IPA: Integrated Predictive Gene Signature from Gene Expression based Breast Cancer Patient Samples

Ashish Saini[1], Jingyu Hou, and Wanlei Zhou
School of Information Technology, Deakin University, Australia
[1]corresponding author (asain@deakin.edu.au)

ABSTRACT

Background: Novel predictive markers are needed to accurately diagnose the breast cancer patients so they do not need to undergo any unnecessary aggressive therapies. Various gene expression studies based predictive gene signatures have generated in the recent past to predict the binary estrogen-receptor subclass or to predict the therapy response subclass. However, the existing algorithms comes with many limitations, including low predictive performances over multiple cohorts of patients and non-significant or limited biological roles associated with the predictive gene signatures. Therefore, the aim of this study is to develop novel predictive markers with improved performances.

Methods: We propose a novel prediction algorithm called IPA to construct a predictive gene signature for performing multiple prediction tasks of predicting estrogen-receptor based binary subclass and predicting chemotherapy response (neoadjuvantly) based binary subclass. The constructed gene signature with considering multiple classification techniques was used to evaluate the algorithm performance on multiple cohorts of breast cancer patients.

Results: The evaluation on multiple validation cohorts demonstrated that proposed algorithm achieved stable and high performance to perform prediction tasks, with consideration given to any classification techniques. We show that the predictive gene signature of our proposed algorithm reflects the mechanisms underlying the estrogen-receptors or response to therapy with significant greater biological interpretations, compared with the other existing algorithm.

Keywords: estrogen receptor, pathologic complete response, chemotherapy, gene expression, prediction, network.

1. INTRODUCTION

Breast cancer is believed as the most common type of cancer and the second leading cause of cancer death among women in the United States. The prime cause of breast cancer death comes from its distant metastasis and recurrence [1]. The diagnosis of breast cancer in the early stage or

the treatment response prediction remains one of the significant challenges, and are yet to achieve. Therefore, to achieve an improved survival rate among breast cancer patients, it is essential to accurately predict the molecular subtypes of breast cancers or the therapy (such as chemotherapy, hormonal therapy) responsiveness outcomes.

The development of microarray gene expression profiling of cancers showed the new strategy that allowed the discovery of new biomarkers for cancer diagnosis, prognosis and treatment [2]. These biomarkers generally consists of a set of informative genes, called gene signatures.

In the recent years, a number of methods [3-10] have been developed using microarray gene expressions to identify the gene signatures that could be used for diagnose, prognose, or treatment response prediction among breast cancer patients. However, some of these existing methods have shown promising results with considering limited number of patient cohorts, but failed to achieve similar performance in additional validation studies [10]. In addition, their gene signatures were not indicative of a gene's role in the biological mechanisms underlying the breast cancer for the tasks of diagnosis, prognosis or treatment response. Evaluating the biological significance from gene expression profiling remains a critical challenge [11]. Thus, predictive markers with robust performance over multiple cohorts and improved biological insights for different predictive tasks are critical factors for translating them into clinical practice, and still remains elusive [10].

For performing the various prediction tasks for breast cancer, [12] indicate that a cancer originates from the driver genes that rapidly alters the expressions in genes that interacts with the driver gene, and is therefore good to consider the network based approach for the reasons, including the networks based approach shows higher reproducibility across different cohorts compared with non-network based approach, and the network based approach shows higher accuracy in performing the prediction tasks. Various network based approach have been developed for performing different prediction tasks [4, 12].

Although, the network based approaches are effective in performing the prediction tasks. These approaches has one major issues, i.e., the existing protein interaction datasets contain many false-positive interactions (the interactions showed in the experimental dataset but never happens biologically), which may cause these interaction datasets biased in discovering the biological knowledge [9]. In other words, the identification of reliable interactions from the experimental protein interaction datasets is one of the most challenging tasks that are yet to be resolve (see [13] for details). While some studies showed the network based approach are effective, the other studies showed the non-network based approaches are effective in performing the prediction tasks. There

is no general consensus on which approach is most effective over the other, and this is one of the potential aims of the current study to discover which approach is better for performing different prediction tasks.

In this paper, we propose a novel computational predictive algorithm, called Integrated Predictive Algorithm (IPA) for performing multiple prediction tasks, including predicting estrogen-receptor based binary subclass and predicting chemotherapy response (neoadjuvantly) based binary subclass. First, from the training cohort, 2-fold cross-validation was performed to extract the predictive gene signatures to avoid over-fitting, and then different classification techniques were incorporated for an algorithm to predict the binary subclasses. Since the performance of the prediction algorithm may vary with different classification techniques, this is the reason that we considered multiple classification techniques, including TreeBoost; decision tree (DT); support vector machine with Linear kernel, Sigmoid kernel, and RBF kernel (SVM-L, SVM-S, SVM-RBF); diagonal linear discriminant analysis (DLDA); the logistic regression model (LR) [14]. Next, we considered the multiple performance measures to further demonstrate the independency of algorithm for the performance measures. Further, the proposed algorithm was constructed using the network based information, called network based IPA algorithm (N-IPA). The performances of the proposed algorithm was evaluated comparing with the *T*test algorithm constructed for performing the binary prediction tasks.

The evaluation of the proposed algorithm demonstrates their robust and stable performances across different cohorts of patients and also outperformed the *T*test algorithm with consideration given to any classification technique. Moreover, the predictive gene signature of the proposed algorithm reflects that the biological meaning of the gene signatures is significant and relates to the mechanisms underlying estrogen-receptor or chemotherapy response based prediction task.

This paper is organized as follows. The proposed algorithm is defined in Section 2. The training and validation cohorts used in this study, statistical validation with patient prediction results, and biological validation are presented in Section 3. Finally, we conclude this paper in Section 4.

2. METHODS

2.1 Integrated Predictive Algorithm

The aim of the proposed algorithm is to construct the gene signature that can predict the binary subclass of estrogen receptors and the neoadjuvant breast cancer patient response to standard chemotherapy regimens. In other words, this problem can be considered as the binary class prediction problem, i.e., predicting estrogen-receptor positive (ERp) and estrogen-receptor negative

(ERn), or predicting pathological complete response (with no residual cancer or lymph-node involved) (pCR) and non-pathological complete response (residual cancer) (npCR).

The ERn subtype is considered as the aggressive form of breast cancer compared with ERp subtype [15, 16]. Also, patients with ERn subtype achieved higher pCR rates to standard chemotherapy regimens (treated neoadjuvantly), compared with their counterpart ERp subtype [17]. Therefore, based on the above ideas, for predicting the binary subclass, we proposed two scoring metrics for each gene in the gene expression dataset (see Table 1 for illustration of gene expression dataset) as:

$$\delta^{g} = \mu^{g}_{pCR,ERn} - \mu^{g}_{pCR,ERp}$$
$$\sigma^{g} = \mu^{g}_{npCR,ERp} - \mu^{g}_{npCR,ERn} \quad (1)$$

here, $\mu^{g}_{pCR,ERn} = \sum_{i \in pCR,ERn} \frac{e^{g}_{i}}{N}$;

e_i^g defines expression of gene g in ith sample, N defines the total number of samples with their labels belongs to pCR and ERn. The prime idea for the equation (1) is to identify the differential expression (DE) pattern of estrogen receptor subclass with considering the chemotherapy response based subclass. Here, δ represents the estrogen-receptor based differential expression effectiveness score with respect to pCR and σ represents the estrogen-receptor based differential expression effectiveness score with respect to npCR.

However, as defined above, we are interested in identifying the genes with their positive or negative δ and σ based two-scoring metric. In other words, If ($\delta > 0$ and$\sigma > 0$) or ($\delta < 0$ and$\sigma < 0$) then gene g has DE strength between subclasses and remained in the dataset; else gene g is removed/filtered from the dataset.

Table 1: Illustration of microarray gene expression dataset incorporated in this study

Chemotherapy response (pCR: 1, npCR: 0)	1	1	0		0
Estrogen receptor status (ERp: 1, ERn: 0)	1	0	1		1
	S_1	S_2	S_3		S_m
g_1	x_{11}	x_{12}	x_{13}		x_{1m}
g_2	x_{21}	x_{22}	x_{23}	$\cdots$	x_{2m}
g_3	x_{31}	x_{32}	x_{33}		x_{3m}
. .	. .	. .	. .		. .
g_n	x_{n1}	x_{n2}	x_{n3}		x_{nm}

Here, g_n defines the n^{th} gene, S_m defines the m^{th} sample, and x_{nm} defines the gene expression of n^{th}gene in m^{th} sample.

The remaining number of genes from the above step with their two-scoring metrics of δ and σ were then retained to from the discriminative score. The discriminative score or *S* measure can be formed by integrating the δ and σ scoring metrics that can evaluate the overall DE strength of any gene g between ERp and ERn with respect to binary chemotherapy response.The equation (2) provides the details for calculating the *S* measure for a gene g (S^g).

$$S^g = \begin{cases} \frac{1}{\theta_1}\delta^g + \frac{1}{\theta_2}\sigma^g \; ; \; \theta_1 \neq 0; \; \theta_2 \neq 0 \\ \frac{1}{\theta_1}\delta^g + \sigma^g \; ; \quad \theta_1 \neq 0; \; \theta_2 = 0 \\ \delta^g + \frac{1}{\theta_2}\sigma^g \; ; \quad \theta_1 = 0; \; \theta_2 \neq 0 \end{cases} \tag{2}$$

Here, θ_1 and $\theta_2$are incorporated to penalize δ andσ, respectively, and represents the number of samples in ERp and ERn groups, respectively. With this equation (2), each gene in a dataset is assigned an*S* measure. Since, the values of θ_1 and θ_2 vary in different cases, the calculation of S^g varies accordingly. The higher the S^g, the higher the DE strength between the two binary classes of estrogen-receptor with considering their binary chemotherapy response. Therefore, based on the *S* measure, the significant genes were then identified with *p*-value <0.05 (using log-rank test), and were extracted. These extracted genes form the gene signature.

Based on the two scoring metrics and the *S* measure, we named this proposed algorithm **I**ntegrated **P**redictive **A**lgorithm for predicting estrogen-receptor and chemotherapy response based binary subclasses (IPA). The pseudo-code is shown in Figure 1.

***Initialize**: $\tau = \emptyset$ // τ is the temporary list of selected genes*
$P_{GS} = \emptyset$ // P_{GS} is the set of predicted gene signature

//Two scoring metrics and S measure evaluation
For any gene g in the dataset D
{
    Evaluate δ andσ
if ($\delta > 0$ and$\sigma > 0$) or ($\delta < 0$ and$\sigma < 0$)
        then generate S measure of gene g

and $\tau = \tau \cup \{g\}$
else
removed/filtered from the dataset D
}

//identifying significant genes to form IDP gene signature
For each gene k contained in τ
{
generate p-Value using log-rank test
if p<0.05
$P_{GS} = P_{GS} \cup \{k\}$
else
removed/filtered from τ
}
return P_{GS}

Fig. 1. Pseudo-code of the IPA algorithm.

2.1.1 Network based IPA Algorithm

Given *n* genes and *m* edges or interactions, the interaction network is defined by G = (V, E), where $|V| = n$ and $|E| = m$. The *S* measure (see equation (2)) was then generated for each gene $g$in the interaction network, as did for IPA algorithm. Therefore, each interaction *y* between genes (g, k) is then assigned a merged *S* measure (φ) from their interacting genes, which is simply the average of *S* measure between genes g and$k$. The generated φ measure for each gene interaction was then used to identify the significant gene interactions with *p*-value <0.05 (using log-rank test), and were extracted. The genes that participate in these extracted significant gene interactions form the gene signature. This algorithm is named network-based integrated predictive algorithm (N-IPA)

2.2 Classification techniques and performance evaluation

The IPA gene signature is then used to perform the prediction tasks for the samples by evaluating their gene signature effectiveness score (*E*) as:

$$E(s) = \sum_{g \in N} E(g, s)/|N| \quad (3)$$

where, *N* defines the total number of genes in the IPA gene signature, $E(s)$ is the gene signature effectiveness score of sample *s*,and $E(g, s)$ is the expression of gene g in sample *s*.

Using equation (3), each sample s can be transformed into$E$, which can be used by any of the existing classification techniques to perform the binary prediction task. For this study, we considered seven widely used classification techniques, including the decision tree (DT); TreeBoost; support vector machine with Linear kernel, Sigmoid kernel, and RBF kernel (SVM-L, SVM-S, SVM-RBF); the logistic regression (LR) model; and the diagonal linear discriminant analysis (DLDA), to evaluate the predictive strength and to show the independency of the algorithm on the classification techniques (see Section 3.3 for details).

Further, if TP represents the number of true positives, TN represents the number of true negatives, FP represents the number of false positives, and FN represents the number of false negatives.

Then, to evaluate the overall performance or the predictive strength of an algorithm, or to demonstrate the algorithm independency on the performance measure, three performance measures were considered, including,

Accuracy, evaluated as: ACC $=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$;

F-measure, evaluated as: F-measure $=\frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$; and

the area under ROC curve (AUC) measure, evaluated from the receiver-operating characteristic (ROC) curve of sensitivity and 1-specificity for different cut-off points [18, 19].

In general, the value of ACC, F-measure, and the AUC is within the range [0, 1], where the value of 1 reflects perfect prediction and 0 reflects false prediction.

3. RESULTS

3.1 Datasets

We retrieved two publicly available breast cancer microarray gene expression datasets (GSE20194 and GSE22226) from the gene expression omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/). The downloaded microarray datasets were normalized as published by the original studies.

These datasets were selected based on the availability of estrogen receptor labels, i.e., ERp or ERn, and the chemotherapy treatment response labels (treated neoadjuvantly), i.e., pCR (pathological complete response with no residual cancer) or npCR (non-pathological complete response with residual cancer). The detailed characteristics of each microarray dataset is shown in Table 2.

Table 2: Microarray datasets used in this study

GEO accession	Platform	Samples	ER (p/n)	Tumour Grade (G1/G2/G3)	Average Age in years (range)	Neoadj. Chemo. regimen	Response rate (pCR/npCR)
GSE20194	GPL96	278	164/114	13/104/150	51.9 (26-79)	TFAC	56/222
GSE22226	GPL1708	130	65/58	7/52/69	48.17 (28-65)	AC or AC/T	32/92

Abbreviations: ER, estrogen receptor; p=positive; n=negative; TFAC, paclitaxel (T), 5-flourouracil (F), doxorubicin (A) and cyclophosphamide (C); AC, doxorubicin (A) and cyclophosphamide (C); AC/T, AC plus taxane (T); pCR, pathologic complete response (disease-free); npCR, non-pathologic complete response (residual disease).

The probe identifier in the cohorts were then converted to gene symbol, as indicated by [20]. In the case when multiple probes mapped the same gene symbol, then the median of the probes was considered to avoid any overestimation that comes with considering the single gene. Following this, the probes with "AFFX" were deleted, as there were no associated genes for these probes.

Since, the samples are limited in size if considering the individual dataset with their binary ER label or their binary chemotherapy response label. Therefore, we integrated the two microarray datasets, as shown in Table 2. However, before integrating the datasets, the following steps were performed. First, the gene expression values of each dataset were normalised (or re-scaled) individually using the formula:

$$\hat{E}_i^{(g)} = \frac{E_i^{(g)} - E^{\min(g)}}{E^{\max(g)} - E^{\min(g)}} \quad (4)$$

where, $E_i^{(g)}$ expresses the g^{th} gene expression value for the ith sample, $E^{\min(g)}$ and $E^{\max(g)}$ represents the minimum and maximum gene expression value for gene g [21]. With this normalization measure (4), the gene expressions generated from varied protocols mapped into a uniform framework to reduce the impact of varied protocols on the data integration. Compared with the original data, we did not observe any significant differences in the normalized gene expressions among the study objects.

Next, a common list of genes from the available microarray datasets with their distinct platforms was extracted by cross-referencing each probe annotation in the dataset, and consists of 8,960

genes. The cross-referencing was done by the UniGene database [22]. Based on these steps, the microarray datasets were then directly integrated to increase the size of the samples and also to make the algorithm independent towards the chosen microarray dataset or their platform types.

From the integrated dataset, the samples with repetitions, missing ER status labels, or missing chemotherapy response labels were excluded. 395 samples remained that consists of 225 ERp patient samples and 170 ERn patient samples, and 87 pCR patient samples and 308 npCR patient samples, respectively.

3.1.1 Network dataset

In this study, we incorporated multiple PPI datasets in order to increase the interactions coverage that are limited with considering the single protein interaction dataset, including Biological General Repository for Interaction Datasets (BIOGRID) [23], INTACT [24], The Molecular Interaction Database (MINT) [25], Database of Interacting Proteins (DIP) [26], The Biomolecular Interaction Network Database (BIND) [27], and Human Protein Reference Database (HPRD) [28]. The gene interaction network were then formed with the genes in the integrated microarray dataset (see Section 3.1) from these multiple PPI datasets by considering the Universal Protein Resource Database [29]. Once constructed, the self-interactions and the duplicate edges were then removed, since they did not have any biological meanings. The resulted gene interaction network contains 75,553 gene interactions involving 7,706 genes.

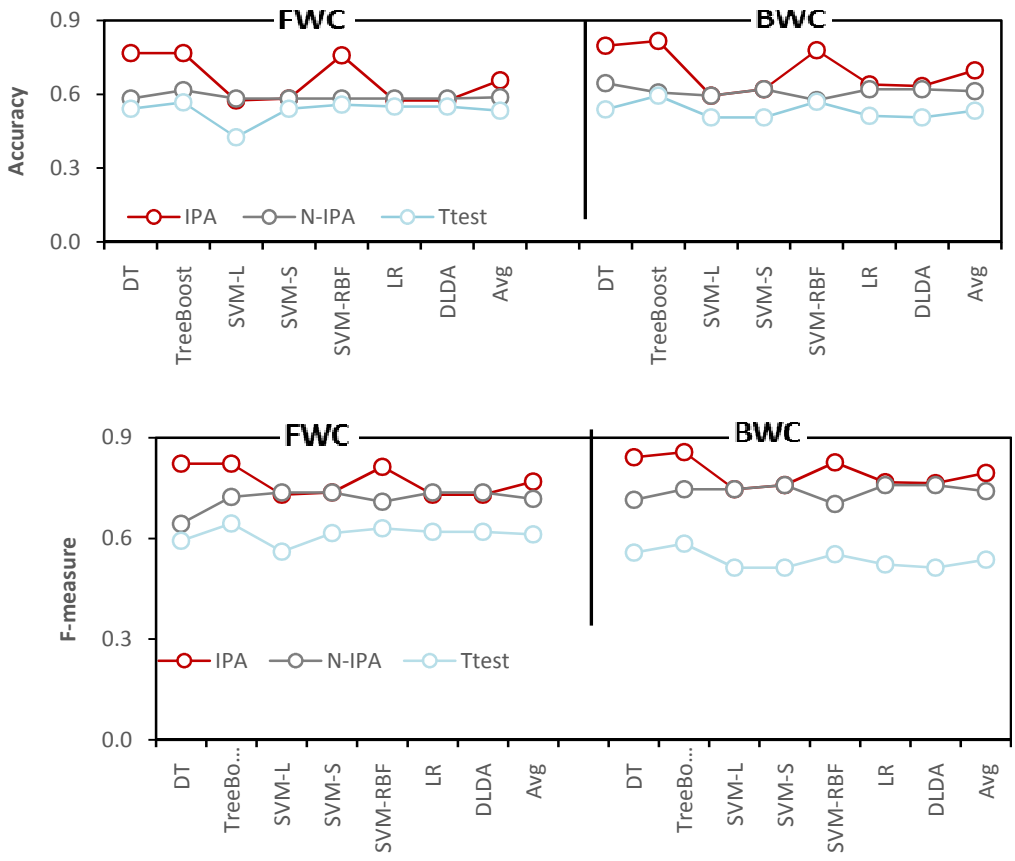
3.2 Two-split dataset to form IPA and N-IPA gene signature

The integrated dataset was split into two different cohorts according to their ER status labels and chemotherapy response labels that contains 197 and 198 samples, respectively, in order to form the IPA and N-IPA gene signature. We named these two distinct cohorts as forward cohort (FWC) and backward cohort (BWC).Two analysis were then performed by applying the IPA algorithm. First, the FWC cohort was used as training set to extract the gene signature, while the BWC cohort was used as validation set to evaluate the algorithm performance. Next, we swapped the cohorts previously used as training and validation. The genes were then extracted that appeared in both the training set based gene signature lists, and formed the IPA gene signature that consists of 18 genes. Similar process was repeated to form the N-IPA gene signature that consists of 106 genes.

3.3 Prediction performance

The performance of our method IPA, N-IPA was evaluated along with their comparison with gene signature generated using *t*-test (denoted as *T*GS). For performing the binary prediction tasks, the algorithms were applied on both FWC cohort and the BWC cohort.

First, we performed the binary prediction task of predicting the ERp or ERn subclass of a sample. For the IPA gene signature, N-IPA gene signature and the *T*GS gene signature, the TreeBoost and SVM-RBF classification technique performed well amongst the other classification techniques. However, performing the binary prediction task with SVM is more computationally time-consuming task. Therefore, the TreeBoost can be chosen as an optimal classifier (being less-time consuming) to build the model compared with SVM, and also showed better prediction performance as SVM-RBF. The Figure 2 shows the IPA algorithm achieved overall best performance measures of Accuracy, F-measure, and AUC. In contrast, the network based IPA algorithm (N-IPA) achieved lesser performance measures, and further, the *T*test achieved worst performance measures. Also, if comparing the algorithms considering the different classification technique, the IPA algorithm achieved nearly the best performance measures, followed by N-IPA and *T*test.



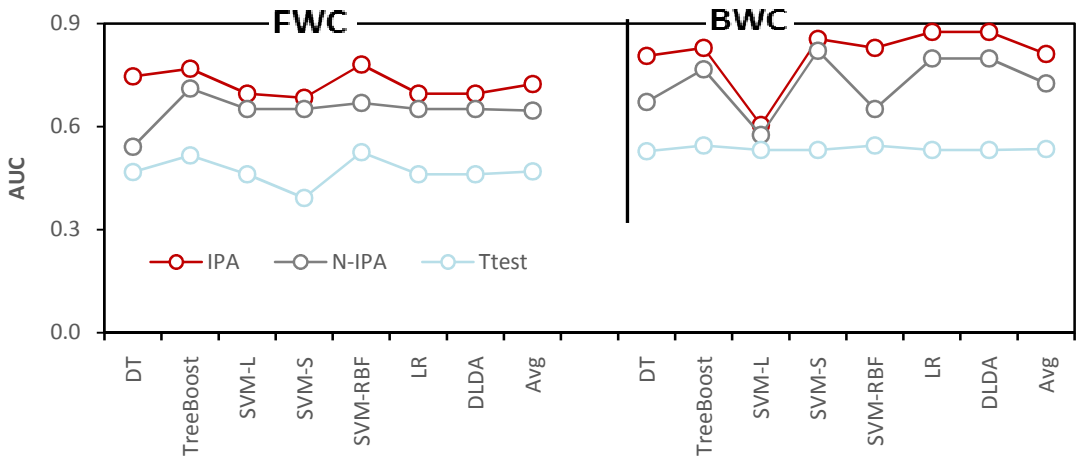
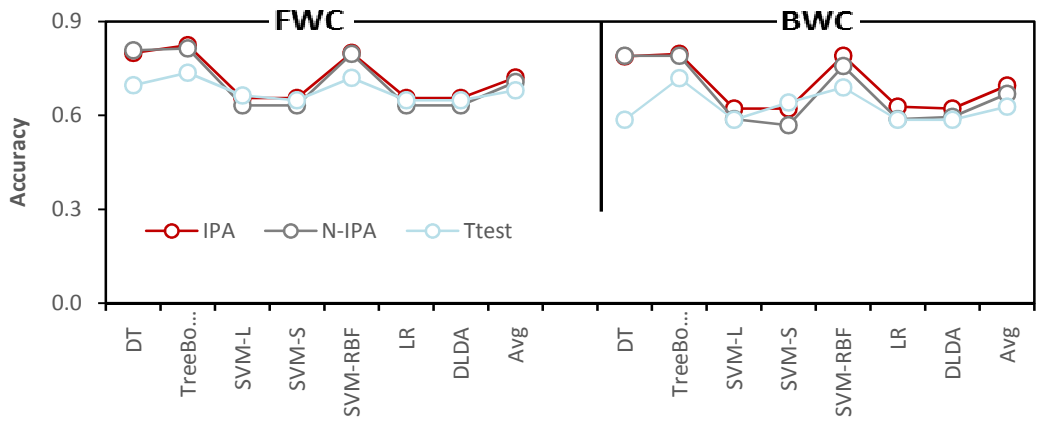


Fig.2. Line charts shows the performance measures of Accuracy, F-measure, and AUC in FWC and BWC cohort for the different prediction algorithms that predict ERp/ERn binary subclass.Here, the Avg for an algorithm denotes the overall average performance, calculated as averaging the performance measure from all the different classification techniques.

Next, we performed the binary prediction task of predicting the pCR or npCR treatment response of a sample. As mentioned previously, the TreeBoost classification technique also performed well in this case amongst the other classification techniques in each of the algorithms. The Figure 3 shows the IPA algorithm still achieved overall best performance measures, followed by N-IPA and *T*test. However, in this case, if considering the individual classification technique, the IPA and the N-IPA algorithm performed nearly similar and contrasting each other. Also, if considering the overall performance measures, the IPA algorithm achieved marginally better performance measures and their difference is statistically non-significant (p-value>0.05).



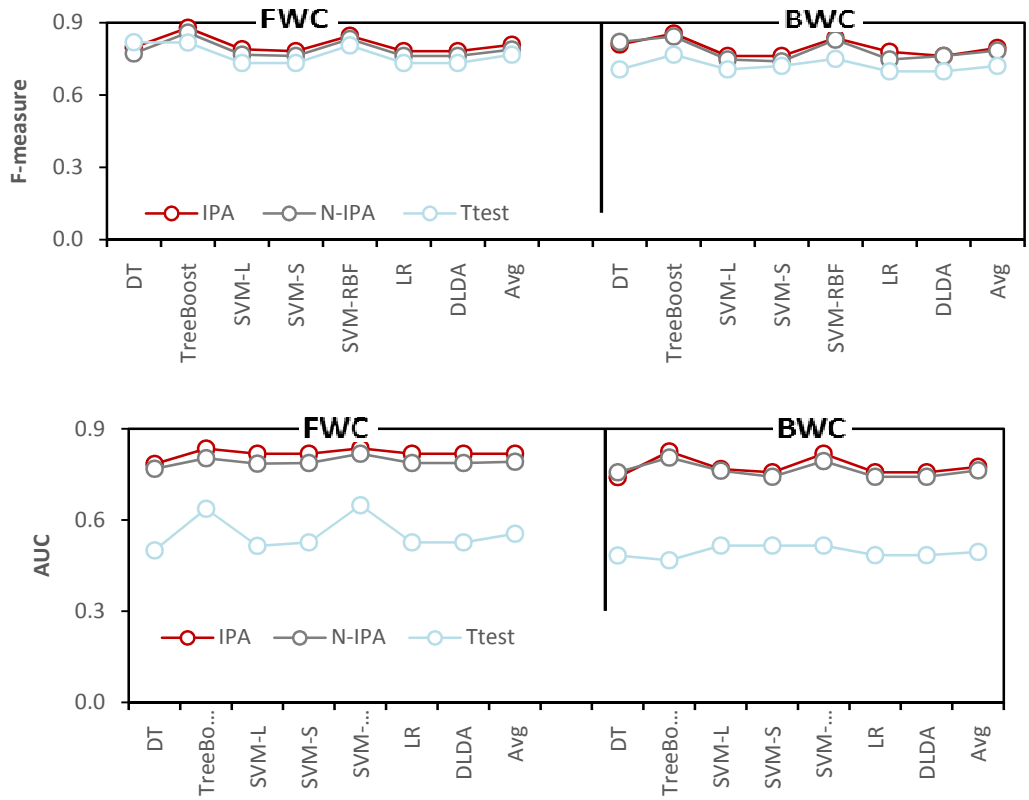


Fig. 3. Line charts shows the performance measures of Accuracy, F-measure, and AUC in FWC and BWC cohort for the different prediction algorithms that predict pCR/npCR binary subclass.

The treatment response prediction of chemotherapy has been shown as complex endpoint to predict [30], due to the differences that comes with the known heterogeneity within the same subclass of cancers or the variation in drug metabolism, dosage, and schedules between each patients [7]. Also, constructing the gene signatures for the prediction of response to chemotherapy has shown to be a more complex problem than predicting the subtypes of cancer [32]. Therefore, the prediction of treatment response is more challenging task than to predict the ERp/ERn breast cancer subclass, and this complexity can be overcome by incorporating interactions network into the gene expressions [10, 33]. Also, from the above experiments, we concluded that considering the network-based information improved the prediction performance measures for performing the treatment prediction task, compared with performing the prediction task of ERp or ERn (see Figure 2 and Figure 3).

However, the protein interactions identified from the experiments usually contain false-positive interactions, i.e., the interactions in the experimental dataset but never happen in real biological processes. As a consequence, the discovered biological knowledge from the interaction networks may be biased. Therefore, the identification of reliable (or biologically valid) interactions is considered to be a challenging issue. The reliable interactions incorporated with gene expressions can significantly lead to improved predictive performance results with improved biological meaning associated with the genes in the gene signature [9].

Since, in the N-IPA algorithm, we directly incorporated the interaction network with the gene expression information without considering the reliability of the interactions. This may leads to biased results and possibly may be one of the potential reasons that N-IPA algorithm showed lesser overall performance measures than IPA algorithm for predicting ERp/ERn subclass and pCR/npCR subclass, respectively. However, if considering the reliability metrics (see [9] for details), then possibly the performance measures may get improved, which leads further investigations and will remain the part of our future work.

Next, as the TreeBoost classification technique achieved best performance measures for each of the algorithm. Therefore for further analysis, we selected the best classification technique to perform the predictive task. The Figure 4 shows the IPA algorithm outperformed the other algorithms on the average (as well as individual cohort) performances in performing the prediction task of ERp/ERn with considering the best classification technique of TreeBoost. Similar results were achieved in Figure 5 when performing the prediction task of pCR/npCR with considering the best classification technique of TreeBoost. From Figure 5, it can be seen that the performance measures of IPA and N-IPA were close together and were consistent with the previously mentioned results as discussed above (Figure 3).

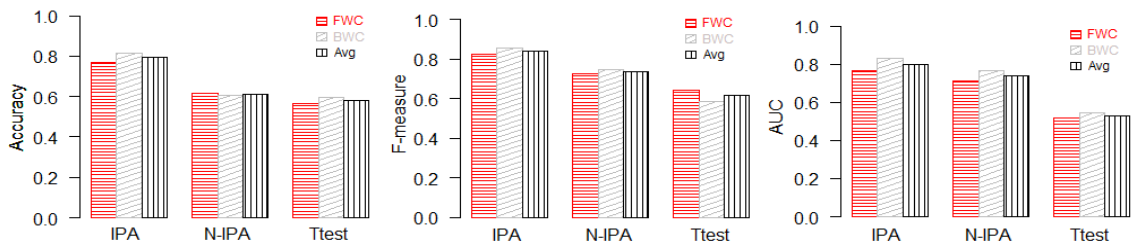


Fig. 4. Bar charts show the performance measures of Accuracy, F-measure, and AUC for different prediction algorithms using the best classification technique of TreeBoost that predict ERp/ERn binary subclass.Here, the Avg for an algorithm denotes the overall average performance measure, calculated as averaging the performance measure from the FWC and BWC.

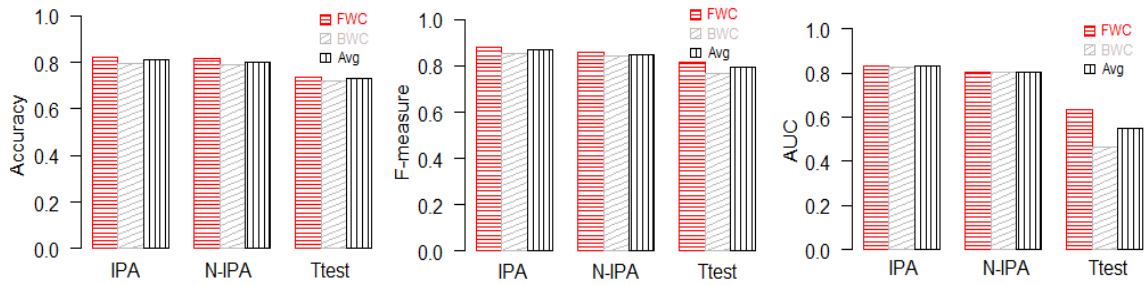


Fig. 5. Bar charts show the performance measures of Accuracy, F-measure, and AUC for different prediction algorithms using the best classification technique of TreeBoost that predict pCR/npCR binary subclass.

In summary, the IPA algorithm based gene signature showed the best predictive performances and higher stability between different cohorts in performing the prediction of binary subclass with demonstrated their independency to the multiple classification technique, and therefore can be more clinically applicable to the patients from other independent cohorts.

3.4 Reproducibility of predictive gene signatures

We performed the reproducibility analysis of gene signatures derived from the IPA, the N-IPA, and the *T*test algorithm, and was performed by calculating the number of overlapping genes between them. There is an overlap of 12 genes (16.7-66.7%) between the IPA gene signature and the other gene signatures. This overlap is greater than the overlap of gene signatures between the N-IPA and the *T*test (2-9%). This reproducibility may contributing to the higher stability that achieved better predictive performances from the different cohorts, compared with the N-IPA and the *T*test algorithms

3.5 Biological analysis of predictive gene signatures

The biological analysis of the gene signatures was performed considering the gene ontology (GO) analysis by using the Ingenuity Pathway Analysis Software (IPA) (http://www.ingenuity.com/). For each gene in the gene signature, GO analysis was performed using the enriched biological process GO terms [34].

Figure 6 lists the enriched biological process associated with the gene signatures along with their *p*-values. The Figure 6 shows many significantly enriched biological process for the N-IPA gene signature, including Cell Signaling, Apoptosis, DNA Replication, Recombination, and Repair, Drug Metabolism, besides others, which are the expected target biological processes that relates

with the treatment response for various anticancer drugs [10] and with the estrogen-receptor subclass.

Further, it can be observed that the N-IPA gene signature is more biologically enriched compared with their counterpart IPA algorithm (see Figure 6). This biological analysis showed that the network based gene signatures are more significantly associated with the phenotype of interest, and the biological meaning of their gene signatures are meaningful and strongly related with the enriched biological processes associated estrogen–receptor subclass or chemotherapy response subclass.

As mentioned previously, incorporating reliability metrics into the interaction network may further enhance the performance results (see Section 3.3), which may also improve the association of their gene signatures with the enriched biological process, significantly.

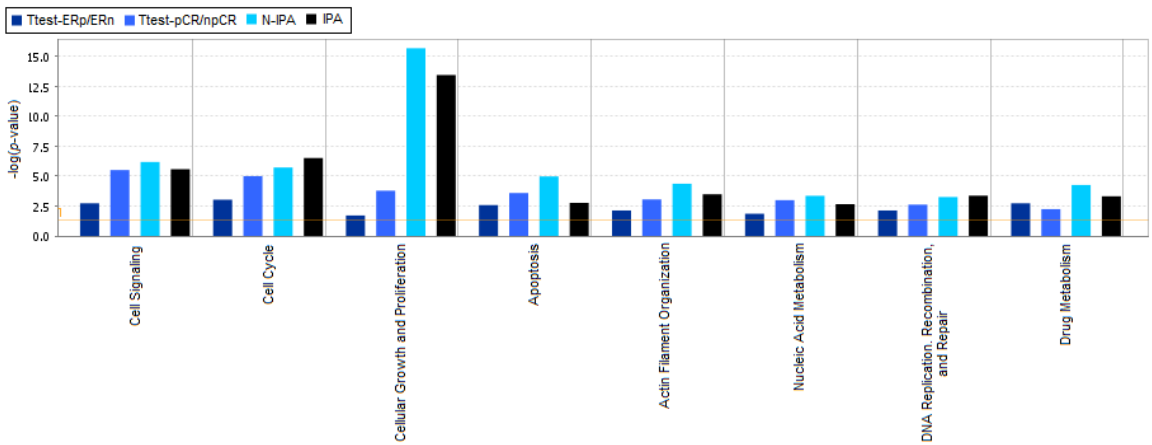


Fig. 6. Enriched biological process based GO functions for the IPA, N-IPA, and *T*test algorithm based gene signatures. Here, *T*test-ERp/ERn stands for the *T*test algorithm based gene signature for predicting the binary estrogen receptor subclass, and *T*test-pCR/npCR stands for the *T*test algorithm based gene signature for predicting the binary chemotherapy response subclass. The *y*-axis shows the logarithm of *p*-values evaluated from the Fisher exact test, with a threshold for statistical significance of *p*-value=0.05 represented by the thin horizontal red line.

4. CONCLUSION

In this study, we derived the gene signatures from our proposed algorithm, IPA and their network-based variant, N-IPA from different cohort of breast cancer patients, which demonstrated the effectiveness of the proposed algorithm in predicting the binary estrogen-receptor subclass and

predicting the binary chemotherapy treatment response subclass with considering multiple classification techniques. Further, we also demonstrated the effectiveness of considering the network-based information in performing the binary prediction task.

From our statistical and biological analyses, we suggest that our proposed algorithm, IPA may serve as better clinical predictors in performing the prediction tasks. While the initial conclusions such as these are motivating, further detailed study is needed with increased samples coverage.

Since, in the N-IPA algorithm we did not consider the reliability metrics that can extract the biologically relevant (or true interactions) by removing the false-positive interactions, which will improve the performance results. This is worthwhile investigating in our future work along with the comparison of existing popular network based algorithms that considered reliability metrics.

5. CONFLICT OF INTEREST

No potential conflicts of interest were disclosed.

REFERENCES

[1] Weigelt, JL, et al. (2005) Breast cancer metastasis: markers and models. *Nature Reviews Cancer*, 5, 591–602

[2] Veer V, (2008) Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, 452, 564-570.

[3] Sotiriou, C. *et al*. (2006) Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade to Improve Prognosis. *Journal of the National Cancer Institute*, 98, 262-272.

[4] Garcia, M. *et al*. (2012) Interactome–transcriptome integration for predicting distant metastasis in breast cancer. *Bioinformatics*, 28, 672-678.

[5] Cun, Y. *et al.* (2012) Prognostic gene signatures for patient stratification in breast cancer - accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. *BMC Bioinformatics*, 13, 69.

[6] Haibe-Kains, B. *et al*. (2012) A three-gene model to robustly identify breast cancer molecular subtypes. *Journal of the National Cancer Institute*, 104, 311–325.

[7] Popovici, et al. (2010) Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Research*, 12, R5.

[8] Miller LD, et al. (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 13550-13555.

[9] Saini, A. *et al*. (2014) RRHGE: A Novel Approach to Classify the Estrogen Receptor Based Breast Cancer Subtypes. *The Scientific World Journal*, 2014.

[10] Dao, P, et al. (2011) Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics*, 27, i205-i213.

[11] Loi S, et al. (2013) CD73 promotes anthracycline resistance and poor prognosis in triple negative breast cancer. *PNAS*.

[12] Chuang, H-Y. *et al*. (2007) Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3, 140.

[13] Hou, J. et al. (2013) Semantically assessing the reliability of protein interactions. *Mathematical Biosciences*, 245, 226-234.

[14] Sherrod, P. H. (2006) DTREG: Classification and Regression Trees and Support Vector Machines for Predictive Modeling and Forecasting. Retrieved from http://www.dtreg.com/

[15] Voduc KD, *et al.* (2010) Breast cancer subtypes and the risk of local and regional relapse. *Journal of Clinical Oncology*, 28, 1684-1691

[16] Millar EK, *et al*. (2009) Prediction of local recurrence, distant metastases, and death after breast-conserving therapy in early-stage invasive breast cancer using a five-biomarker panel. *Journal of Clinical Oncology*, 27, 4701-4708.

[17] Fasching PA, et al. (2011) Ki67, chemotherapy response, and prognosis in breast cancer patients receiving neoadjuvant treatment. *BMC Cancer*, 11, 486.

[18] Zhu W, et al. (2010) Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations. *NESUG 2010 proceedings: Health Care and Life Sciences, November 14-17, 2010- Baltimore, Maryland.*

[19] Powers DMW. (2011) Evaluation: From Precision, Recall and F-Factor to ROC, Informedness Markedness & Correlation. *Journal of Machine Learning Technologies*, 2, 37-63.

[20] Reyal F, et al. (2005) Visualizing Chromosomes as Transcriptome Correlation Maps: Evidence of Chromosomal Domains Containing Co-expressed Genes - A Study of 130 Invasive Ductal Breast Carcinomas. *Cancer Research*, 65, 1376-1383.

[21] Sun Y, et al. (2007) Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics*, 23, 30-37.

[22] Wheeler DL, et al. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **33,** D39-D45.

[23] Stark, C. *et al*. (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Research*, 39, D698-D704.

[24] Kerrien, S. *et al*. (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 40, D841-D846.

[25] Licata, L. *et al*. (2011) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research*, 40, D857-D861.

[26] Xenarios, I. *et al*. (2000) DIP: the Database of Interacting Proteins. *Nucleic Acids Research*, 28, 289-291.

[27] Bader, G.D. *et al*. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Research*, 31, 248-250.

[28] Prasad, T.S.K. *et al*. (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Research*, 37, D767–D772.

[29] The Universal Protein Resource (UniProt). (2007), *Nucleic Acids Research*, 35, D193-D197.

[30] Shi L. (2010) The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, 28, 827-838.

[31] Veer V, (2008) Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, 452, 564-570.

[32] Saini, A. *et al*. (2013) Hub-Based Reliable Gene Expression Algorithm to Classify ER+ and ER- Breast Cancer Subtypes. *International Journal of Bioscience, Biochemistry and Bioinformatics*, 3, 20-26.

[33] Ashburner M, et al. (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25, 25-29.