INTEGRATIVE NETWORK ANALYSIS TO IDENTIFY ABERRANT PATHWAY NETWORKS IN OVARIAN CANCER

LI CHEN[1,2], JIANHUA XUAN[1,*], JINGHUA GU[1], YUE WANG[1],
ZHEN ZHANG[2], TIAN-LI WANG[2], IE-MING SHIH[2]

[1]*The Bradley Department of Electrical and Computer Engineering, Virginia Tech*
900 N. Glebe Road, Arlington, VA 22203, USA
Email: {lchen06, xuan, gujh, yuewang}@vt.edu

[2]*Department of Pathology, Johns Hopkins Medical Institutions*
1550 Orleans Street, Baltimore, MD 21231, USA
Email: {lchen93, zzhang7}@jhmi.edu, shihie@yahoo.com, tlw@welch.jhu.edu

Ovarian cancer is often called the 'silent killer' since it is difficult to have early detection and prognosis. Understanding the biological mechanism related to ovarian cancer becomes extremely important for the purpose of treatment. We propose an integrative framework to identify pathway related networks based on large-scale TCGA copy number data and gene expression profiles. The integrative approach first detects highly conserved copy number altered genes and regards them as seed genes, and then applies a network-based method to identify subnetworks that can differentiate gene expression patterns between different phenotypes of ovarian cancer patients. The identified subnetworks are further validated on an independent gene expression data set using a network-based classification method. The experimental results show that our approach can not only achieve good prediction performance across different data sets, but also identify biological meaningful subnetworks involved in many signaling pathways related to ovarian cancer.

1. Introduction

Ovarian cancer is the fifth leading cause of cancer-related deaths among women in the United States [1]. It has been estimated that 21,990 women will be newly diagnosed and 15,460 women will die of ovarian cancer in 2011 [1]. Most ovarian cancers are serous ovarian carcinomas and only less than 20% of them can be early detected. Prognosis for high grade serous carcinoma patients remains unsatisfactory because most patients develop resistance to chemotherapy after surgery and eventually die [2]. Therefore, chemoresistance has been a critical clinical problem and it is important to understand the biological mechanism of ovarian cancer to overcome the resistance to chemotherapy. Many research topics, such as gene mutation analysis [3], biomarker identification [4], etc., have been carried out to study the chemotherapy resistance. Among them, identification of pathway networks in ovarian cancer becomes an important topic in study.

* To whom correspondence should be addressed.

New technologies have generated large amounts of high-throughput genomic and proteomic data related to ovarian cancer, which make it possible to conduct a comprehensive study from the computational point of view to better examine the cancer genome. The Cancer Genome Atlas (TCGA) project is the one of the studies that collects various biological data for ovarian cancer using genome analysis techniques. It provides opportunities and challenges to develop computational methods to study cancers based on multiple biological data, which can reveal different aspects and levels of biological system function. Traditional computational or statistical approaches, mainly focusing on one type of data source, cannot provide a system view of complex biological system. Current and future needs require sophisticated integration of diverse sets of data, aiming to better understand the main features (e.g., components and their interactions) of biological processes or systems [5, 6]. Many integrative approaches have been proposed to study glioblastoma based on TCGA data portal [7-9]. A recent study explored mRNA expression, microRNA expression, promoter methylation and DNA copy number data on TCGA ovarian cancer samples and provided biologically meaningful results successfully [10]. However, the paper [10] has not conducted a sophisticated integrative analysis across different data sources. Here, we propose a new integrative framework for pathway network identification in ovarian cancer.

Our integrative approach is based on the hypothesis that the cancer phenotype can be reflected by gene expression profiles, which are driven by genomic changes at the copy number level. It is also based on the hypothesis that the highly conserved copy number altered genes might not be differentially expressed at the gene expression level. Therefore, our approach first detects the consensus regions in the DNA copy number data in high-grade ovarian cancer patients and regards the genes with conserved copy number altered as the seed genes. Then a network identification method, based on gene expression profiles and protein-protein interaction network [11], is used to identify significant subnetworks from seed genes that could differentiate two different phenotypes among patients according to the overall survival time. Finally, the identified subnetworks are cross validated on a public gene expression data set using a network-based prediction model. Our results show that the proposed integrative approach can achieve good prediction performance with a high reproducibility across different data sets. Moreover, it also identifies several important pathway related networks, such as ErbB signaling pathway and Notch signaling pathway, which are likely associated with the development of ovarian cancer.

2. Materials and method

2.1. *Integrative framework*

Fig.1 illustrates an integrative framework for copy number analysis, subnetwork identification and prediction by integrating DNA copy number data, mRNA gene expression profiles, protein-protein interaction network and clinical information. From DNA copy number data, we detect significant consensus amplified and deleted regions using Genomic Identification of Significant Targets in

Cancer (GISTIC) algorithm [12] and then extract the genes located in these regions. We consider these genes are highly related to the ovarian cancer mechanism and may function as 'drivers' to form different phenotypes. We also curate ovarian cancer related genes from literature [2] and combine them with the consensus genes from copy number data as seed genes for subnetwork identification. Then we identify subnetworks based on the seed genes using bootstrapping Markov random field-based method (BMRF), which integrates protein-protein interaction networks and the gene expression profiles. For the purpose of evaluation, we adopt a public gene expression data set in the study. We train a classifier on the TCGA gene expression data set on the identified significant subnetworks and then test on the public data set using network-constrained support vector machines (netSVM). Finally we measure the prediction performance and conduct survival analysis on these two gene expression profiles.

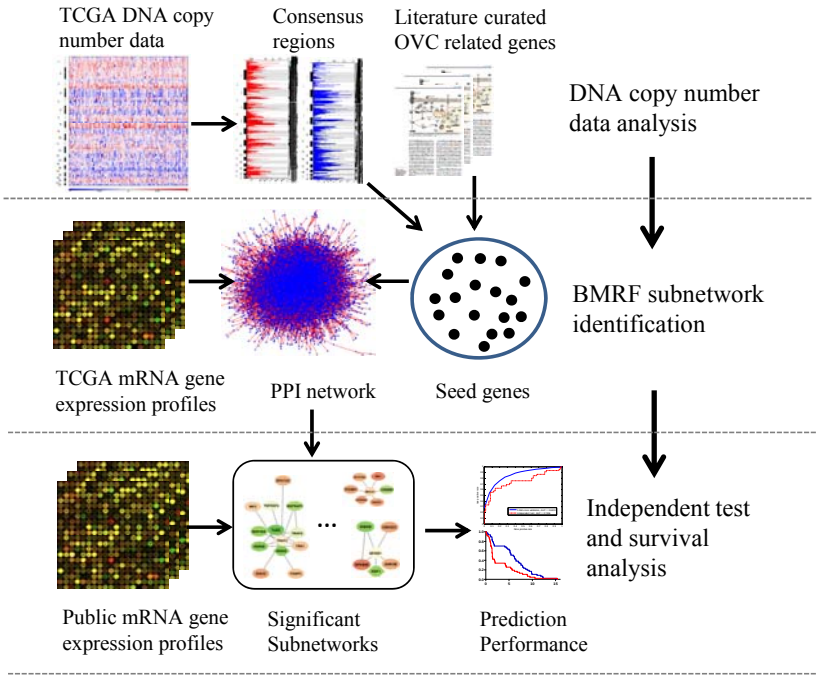


Fig. 1. Illustration of an integrative framework for TCGA ovarian cancer data analysis.

2.2. *Data description*

DNA copy number data and mRNA gene expression data for ovarian serous cystadenocarcinoma cases are obtained from TCGA data portal. 157 patients have DNA copy number expression, mRNA gene expression, as well as the clinical survival information. Among them, 58 patients have overall survival time less than 2 years and 45 patients have overall survival time larger than 4 years. We categorize these two groups as 'high risk' and 'low risk', respectively. An independent public mRNA gene expression data set (GSE3149) for ovarian cancer cases is obtained from Bild

et al. [13]. Accordingly, 45 patients are grouped in 'high risk' and 49 patients are grouped in 'low risk'. The copy number data are arrayed by Agilent Human Genome CGH microarray 244A chips and both mRNA gene expression profiles are arrayed by Affymetrix HG U133A chips. The copy number data are normalized using dChip software [14] and mRNA gene expression data are normalized by Plier and quantile normalization methods [15]. Subnetworks are identified from protein-protein interaction (PPI) network obtained from the HPRD database [16], which contains about 9,000 genes and 35,000 interactions. We convert probe set IDs used in gene expression data to Entrez gene IDs. The probe set ID with the largest variance across patients' samples is used where multiple probe set IDs are linked to one Entrez gene ID. By mapping the PPI network and two data sets we obtain 7,249 genes in 27,885 interactions to be investigated.

2.3. *DNA copy number consensus region detection*

The segmentations on the DNA copy number data are detected by Circular Binary Segmentation (CBS) method [17]. The CBS is a modified binary segmentation method that splits the chromosome into regions of equal copy number to reduce noises and estimates parameters through permutation distribution. The significantly amplified or deleted genomic regions across the ovarian cancer samples are detected by GISTIC algorithm [12] on the segmented DNA copy number data. The GISTIC algorithm takes segmented copy number data and identifies regions of the genome that are significantly amplified or deleted across a set of samples. A G-score is assigned to each aberration that considers the amplitude of the aberration as well as the frequency of its occurrence across all samples. False Discovery Rate q-values are then calculated for the aberrant regions, and regions with q-values below a user-defined threshold are considered significant. Here we set both amplification threshold and deletion threshold as 1 (4 copies for amplification and 1 copy for deletion), and the false discovery rate q-value threshold as 0.01.

2.4. *Network identification by bootstrapping MRF (BMRF)*

We apply a bootstrapping Markov random file (BMRF) method to the TCGA gene expression data by integrating protein-protein interaction network to identify significant subnetworks that could distinguish the expression patterns between 'high risk' and 'low risk'. BMRF method follows a maximum a posteriori (MAP) principle to form a novel network score that explicitly considers pairwise gene interactions in PPI networks.

Let's first define a random variable vector $\mathbf{f} = \{f_1, \cdots, f_m\}$ to represent a set of discriminative scores of m genes (or proteins) between two phenotypes. In the context of a PPI network, Let S represent a gene set of m genes in a network and N_i represent connected neighbors of gene i. We define a pairwise clique C_2 on N_i and S as $C_2 = \{\{i, i'\} \mid i' \in N_i, i \in S\}$.

The random variable vector **f** is said to form a Markov random field on S with respect to N_i and subject to the following conditions:

$$P(\mathbf{f}) > 0,\ \forall \mathbf{f} \in \mathbf{F}$$
$$P(f_i \mid f_{S-\{i\}}) = P(f_i \mid f_{N_i}) \quad (1)$$

The second criterion in Equation (1) is the Markov property of a random field, which states that the probability of a certain configuration at gene *i* is statistically independent of the configurations of all other genes ($j \in S$) given configuration N_i.

The possible configuration **f** of a set of random variable vector **F** obeys a Gibbs distribution if the joint distribution takes the following form:

$$P(\mathbf{f}) = \frac{1}{Z} \times e^{-\frac{1}{T}U(\mathbf{f})}, \quad (2)$$

where *Z* is a normalizing constant given by $Z = \sum_{\mathbf{f} \in \mathbf{F}} e^{-\frac{1}{T}U(\mathbf{f})}$ and U is given by $U(\mathbf{f}) = \sum_{c \in C} V_c(\mathbf{f})$.

U is an energy function that is determined by a sum of clique potentials $V_c(\mathbf{f})$ over all cliques. Clique potentials allow the modeling of knowledge (*a priori*) about the contextual interactions between genes at neighboring sites. For simplicity, we usually assign 0 potential to all cliques of size greater than 2. The energy $U(\mathbf{f})$ corresponds to the probability of that configuration. From Equation (2), we can see that lower energies correspond to more likely configurations. The parameter *T* is often referred to as 'temperature' that controls the sharpness of the distribution. *Z* is a normalization constant and does not need to be calculated.

Denote the observed discriminative scores of genes between two phenotypes as $\mathbf{z} = \{z_1, \cdots, z_m\}$. Here, we define z_i as the z-score of its corresponding p-value p_i using $z_i = \Phi^{-1}(1 - p_i)$, where Φ^{-1} is the inverse normal cumulative density function (CDF) [18]. We assume that the observed discriminative score is a result of the addition of independent zero mean Gaussian noise to the underlying discriminative score; $\mathbf{z} = \mathbf{f} + \mathbf{e}$, $\mathbf{e} \sim N(0,1)$. One possible estimate of the underlying discriminative score **f** is the MAP estimate $\hat{\mathbf{f}}$ that maximizes the likelihood of posterior probability ($\log P(\mathbf{f} \mid \mathbf{z})$); with the help of Bayes' rules and Gibbs distribution, it is equivalent to state that the MAP estimate $\hat{\mathbf{f}}$ minimizes the following posterior potential function: $\hat{\mathbf{f}} = \arg\min_{\mathbf{f}}(U(\mathbf{f}) + U(\mathbf{z} \mid \mathbf{f}))$. The first term in the posterior potential function is the prior potential given by:

$$U(\mathbf{f}) = \sum_{i \in S} V_1(f_i) + \sum_{i \in S} \sum_{i' \in N_i} V_2(f_i, f_{i'}) = \frac{-1}{m} \sum_{i \in S} f_i + \frac{\lambda}{k} \sum_{(i,i') \in E} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_{i'}}{\sqrt{d_{i'}}} \right)^2, \quad (3)$$

where d_i is the degree of gene *i* in the PPI network, *k* is the number of interactions (or edges), and λ is a trade-off parameter. The first term in Equation (3) is the average discriminative score in a subnetwork; the second term in Equation (3) imposes the smoothness across the subnetwork, while putting more weights on the genes with large degrees. Note that the posterior potential function is normalized by the number of genes and the number of edges in the subnetwork, hence, independent of the subnetwork size.

The second term in the posterior potential function is the likelihood potential given by:

$$U(\mathbf{z} \mid \mathbf{f}) = \frac{\gamma}{m} \sum_{i \in S} (z_i - f_i)^2 / 2, \tag{4}$$

where γ is a trade-off parameter. The likelihood potential gives the average square of difference between observed and underlying discriminative scores, given the assumption of a Gaussian distribution of the noise signal with 0 mean and 1 standard deviation.

Thus, we can define the subnetwork score as the negative posterior potential function that takes into account the dependency among the genes of a subnetwork, which, in the form of estimated discriminative scores, can be defined as follows:

$$NetScore(G) = -U(\hat{\mathbf{f}} \mid \mathbf{Z}) = \frac{1}{m} \sum_{i \in S} \hat{f}_i - \frac{\lambda}{k} \sum_{(i,i') \in E} \left(\frac{\hat{f}_i}{\sqrt{d_i}} - \frac{\hat{f}_{i'}}{\sqrt{d_{i'}}} \right)^2 - \frac{\gamma}{m} \sum_{i \in S} (z_i - \hat{f}_i)^2 / 2. \tag{5}$$

Once we define the MRF-based network score, a modified simulated annealing search algorithm is then developed to efficiently find optimal or suboptimal subnetworks with maximal network scores. Finally, to improve their reproducibility across data sets, a bootstrapping scheme is implemented to statistically select confident subnetworks. BMRF method has the advantage in identifying hub genes that usually express little changes among different phenotypes and are hard to detect, therefore improves the mechanism study of ovarian cancer. In this experiment, we determine the significant subnetworks according to network size and network score. A network is considered as significant if its size is greater than 5 and the network score is larger than 1.65 ($p \leq 0.05$; normal distribution).

2.5. *Network constrained support vector machines (NetSVM)*

Given a training sample set ($\mathbf{x}_1$, y_1), …, ($\mathbf{x}_l$, y_l) with p features and l samples, where $\mathbf{x}_i \in R^p$ and $y_i \in \{-1,+1\}$, the SVM learning algorithm aims to find a linear function of the form $f(\mathbf{x}) = \boldsymbol{\beta} \cdot \mathbf{x} + b$, with $\boldsymbol{\beta} \in R^p$ and $b \in R$ such that a data point $\mathbf{x}$ is assigned to a label +1 if $f(\mathbf{x}) > 0$, and a label -1 otherwise. Consider a gene network that is represented by a graph $G = (V, E, W)$, where V is a set of vertices that correspond to p genes, $E = \{u \sim v\}$ is a set of edges indicating that gene u and v are linked on the network and W is the weights of the edges. The degree of a vertex v is defined as $d_v = \sum_u w(u,v)$, where $w(u, v)$ indicates the weight of edge u~v. For this application, the weights could represent the probabilities of having edges between two vertices. Following Chung *et al.* [19], we define the Laplacian matrix $\mathbf{L}$ of G with the uv^{th} element to be:

$$\mathbf{L}(u,v) = \begin{cases} 1 - w(u,v)/d_u & \text{if } u = v \text{ and } d_u \neq 0 \\ -w(u,v)/\sqrt{d_u d_v} & \text{if } u \text{ and } v \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases}. \tag{6}$$

This matrix is symmetric and non-negative definite and its corresponding eigenvalues or spectra reflect many properties of the graph as detailed in [19].

We define the network-constrained SVM given non-negative parameter η as follows:

$$\min_{\beta,b,\xi} \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\beta} + \eta\boldsymbol{\beta}^T L\boldsymbol{\beta} + C\sum_{i=1}^{l}\xi_i \quad s.t.\ y_i\left(\boldsymbol{\beta}\cdot\mathbf{x}_i + b\right) \geq 1-\xi_i, \xi_i \geq 0\,. \tag{7}$$

Note that **L** can be written as $\mathbf{L} = \mathbf{S}\mathbf{S}^T$, where **S** is the matrix whose rows are indexed by the vertices and whose columns are indexed by the edges of G such that each column (corresponding to an edge e = {*u*, *v*}) has an entry $\sqrt{w(u,v)}/\sqrt{d_u}$ in the row corresponding to *u*, an entry $-\sqrt{w(u,v)}/\sqrt{d_u}$ in the row corresponding to *v*, and zero entries elsewhere. Therefore we can see that $\boldsymbol{\beta}^T\mathbf{L}\boldsymbol{\beta}$ can be re-written as

$$\boldsymbol{\beta}^T\mathbf{L}\boldsymbol{\beta} = \sum_{u\sim v}\left(\frac{\beta_u}{\sqrt{d_u}} - \frac{\beta_v}{\sqrt{d_v}}\right)^2 w(u,v)\,. \tag{8}$$

From this representation we can understand that the added regularization term $\eta\boldsymbol{\beta}^T\mathbf{L}\boldsymbol{\beta}$ imposes the smoothness of parameters (coefficients) $\boldsymbol{\beta}$ over the network via penalizing the weighted sum of squares of the scaled difference of coefficients between neighboring vertices in the network.

The solution of Equation (7) could be obtained by reducing it to a conventional SVM optimization problem based on the property of **L** that is symmetric and semi-positive definite. We set equal weight 1 for all connections in PPI database in our experiment.

2.6. *Classification performance merits and survival analysis*

For the identified subnetworks, we conduct three-fold cross validation on TCGA data set and independent test on the public data set. During the cross validation iteration, each time we leave one fold as validation set and the others as training set. Note that the folds are stratified so that they contain the approximately same proportions of labels as the original data. The three-fold cross validation procedure is repeated 100 times in order to get more reliable performance estimation by different randomizations. The average validation performance is reported.

We evaluate the prediction performance through several statistical analyses. Given the true labels of samples and prediction results, we use the Receiver Operating Characteristic (ROC) curve [20] and the area under the curve (AUC) to measure the prediction accuracy of the classifier. ROC curve is a graphical plot of true positive rate (TPR) vs. false positive rate (FPR). AUC is an important performance measure that provides an overall measure of accuracy for the prediction. Furthermore, accuracy, sensitivity and specificity are calculated as well.

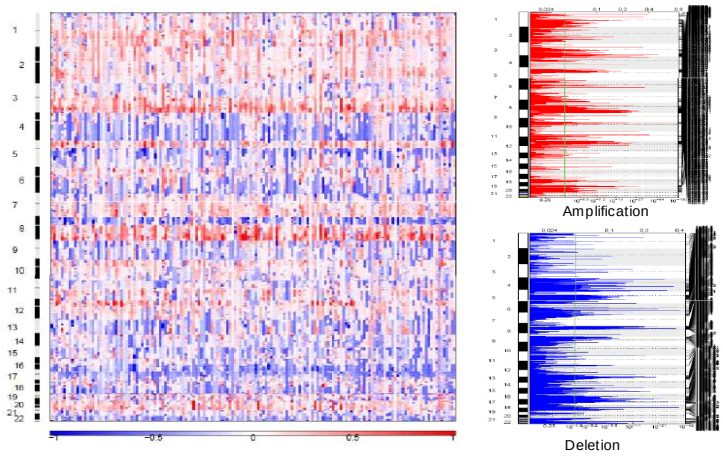
Also, given the sample survival time information, we conduct the Kaplan-Meier survival analysis [21] for the prediction results to generate plots for overall survival time. To compare the difference of two survival curves, p-value and hazard ratio are reported. P-value is calculated by using the log-rank test and hazard is calculated using the Cox proportional model.

3. Results and discussion

Fig. 2 shows the heatmap of the TCGA copy number data and detected consensus regions for both amplification and deletion. There are 869 amplification regions and 979 deletion regions that are

significantly consensus across all the high grade ovarian cancer tumor samples. 751 and 816 genes are located in the amplification regions and deletion regions, respectively. Some of these genes are functionally related to kinase, transcription factor, oncogenes, etc. For example, CCNE1 and MYC are located in the focal amplification regions and they are oncogenes. SMAD4 is located in the focal deletion region and it is a tumor suppressor gene. These findings are biologically interesting; however, many of them do not have clear functional annotation because of the limited knowledge. Most of genes are isolated in the context of PPI network; and it is difficult to understand the underlying biological mechanism even though they are important to sever as the potential 'drivers' in ovarian cancer. Therefore we treat these genes as the seed genes in the subnetwork identification methods.

We also collect 74 genes associated with ovarian cancer pathways from literatures [2], which are functionally related to tumor suppressor, oncogenes, signaling and tumor biology. Finally we obtain 511 genes as the seed genes from DNA copy number data after mapping the genes to PPI network.



(a) (b)

Fig. 2. (a) Heatmap of TCGA copy number data and (b) detected significant consensus regions.

We identified subnetworks from TCGA gene expression data and the PPI network, using BMRF method for all seed genes. Based on the network score and network size, 36 subnetworks are significantly ($p < 0.05$) showing different expression patterns between 'high risk' and 'low risk' groups on the TCGA data set. We trained a classifier using netSVM based on these subnetworks with cross validation. The accuracy (standard deviation) of three-fold cross validation is 79.53% (0.0313) with 76.02% (0.0584) sensitivity and 82.26% (0.0312) specificity. The classifier is further validated on an independent gene expression data set [13] and achieves 74.47% accuracy with 62.22% sensitivity and 85.71% specificity. The ROC curves for cross validation and independent test are shown in Fig. 3. Kaplan-Meier analysis of independent test in Fig. 4 also shows significant different ($p = 0.0003$) in overall survival between two groups predicted as 'high risk' and 'low risk'. We have also performed prediction using traditional gene selection method T-

test (220 genes with $p < 0.01$) and conventional SVM and achieved 86.92% accuracy for cross validation and 69.15% accuracy for independent test. As a comparison, our proposed method achieves better reproducibility across different data sets. Moreover the identified genes by the proposed method are more biologically meaningful than the ones selected by T-test in terms of GO functional annotation, where many signaling pathway related genes are identified by the proposed method (see below), while no significant pathway is enriched in the genes selected by T-test.

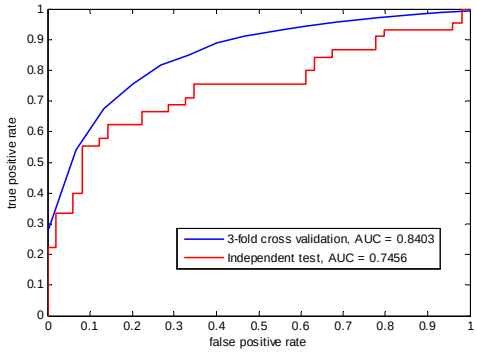
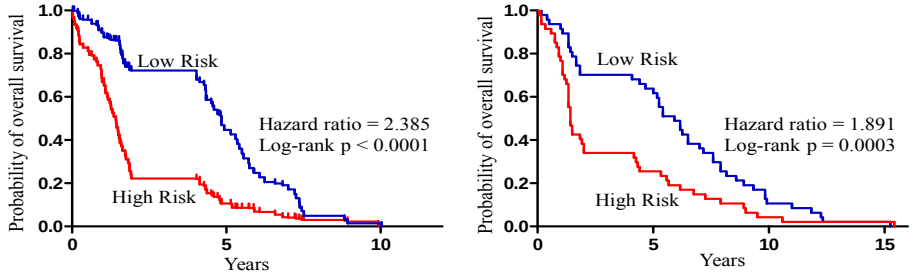


Fig. 3. ROC curves of three-fold cross validation on the TCGA data set and independent test on the public data set.

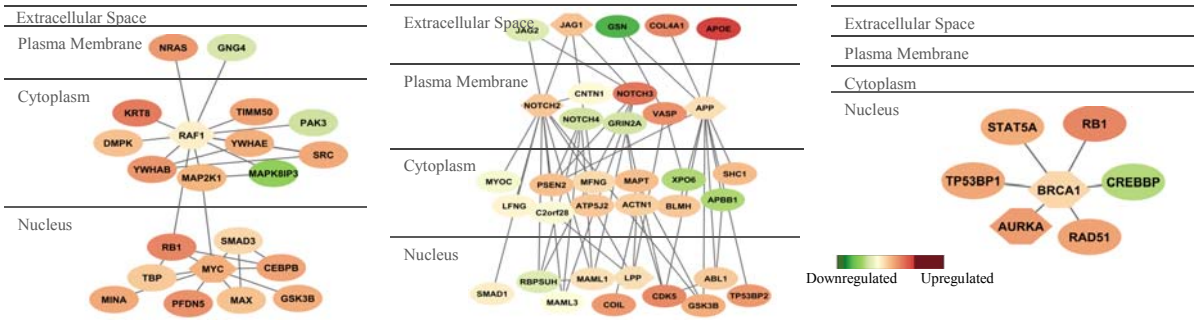


(a) (b)

Fig. 4. Kaplan-Meier overall survival analysis of three-fold cross validation on the (a) TCGA data set and (b) independent test on the public data set.

There are in total 224 genes in the identified subnetworks, grown from 36 seed genes. Among these seed genes, 29 genes are from copy number altered genes and 11 genes are from literature collection (four genes: CCNE1, JAK2, SMAD4 and MYC are overlapped). In terms of gene family of identified seed genes, seven are oncogenes (EGFR, JAK2, JUN, LPP, MYC, NOTCH2 and RAF1); two are tumor suppressors: BRCA1 and SMAD4; and other genes are belonging to transcription factors, cell differentiation markers, protein kinases, etc. Note that CCN1, MYC, BRCA1 are also reported in the study of [10], which indicates that our proposed method could identify biological meaningful genes.

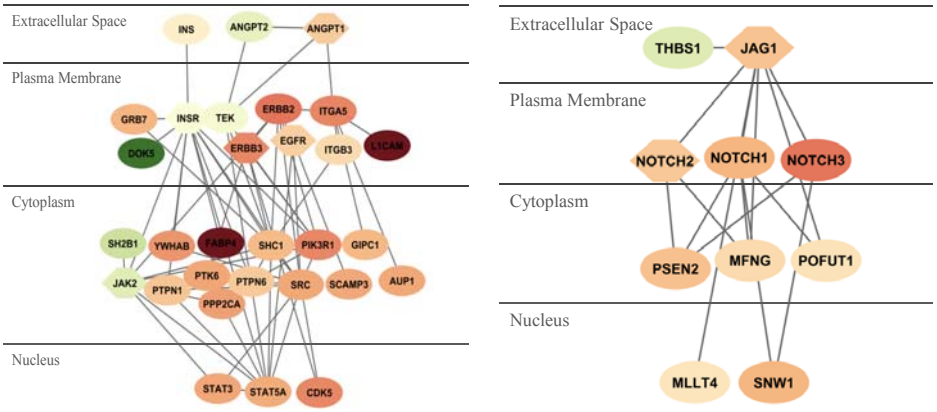
We then conducted functional annotation and pathway analysis using MsigDB database [22] for the identified subnetworks. The functions of 'Cell cycle', 'Apoptosis', 'Nucleus' and 'DNA repair' are significantly enriched in some of the subnetworks, shown in Fig. 5. These findings are consistent with our understanding of the cancer development.



(a) (b) (c)

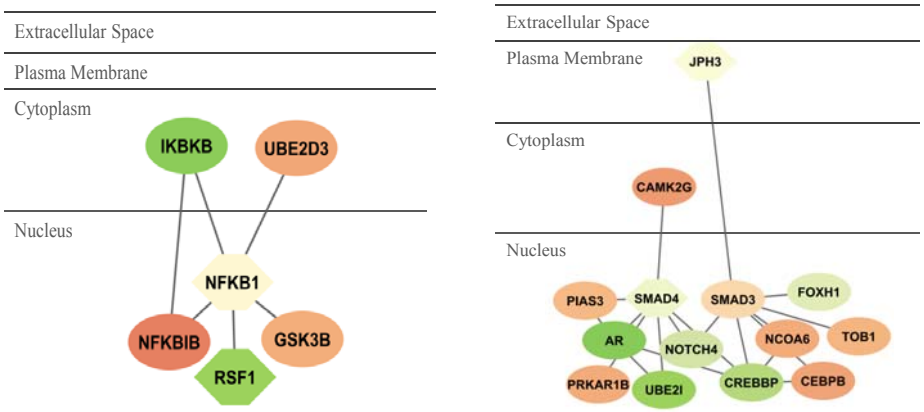
Fig. 5. Subnetworks identified from TCGA ovarian cancer gene expression data set. Subnetworks are merged if more than 2 genes are common. Node shape indicates the seed gene (hexagon) or non-seed gene (ellipse). Node color indicates the fold change between 'high risk' and 'low risk' groups. Red represents over-expressed in 'high risk' group and green reflects over-expressed in 'low risk' group. Enriched pathways and GO functional annotations are: (a) ErbB signaling pathway: p=2.43e-10; Cell cycle: p=1.47e-07. (b) Notch signaling pathway p=1.11e-16; Apoptosis p=3.76e-04. (c) Genes involved in Homologous Recombination Repair p=8.56e-08; Nucleus p=6.61e-04.

Interestingly, many signaling pathways are also significantly enriched in several subnetworks, for example, ErbB signaling pathway (Fig. 5(a), Fig. 6(a)), Notch signaling pathway (Fig. 5(b), Fig. 6(b)), NFκB signaling pathway (Fig. 7(a)), and TGF beta signaling pathway (Fig. 7(b)). Signaling pathway is more complicated and diverse. Epidermal growth factor receptor (EGFR) and ERBB2/HER-2 are members of the ErbB family of tyrosine kinase receptors. The studies have shown that the aberrant activity of EGFR and ERBB2 are important in tumor growth and development. Moreover, the overexpression of EGFR and ERBB2 and their downstream targets is associated with resistance to ovarian cancer chemotherapy [23]. Notch signaling pathway has been studied in many papers showing that it is active in ovarian cancer [24-27]. It is suggested that the inhibition of Notch signaling may be a therapeutic strategy for ovarian cancer [27]. NFκB transcription factors are key regulators of cell proliferation and apoptosis [28]. It is believed that changes in the upstream pathways will deregulate NFκB activation in cancer. Notice that many studies have focused on gene RSF-1 in NFκB network (Fig.7 (a)) and shown that it is involved in paclitaxel resistance in ovarian cancer [29, 30]. Transforming growth factor-beta (TGF-beta) is a tumor suppressor, which is involved in many types of human cancer, including ovarian cancer [31]. Recent study also shows that the activated TGF-b signaling pathway in omental metastases of ovarian cancer is a potential therapeutic target [32].



(a) (b)

Fig. 6. Enriched pathways and GO functional annotations are: (a) ErbB signaling pathway: p=9.10e-13; Signal transduction p=2.33e-07. (b) Notch signaling pathway p=2.86e-15. Figure legends are same as the ones in Fig. 5.



(a) (b)

Fig. 7. Enriched pathways and GO functional annotations are: (a) NFκB signaling pathway p=7.03e-05. (b) TGF beta signaling pathway p=1.88e-06. Figure legends are same as the ones in Fig. 5.

4. Conclusion

We have proposed an integrative framework to analyze TCGA ovarian cancer data by integrating DNA copy number data, microarray data, protein-protein interaction data and patient clinical information. After detecting highly conserved copy number altered genes, we have applied network-based methods to identify pathway networks that can differentiate different expression patterns between different phenotypes. The experimental results have shown that our identified subnetworks could achieve good prediction performance and generalizability on independent data set using a network-constrained classifier. The results also show that our integrative approach can identify biological meaningful subnetworks that related to the development of ovarian cancer and drug resistance.

For the future work, the proposed method will be further enhanced through optimal parameter selection, statistical assessment and multiple hypothesis testing. Moreover, more ovarian cancer samples from TCGA portal and public data sets will be investigated.

5. Acknowledgments

This research was supported in part by NIH Grants (CA139246, CA149653, CA149147, and NS29525-18A1) and a DoD/CDMRP grant (BC030280).

References

1. Jemal, A., F. Bray, *et al.*,CA Cancer J Clin, 61(2): p. 69-90,(2011).
2. Bast, R.C., Jr., B. Hennessy, and G.B. Mills,Nat Rev Cancer, 9(6): p. 415-28,(2009).
3. Norquist, B., K.A. Wurz, *et al.*,J Clin Oncol,(2011).
4. Tcherkassova, J., C. Abramovich, *et al.*,Tumour Biol,(2011).
5. Hanash, S.,Nat Rev Cancer, 4(8): p. 638-44,(2004).
6. Taylor, I.W., R. Linding, *et al.*,Nat Biotechnol, 27(2): p. 199-204,(2009).
7. Cooper, L.A., J. Kong, *et al.*,IEEE Trans Biomed Eng, 57(10): p. 2617-21,(2010).
8. Ovaska, K., M. Laakso, *et al.*,Genome Med, 2(9): p. 65,(2010).
9. Gundem, G., C. Perez-Llamas, *et al.*,Nat Methods, 7(2): p. 92-3,(2010).
10. TCGA,Nature, 474(7353): p. 609-615,(2011).
11. Clarke, R., N. Brunner, *et al.*,Proc Natl Acad Sci, 86: p. 3649-3653,(1989).
12. Beroukhim, R., G. Getz, *et al.*,Proc Natl Acad Sci U S A, 104(50): p. 20007-12,(2007).
13. Bild, A.H., G. Yao, *et al.*,Nature, 439(7074): p. 353-7,(2006).
14. Zhao, X., C. Li, *et al.*,Cancer Res, 64(9): p. 3060-71,(2004).
15. Affymetrix,Edited by Affymetrix I. Santa Clara, CA,(2005).
16. Mishra, G.R., M. Suresh, *et al.*,Nucleic Acids Res, 34(Database issue): p. D411-4,(2006).
17. Olshen, A.B., E.S. Venkatraman, *et al.*,Biostatistics, 5(4): p. 557-72,(2004).
18. Ideker, T., O. Ozier, *et al.*,Bioinformatics, 18 Suppl 1: p. S233-40,(2002).
19. Chung, F., *Spectral Graph Theory*. CBMS Reginal Conferences Series. Vol. 92. 1997: American Mathematical Society, Providence.
20. Witten I. and Frank E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. 2000: Morgan Kaufmann.
21. Kaplan, E. and P. Maier,J AM Stat Assoc, 53: p. 457-481,(1958).
22. Subramanian, A., P. Tamayo, *et al.*,Proc Natl Acad Sci U S A, 102(43): p. 15545-50,(2005).
23. de Graeff, P., A.P. Crijns, *et al.*,Br J Cancer, 99(2): p. 341-9,(2008).
24. Rose, S.L.,Int J Gynecol Cancer, 19(4): p. 564-6,(2009).
25. Choi, J.H., J.T. Park, *et al.*,Cancer Res, 68(14): p. 5716-23,(2008).
26. Park, J.T., X. Chen, *et al.*,Am J Pathol, 177(3): p. 1087-94,(2010).
27. Shih, I.M. and T.L. Wang,Cancer Res, 67(5): p. 1879-82,(2007).
28. Rayet, B. and C. Gelinas,Oncogene, 18(49): p. 6938-47,(1999).
29. Sheu, J.J., B. Guan, *et al.*,J Biol Chem, 285(49): p. 38260-9,(2010).
30. Choi, J.H., J.J. Sheu, *et al.*,Cancer Res, 69(4): p. 1407-15,(2009).
31. Zhang, Y.Y., X. Li, *et al.*,Yi Chuan Xue Bao, 31(8): p. 759-65,(2004).
32. Yamamura, S., N. Matsumura, *et al.*,Int J Cancer,(2011).