Hub-Based Reliable Gene Expression Algorithm to Classify ER+ and ER- Breast Cancer Subtypes

Ashish Saini, Jingyu Hou, and Wanlei Zhou

***Abstract*—Identifying gene signatures that are associated with the estrogen receptor based breast cancer samples is a challenging problem that has significant implications in breast cancer diagnosis and treatment. Various existing approaches for identifying gene signatures have been developed but are not able to achieve the satisfactory results because of their several limitations. Subnetwork-based approaches have shown to be a robust classification method that uses interaction datasets such as protein-protein interaction datasets. It has been reported that these interaction datasets contain many irrelevant interactions that have no biological meaning associated with them, and thus it is essential to filter out those interactions which can improve the classification results. In this paper, we therefore, proposed a hub-based reliable gene expression algorithm (*HRGE)* that effectively extracts the significant biologically-relevant interactions and uses hub-gene topology to generate the subnetwork based gene signatures for ER+ and ER- breast cancer subtypes. The proposed approach shows the superior classification accuracy amongst the other existing classifiers, in the validation dataset.**

***Index Terms*—Breast cancer diagnosis, estrogen-receptor, gene signature, hub-gene.**

I. Introduction

With the rapid accumulation of high-throughput technologies, researchers have generated a large amount of data at different levels such as gene expression profiles using microarrays [1], protein-protein interactions (PPI) [2], [3], and many more. These biological data plays a significant role in performing various biological analyses such as to prognose or diagnose the specific cancers.

It has been believed that the breast cancer is the most common types of cancer among the females that has high mortality rate. The subtypes of breast cancer subtypes behaved heterogeneously to different treatment approach and have different survival or death rate. The existing gene signature for the breast cancer classification does not provide the significant results and consistently varies across the datasets [4]. This heterogeneity nature of these gene signatures can classify the patients into irrelevant subtypes, and as a result irrelevant therapy or drug combinations can be given, which has adverse consequences such as early death of a patient. In the last several years, various classification methods for breast cancer have been proposed, such as [5] developed the 70-gene signature (Mammaprint) that classifies the breast cancer patients into good or poor prognosis groups. [6] developed a 76-gene signature that consists of 60 genes for ER+ (Estrogen Receptor Positive) group and remaining 16 genes for ER- (Estrogen Receptor negative) group, in order to classify and to predict the distant metastasis of breast cancer. The gene signatures generated from these existing studies are not stable and heavily depended on the datasets chosen for study [7]. However, to validate the gene signature that can effectively classify the cancer classes, their classification accuracy, robustness and biological meaning is highly essential [4].

It has been believed that the cancers originate from the driver genes that alter the expressions of greater amplitude for the genes that interacts with the driver genes [8]. Since, the gene interactions in the subnetworks provides the models of the molecular mechanisms underlying breast cancer [8], it is therefore essential to incorporate the subnetwork based approach to effectively draws out the biological conclusion, such as to classify the breast cancer subtypes. Many subnetwork-based approaches for breast cancer classification have been developed, such as, [9] is based on the condition-dependent networks from differential expression and no prior interaction information are used, [10] is based on SVM framework that directly incorporates the interaction data in an algorithm for the microarray classification, [4] uses protein interaction data that incorporates with the gene expression data to detect the subnetworks and validates them by randomly shuffling the interactions with their gene expressions.

The existing subnetwork-based gene signatures have various issues associated with them, such as the classification performance is largely correlated with the datasets. Also, the existing protein-protein interaction datasets (PPI) contains several irrelevant (false-positive) interactions that are not real in biological processes, and is believed that only 30-50% of the interactions are biological validated [11]. Therefore, the identification of the reliable interactions from the original PPI datasets is one of the challenging tasks, when using the PPI data for several biological analyses.

In this study, we propose a subnetwork-based approach to classify the estrogen-receptor based breast cancer subtypes that overcomes the above issues associated with the existing approaches. The proposed *HRGE* approach incorporates the three reliability metrics to extract the biologically-relevant interactions and the hub-gene topology to extract the significant genes associated with the two estrogen-receptor based breast cancer subtypes (ER+ and ER-), by using larger compendium of six training datasets. The gene signature of *HRGE* algorithm is then compared with other existing algorithms [4]-[6], [12]. The experimental results demonstrated that the *HRGE* significantly improves the classification performance, as compared with the above

Manuscript received October 15, 2012; revised December 5, 2012.

Ashish Saini is with Deakin University, Victoria, Australia (e-mail: asain@deakin.edu.au).

DOI: 10.7763/IJBBB.2013.V3.156

existing classifiers. The *HRGE* algorithm and their statistical validation results are defined in Section III and Section IV, respectively.

II. Datasets and Data Preprocessing Steps

Four microarray gene expression datasets and Six PPI datasets are downloaded, and transform the proteins to the genes in the gene expression dataset, in order to construct the gene interaction network. Six training datasets are then generated for the extraction of subnetwork-based gene signatures, and are explained below.

A. Microarray Gene Expression Dataset

We incorporated four publicly available breast cancer gene expression datasets, by considering the factors such as estrogen-receptor status (ER+ and ER-), histologic grade (Grade 1, Grade 2, Grade 3), overall survival (OS), distant metastasis free survival (DMFS). In our study, 703 ER+ samples and 255 ER- samples is used for experimental analysis, which makes the total of 958 samples. The detailed information of the sample sizes is shown in Table I. The raw microarray gene expression datasets for breast cancer were downloaded from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) [13], on April 1st, 2012. After downloaded, each dataset is normalized by using Equation 1.

$$\hat{g}_n^i = \frac{\left(g_n^i\right) - \left(\min_p g_p^i\right)}{\left(\max_q g_q^i\right) - \left(\min_p g_p^i\right)} \qquad (1)$$

where g_n^i defines the *i*th feature in the sample *n*, *p* is an sample that shows the minimum gene expression for the *i*th feature, and *q* is an sample that shows the maximum gene expression for the *i*th feature. In this way, all the genes in a dataset can be normalized across the samples.

B. Protein-Protein Interaction Dataset

The protein interactions play an essential part in number of biological processes where the physiological interactions of several proteins are indulged in the construction of biological pathways, such as signal transduction pathways or metabolic pathways. In this study, We incorporated six protein-protein interaction (PPI) datasets, namely, Biological General Repository for Interaction Datasets (BIOGRID) [14], INTACT [15], The Molecular Interaction Database (MINT) [16], Database of Interacting Proteins (DIP) [17], The Biomolecular Interaction Network Database (BIND) [18], and Human Protein Reference Database (HPRD) [19]. The gene interaction network is then constructed from these PPI datasets, by Universal Protein Resource database (UniProt) [20] that transforms the proteins to the genes in the microarray dataset. The self-interactions and the duplicate edges within the constructed gene interaction network are removed, as they did not have any significant meaning in terms of interaction with the other genes. The resulting gene interaction network from the above mentioned six PPI datasets contains 13,012 unique genes and having 69,914 unique interactions among them.

C. Training and Validation Dataset

We have used four microarray gene expression datasets across three distinct platforms, in order to increase the sample size and to balance the other factors, as mentioned in Table I. The integrated dataset is constructed by merging four datasets namely GSE7390, GSE6532, GSE21653, and GSE11121 that contains 958 samples, and the detailed information of each dataset is defined in Table I. Six training datasets is then constructed from the integrated dataset, in order to balance estrogen receptor and the histologic grade status.

TABLE I: Microarray Datasets

	*Desmedt [17] (GSE7390) †	*Loi [24] (GSE6532)	*Sabatier [25] (GSE21653)	*Schmidt [26] (GSE11121)
Platform	HG-U133A	HG-U133A, HG-U133B	HG- U133Plus2.0	HG-U133A
Samples	198	327	255	200
ER				
ER+ (# of samples)	134	263	150	156
ER- (# of samples)	64	45	102	44
Tumour Grade				
Grade 1 (# of samples)	30	52	44	29
Grade 2 (# of samples)	83	158	88	136
Grade 3 (# of samples)	83	57	116	35
Metastasis Free Survival				
Yes (# of samples)	62	70	81	46
No (# of samples)	136	224	160	154
Total Samples Selected (On the basis of histologic grade and receptor status)	**958 (703 (ER+) and 255 (ER-))**			

Microarray gene expression datasets used in this study. Patients with missing histologic grade and estrogen receptor status based information, are excluded from the training datasets. Here, # denotes 'number', * and † represents the training datasets and validation dataset, respectively.

Each of the six training datasets is used in our algorithm to construct the effective gene signatures for two estrogen-receptor based samples, which is presented in Section III. The subnetwork-based gene signatures generated from the training datasets is then tested on a validation dataset, and the result is presented in Section IV.

III. Algorithm

Our main focus is on to extract the subnetwork based gene signature that shows highly correlated gene expressions with the estrogen receptor status.

For the construction of reliable gene interaction maps, the reliability metrics (*WR*) and gene expression metrics (*MGE*) are incorporated which detects the reliable gene interactions that are real in biological processes. The generated reliable gene expressions (*RGE*) of interactions are then used by the hub-gene based approach and presented later. Six analyses were then performed (three for ER+ and three for ER- (i.e., Grade 1, Grade 2 and Grade 3, respectively)) on the training datasets in order to extract the optimal size of subnetwork based gene signature, and the classification analysis was done on a validation dataset to evaluate the gene signature accuracy and its stability. In the below sub-sections, the proposed algorithm is defined.

A. Reliability Metrics

The PPI datasets contain large amount of protein interaction data and is considered as a rich information source from which biological knowledge of interest and facts can be discovered, such as classifying the breast cancer classes or to classify the patients according to their survival rate. However, the analyses of high-throughput protein interaction data signify that the protein interactions identified by the experiments usually contains several irrelevant interactions that never takes place in the real biological processes. As a matter of fact, the discovered biological knowledge or inferred facts from the protein interaction database may be distorted or biased. Therefore, the identification of the reliable protein interactions from the original protein interaction datasets is considered as one of the essential and challenging issues, which can significantly improves the quality of protein interaction datasets and as a result increase the reliability of the discovered biological knowledge and facts.

TABLE II: THREE PROPOSED RELIABILITY MEASURE: (1). *RW1*, (2). *RW2*, (3). *RW3*

	Data Sources **(*RW1*)**	**Experimental Methods** **(*RW2*)**	**Level-based Interaction Partners** **(*RW3*)**
Definition:	It evaluates the reliability of an interaction on the basis of data sources.	It evaluates the reliability of an interaction on the basis of the experimental methods.	It evaluates the reliability in respect of the interacting neighbours of a gene.
Evaluation:	*RW1* is calculated as counting the number of data sources *n* that contains an interaction (*a*, *b*), i.e, $RW1^{(a,b)} = \sum_{n=1}^{6} d_n^{(a,b)}$ (2) Where, $d_n^{(a,b)} = \begin{cases} 1, if\ n\ contains(a,b) \\ 0, otherwise \end{cases}$	*RW2* is defined as the reliability measure which evaluates the reliability of any interaction (*a, b*) by counting the number of experimental methods *n* that identified (*a*, *b*), i.e, $RW2^{(a,b)} = \sum_{n=1}^{N} E_n^{(a,b)}$ (3) Where, $E_n^{(a,b)} = \begin{cases} 1, if\ n\ \text{identified}(a,b) \\ 0, otherwise \end{cases}$ and, *N* defines the total number of experimental methods.	*RW3* is defined as the reliability measure that evaluates the reliability of any interaction (*a, b*) by counting the number of level- *p* neighbours, where $p \geq 2$ i.e, $RW3^{(a,b)} = \sum_{p=1}^{n} I_p^a + \sum_{p=1}^{m} I_p^b$ (4) Where, $I_p^a = \begin{cases} N_p, if\ p \geq 2 \\ 0,\ otherwise \end{cases}$, *n, m* defines the highest level neighbours of gene *a, b* respectively, and N_p defines the number of level-*p* neighbours of a gene.
Conclusion:	Higher the *RW1* of an interaction, more strong the interaction strength is, and thus more reliable the interaction is.	Higher the *RW2* more will be the reliability of a gene interaction.	From our experimental analysis, higher the value of *RW3* more will be the reliability of a gene interaction.

For the interaction between any two genes, we incorporated three reliability measures to assess the reliability on the basis of three distinct factors, i.e, Data Sources (e.g., HPRD, DIP, MINT), Experimental Methods (e.g., two hybrid, affinity chromatography), and Level-based Interaction Partners (e.g., level-1, level-2 interaction partners). Each of the three reliability measures is assigned a reliability weight to an interaction and called as *RW1*, *RW2* and *RW3*. These proposed reliability measures are defined above in Table II.

Once evaluated the reliability measures from data sources (*RW1*), experimental methods (*RW2*), and level-based interaction partners (*RW3*), we performed two major steps. First, each of the reliability measure is normalized (by using Equation 1) across the gene interactions. The essentiality of normalization is to propose a global scale of reliability that defines the reliability strength of each reliability measure within that scale. Once, it is done, the second step is to integrate the three reliability measure to form Weighted Reliability Measure (*WR*), using Equation 5, i.e,

$$WR^{(a,b)} = \sqrt{\sum_{i=1}^{3}\left(RWi^{(a,b)}\right)^2 \times w_i} \quad (5)$$

where $RWi^{(a,b)}$ (*i=1, 2, 3*) defines $RW1^{(a,b)}$, $RW2^{(a,b)}$ and $RW3^{(a,b)}$ respectively for any interaction (*a*, *b*), and w_i defines the weight that is assigned to each reliability measure. From our experimental observations, we assigned a weight of 0.40 to *RW1*, 0.35 to *RW2*, and 0.25 to *RW3*. The possible range of *WR* for any interaction ranges between 0 and 1, i.e, *WR* $\in$ [0, 1].

Therefore, using Equation 5, the *WR* measure evaluates the reliability of all gene interactions in the gene interaction dataset, in terms of three essential criterions i.e., data information sources, experimental evidences, and level-based interaction partners of an interaction.

B. Gene Expression Metrics

For gene expression metrics, we used our six training datasets, as defined in Section II. From a training dataset, each gene is summarized by evaluating the generalized mean (*GE*) across the samples by using Equation 6.

$$GE^{(a)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(g_i^{(a)}\right)^2} \quad (6)$$

where *n* defines the total number of samples, and $g_i^{(a)}$ defines the gene expression of gene *a* in the *i*th sample. Therefore, using Equation 6, each gene in the six training datasets can be summarized across the samples. Each gene in our gene interaction network is assigned a value from each summarized gene (Equation 6), for each training dataset, leading to a total of 6 gene interaction networks (from 6 training datasets). In each gene interaction network, each interaction is then assigned a merged gene expression value (*MGE*) from the two interacting genes, as defined in Equation 7, i.e.,

$$MGE^{(a,b)} = \frac{2 \times GE^{(a)} \times GE^{(b)}}{GE^{(a)} + GE^{(b)}} \quad (7)$$

Therefore, by using Equation 7, each gene interaction is assigned a *MGE* value, which is the harmonized gene expression mean value of two interactors. The possible range of *MGE* for any interaction in the training datasets ranges between 0 and 1, i.e, $MGE \in [0, 1]$.

Every interaction in each gene interaction network, can be evaluated, where the value closer to 1 relates more chances of being a differentially expressed gene to identify the estrogen receptor status based breast cancers, and for 0, vice-versa. However, as mentioned above, these interactions are not reliable and contain many false-positive interactions that do not take place in real biological processes. Therefore in the gene interaction networks of each six training datasets, their *MGE* value is incorporated with the proposed reliability measure i.e., *WR*.

C. *Reliable Gene Expression Metrics*

To construct the gene network that signifies the reliability of each gene interactions with their associated gene expressions, we integrate the proposed reliability measure (*WR)* with the merged gene expression value (*MGE*) of gene interactions. As the *WR* measure assesses the reliability of each gene interaction on the basis of three vital criterions, *MGE* measure assesses the integrated gene expression of each gene interaction. Therefore, in order to get the significance of the gene expression of an interaction in terms of the reliability, we combine *WR* and *MGE,* and we called it as the Reliable Gene Expression (*RGE*).

However, before performing the *RGE* metrics, correlation between *WR* and *MGE* is evaluated. To get the influence, if both *WR* and *MGE* of any interaction (*a*, *b*) is positively correlated or not, we evaluate the correlation coefficient (δ), using Equation 8, i.e,

$$\delta^{(a,b)} = \frac{\left(WR^{(a,b)} - \overline{WR}\right)\left(MGE^{(a,b)} - \overline{MGE}\right)}{\sqrt{\left(WR^{(a,b)} - \overline{WR}\right)^2 \left(MGE^{(a,b)} - \overline{MGE}\right)^2}} \tag{8}$$

where $\overline{WR}$ and $\overline{MGE}$ represents the mean of *WR* and *MGE* in a training dataset, respectively.

We are interested in extracting for the positively correlated terms, as it is more strongly related to the patterns that can extract the gene signatures to accurately classify the subtypes of breast cancer classes. In other words higher the strength of the relationship between *WR* and *MGE*, more chances of a gene interaction to be related to the phenotype.

Once the positively correlated terms are extracted for each of the six training datasets, then the *RGE* can be evaluated. *RGE* of any interaction *y* between two genes (*a*, *b*) can be evaluated as:

$$RGE^{y} = \alpha\left(\sqrt{\sum_{i=1}^{3}\left(RWi^{(y)}\right)^{2} \times w_i}\right) + \beta\left(\frac{2 \times GE^{(a)} \times GE^{(b)}}{GE^{(a)} + GE^{(b)}}\right) \tag{9}$$

where α and β are the weights associated with *WR* and *MGE*, respectively. From our experiments, we define α and β to be 0.4 and 0.6, respectively. In nutshell, for any cancer classifier that uses gene interaction network information, both the measures (gene expression and reliability) are relatively essential for identifying the gene signatures that are highly associated to the phenotype of interest.

Using Equation 9, *RGE* of each positively correlated gene interactions in each training dataset can be evaluated. Once evaluated, the significant gene interactions ($p < 0.05$) is extracted for each training datasets, which is then used to construct the reliable gene signature to classify the estrogen receptor based samples, and is defined in the next section.

D. *Hub-Based Reliable Gene Expression Algorithm (HRGE)*

To detect the discriminative subnetworks for each training dataset, the significant positively correlated reliable gene expression of interactions was used. The significant reliable gene interactions for each training dataset (as evaluated from previous subsection) are taken for the hub- gene evaluation, where the hub-gene is the gene in the interaction network that contains maximal interactions amongst the other genes. For each training dataset, it may be possible that several subnetworks exist, and each subnetwork is used to detect the hub-gene. Fig. 1 illustrates this concept.

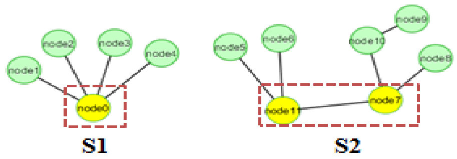


Fig. 1. Shows the two subnetworks for any training dataset *d*. The dashed square box shows the hub-gene, which has highest number of interactions among the other genes in each subnetwork. In Subnetwork S_2, two hub-genes are identified since they both have the maximal and equal number of interactions, i.e, each gene has 3 interactions.

For each of the six training datasets, once the subnetwork based hub-genes are identified, two major steps are performed. First, the subnetwork score (*SS*) is calculated for each subnetwork in a training dataset, by using Equation 10, i.e,

$$(SS)_{S_n} = \sum_{y=1}^{Y}\left(RGE^{(y)}\right) \tag{10}$$

where s_n defines the *n*th subnetwork of a training dataset, $RGE^{(y)}$ signifies the reliable gene expression (Equation 9) for any gene interaction *y* in subnetwork s_n, and *Y* defines the total number of gene interactions in s_n. Using Equation 10, the subnetwork score of each subnetwork in a given training dataset, can be evaluated. A subnetwork with the maximum subnetwork-score (*SS*) is chosen and retained for the further analysis. A maximum *SS* based subnetwork is chosen, because that subnetwork shows highly connected reliable gene interactions amongst the other subnetworks for a given training dataset, and believed to involve in the essential real biological processes. Using this step, only the subnetwork with maximum *SS* is chosen, and other subnetworks are ignored. However, the other subnetworks might contain essential genes that are highly associated in discriminating the two estrogen-receptor based breast cancer subtypes. Therefore, to identify those significant genes, step 2 is performed, which uses the hub-gene topology.

For each training dataset, a hub-gene with their associated gene interactions of each subnetworks are identified (as shown in Fig. 1). Then, the hub-gene score (*HS*) is evaluated as:

$$(HS)_{S_n} = \sum_{h=1}^{H} \frac{RGE^{(h)}}{H} \tag{11}$$

where *h* defines the gene interaction of any gene that interacts to the hub gene, and *H* defines the total number of such interactions in subnetwork s_n.

The associated *HS* of the chosen subnetwork that has maximum *SS* (from step 1) is then used as a threshold for extracting the relevant gene interactions. The chosen *HS* works as a threshold that can filter out the irrelevant gene interactions in regards to the real biological processes associated with breast cancer. This can be done by comparing the *HS* of the chosen subnetwork with all the other *HS* of the subnetworks, in a given training dataset. If any other subnetwork/s that has *HS* greater than the *HS* of the chosen subnetwork, then their hub-gene with their interactions is chosen, and ignored for other subnetworks. The reason to select the hub-gene as the benchmark is that a hub-gene has the maximum number of interactions in a given subnetwork and they have higher chances to be as a reliable gene signature, because hubs are highly connected nodes which have high probabilities to act as driver genes for cancers and may indulge in several essential biological functions and processes [21], [22]. Therefore, the hub-gene topology is used to extract the significant gene interactions from the subnetworks other than the chosen subnetwork (Step1) of a given training dataset, which has high probability to be as the reliable and stable gene marker for classifying ER+ and ER- subtypes.

Finally, our subnetwork based gene signature contains the subnetwork chosen from step 1, and the hub genes with their interactors chosen from step 2.

E. Classification of ER+ and ER- Subtypes Using HRGE Gene Signature with Expression-Mean Methodology

The six training dataset that contains three ER+ sample datasets and three ER- sample datasets, generates six subnetwork lists, from previous subsection. For each training dataset, a subnetwork list (that contains number of subnetworks) was used to construct the gene signatures to distinguish ER+ and ER- breast cancer samples, effectively.

Next, the three subnetwork lists of three ER+ training datasets were combined and removed the duplicates, similar for the three ER- training datasets. Therefore, after combining the subnetwork lists, and removing the duplicates in ER+ and ER- datasets respectively, the final gene signature set consists of 300 genes i.e., 159 distinct genes for ER+ subtypes and 141 distinct genes for ER- subtypes.

For any sample *s*, in the validation dataset, we considered Leave-one out Cross-Validation technique (LOOCV) that assumes a given labeled sample *s* to be unlabeled and all the other samples are labeled in a dataset. Next, we apply our ER+ and ER- gene signatures to *s*, and evaluate the corresponding expression-mean of the genes in ER+ gene signature and similarly for ER- gene signature. Once evaluated, the expression-means of ER+ and ER- gene signatures are compared and assigned an estrogen-receptor based class label to *s*, whose expression-mean dominates the other. Therefore, all the samples in a validation dataset can be classified by applying our *HRGE* gene signature, with expression- mean methodology.

IV. Results

In the validation dataset, we first performed the experiments to compare the results of other classifiers with *HRGE*. Then we performed the heat-map analysis.

We generated two sets of subnetwork based gene signatures for ER+ and ER- breast cancer subtypes, i.e., 159 genes for ER+ and 141 genes for ER-, respectively. To test the classification accuracy of the gene signatures, we applied them on the Desmedt dataset [23] (validation dataset). We then compared the *HRGE* gene signature with four other previously existing gene signatures i.e., the 70-gene signature [5], the 76-gene signature [6], the Genomic Grade Index (GGI) [12], and the Interactome-Transcriptome Integration (ITI) [4]. The evaluation of our approach and the experimental comparisons with other existing approaches (as mentioned above) signifies that *HRGE* significantly increased the classification performance, as compared with others. The result shows that our *HRGE* approach achieves the classification accuracy of 89% and 59% for ER+ and ER- samples respectively, in the validation dataset of Desmedt [23]. Table III shows the detailed classification result of our gene signature and other existing gene signatures.

From Table III, *HRGE* gave better classification accuracy (in ER+ and ER-) as compared to, ITI that gives 74% and 54% (in ER+ and ER- samples, respectively), GGI that gives 65% and 48% (in ER+ and ER- samples, respectively), 76g that gives 61% and 38% (in ER+ and ER- samples, respectively), and 70g that gives 41% and 44% (in ER+ and ER- samples, respectively). We believe that the classification accuracy can be further strengthened, by increasing the training compendia with increased multiple platforms across multiple datasets.

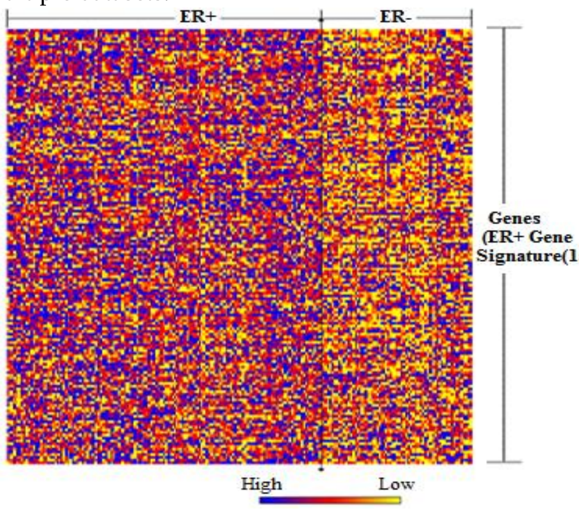


Fig. 2. The Heatmap of the *HRGE* gene signature by using the expression values of the genes from the Desmedt dataset, and is drawn by using R statistical package [24], which is freely available to download. Here, each row corresponds to the genes and each column corresponds to the samples arranged from ER+ to ER-. The Expression levels of each gene are normalized across the samples with zero mean and standard deviation equals one, where blue and yellow represents high and low expression levels than the mean, respectively. The heatmap shows the genes highly expressed in ER+ samples i.e., by applying our ER+ gene signature (159 genes) on the Desmedt dataset visualizes two groups, one is highly expressed (ER+) and other, vice-versa. Similar visualization pattern can be observed from ER- gene signature (data not shown).

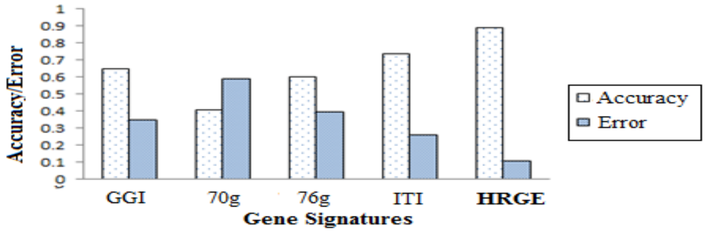
In order to illustrate the behavioral pattern of *HRGE* gene signature on the Desmedt dataset, heat maps are drawn that shows the differential expression of genes in ER+ samples (and ER- samples). Heat maps are the rectangular grids with colours that represent the corresponding expression value of the genes, with high expression value than the mean, represent blue and low represent yellow. The rows of Heatmap correspond to the *HRGE* gene signature (ER+/ER- gene signature) and the columns correspond to the samples in the Desmedt data (arranged from ER+ to ER-). Fig. 2 shows the heatmap of ER+ gene signature (similar patterns can be observed for ER- gene signature, data not shown).

From Fig. 2, although the genes in a gene signature group seem correlated with ER+ subtypes, no single gene can be seen that shows uniformity of expressions across the samples. This illustrates the significance of the gene signatures as a multigene classification method. From Fig. 2, it is true to say that the *HRGE* gene signature is significant in distinguishing ER+ and ER- subtypes.

TABLE III: CLASSIFICATION RESULTS OF *HRGE* AND OTHER EXISTING GENE SIGNATURES

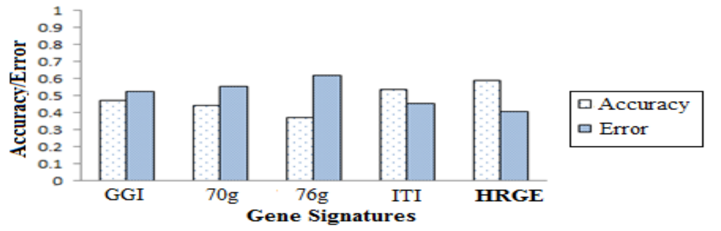
(A) ER+ Samples

Gene Signatures	GGI	70g	76g	ITI	**HRGE**
Number:	129	129	129	129	**129**
Positives:	84	53	78	95	**115**
Negatives:	45	76	51	34	**14**
Accuracy:	0.651	0.411	0.605	0.736	**0.891**
Error:	0.349	0.589	0.395	0.264	**0.108**



(B) ER- Samples

Gene Signatures	GGI	70g	76g	ITI	**HRGE**
Number:	61	61	61	61	**61**
Positives:	29	27	23	33	**36**
Negatives:	32	34	38	28	**25**
Accuracy:	0.475	0.443	0.377	0.541	**0.590**
Error:	0.525	0.557	0.623	0.459	**0.410**



The left-part of the figure shows the table that represents the statistics of classification results and the right-part of the figure shows the bar-graph of the classification accuracy and error of *HRGE* gene signature, and four other existing gene signatures. Here, Number defines the total number of ER+/ER- Samples in Desmedt dataset, Positives defines the total number of samples accurately classified, and Negatives defines the total number of samples classified inaccurately. The subnetwork based *HRGE* gene signature give superior performance for both, (A) ER+ and (B) ER- samples in Desmedt dataset. For simplicity, we represents Genomic Grade Index as GGI, 70 gene signature as 70g, 76 gene signature as 76g, and Interactome-Transcriptome Integration as ITI.

V. CONCLUSIONS

As the genes perform its function by interacting with other genes, therefore subnetwork based approach is highly relevant to extract the specific genes whose processes or functions seems to be disrupted in cancers. In this paper, we therefore proposed a subnetwork-based *HRGE* algorithm that effectively extracts the significant biologically-relevant interactions, and is based on the hub-gene topology that construct the two estrogen receptor based gene signatures to distinguish ER+ and ER- breast cancer subtypes, as defined in Section III.

To make the *HRGE* algorithm independent towards the particular dataset, we incorporated multiple datasets with distinct platforms, in order to improve the classification performance in terms of accuracy. Six training datasets is constructed on the basis of estrogen-receptor status and histologic grade, and by applying our algorithm, six subnetwork lists can be generated which gives the robust and effective gene signature of 300 genes that constitutes of 159 genes for ER+ subtypes and 141 genes for ER- subtypes. The classification results of *HRGE* gene signature shows its superiority amongst other previously published gene signatures, as shown in Table III. This illustrates the effectiveness of *HRGE* gene signature.

REFERENCES

[1] R. J. Cho, M. J. Campbell, E. A. Winzeler *et al.*, “A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle,” *Molecular Cell*, vol. 2, no. 1, pp. 65-73, 1998.

[2] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chiba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki, “Toward a protein–protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins,” in *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 3, 2000, pp. 1143-1147.

[3] P. Uetz, L. Giot, G. Cagney *et al.*, “A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae,” *Nature*, vol. 403, no. 6770, pp. 623-627.

[4] M. Garcia, R. Millat-Carus, F. Bertucci, P. Finetti, D. Birnbaum, and G. Bidaut, “Interactome–transcriptome integration for predicting distant metastasis in breast cancer,” *Bioinformatics*, vol. 28, no. 5, pp. 672-678, 2012.

[5] M. J. van de Vijver, Y. D. He, L. J. van 't Veer *et al.*, “A Gene-Expression Signature as a Predictor of Survival in Breast Cancer,” *The New England Journal of Medicine*, vol. 347, no. 25, pp. 1999-2009, 2002

[6] Y. Wang, J. G. M. Klijn, Y. Zhang *et al.*, “Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer,” *The Lancet*, vol. 365, no. 9460, pp. 671-679, 2005.

[7] S. Michiels, S. Koscielny, and C. Hill, “Prediction of cancer outcome with microarrays: a multiple random validation strategy,” *The Lancet*, vol. 365, no. 9458, pp. 488-492, 2005.

[8] H-Y. Chuang, E. Lee, Y-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology*, vol. 3, no. 140, 2007.

[9] R. Gill and S. Datta "A statistical framework for differential network analysis from microarray data," *BMC Bioinformatics*, vol. 11, no. 1, pp. 95, 2010.

[10] Y. Zhu, X. Shen, and W. Pan, "Network-based support vector machine for classification of microarray samples," *BMC Bioinformatics*, vol. 10, no. suppl 1, pp. 1-11, 2009.

[11] C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg, "Protein interactions: two methods for assessment of the reliability of high throughput observations," *Molecular and Cellular Proteomics*, vol. 1, no. 5, pp. 349-356, 2002.

[12] C. Sotiriou, P. Wirapati, S. Loi *et al.*, "Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis," *Journal of the National Cancer Institute*, vol. 98, no. 4, pp. 262-272, 2006.

[13] T. Barrett, D. B. Troup, S. E. Wilhite *et al.*, "NCBI GEO: archive for high-throughput functional genomic data," *Nucleic Acids Research*, vol. 37, no. suppl. 1, pp. D885-D890, 2009.

[14] C. Stark, B.-J. Breitkreutz, A. Chatr-aryamontri, L. Boucher *et al.*, "The BioGRID Interaction Database: 2011 update," *Nucleic Acids Research*, vol. 39, no. suppl. 1, pp. D698-D704, 2011.

[15] S. Kerrien, B. Aranda, L. Breuza *et al.*, "The IntAct molecular interaction database in 2012," *Nucleic Acids Research*, vol. 40, no. D1, pp. D841-D846, 2012.

[16] L. Licata, L. Briganti, D. Peluso *et al.*, "MINT, the molecular interaction database: 2012 update," *Nucleic Acids Research*, vol. 40, no. D1, pp. D857-D861, 2012.

[17] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, "DIP: the Database of Interacting Proteins," *Nucleic Acids Research*, vol. 28, no. 1, pp. 289-291, 2000.

[18] G. D. Bader, D. Betel, and C. W. V. Hogue, "BIND: the Biomolecular Interaction Network Database," *Nucleic Acids Research*, vol. 31, no. 1, pp. 248-250, 2003.

[19] T. S. K. Prasad, R. Goel, K. Kandasamy *et al.*, "Human Protein Reference Database–2009 update," *Nucleic Acids Research*, vol. 37, no. suppl 1, pp. D767–D772, 2009.

[20] The Universal Protein Resource (UniProt)., *Nucleic Acids Research*, vol. 35, no. suppl 1, pp. D193-D197, 2007.

[21] X. He and J. Zhang, "Why Do Hubs Tend to Be Essential in Protein Networks?," *PLoS Genetics*, vol. 2, no. 6, pp. e88, 2006.

[22] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41-42, 2001.

[23] C. Desmedt, B. Haibe-Kains, P. Wirapati, M. Buyse, D. Larsimont, G. Bontempi, M. Delorenzi, M. Piccart, and C. Sotiriou, "Biological Processes Associated with Breast Cancer Clinical Outcome Depend on the Molecular Subtypes," *Clinical Cancer Research*, vol. 14, no. 16, pp. 5158-5165, 2008.

[24] R Development Core Team., "R: A Language and Environment for Statistical Computing," *Vienna, Austria: R Foundation for Statistical Computing*, 2008.

Ashish Saini did the Bachelor (Honors) degree in the area of Bioinformatics from School of Information Technology, Deakin University, Australia, in 2011. He is currently pursuing a Ph.D. degree at the School of Information Technology, Deakin University, Australia. His research focusses in the area of bioinformatics and pattern recognition.