E-11) and "cell cycle" (15.5%, P-value 4.9E-9, Table 4). In general, the enriched ratio of cancer-specific signatures was higher than ensemble signatures. The most

Table 2. Protein interaction modules predicted to be the most discriminating markers of cancer phenotype.

Gene module					N_0	N_1	H	Prediction accuracy
Id	Gene1	Gene2	Gene1 state	Gene2 state				
1	CANX	FAM107A	0	1	27	1	0.222285	0.94
2	ABCB1	CAV1	1	1	27	1	0.222285	1.00
3	COL10A1	P4HB	0	0	27	1	0.222285	0.98
4	PAICS	CHD3	0	0	26	1	0.228538	0.98
5	CAV1	SRC	1	1	26	1	0.228538	0.89
6	TNFRSF1B	SSR4	1	0	26	1	0.228538	0.91
7	LMO2	MAPRE3	1	0	26	1	0.228538	0.96
8	SMAD3	EPAS1	0	1	26	1	0.228538	0.94
9	NOS3	CAV1	0	1	26	1	0.228538	0.94
10	COL10A1	P4HB	0	0	26	1	0.228538	0.96
11	PDK1	EPAS1	0	1	26	1	0.228538	0.96
12	LMO4	TCF21	0	1	26	1	0.228538	0.96
13	SKIL	SASH1	1	1	26	1	0.228538	0.96
14	PAFAH1B1	NDEL1	0	0	1	26	0.228538	0.89
15	CAV1	SRC	0	1	1	26	0.228538	0.89
16	NOS3	CAV1	0	0	1	26	0.228538	0.94

N_0, number of times that the state S appeared in normal samples.N_1, the number of times it appeared in cancerous samples. H, Calculated Entropy, Prediction Accuracy is calculated applying Leave-one-out cross-validation on SVM classifier (see Materials and Methods for more details).

Table 3. Association of gene signatures with diseases.*

Signatures	Term	Count	Enrichment ratio[a]	*P* value[b]	Genes
Ensemble Signatures (354 genes)	CANCER	25	7.06%	5.60E-06	TP53, PTGS2, CDKN1B, ABCB1, SFN, CCND1, AR, TGFA, ESR1, CDKN1A, EGFR, IL6, VDR, CBFB, AGER, BCL2, FAS, ALDH2, ERBB2, CDK4, NME1, HRAS, MC1R, CTNNB1, IL8
	LUNG CANCER	4	1.13%	0.036935	TP53, PTGS2, CCND1, CDKN1A
Cancer-specific Signatures (187 genes)	CANCER	18	9.63%	6.70E-05	IL6, CTNNB1, ALDH2, CDKN1A, ABCB1, CBFB, BCL2, TP53, TGFA, HRAS, AGER, ERBB2, SFN, ESR1, NME1, EGFR, PTGS2, AR
	LUNG CANCER	3	1.60%	0.09121	CDKN1A, TP53, PTGS2

*354 genes involved in gene pairs which Entropy <0.3 were selected for further analysis using the DAVID tool (http://david.abcc.ncifcrf.gov) which considers the functional assignment of the genes according to the Gene Ontology Index. These genes were defined to be the "ensemble signatures", and the 187 genes that showed an "high-expression" status in cancer samples were defined as "cancer-specific signatures". [a]Enrichment ratio means the percentage of input genes are annotated on given term. [b]*P* value is calculated by DAVID tool.

enriched signaling pathway was the MAPK pathway (9.1%, *P*-value 1.1E-4, Table 5).

An important advantage of our method is that it might reveal cancer-associated expression pattern of gene pairs involved in particular protein-protein interactions. For example, it is widely accepted that Cav-1 might play an important role in oncogenic transformation and metastasis.[13] Cav-1 normally functions as a tumor suppressor gene candidate and could act as a negative regulator of the Ras-p42/44 MAP kinase cascade.[14,15] Here we show that Cav-1 is involved in five gene pairs which is "high-expression" in normal samples (ID = 2, 5, 9, Table 2) and "low-expression" in cancer samples (ID = 15, 16, Table 2). More significantly, the combination of its status with Src or NOS3 (eNOS) could discriminate between cancer and normal phenotypes (Table 6). Src is an oncogene which can down-regulate Cav-1 expression through transcriptional mechanisms.[16,17] Our results clearly demonstrated this pattern: "*If Src high-expression, and Cav-1 low-expression, then leads to cancer*", and "*If Src high-expression, and Cav-1 (still) high-expression, then leads to normal*" (Table 6). It suggests that different outcomes of the down-regulation action of Src on Cav-1 might determine the phenotype discrimination. This is summarized concisely in Table 6 and suggests that the discovery of novel relationships between

Table 4. Gene ontology enriched in gene signatures.

Biological process	Ensemble signature			Cancer specific signatures		
	Count	ratio	*P*-value[b]	Count	ratio	*P*-value[b]
Signal transduction	128	36.16%	7.10E-14	74	39.57%	5.05E-11
Cell cycle	52	14.69%	2.55E-14	29	15.51%	4.92E-09
Cell proliferation	40	11.30%	2.53E-11	22	11.76%	5.46E-07
Protein kinase cascade	20	5.65%	1.10E-05	15	8.02%	3.09E-06
Regulation of metabolism	93	26.27%	1.10E-07	50	26.74%	4.31E-05
Apoptosis	34	9.60%	1.36E-07	17	9.09%	4.85E-04
Mitotic cell cycle	19	5.37%	5.41E-07	10	5.35%	5.18E-04
Regulation of transcription	77	21.75%	6.31E-05	40	21.39%	4.10E-03

[b]*P* value is calculated by DAVID tool.

Table 5. KEGG pathway enriched in gene signatures.

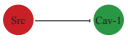
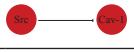
Term	Count	%	*P* value[b]	Genes
MAPK SIGNALING PATHWAY	17	9.09	1.07E-04	TRAF6, IKBKG, TP53, GADD45B, AKT3, MAP3K1, MAP3K3, HRAS, CHUK, MAP3K14, NFKB2, EGFR, MAP3K7IP1, TNFRSF1A, IKBKB, PRKCG, IKBKE,
FOCAL ADHESION	14	7.49	3.62E-04	CTNNB1, BCL2, SRC, AKT3, CAV2, HRAS, ERBB2, CAV1, FYN, EGFR, LAMB2, PRKCG, SHC1, VCL,
APOPTOSIS	12	6.42	1.98E-06	BCL2L1, MAP3K14, CHUK, IKBKG, BCL2, TP53, NFKB2, AKT3, IKBKB, TNFRSF1A, IRAK1, TRADD,
CELL CYCLE	12	6.42	1.35E-05	YWHAZ, CDK2, CDKN1A, MAD2L1, SFN, PCNA, TP53, SMAD3, GADD45B, CCNE1, CREBBP, MCM6,
ADHERENS JUNCTION	11	5.88	3.61E-06	TJP1, CTNNB1, ERBB2, INSR, FYN, SMAD3, SRC, EGFR, PARD3, CREBBP, VCL

[b]*P* Value is calculated by DAVID tool.

Cav-1 and a variety of signaling pathways will offer novel opportunities to develop anti-cancer therapies that target Cav-1.[13]

The idea of extracting synergistic gene pairs for biomarker identification is not new, but our method has several advantages: (1) Interpretability. Compared to methods which search all possible synergistic gene pairs without biological evidence,[18] the cancer signatures identified in the present study are based on protein-protein interactions, which is recognized as the molecular basis of signaling pathways. Furthermore, phenotype discrimination based on protein-protein interactions could contribute to elucidation of the tumorigenesis mechanism. (2) Efficiency. Compared to other global search methods, the use of protein-protein interaction data optimizes exploration of the protein-protein interaction space by focusing on regions which are more likely to yield synergistic gene pairs. (3) Application. Our approach for describing synergistic phenotype discrimination suggests that our method might play a useful role in the identification of combinatory drug targets.

Table 6. The status of protein interaction modules lead to cancer phenotype switch.

Module logic		Phenotype	The mechanism
Src	Cav-1		
High	Low	Cancer	
High	High	Normal	

Acknowledgements

We thank our colleagues for their suggestions on the manuscript. This work was partially supported by the National Natural Science Foundation of China to J.X. (30600759) and the Advanced Space Medico-Engineering Research Project of China to J.X. (01105015, 01104099).

Disclosures

This manuscript has been read and approved by all authors. This paper is unique and is not under consideration by any other publication and has not been published elsewhere. The authors and peer reviewers of this paper report no conflicts of interest. The authors confirm that they have permission to reproduce any copyrighted material.

References

1. Nevins JR, Potti A. Mining gene expression profiles: expression signatures as cancer phenotypes. *Nat Rev Genet*. 2007;8(8):601–9.
2. Wang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*. 2005;365(9460):671–9.
3. van de Vijver MJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*. 2002;347(25):1999–2009.
4. Sotiriou C, et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst*. 2006;98(4):262–72.
5. Paik S, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med*. 2004;351(27):2817–26.
6. Yu K, et al. A molecular signature of the Nottingham prognostic index in breast cancer. *Cancer Res*. 2004;64(9):2962–8.
7. Yu JX, et al. Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC Cancer*. 2007;7:182.
8. Bild AH, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*. 2006;439(7074):353–7.
9. Salter KH, et al. An integrated approach to the prediction of chemotherapeutic response in patients with breast cancer. *PLoS ONE*. 2008;3(4):e1908.

10. Su LJ, et al. Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme. *BMC Genomics*. 2007;8:140.
11. Varadan V, Anastassiou D. Inference of disease-related molecular logic from systems-based microarray analysis. *PLoS Comput Biol*. 2006;2(6):e68.
12. Statnikov A, et al. GEMS: a system for automated cancer diagnosis and biomarker discovery from microarray gene expression data. *Int J Med Inform*. 2005;74(7–8):491–503.
13. Williams TM, Lisanti MP. Caveolin-1 in oncogenic transformation, cancer, and metastasis. *Am J Physiol Cell Physiol*. 2005;288(3):C494–506.
14. Williams TM, et al. Caveolin-1 gene disruption promotes mammary tumorigenesis and dramatically enhances lung metastasis in vivo. Role of Cav-1 in cell invasiveness and matrix metalloproteinase (MMP-2/9) secretion. *J Biol Chem*. 2004;279(49):51630–46.
15. Capozza F, et al. Absence of caveolin-1 sensitizes mouse skin to carcinogen-induced epidermal hyperplasia and tumor formation. *Am J Pathol*. 2003;162(6):2029–39.
16. Engelman JA, et al. Reciprocal regulation of neu tyrosine kinase activity and caveolin-1 protein expression in vitro and in vivo. Implications for human breast cancer. *J Biol Chem*. 1998;273(32):20448–55.
17. Engelman JA, et al. p42/44 MAP kinase-dependent and -independent signaling pathways regulate caveolin-1 gene expression. Activation of Ras-MAP kinase and protein kinase a signaling cascades transcriptionally down-regulates caveolin-1 promoter activity. *J Biol Chem*. 1999;274(45):32333–41.
18. Hanczar B, et al. Feature construction from synergic pairs to improve microarray-based classification. *Bioinformatics*. 2007;23(21):2866–7.