20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Gene expression based Computer Aided diagnostic system for Breast Cancer: A novel biological filter for biomarker detection

A. Al-yousef[1], S. Samarasinghe[2], and D. Kulasiri[2]

[1]Jerash University, Jerash, Jordan
2 *Centre for Advanced Computational Solutions (C-fACS), Lincoln University, Canterbury, New Zealand*
Email:Sandhya.samarasinghe@lincoln.ac.nz

Abstract: Cancer is a complex disease because it makes complex cellular changes. Therefore, microarrays have become a powerful way to analyse cancer and identify what changes are produced within a cell. Through DNA microarrays, it has become possible to look at the expression of thousands of genes in one sample and this is called gene expression profiling. Gene expression profiling is important to capture a set of expressed genes that determines a cell phenotype.

However, analysing microarray data is challenged by the high-dimensionality of the data compared with the number of samples. The aim of this study was to enhance the diagnostic accuracy of Breast Cancer Computer Aided Diagnostic Systems (CADs) that use gene expression profiling of peripheral blood cells, by introducing a novel feature selection method called Bi-biological filter that was further refined by Best First Search with Support Vector Machines SVM (BFS-SVM) to select a small set of the most effective genes predictive of breast cancer. From each patient's gene expression profiles, a gene co-expression network was built and divided into functional groups or clusters using Topological Overlap Matrix (TOM) and Spectral Clustering (SC) in the design of the Bi-Biological filter to obtain the preliminary set of gene markers. BFS-SVM was used to further filter a smaller set of best gene markers, and Artificial Neural Networks (ANN), SVM and Linear Discriminant Analysis (LDA) were used to assess their classification performance. The study used 121 samples – 67 malignant and 54 benign cases as input to for the system. The Bi-biological filter selected 415 genes as mRNA biomarkers and BFS-SVM was able to select just 13 out of 415 genes for classification of breast cancer. ANN was found to be the superior classifier with 93.4% classification accuracy which was a 14% improvement over the past best CAD system developed by Aaroe et al. (2010).

Keywords: *gene expression, mRNA, blood, feature selection, neural networks, breast cancer, Computer Aided Diagnosis, early detection*

1. INTRODUCTION

Gene expression profiling of peripheral blood cells has been used for early detection of breast cancer. However, analysing microarray data is challenged by the high-dimensionality of the data compared with the number of samples. Most of the previous studies in this field have used either filters or wrappers to select a subset of genes that differentiate cancer from control cases (Aaroe, et al., 2010; Fan et al., 2010; Kretschmer et al., 2011; Ma, Kosorok, Huang, & Dai, 2011; Schrauder et al., 2012). However, both of the above methods have disadvantages; for example, the wrapper methods are challenged by the high dimensionality of the data whereas the filters ignore the relation between the features. For example, mRNA that is extracted from peripheral blood cells has been used by a Norwegian group for analysing the expression of 1,368 genes extracted from peripheral blood cells of 56 women: 24 breast cancer and 32 healthy, for early detection of breast cancer. The study used wrapper method for feature selection and correctly predicted 82% of the samples using 37 probes (29 genes) (Sharma et al., 2005). To confirm the results of this study, a larger sample of 11,217 genes extracted from 130 women was analysed (Aaroe, et al., 2010). Aaroe, et al. (2010) used Partial Least Square Regression (PLSR) and Jackknife testing (Wu., 1986) with dual leave one out cross validation for feature selection and classification. The PLSR with leave one out cross validation was used to reduce the dimensionality of the data by selecting the optimum number of latent variables. This is a set of variables that are selected by analysing the covariance between the gene expression vectors and the class label. The regression also returns the regression coefficient for each gene. The Jackknife testing method is used to select the variables that have regression coefficient different from 0 with p-value <0.05. The study obtained a set of 738 probes that differentiated healthy from cancer samples with 79.5% prediction accuracy, 80.6% sensitivity, and 78.3% specificity. The study compared the 738 genes with the 29 genes obtained in the previous study (Sharma, et al., 2005) and found that 20 out of 29 genes were not significant in relation to the disease status in the present study (Aaroe, et al., 2010). The hypothesis that a larger sample could increase prediction accuracy was not shown in this case. It used filtering methods for feature selection that ignored the biological relation between genes and selected the genes based on the ability of individual genes in differentiating cancer form control cases. In addition, the method selected a large number of genes compared with the number of samples as inputs for the classifier which can negatively affect the outcomes. Furthermore, the accuracy of current CAD systems based on mRNA is about 79% and needs further enhancement in order to save the lives of the undetected.

Gene co-expression provides key information to understand living systems, where the co-expressed genes are often involved in the same biological pathway (Yip & Horvath, 2007). Therefore, similarity between genes may reflect the biological relation between them, where the genes with high similarity may have similar biological function or may be a part of the same (Yip & Horvath, 2007; Zhang. & Horvath, 2005). There are several ways to compute the similarity between genes. The most common for gene expression data is Pearson Correlation Coefficient (PCC) and it is a measure that reflects the linear relationship between a pair of genes. The PCC takes values in the range from +1 to -1. A correlation of +1 means that there is a perfect positive linear relationship between variables and -1 means that there is a perfect negative linear relationship between variables. However, PCC considers each pair of variables in isolation to other variables. From a biological perspective, the relationships with other variables, genes or proteins, should be taken into account. This is because if a pair of genes shares relations with other genes, they all may be similar and may belong to the same pathway or may have similar biological functions (Yip & Horvath, 2007; Zhang. & Horvath, 2005).

A meaningful approach to increase the diagnostic accuracy of mRNA CAD systems is to incorporate biological relations between genes that would require novel gene selection methods that consider biologically significant relations between genes. Then, incorporate the selected genes into a CAD system. In this paper, we aim to introduce a novel feature selection method called Bi-biological filter enhanced through Best First Search with Support Vector Machines (BFS-SVM).

2. MATERIALS AND METHODS

In this study we used 121 samples (67 malignant and 54 benign) that were collected by (Ullevål University Hospital and Haukeland University Hospital in Norway) between 2002-2004. The malignant group contains: 10 samples with ductal carcinoma *in situ* and 57 invasive carcinoma samples spread across a number of severity Grades from I to III. Invasive carcinoma samples include 49 Ductal, 4 Lobular and 4 other invasive types. The dataset also contains 12 benign cases and 42 normal cases with neither benign nor malignant findings. Each sample contains 11, 217 genes (7351 known genes and 3866 unknown genes). The known genes are genes that have gene symbols and gene IDs where the unknown genes are those with no symbols and IDs at this stage. In this study, we only used the 7351 known genes. The samples are publicly available in the NCBI's Gene Expression Omnibus through GEO: GSE16443 accession number. The data is already pre-processed, where the effect of the background has been removed, normalised and summarised.

Learning tasks, such as classification and clustering are challenged by high dimensional data. Such data may have a lot of noisy features which make learning task very complex. The process of removing noisy data (irrelevant and redundant) or choosing a subset of features (relevant) from a given set of features is called feature selection (Blum and Langley, 1997; Gilad-Bachrach et al., 2004). In this research we used a new method of feature selection called Bi-Biological Filter with further enhancements through Best First Search with SVM wrapper (BiBio-BFSS). This method contains two main steps: bi-biological filter and Best First Search with SVM wrapper. The bi-biological filter also contains two steps; neither cancer nor healthy biomarker elimination followed by healthy biomarker removal. The output genes from Bi-Biological filter are used as inputs to BFS-SVM wrapper to select a smaller set of genes for classification.

2.1 Bi-biological filter

This step is responsible for selecting a group of genes that are strongly related to breast cancer by using gene co-expression networks as described by Zhang and Horvath (2005). The first step is designed for removing the biomarkers or groups of genes that are not shared between the cancer cases (neither cancer nor healthy groups). This type of groups may result from noise in the dataset or other disease biomarkers shared between cancer cases in the study. The second step is to filter out the healthy biomarkers from the selected biomarkers found in the first step.

2.1.1 Neither cancer nor healthy biomarker filtering

This is to remove the genes that are not shared between all cancer cases. To do this, we randomly divided the malignant dataset into two sub-datasets, 33 samples assigned to M1 and 34 samples to M2. Then, the genes of each subset were divided into functional groups as follows:

1- Build Co-expression Network described by Zhang and Horvath (2005) as follows:

a. Find the correlation matrix (similarity matrix) using Pearson Correlation Coefficient (PCC) between gene i and gene j (Equation 1) for all pairs of genes in the set.

$$\mathrm{PCC}_{ij}=\frac{n\ \sum_{u=1}^{n} x_{ju}x_{iu}-\sum_{u=1}^{n} x_{iu}\ \sum_{u=1}^{n} x_{ju}}{\sqrt{n\ \sum_{u=1}^{n} x_{iu}^{2}-(\sum_{u=1}^{n} x_{iu})^{2}}\sqrt{n\ \sum_{u=1}^{n} x_{ju}^{2}-(\sum_{u=1}^{n} x_{ju})^{2}}} \qquad (1)$$

where n is the number of samples, x_{ju} represents the expression value of gene j in sample u.

To highlight strong relations, dampen weak correlations and to convert the network to scale-free topology we powered the absolute value of PCC to β where β is an integer number (Equation 2).

$$w_{ij}=\left|PCC_{ij}\right|^{\beta} \qquad (2)$$

2- Obtain the Topology overlap between all pairs of genes:

a. Find the degree or the connectivity value c_i of gene *i* (Equation 3)

$$c_i=\sum_{\substack{j=1\\ i\neq j}}^{n} w_{ij} \qquad (3)$$

where w_{ij} is the similarity value between gene i and j from step 1.

b. For each pair of genes we find $I_{ij,}$ where I_{ij} is a number between 0 and n as in Eq. 4 where w_{ui} is the similarity value; a high value of I_{ij} means that there is a high number of shared genes between i and j and low value means that there is no or few shared genes between i and j.

$$I_{ij}=\sum_{\substack{u=1\\ u\neq j,i}}^{n} w_{uj}\, w_{ui} \qquad (4)$$

c. Now, we use the topological overlap similarity between each gene pair (Equation 6)

$$t_{ij}=\frac{I_{ij}+w_{ij}}{\min(c_i,c_j)+1-w_{ij}} \qquad (6)$$

The matrix containing all t_{ij} is called Topological Overlap Matrix (TOM) and it is plotted in a weighted undirected graph where each gene is represented as a vertex and the edge between the pair of genes is weighted by the topological overlap similarity t_{ij} value.

Module extraction

Clustering plays an important role in data analysis. In biology, especially with high dimensional data, clustering has been used to reduce the dimensionality of data by grouping the similar dimensions. The module is a group of genes working together to do a specific biological function. From the above definition, the genes that are strongly correlated to each other may belong to one pathway or do similar functions. To find the clusters of genes, this work uses Spectral Clustering. To apply spectral clustering on TOM graph we follow the following steps as described by Ng et al. (2001):

1. Find the degree of each vertex on the graph (Equation 7):

$$d_i = \sum_{j=1}^{n} t_{ij} \tag{7}$$

2. Next find the graph's Normalised Laplacian matrix (L) (Equation 8):

$$l_{ij} = \begin{cases} 1 & i = j \\ \dfrac{t_{ij}}{\sqrt{d_i d_j}} & i \neq j \end{cases} \tag{8}$$

The normalised spectral clustering is known to outperform the non-normalised version in high dimensional data (Ng. et al., 2001). So we use normalised spectral clustering.
Now we perform spectral clustering as follows:

a- Find all eigenvalues and eigenvectors for the normalised Laplacian matrix.

b- Next step is to determine the number of sub-graphs or clusters in the gene co-expression graph. This is done by selecting K representative eigenvectors from the n eigenvectors of the graph, where K is much smaller than n, in order to reduce the dimensionality of the data and to select the number of clusters.

c- Let V ϵ R^{nxk} be the matrix containing the selected eigenvectors v_1, v_2, v_3, …, v_k as columns. Form the matrix U by normalizing the row sums to 1 (Equation 9). Now, genes of the original adjacency matrix is represented by lower dimensional matrix U, where row i in U represents gene i in original matrix:

$$\mathrm{u_{ij}} = \frac{\mathrm{v_{ij}}}{\left(\sum_{\mathrm{m=1}}^{\mathrm{k}} \mathrm{v_{im}^2}\right)^{\frac{1}{2}}} \tag{9}$$

d- Gene clustering: for i=1, …, n , let y_i ϵ R^k be the vector corresponding to i^th row of U. Cluster the points y_i with k-means algorithm into clusters C_1,…,C_k.

Selection of shared clusters

By the end of the above step, the genes in M1 and M2 datasets have been divided into functional groups. In the current step, we select the shared groups of genes between M1 and M2 (Figure 1).

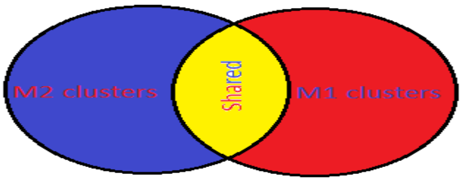


Figure 1. The shared clusters between M1 and M2 cancer subsets. Red colour represents the clusters that were in M1 only, blue represents the clusters that were in M2 only and yellow represents the shared clusters between M1 and M2

A direct comparison between all genes in the groups is difficult; therefore, we select the hub genes for each cluster and the comparison takes place on hub gene level instead of all genes. To select the hub genes of a cluster, we find the within cluster connectivity for each gene (Equation 3) and select the m genes with the highest connectivity, a relatively small number, as hub genes. Selecting more than one hub gene allows partial similarity between a pair of clusters where clusters are considered shared if one or more hub genes are shared between two clusters.

By the end of this step, neither cancer nor healthy clusters are removed and we keep only the shared groups that represent 'cancer and healthy biomarkers'.

2.1.2 Removal of healthy biomarkers

This step is for filtering out the healthy biomarkers from the selected biomarkers in the shared gene set and to keep only the probable cancer biomarkers by applying the following steps:

1. Build the co-expression gene network for the healthy dataset as described previously using the method of Zhang and Horvath (2005) and apply the spectral clustering to extract the healthy biomarkers as described above.
2. Extract the hub genes for the 'healthy' clusters.
3. Find the shared clusters between the 'healthy' clusters found here and the clusters containing 'cancer and healthy biomarkers' found previously.
4. Delete the shared clusters found in step 3 from the 'cancer and healthy biomarkers' set. This provides a set of clusters that are potential cancer biomarkers.

2.2 Best First Search and SVM with 5-fold cross validation wrapper

In this step we use the forward BFS and the accuracy of SVM with K-fold cross validation to find the subset of genes that strongly relate to breast cancer from the output genes of the bi-biological filter.

2.3. Classification and evaluation

This paper applies three supervised classifiers: Multilayer Feed Forward Neural Network (MFFNN) optimised/pruned by clustering correlated weighted hidden neuron outputs developed by Samarasinghe (2010), Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA). The results of each classifier are evaluated using False Positive rate (FP), False Negative rate (FN), Sensitivity, Specificity and Accuracy measures.

3. RESULTS AND DISCUSSION

3.1. Feature selection

As mentioned earlier, this step is divided into two main steps; Bi-biological filter and Best First Search supported SVM with 5-fold cross validation wrapper method.

Table 1. The shared clusters between cancer1 and cancer2 datasets. For simplicity, each shared pair was given a new Id.

Shared Clusters	Cancer1 Cluster Id	Cancer2 Cluster Id	Shared hub Gene Id	Gene Symbol
1	11	1	22827	SIAHBP
2	13	2	6139	RPL17
3	6	4	8786	RGS11
4	8	6	150372	NFAM1
5	15	15	79143	LENG4

3.1.1 Neither cancer nor healthy biomarker filter:

The cancer dataset is divided randomly into two subsets; cancer1 (M1) and cancer2 (M2). The cancer1 subset contains 33 samples and cancer2 contains 34 samples; both subsets contain 7351 mRNA of genes. By applying spectral clustering on the TOM of cancer1 and TOM of cancer2 we found that cancer1 dataset was divided into 17 clusters and cancer2 into 16 clusters. Then for each cluster we selected two hub genes and found five clusters shared between cancer1 and cancer2 datasets (Table 1). This removes all biomarkers which were not shared between cancer1 and cancer2 and only the shared ones are carried to the next step for filtering out healthy ones.

3.1.2 Removal of healthy biomarkers:

In this step, we analysed the healthy dataset to find all healthy biomarkers. The analysis revealed that the genes in the healthy dataset were spread over 15 different clusters. Then, for each healthy cluster we also extracted two representative hub genes. By comparing the healthy hub genes and the hub genes that were shared between cancer1 and cancer2 datasets (Table 1), we found that 2 out of 5 clusters (2 and 5) in Table 1 were found in the healthy dataset and considered to be healthy biomarkers. Thus, the remaining three clusters (1, 3 and 4) in Table 1 are considered to be the breast cancer biomarkers. Because we allowed partial similarity, the shared genes between pairs of clusters (cancer1-cancer2) were selected and considered to be breast cancer biomarkers. The total number of selected genes was 415.

Now, for genes in each cluster, we investigated the biological processes related to the genes of that cluster using The Database for Annotation, Visualization and Integrated Discovery (DAVID) (Dennis et al., 2003) and selected the processes that had False Discovery Rate (FDR) less than 20% and (p-value<0.05).

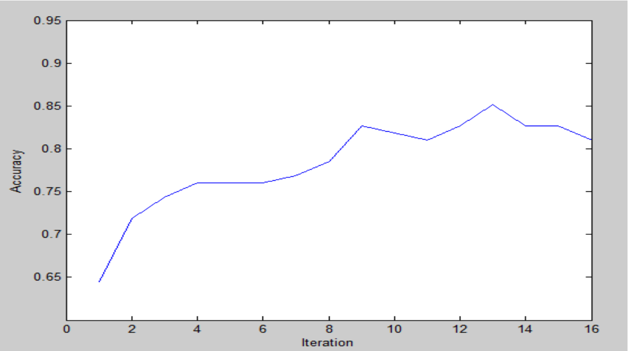
For the first cluster (1) we found that apoptosis (GO:0006915) and regulation of apoptosis (GO:0042981) were the most important processes that differentiated normal and cancer cells. Several previous studies found apoptosis in relation to breast cancer where the apoptosis process in cancer patients was weaker compared to healthy ones (Feng et al., 1995; Graham & Clarke, 1997; Haldar, Negrini, Monne, Sabbioni, & Croce, 1994; Parton, Dowsett, & Smith, 2001). We found that the second cluster (3) relates to the cellular respiration process (GO:0045333), which is the enzymatic release of energy from organic compounds that either requires oxygen (aerobic respiration) or does not (anaerobic respiration) (European Molecular Biology Laboratory, 2011) (European Molecular Biology Laboratory, 2011), and several studies have found its relation to breast cancer (Simonnet et al., 2002; Warburg, 1956). Another process related to the genes of cluster (3) was epithelial cell differentiation process (GO:0030855) where a relatively unspecialised cell acquires specialised features of an epithelial cell (European Molecular Biology Laboratory, 2011), and it was found in relation to breast cancer in Beitsch and Clifford (2000) study. In the last cluster (4), we found processes in relation to producing energy necessary for cellular processes. Glucose is an important source of energy in the body and it is considered as fuel for a cell. Without energy, cells cannot perform their natural processes. Cancer cells are characterised by uncontrolled and rapid division so there is a need for providing cancer cells with a larger amount of glucose and speed up the process of producing energy from it (Annibaldi & Widmann, 2010).

By the end of this investigation we found that genes in the same cluster work together to carry out the same biological processes which provides support for the biological relation between the genes. Furthermore, we found that the relationship between the biological processes of the groups and breast cancer are strong which also provides support for the selected group of biomarkers.

3.1.3 Best First Search and SVM with 5-fold cross validation wrapper

The Bi-Biological filter selected 415 genes that were spread over 3 clusters; these clusters are potential breast cancer biomarkers. But we still needed to reduce the dimensionality of the data and select a subset of genes from the 415 genes for classification. To do this, we applied the BFS and SVM with 5-fold cross validation wrapper for gene selection as described previously. After 16 iterations, the algorithm was stopped because there were no improvements in iterations 14, 15 and 16 (number of fails m=3). The highest accuracy that was obtained by the wrapper was at iteration number 13 with 85.1% classification accuracy; 88.05% sensitivity and 81.4% specificity (Figure 2) using 13 genes (DAPK3, CTDSP1, CXX1, RCOR3, MYL4, YWH, ABGMEB2, GPR78, ILK, HSPC171, ACAT2, PRKRIP1, PP3856). Also, we found DAPK3 (Death-Associated Protein Kinase 3) obtained the highest individual classification accuracy (iteration1).

Figure 2. The output accuracy from the Best First Search and SVM with 5-fold cross validation wrapper for 16 iterations. The X-axis represents the iteration number and the Y-axis represents the corresponding accuracy value.



3.2 Classifications

The thirteen selected genes were used as input for the three classifiers, Multilayer Feed Forward Neural Network (MFFNN) optimised with hidden network pruning by a novel method with 5-fold cross validation, SVM with Leave One Out Cross Validation (LOOCV) and Linear Discriminant Analysis (LDA) with LOOCV. Then the output results of each classifier were evaluated using the sensitivity, specificity, FP, FN and accuracy measures (Table 2). From the outputs of different classifiers we found the MFFNN was the superior classifier with 94.02% sensitivity, 92.6% specificity and 93.4% accuracy.

Table 2. The performance of different classifiers using the 13 selected genes (SN- sensitivity, SP- specificity and Ac-accuracy).

Classifier	SN	SP	FN rate	FP rate	Ac
SVM	86.6%	81.5%	16.98%	14.7%	84.3%
LDA	82.1%	79.6%	21.8%	16.7%	80.9%
MFFNN	94.02%	92.6%	7.4%	5.97%	93.4%

Now, reviewing the false positive rate and FN cases of the best classifier, MFFNN, we found 2 out of 4 FP cases had a cyst or benign tumour which means that the presence of benign findings in the breast may reduce the specificity of our system. The grades of FN cases were: one case of grade one, two cases of grade two and one case not classified. From these FN cases we found that our system correctly classified 92.8% of grade one cases and 90.9% of grade two cases and the local sensitivity values of both grades are close to the overall sensitivity of the system. This means that our system successfully predicts breast cancer in early stages.

From the literature review we found that there was only one blood based mRNA CAD for early detection of breast cancer (Aaroe, et al., 2010). This CAD system extracted 738 mRNA probes as breast cancer biomarkers using a filtering method. These biomarkers were used to classify the samples into healthy and cancer cases. The accuracy, sensitivity and specificity value for Aaroe et al. (2010) CAD system were 79.5%, 80.6%, 78.3%, respectively. By comparing our CAD system results with their system we found that the accuracies of all classifiers in our CAD systems are better. Specifically, the diagnostic accuracy of LDA classifier was enhanced by 2% and the SVM improved the diagnostic accuracy by 5.5%. The significant improvement of our BC-CAD over Aaroe et al. (2010) system was in the accuracy, sensitivity and specificity values of MFFNN where our system obtained 93.4%, 94.02% and 92.6%, respectively, which means that about 14% of cancer cases misdiagnosed in Aaroe et al. (2010) system are diagnosed correctly by our system and hence more lives could be saved. Furthermore, using 738 probes for classification reduces the performance of classifiers due to high dimensionality of data compared with 13 genes in our study, 56 times less than in their study.

4. CONCLUSION

In this study we introduced a new method for breast cancer biomarker selection and we called this Bi-Biological filter and Best first Search (BFS) supported SVM with K-fold cross validation and we successfully identified 13 genes as breast cancer biomarker in the blood. Also, we evaluated the diagnostic accuracy of three classifiers MLFFNN, LDA and SVM. The best results were obtained using MFFNN with 93.4% classification. Furthermore, about 14% of cancer cases misdiagnosed in the previous CAD system (Aaroe, et al., 2010) were diagnosed correctly in our system and hence more lives could be saved. In future, a larger dataset will be used for validation of the performance of our method.

REFERENCES

Aaroe, J., Lindahl, T., Dumeaux, V., Saebo, S., Tobin, D., Hagen, N., et al. (2010). Gene expression profiling of peripheral blood cells for early detection of breast cancer. *Breast Cancer Research, 12*(1), R7.

Annibaldi, A., & Widmann, C. (2010). Glucose metabolism in cancer cells. *Current Opinion in Clinical Nutrition & Metabolic Care, 13*(4), 466-470 410.1097/MCO.1090b1013e32833a35577.

Beitsch, P. D., & Clifford, E. (2000). Detection of carcinoma cells in the blood of breast cancer patients. *American journal of surgery, 180*(6), 446-449.

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence, 97*(1-2), 245-271.

Dennis, G., Sherman, B., Hosack, D., Yang, J., Gao, W., Lane, H. C., et al. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biology, 4*(5), P3.

European Molecular Biology Laboratory. (2011). EMBL-EBI. from European Bioinformatics Institute: http://www.ebi.ac.uk/

Fan, X., Shi, L., Fang, H., Cheng, Y., Perkins, R., & Tong, W. (2010). DNA Microarrays Are Predictive of Cancer Prognosis: A Re-evaluation. *Clinical Cancer Research, 16*(2), 629-636.

Feng, Z., Marti, A., Jehn, B., Altermatt, H. J., Chicaiza, G., & Jaggi, R. (1995). Glucocorticoid and progesterone inhibit involution and programmed cell death in the mouse mammary gland. *The Journal of Cell Biology, 131*(4), 1095-1103.

Gilad-Bachrach, R., Navot, A., & Tishby, N. (2004). *Margin based feature selection - theory and algorithms*. Paper presented at the Proceedings of the twenty-first international conference on Machine learning.

Graham, J. D., & Clarke, C. L. (1997). Physiological Action of Progesterone in Target Tissues. *Endocrine Reviews, 18*(4), 502-519.

Haldar, S., Negrini, M., Monne, M., Sabbioni, S., & Croce, C. M. (1994). Down-Regulation of bcl-2 by p53 in Breast Cancer Cells. *Cancer Research, 54*(8), 2095-2097.

Kretschmer, C., Sterner-Kock, A., Siedentopf, F., Schoenegg, W., Schlag, P., & Kemmner, W. (2011). Identification of early molecular markers for breast cancer. *Molecular Cancer, 10*(1), 15.

Ma, S., Kosorok, M. R., Huang, J., & Dai, Y. (2011). Incorporating higher-order representative features improves prediction in network-based cancer prognosis analysis. *BMC medical genomics, 4*, 5.

Ng., A. Y., Jordan., M. I., & Weiss, Y. (2001). On Spectral Clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems, 14*.

Parton, M., Dowsett, M., & Smith, I. (2001). Studies of apoptosis in breast cancer. *BMJ, 322*(7301), 1528-1532.

Samarasinghe, S. (2010). *neural networks for water system analysis:from fundamentals to complex pattern recognition*. Paper presented at the Hydrocomplexity:New Tools for Solving Wicked Water Problems.

Schrauder, M. G., Strick, R., Schulz-Wendtland, R., Strissel, P. L., Kahmann, L., Loehberg, C. R., et al. (2012). Circulating Micro-RNAs as Potential Blood-Based Markers for Early Stage Breast Cancer Detection. *PLoS ONE, 7*(1), e29770.

Sharma, P., Sahni, N., Tibshirani, R., Skaane, P., Urdal, P., Berghagen, H., et al. (2005). Early detection of breast cancer based on gene-expression patterns in peripheral blood cells. *Breast Cancer Research, 7*(5), R634 - R644.

Simonnet, H., Alazard, N., Pfeiffer, K., Gallou, C., Béroud, C., Demont, J., et al. (2002). Low mitochondrial respiratory chain content correlates with tumor aggressiveness in renal cell carcinoma. *Carcinogenesis, 23*(5), 759-768.

Warburg, O. (1956). On the Origin of Cancer Cells. *Science, 123*(3191), 309-314.

Wu., C. F. J. (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. *The Annals of Statistics, 14*(4), 1261-1295.

Yip, A., & Horvath, S. (2007). Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics, 8*(1), 22.

Zhang., B., & Horvath., S. (2005). A General Framework for Weighted Gene Co-expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*.