



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/authorsrights>



Contents lists available at ScienceDirect

# Computers in Biology and Medicine

journal homepage: [www.elsevier.com/locate/cbm](http://www.elsevier.com/locate/cbm)



## Impact of TGF- $\beta$ on breast cancer from a quantitative proteomic analysis



Jaegyeon Ahn<sup>a</sup>, Youngmi Yoon<sup>b</sup>, Yunku Yeu<sup>a</sup>, Hookuen Lee<sup>c</sup>, Sanghyun Park<sup>a,\*</sup>

<sup>a</sup> Department of Computer Science, Yonsei University, 134 Sinchon-dong, Seodaemun-gu, Seoul 120-749, South Korea

<sup>b</sup> Department of Computer Engineering, Gachon University, Seongnam, South Korea

<sup>c</sup> Department of Pharmacy, Gachon University, Seongnam, South Korea

### ARTICLE INFO

#### Article history:

Received 11 October 2012

Accepted 25 September 2013

#### Keywords:

Quantitative proteomic analysis

Phosphorylated protein pathway

Data integration

### ABSTRACT

There has been much active research in bioinformatics to support our understanding of oncogenesis and tumor progression. Most research relies on mRNA gene expression data to identify marker genes or cancer specific gene networks. However, considering that proteins are functional molecules that carry out the biological tasks of genes, they can be direct markers of biological functions. Protein abundance data on a genome scale have not been investigated in depth due to the limited availability of high throughput protein assays. This hindrance is chiefly caused by a lack of robust techniques such as RT-PCR (real-time polymerase chain reaction). In this study, we quantified phospho-proteomes of breast cancer cell lines treated with TGF- $\beta$  (transforming growth factor beta). To discover biomarkers and observe changes in the signaling pathways related to breast cancer, we applied a protein network-based approach to generate a classifier of subnet markers. The accuracy of that classifier outperformed other network-based classification algorithms, and current feature selection and classification algorithms. Moreover, many cancer-related proteins were identified in those sub-networks. Each sub-network provides functional insights and can serve as a potential marker for TGF- $\beta$  treatments. After interpreting the roles of proteins in sub-networks with various signaling pathways, we found strong candidate proteins and various related interactions that are expected to affect breast cancer outcomes. These results demonstrate the high quality of the quantified phospho-proteomes data and show that our network construction and classification method is appropriate for an analysis of this type of data.

© 2013 Elsevier Ltd. All rights reserved.

### 1. Introduction

Understanding oncogenesis and tumor progression at the gene or protein level is fundamentally related to cancer diagnosis and prognosis, treatment, and drug discovery related to cancer. Active study in bioinformatics supports this understanding. Currently, most research in this area aims to identify marker genes or build gene networks by comparing the differential expressions of mRNA. A latter group [1–3] commonly integrates interactome and transcriptome data to build cancer specific gene networks. These networks contain highly relevant cancer genes and provide improved diagnostic and prognostic predictive functions.

Chuang et al. [1] found that significantly up-regulated or down-regulated gene sub-networks were more accurate predictors than single genes. Those sub-networks contain many known breast cancer genes which were undetectable in previous studies, which analyzed only transcriptome data. Dutkowski and Ideker [2] found protein modules related to tissue development, breast cancer metastasis, and

the progression of brain cancer. They suggested a network-guided forest method for identifying protein modules and decision logic functions that are composed of proteins in those modules. Taylor et al. [3] suggested that differential gene expression in breast cancer changes the organization of interactome data and eventually affects breast cancer outcomes. In order to search for changes in the global modularity of protein interaction networks, they calculated the average Pearson correlation coefficient (PCC) of a hub protein and its interacting partners. The results revealed that many interactions with altered PCCs were closely related to disease outcomes. Current studies rely on mRNA gene expression data to identify marker genes or cancer specific gene networks. However, because proteins are functional molecules that carry out the biological tasks of genes, they also can be direct markers related to biological functions. Frequently, mRNA abundance changes have not coincided with protein abundance changes and have differed depending on the phenotype [4]. Nonetheless, protein abundance data on a genome scale have not been actively studied because the availability of high throughput protein assays is hindered by the lack of robust techniques, such as RT-PCR.

In our study, we quantified phospho-proteomes of breast cancer cell lines that were treated with TGF- $\beta$ . Next, we applied to the data a protein network-based method to generate

\* Corresponding author. Tel.: +82 2 2123 5714; fax: +82 2 365 2579.  
E-mail address: [sanghyun@cs.yonsei.ac.kr](mailto:sanghyun@cs.yonsei.ac.kr) (S. Park).

a classifier of subnet markers. To build a human omics network, KEGG pathways were added to protein–protein interactions (PPIs). This was done because PPI alone missed many signaling interactions in the KEGG database [5]. The sole use of KEGG has been found to be problematic due to sparsity and the lack of the assignment of the majority of human genes to a definitive pathway [1]. Interaction data collected from various databases were integrated based on the NCBI Entrez gene name. From the integrated omics network, we extracted sub-networks that differentiated TGF-beta treatments in breast cancer and demonstrated the predictive performance of the sub-network markers. This classification outperformed network-based classification algorithms as well as current feature selection and classification algorithms in terms of accuracy. Many cancer-related proteins were involved in those sub-networks. Each sub-network provides functional insights and can be a potential marker for TGF-beta treatments. Interpreting the roles of proteins in a sub-network with various signaling pathways, we found strong candidate proteins that elicited an opposing breast cancer outcome. These results imply that the quality of our phosphorylated protein abundance data is high and that our network construction and classification method is fit for an analysis of quantified phosphoproteomes data.

## 2. Methods

### 2.1. Protein selection

The phospho-proteomes of breast cancer cell lines at four different progression stages (MCF10A1 (M1), MCF10AT1k (M2), MCF10CA1h (M3), and MCF10CA1a (M4)) [6–9] were treated with TGF-beta for 1 h. For each cell, two samples (a control and the sample treated for 1 h) were prepared (eight samples in total). We used control (labeled as *tumor*) samples and the treated samples (labeled as *treated*) for the analysis. After proteolysis of the samples, phospho-proteome was selectively captured from whole cell lysates using TiO<sub>2</sub> chromatography and immobilized metal affinity chromatography (IMAC) methods. Each sample was analyzed by microcapillary liquid chromatography–tandem mass spectrometry (uLC-MS/MS). In all, 1229 proteins were quantified and analyzed. This data is available at <http://embio.yonsei.ac.kr/~Ahn/tc2.php>. A detailed description of this process can be found in Supplementary Table S1.

### 2.2. Additional datasets

In February of 2012, we downloaded 319,308 human PPIs from the I2D database [10], which includes known, experimental and predicted PPIs for humans and five other organisms. The proteins in those PPIs were mapped into protein IDs using UniPROT.

At the same time, we downloaded 549 interactions from the KEGG database [11]. Those interactions included phosphorylation, dephosphorylation, binding/association and indirect effect, or activation/inhibition interactions. We also downloaded a cancer related gene list from the Cancer Genome Project as a reference gene set. This included 474 cancer genes and 19 breast cancer genes.

### 2.3. Identifying TGF-beta affected pathways

Among the numerous interactions from various network datasets, we needed to identify the interactions included in pathways that were impacted by the TGF-beta treatment of breast cancer. To do this, we applied our previous work [12], which exploits the changes in interaction levels from a normal to a tumor state.

The interaction level of the two proteins can be represented by the PCC of the abundance level. Given *tumor* and *treated* groups of samples, a large difference in the PCC value means that there are significant changes in the correlation of two proteins between these two groups of samples. Based on this rationale, each interaction of the network datasets in Section 2.2 was tested as to whether its two PCC values were significantly different. The interaction of two proteins *a* and *b* has 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 the abundance values of protein *a* for the *tumor* and *treated* samples, respectively, and  $v_{bt}$  and  $v_{bn}$  are vectors of the abundance values of protein *b* on *tumor* and *treated* samples, respectively.

Because the threshold is the minimal difference between two groups of samples, fewer, more definite interactions and regulations were selected as the threshold increases. When the TGF-beta affected network is used to predict whether or not a given sample is TGF-beta treated, the false positive rate decreases and the false negative rate increases as the threshold increases. It then becomes necessary to find the optimal threshold yielding the best result in the tradeoffs between the false negative and false positive rate. The optimal threshold value was identified using the leave-one out cross validation (LOOCV) method while lowering the threshold by 0.005.

We presume that true positive interactions help correct predictions of a sample's label. Therefore, we use the network as a classifier with the LOOCV method. Each edge of the 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 procedure shown in Fig. 1.

In Fig. 1, the values *t* and *r* represent the changed values of PCC when an unknown sample *s* is added to the *tumor* and *treated* sample set, respectively. For each interaction, *t* tends to be greater than *r* if *s* is *tumor*. Therefore, if the proper interactions are selected as a classifier, *score<sub>t</sub>* would increase and *s* would be labeled as *tumor*.

## 3. Experimental results

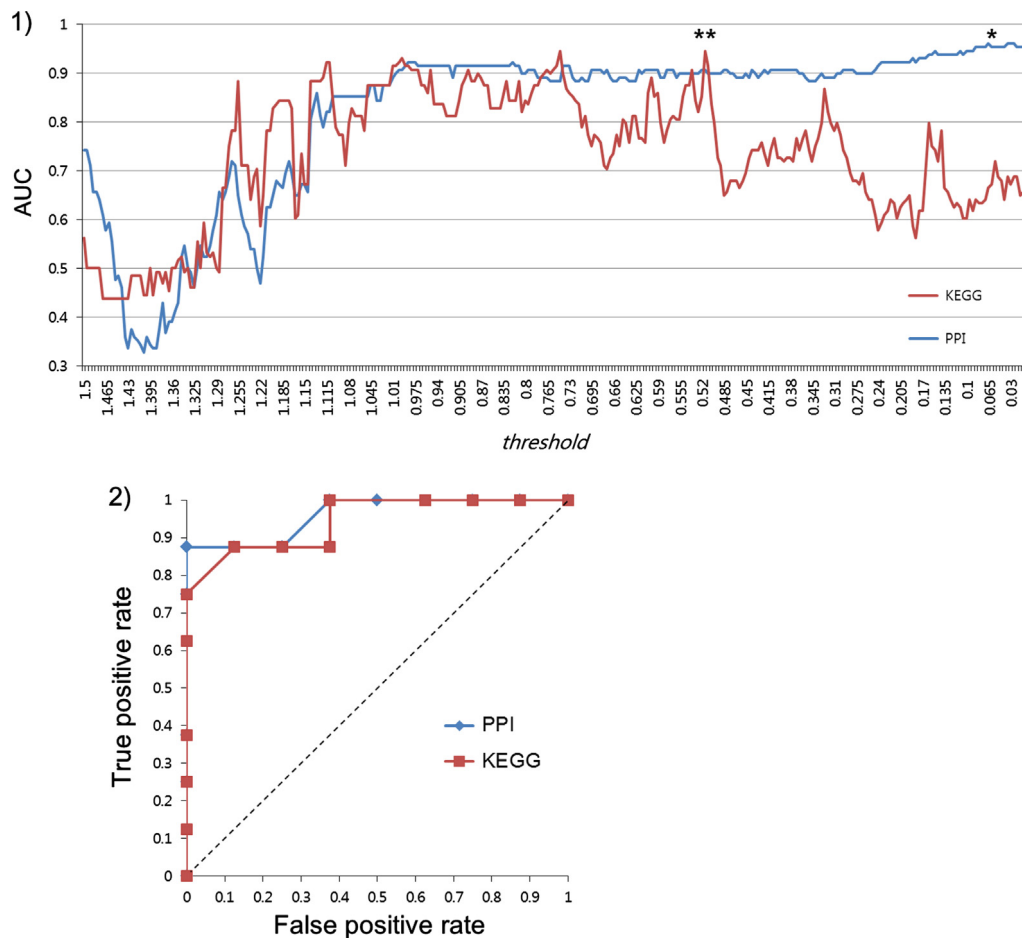
### 3.1. Classification accuracy

We performed a LOOCV test to obtain the optimal thresholds for the PPI and KEGG interactions, lowering the threshold by 0.005. This process is shown in Fig. 2. After we selected the optimal thresholds, we performed the LOOCV again for the integrated network. These results are shown in Table 1, which shows that the optimal threshold for the PPI is low. This means nearly all PPIs contribute to the correct classification.

We also tested two network-based classification algorithms [1,3] and the relief-F [13], correlation-based filter [14], information gain

1. Assign two scores, *score<sub>t</sub>* and *score<sub>r</sub>*, to unknown sample *s* and initialize to zero
2. For each edge  $e = (a, b)$  in the network,
  - (1) Calculate  $\text{PCC}(v'_{at}, v'_{bt})$  and  $\text{PCC}(v'_{an}, v'_{bn})$ , where
 
$$\begin{aligned} v'_{at} &= v_{at} + x_a \\ v'_{bt} &= v_{bt} + x_b \\ v'_{an} &= v_{an} + x_a \\ v'_{bn} &= v_{bn} + x_b \end{aligned}$$
 $x_a$  and  $x_b$  be *s*'s two protein abundance values of *a* and *b*, respectively
  - (2) Calculate *n* and *t* where
 
$$\begin{aligned} t &= |\text{PCC}(v'_{at}, v'_{bt}) - \text{PCC}(v_{at}, v_{bt})| \\ n &= |\text{PCC}(v'_{an}, v'_{bn}) - \text{PCC}(v_{an}, v_{bn})| \end{aligned}$$
  - (3) If  $t \geq n$ , *score<sub>t</sub>* = *score<sub>t</sub>* + 1, else *score<sub>r</sub>* = *score<sub>r</sub>* + 1.
3. If *score<sub>t</sub>* ≥ *score<sub>r</sub>*, *s* is labeled as *tumor* and otherwise, *s* is labeled as *treated*

Fig. 1. Class label prediction procedure.



**Fig. 2.** Determination of the optimal threshold. (1) AUC measured for the PPI network (blue line) and the KEGG network (red line). The threshold that shows the best AUC is selected as the optimal threshold. The optimal thresholds 0.515 and 0.065 are marked with \*\* and \*, respectively. (2) ROC curves for two optimal thresholds. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 1**  
Optimal parameters and classification accuracy of the proposed algorithm.

Interaction type	Optimal threshold	Number of proteins in the network	Number of interactions in the network	AUC
PPI	0.065	592	2360	0.9609
KEGG	0.515	95	87	0.9453
Integrated	–	621	2447	0.9609

[15], gain ratio [16], and Chi-squared [17] gene-based feature-selection algorithms for comparison. The lone exception came when we tested the algorithm of Taylor et al., SMO [18]. For those tests, a sequential minimal optimization algorithm used for the training of a support vector machine (SVM) was used as the classification algorithm. Taylor et al. provide their own classification system. For all tests, LOOCV was utilized. The classification results are shown in Table 2. Also, ROC curves are presented in Fig. 2.

The higher AUC value of the proposed algorithm indicates that it outperformed the other tested algorithms. The results show that the proposed algorithm has a false positive rate of 0, while maintaining a true positive rate of 0.875, as shown in Fig. 3. The algorithm of Chuang et al., which shows good performance with mRNA microarray data [1], showed a low level of classification accuracy here, possibly because the characteristics of protein abundance data are very different from those of mRNA measurement data. Therefore, the algorithm of Chuang et al. is not suitable for protein abundance data. The same explanation can be applied

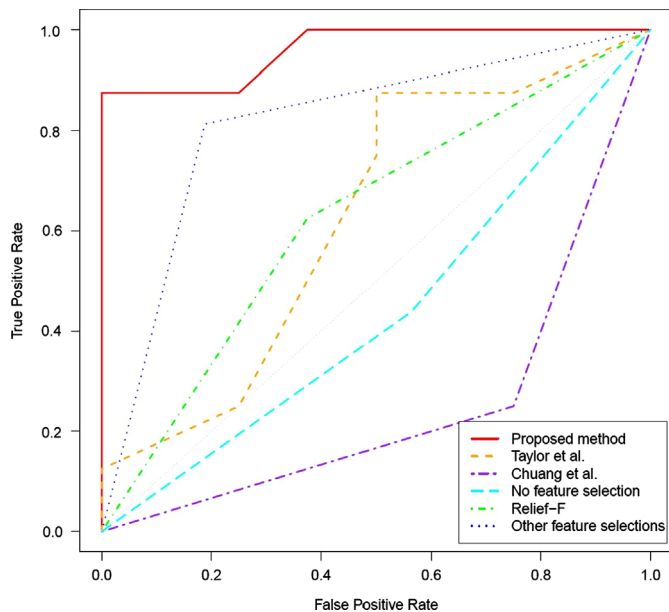
**Table 2**  
LOOCV test results.

Algorithm	Interaction type	Number of proteins in the network	Number of interactions in the network	AUC
Proposed	Integrated	621	2213 <sup>a</sup> 234 <sup>b</sup>	0.9609
Chuang et al.	PPI	14	27	0.310
	KEGG	5	5	0.375
	Integrated	12	27	0.250
Taylor et al.	PPI	53	49	0.625
No feature selection	–	1229	–	0.438
Relief-F	–	361	–	0.625
Gain ratio	–	12	–	0.813
Information gain	–	12	–	0.813
Chi-squared	–	12	–	0.813
Correlation-based filter	–	12	–	0.813

<sup>a</sup> Interactions that are active only in tumor samples.

<sup>b</sup> Interactions that are active only in TGF-beta treated samples.

to the algorithm of Taylor et al. When no feature selection algorithms are applied, the classification accuracy decreases. Therefore, nearly all of the protein abundance signatures are interpreted as noise in the SVM classification system. We found that relief-F, one of the most effective feature selection algorithms for mRNA microarray data, appears to be less effective than other feature



**Fig. 3.** ROC curves of the proposed methods and comparison methods. Other feature selection methods are the gain ratio, information gain, Chi-squared and the correlation-based filter algorithms. These four algorithms selected the same set of genes, causing SVM to result in the same ROC curve. The curves of the proposed method and that by Chuang et al. are drawn using integrated network. The false positive rate is defined as the ratio of incorrectly predicted samples out of all treated samples, and the true positive rate is defined as the ratio of correctly predicted samples out of all tumor samples.

selection algorithms. These results offer a reasonable explanation of the different characteristics of protein abundance and mRNA measurement data.

### 3.2. Analysis of breast cancer specific sub-pathways

We found that many proteins and their interactions have different levels of strength in the tumor and TGF-beta treated state of the cell within TGF-beta related pathways. These pathways contain 621 proteins, as shown in [Supplementary Table S1](#). The number of interactions that are active only in the TGF-beta-treated samples is 234, while the number of interactions that are active only in the tumor samples is 2213—nearly a tenfold increase compared to the TGF-beta treated samples. This result indicates that most portions of TGF-beta related pathways are abnormally activated in the tumor state.

We functionally enriched 621 proteins of our breast cancer specific sub-pathways and proteins as detected by other methods and used for the comparison shown in [Section 3.1](#). The protein lists were annotated using g:Profiler [19], with a  $p$ -value of less than 0.05. The annotation results were quantified as the number of terms that were successfully annotated, as shown in [Table 3](#). Because several terms can be grouped by a hierarchical structure of databases, we also counted the number of groups.

A TGF-beta treatment can affect upstream pathways, including various signaling pathways, leading to a dynamic change in the large downstream pathways. From [Table 3](#), we can confirm that the proposed algorithm is capable of constructing a network which meets this expectation. It is interesting that proteins selected using the gain ratio, information gain, Chi-squared and the correlation-based filter algorithms was not annotated at all, whereas they were shown to be effective features for sample classifications. In addition to these feature selection algorithms, other methods for comparison identified fewer significant marker proteins and cellular processes with a discernible trend.

**Table 3**  
GO annotation result.

Algorithm	Number of groups	Total number of terms	Number of proteins
Proposed	71	689	612
Chuang et al.	22	57	12
Taylor et al.	45	234	53
Relief-F	18	61	361
Other feature selection algorithms*	0	0	12

\* Gain ratio, information gain, Chi-squared and correlation-based filter.

The data used a measurement of the phosphorylated protein abundance with many driver proteins in various signaling pathways. Changes in the abundance of driver proteins tend to be subtle, isolated, and small in amplitude. Because our method is not based on a single protein but on the concordant behavior of proteins connected in a functional network, subtle changes in two interacting driver proteins are detectable. We provide the entire set of functional analysis results in [Supplementary Table S2](#).

We identified 18 genes (21 proteins) and 105 genes (68 proteins) which are specific to breast cancer from the Cancer Genome Project database (CGP dataset) and from the Catalog of Published Genome-Wide Association Studies (the GWAS dataset, [20]), respectively. Our data is composed of 2257 proteins, which includes one protein (P06400 (RB1)) and two proteins (P54727 (RAD23B) and P46663 (BCKR1)) from the CGP and GWAS datasets, respectively. Among these three proteins, our network, which is composed of 621 proteins, contains two proteins (P06400 and P54727). Although the sample size is not sufficiently large, we can say that the proposed algorithm has statistical power to select experimentally validated proteins ( $p$ -value=0.0708, hypergeometric test).

### 3.3. Biological discussion

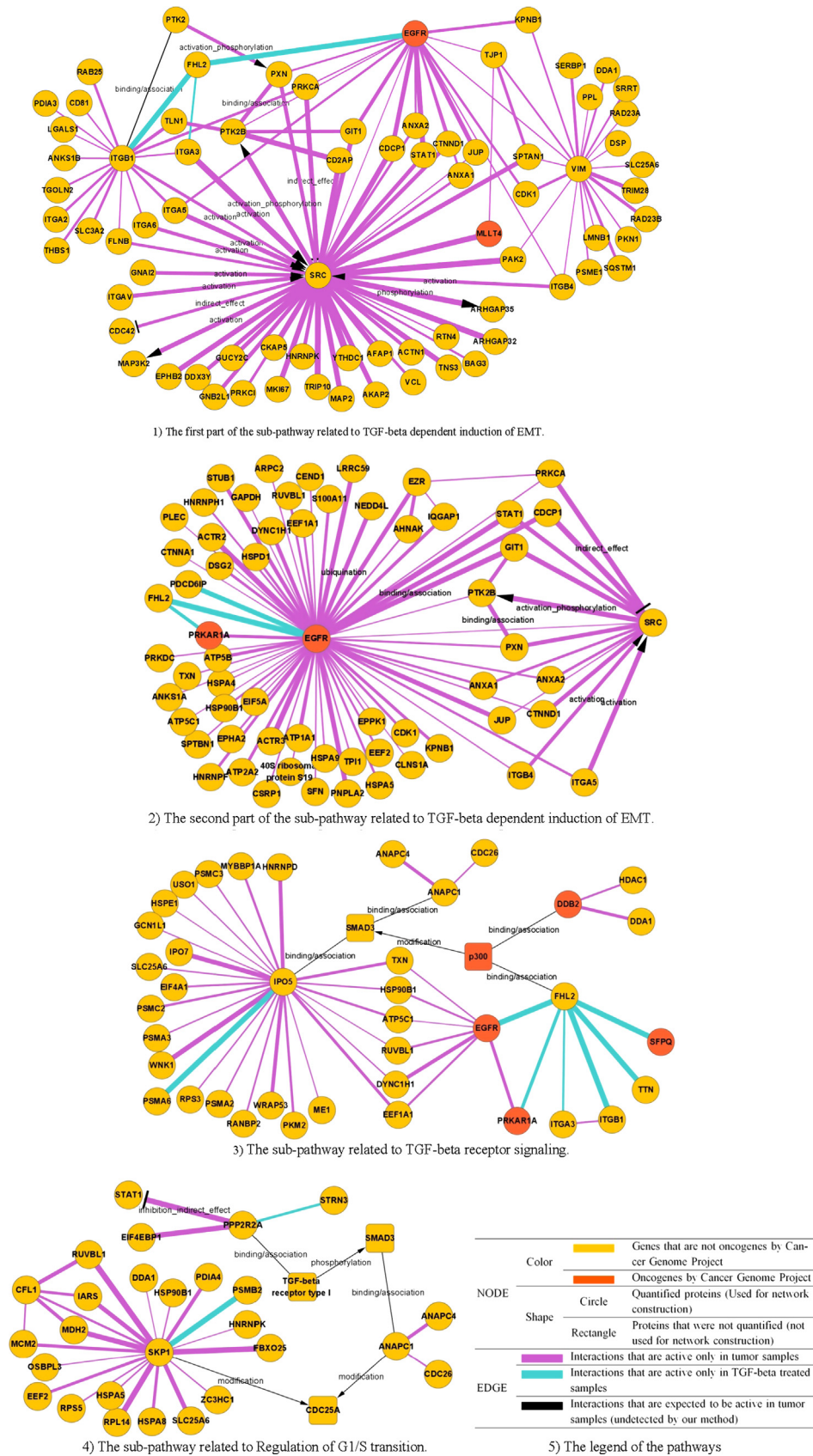
For a more detailed analysis of our network, we used Metacore™ (version 6.5; GeneGo, St. Joseph, MO, USA). Several related works within Metacore™ supported the proteins in our network, and some of them were directly or indirectly related to breast cancer. We selected four annotated sub-networks with the greatest level of significance, as shown in [Fig. 4](#). All edges of the figures are protein–protein interactions unless otherwise specified. The edge thickness correlates with the *Score of an interaction*. Other notations can be found in [Fig. 4](#)–(5). We used Cytoscape [21] to create the graphic presentation.

TGF-beta is one of the major inducers of EMT (epithelial to mesenchymal transition) [22] and can also induce the SMAD-independent pathways involved in an EMT program. Also, it has been reported that TGF-beta receptor-mediated p38 MAPK activation and EMT are enhanced by ITGB1 [23], as shown in [Fig. 4](#)–(1). Sarrio et al. reported that EMT likely occurs within a specific genetic context of the basal phenotype in breast tumors. Proclivity to mesenchymal transition may be related to the high level of aggressiveness and the characteristic metastatic spread of these tumors [24].

Moreover, several integrin subunit proteins, such as ITGB1, ITGB2, ITGB3, ITGAV and ITGB4, play a role in cellular adhesion. Integrin receptors may play a role in breast cancer invasion and metastasis by modifying cells' ability to adhere to surrounding cells and the extracellular matrix [25]. [Fig. 4](#)–(1) shows that ITGB1 has many interactions that are mostly active only in breast tumor samples.

Furthermore, [Fig. 4](#)–(1) shows that ITGB1 and its interacting partners ITGA3, ITGA5 and ITGA6 activate SRC (c-SRC). These





**Fig. 4.** The sub-networks dynamically changed when breast cancer cells were treated with TGF-beta. (1) The first part of the sub-pathway related to TGF-beta dependent induction of EMT, (2) The second part of the sub-pathway related to TGF-beta dependent induction of EMT, (3) The sub-pathway related to TGF-beta receptor signaling, (4) The sub-pathway related to Regulation of G1/S transition and (5) The legend of the pathways.

activations occur only in tumor samples, indicating the important role of SRC in breast cancer. Our deduction is supported by related research which found the activation and suspected key roles of SRC in many breast cancers [26–28]. SRC's involvement in interactions active only in tumor samples include interactions involving proteins EGFR (EGF) and MLLT4, which are reportedly oncogenes according to the Cancer Genome Project. These findings suggest that ITGB1, ITGA3, ITGA5 and ITGA6 interdiction strategies may represent an innovative approach to reestablishing TGF-beta mediated tumor suppression in progressing human breast cancers.

Together with ITGB1, EGFR also modulates EMT. During tumor development, excessive or inadequate EGFR stimulation leads to EMT [29,30]. Fig. 4-(2) shows the detailed interactions of EGFR. Except for FHL2 and PDCD6IP, all proteins have strong interaction with EGFR only in tumor samples. These interactions may support various roles of EGFR during the tumor development process.

EGFR has several common interacting partners with IPO5 (importin (karyopherin)-beta), as shown in Fig. 4-(3). Research shows that the binding of IPO5 to SMAD3 stimulates the translocation of SMAD3 to the nucleus [31]. SMAD3 serves as an intermediate effector, transducing the TGF-beta signal from the plasma membrane to the nucleus, where it participates in the transactivation of downstream target genes [32].

Transcription mediated by SMAD3 is enhanced by p300 [33,34]. Transcriptional co-activator p300 can play a variety of roles in the transcription process, and the mutation of p300 has been found in certain types of human cancers [32].

Fig. 4-(3) also shows several interacting partners of p300, such as DDB2 and FHL2. DDB2 takes part in global genomic repair by recruiting p300 to the damaged chromatin [35]. FHL2 and p300 directly interact and synergistically enhance beta-catenin transcriptional activity [36]. When translocated to the nucleus, beta-catenin can function as an oncogene. High beta-catenin activity significantly correlates with poor prognosis in breast cancer [37]. It follows that Fig. 4-(3) reveals interactions between FHL2 and many proteins, including several oncogenes, such as SFPQ, PRKAR1A and EFGR, which are active only in TGF-beta treated samples. In other words, the processes implied by these interactions are deactivated in the tumor state. Protein p300 was not quantified in our data and is thus not included in our network. However, we were able to detect numerous proteins and their interactions that were included in the TGF-beta receptor signaling pathway. Therefore, the interactions of p300 are expected to affect the binding of p300 and FHL2. For the same reason, we can infer that several tumor-only interactions of IPO5 affect the binding of IPO5 and SMAD3, as described above.

The binding of SMAD3 and ANAPC1 (an anaphase-promoting complex) is also shown in Fig. 4-(3), and Fig. 4-(4) provides more details regarding the binding and related pathways. Fig. 4-(4) describes the regulation of the G1/S checkpoint by TGF-beta and shows how TGF-beta signaling leads to the excitation of various pathways. SMAD3-mediated ANAPC1 activation leads to the degradation of SnoN, thought to play an important role in the transactivation of TGF-beta responsive genes [38]. Phosphorylated CDC25A may be exposed to ubiquitination by ANAPC1 and/or SKP1 (Cul1/Rbx1 E3 ligase) in a Smad3-dependent manner [39]. Although CDC25A was not quantified in our data, or included in our network, many proteins and their interactions were detected in the pathway for the TGF-beta regulation of the G1/S transition. It follows that the ubiquitination of CDC25A by SKP1 affects several tumor-only interactions involving SKP1.

Lastly, PP2A regulatory subunits differentially regulate TGF-beta signaling, eliciting opposing biological outcomes [40], as illustrated in Fig. 4-(4). PP2A is a ubiquitously expressed member of the serine-threonine phosphatase family that is involved in the regulation of many cellular processes, including transcription, translation, cellular

metabolism, and apoptosis [40]. Due to the correlation between PP2A and estrogen receptor alpha (ER) expression in several human breast cancer cell lines, the effect of PP2A on the regulation of ER expression in the human breast cancer cell line MCF-7 was studied [41].

#### 4. Conclusion

In this study, we quantified phospho-proteomes of breast cancer cell lines that were treated with TGF-beta and then applied an integrated protein network construction method to the proteome data. To construct the protein network, we used KEGG pathway data and a PPI dataset. As a result, we identified breast cancer specific protein pathways in which interactions show different activity characteristics between tumor and TGF-beta treated samples. The performance in predicting sample labels using this pathway was superior to that of existing feature selection and classification algorithms and network-based classification algorithms. Moreover, the proteins in this pathway were more functionally enriched using various biological databases.

Many cancer-related proteins supported by the Cancer Genome Project or existing studies were involved in our pathway. Our method is not based on a single protein but rather on the concordant behavior of proteins connected in a functional network. Thus, we could detect subtle changes of driver proteins, enabling the construction of plentiful and accurate protein pathways. In addition, we found that many sub-pathways can be potential markers for TGF-beta treatments. In our discussion, we interpreted the roles of proteins in the sub-network, including various signaling pathways, and found candidate proteins that can be expected to affect breast cancer outcomes. We therefore conclude that our network construction and classification method is appropriate for analyses of quantified phospho-proteomes data.

#### Conflict of interest statement

None declared.

#### 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 work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2012R1A2A1A01010775).

#### Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at <http://dx.doi.org/10.1016/j.compbiomed.2013.09.022>.

#### References

- [1] H. Chuang, E. Lee, Y. Liu, D. Lee, T. Ideker, Network-based classification of breast cancer metastasis, *Mol. Syst. Biol.* 3 (2007) 140.
- [2] J. Dutkowski, T. Ideker, Protein networks as logic functions in development and cancer, *PLoS Comput. Biol.* 7 (2011) e1002180.
- [3] I.W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, et al., Dynamic modularity in protein interaction networks predicts breast cancer outcome, *Nat. Biotechnol.* 27 (2009) 199–204.
- [4] J. Ling, Translation of human genome, *Biochem. Anal. Biochem.* 1 (2011) 101e.
- [5] S. Hwang, S. Kim, H. Shin, D. Lee, Context-dependent transcriptional regulations between signal transduction pathways, *BMC Bioinf.* 12 (2011) 19.

- [6] F.R. Miller, Xenograft models of premalignant breast disease, *J. Mammary Gland Biol. Neoplasia* 5 (4) (2000) 379–391.
- [7] S.J. Santner, P.J. Dawson, L. Tait, H.D. Soule, J. Eliason, et al., Malignant MCF10CA1 cell lines derived from premalignant human breast epithelial MCF10AT cells, *Breast Cancer Res. Treat.* 65 (2001) 101–110.
- [8] L.B. Strickland, P.J. Dawson, S.J. Santner, F.R. Miller, Progression of premalignant MCF10AT generates heterogeneous malignant variants with characteristic histologic types and immunohistochemical markers, *Breast Cancer Res. Treat.* 64 (2000) 235–240.
- [9] B. Tang, M. Vu, T. Booker, S.J. Santner, F.R. Miller, et al., TGF-switches from tumor suppressor to prometastatic factor in a model of breast cancer progression, *J. Clin. Invest.* 112 (2003) 1116–1124.
- [10] K.R. Brown, I. Jurisica, Unequal evolutionary conservation of human protein interactions in interologous networks, *Genome Biol.* 8 (2007) R95.
- [11] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, M. Tanabe, KEGG for integration and interpretation of large-scale molecular datasets, *Nucleic Acids Res.* 40 (2012) D109–D114.
- [12] J. Ahn, Y. Yoon, C. Park, E. Shin, S. Park, Integrative gene network construction for predicting a set of complementary prostate cancer genes, *Bioinformatics* 27 (2011) 1846–1853.
- [13] K. Kira, L.L. Rendell, A practical approach to feature selection, in: *Proceedings of Ninth International Workshop on Machine Learning*, 1992, pp. 249–256.
- [14] L. Yu, H. Liu, Feature selection for high-dimensional data: a fast correlation-based filter solution, in: *Proceedings of 12th International Conference on Machine Learning*, 2003, pp. 856–863.
- [15] T. Mitchell, *Machine Learning*, McGraw Hill, 1996.
- [16] R. Pano, *Transmission of Information*, MIT Press, Cambridge, MA, 1961.
- [17] H. Liu, R. Setiono, Chi2: feature selection and discretization of numeric attributes, in: *Proceedings of IEEE Seventh International Conference on Tools with Artificial Intelligence*, 1995, pp. 338–391.
- [18] J.C. Platt, Fast Training of Support Vector Machines Using Sequential Minimal Optimization, *Advances in Kernel Methods: Support Vector Learning*, MIT Press, Cambridge, MA, 1999.
- [19] J. Reimand, T. Arak, J. Vilo, g:Profiler—a web server for functional interpretation of gene list (2011 update), *Nucleic Acids Res.* 39 (suppl. 2) (2011) W307–W315.
- [20] L.A. Hindorff, J. MacArthur, J. Morales, H.A. Junkins, P.N. Hall, A.K. Klemm, T.A. Manolio, A Catalog of Published Genome-Wide Association Studies, Available at: ([www.genome.gov/gwastudies](http://www.genome.gov/gwastudies)). Accessed [June 09, 2013].
- [21] P. Shannon, A. Markiel, O. Zier, N.S. Baliqa, J.T. Wang, et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks, *Genome Res.* 13 (2003) 2498–2504.
- [22] J. Zavadil, E.P. Böttinger, TGF-beta and epithelial-to-mesenchymal transitions, *Oncogene* 24 (2005) 5764–5774.
- [23] N.A. Bhowmick, R. Zent, M. Ghiassi, M. McDonnell, H.L. Moses, Integrin beta 1 signaling is necessary for transforming growth factor-beta activation of p38MAPK and epithelial plasticity, *J. Biol. Chem.* 276 (2001) 46707–46713.
- [24] D. Sarrió, S.M. Rodríguez-Pinilla, D. Hardisson, A. Cano, G. Moreno-Bueno, J. Palacios, Epithelial–mesenchymal transition in breast cancer relates to the basal-like phenotype, *Cancer Res.* 68 (4) (2008) 989–997.
- [25] H. Shimizu, N. Koyama, M. Asada, K. Yoshimatsu, Aberrant expression of integrin and erbB subunits in breast cancer cell lines, *Int. J. Oncol.* 21 (2002) 1073–1079.
- [26] L.C. Kim, L. Song, E.B. Haura, Src kinases as therapeutic targets for cancer, *Nat. Rev. Clin. Oncol.* 6 (2009) 587–595.
- [27] H. Kim, R. Chan, D.L. Dankort, D. Zuo, M. Najoukas, et al., The c-Src tyrosine kinase associates with the catalytic domain of ErbB-2: implications for ErbB-2 mediated signaling and transformation, *Oncogene* 24 (2005) 7599–7607.
- [28] R. Marcotte, L. Zhou, H. Kim, C.D. Roskelley, W.J. Muller, c-Src associates with ErbB2 through an interaction between catalytic domains and confers enhanced transforming potential, *Mol. Cell. Biol.* 29 (2009) 5858–5871.
- [29] H.W. Lo, S.C. Hsu, W. Xia, X. Cao, J.Y. Shih, Y. Wei, et al., Epidermal growth factor receptor cooperates with signal transducer and activator of transcription 3 to induce epithelial-mesenchymal transition in cancer cells via up-regulation of TWIST gene expression, *Cancer Res.* 67 (19) (2007) 9066–9076.
- [30] S. Barr, S. Thomson, E. Buck, S. Russo, F. Petti, et al., Bypassing cellular EGF receptor dependence through epithelial-to-mesenchymal-like transitions, *Clin. Exp. Metastasis* 25 (6) (2008) 685–693.
- [31] A. Kurisaki, S. Kose, Y. Yoneda, C.H. Heldin, A. Moustakas, Transforming growth factor-beta induces nuclear import of Smad3 in an importin-beta1 and Ran-dependent manner, *Mol. Biol. Cell* 12 (4) (2001) 1079–1091.
- [32] X. Shen, P.P. Hu, N.T. Liberati, M.B. Datto, J.P. Frederick, X.F. Wang, TGF-beta-induced phosphorylation of Smad3 regulates its interaction with coactivator p300/CREB-binding protein, *Mol. Biol. Cell* 9 (1998) 3309–3319.
- [33] R. Janknecht, N.J. Wells, T. Hunter, TGF-beta-stimulated cooperation of smad proteins with the coactivators CBP/p300, *Genes Dev.* 12 (14) (1998) 2114–2119.
- [34] A.W. Tu, K. Luo, Acetylation of Smad2 by the co-activator p300 regulates activin and transforming growth factor beta response, *J. Biol. Chem.* 282 (29) (2007) 21187–21196.
- [35] A. Datta, S. Bagchi, A. Nag, P. Shivanov, G.R. Adami, et al., The p48 subunit of the damaged-DNA binding protein DDB associates with the CBP/p300 family of histone acetyltransferase, *Mutat. Res.* 486 (2) (2001) 89–97.
- [36] Y. Wei, C. Labalette, C. Renard, C. Neuveut, M. Buendia, Study of the role of FHL2 in beta-catenin-associated transcription regulation and oncogenesis, in: *Proceedings of the American Association for Cancer Research* 46, Abstract #1853, 2005.
- [37] A. Lin, R.T. Wang, S. Ahn, C.C. Park, D.J. Smith, A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes, *Genome Res.* 20 (2010) 1122–1132.
- [38] Y. Wan, X. Liu, M.W. Kirschner, The anaphase-promoting complex mediates TGF-beta signaling by targeting SnoN for destruction, *Mol. Cell* 8 (5) (2001) 1027–1039.
- [39] D. Ray, Y. Terao, D. Nimbalkar, L.H. Chu, M. Donzelli, T. Tsutsui, et al., Transforming growth factor beta facilitates beta-TrCP-mediated degradation of Cdc25A in a Smad3-dependent manner, *Mol. Cell. Biol.* 25 (8) (2005) 3338–3347.
- [40] J.C. Keen, Q. Zhou, B.H. Park, C. Pettit, K.M. Mack, et al., Protein phosphatase 2A regulates estrogen receptor alpha (ER) expression through modulation of ER mRNA stability, *J. Biol. Chem.* 280 (2005) 29519–29524.
- [41] J. Batut, B. Schmierer, J. Cao, L.A. Raftery, C.S. Hill, M. Howell, Two highly related regulatory subunits of PP2A exert opposite effects on TGF-beta/activin/nodal signaling, *Development* 135 (2008) 2927–2937.