Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Fusion of FNA-cytology and Gene-expression Data Using Dempster-Shafer Theory of Evidence to Predict Breast Cancer Tumors

Mansoor Raza[1*], Iqbal Gondal[1, 3], David Green[1] and Ross L. Coppel[2, 3]

[1]GSIT, Faculty of IT, [2]Deparment of Microbiology, [3]Victorian Bioinformatics Consortium

Monash University, Australia; Mansoor Raza* - Email: Mansoor.Raza@Infotech.monash.edu.au; * Corresponding author

received May 18, 2006; revised July11, 2006; accepted July 17, 2006; published online July 19, 2006

Abstract:

Decision-in decision-out fusion architecture can be used to fuse the outputs of multiple classifiers from different diagnostic sources. In this paper, Dempster-Shafer Theory (DST) has been used to fuse classification results of breast cancer data from two different sources: gene-expression patterns in peripheral blood cells and Fine-Needle Aspirate Cytology (FNAc) data. Classification of individual sources is done by Support Vector Machine (SVM) with linear, polynomial and Radial Base Function (RBF) kernels. Out put belief of classifiers of both data sources are combined to arrive at one final decision. Dynamic uncertainty assessment is based on class differentiation of the breast cancer. Experimental results have shown that the new proposed breast cancer data fusion methodology have outperformed single classification models.

Keyword: Data fusion; Dempster-Shafer Theory; Classification; SVM; Breast cancer; Microarray; FNAc

Background:

Medical practitioners diagnose on the basis of information collected from different sources, effectively fusing the information to reach the decision. Information fusion refers to the combination of data originating from multiple sources and improving decision tasks, such as classification, estimation and prediction. Ultimately it provides a better understanding of the phenomena under consideration. In case of breast cancer, number of factors such as heterogeneity in diet, age, race, environmental factors, geographic location, number of pregnancies, as well as genetic makeup determines the risk of malignancy. **[1, 2]** The degree of complexity of the disease is further enhanced by chromosomal rearrangements frequently associated with the pre-malignant disease. The cellular pathways that are altered by these aberrations have been difficult to evaluate in patients, especially during early stages of the disease process. **[1, 2]**

Since there are number of factors that determined the risk of breast cancer, so it is not advisable to rely on just one source of information for diagnosis. There is a well established association between different symptoms of breast cancer e.g. germline BRCA1 or BRCA2 mutations are associated with increased lifetime risk of developing breast cancer **[3]** but not all mutation carriers develop breast cancer and the age of onset of breast cancer remains unpredictable. **[4]** There is a well established association between atypical ductal epithelium identified by histological biopsy, nipple aspiration (NA) or fine needle aspiration (FNA) and an increased risk of future breast cancer. **[4]** The relative risk of developing invasive breast carcinoma for women found to have atypical ductal hyperplasia on breast biopsy is 4.3 times that of the general population and, when combined with a positive family history, the relative risk of invasive breast cancer rises to 9.7 times that of the general population. **[5]**

Association between the different symptoms are not only factor that contribute the idea of fusing information from different resources but the limitation of diagnostic methods are one of the major fact as well e.g. mammographic screening is the most reliable method but often fails to detect tumors that are less then 5mm in size and also dense breast tissue are difficult to interpret. **[6]** The limitation of FNA can either be technical or related to the nature of the lesion itself. **[6]**

Medical information fusion has been demonstrated by Azuaje *et al.*, **[9]** an information fusion technique based on a knowledge discovery model and the case-based reasoning decision framework using data from heart disease risk estimation domain. Fusion techniques combine information at the retrieval-outcome level and data at the discovery-input level. **[9]** Paquerault et al., proposed a technique based on the fusion of one-view and two-view information to improve the performance of mammography mass detection of breast cancer. A classifier was trained to differentiate the true mass pairs from the false pairs. A final fusion stage combined the two-view object pair information with the one-view object scores. **[10]**

There is wide recognition of Fine Needle Aspiration (FNA) and microarray analysis as the principal diagnostic methods. **[1, 18]** Microarray methodology involves placing a large number of DNA fragments corresponding to the different genes to be studied on a glass slide or glass wafer. **[3]** Microarray analysis determined the level of expression in a tissue sample of many genes simultaneously. Microarray experiments generate large datasets with expression values for thousands of genes **[4]**, but usually not more than a few dozens of array, that gives rise to the issue of the curse of dimension. FNA cytology is the technique that involves the insertion of a fine needle (between 21 and 25 gauge) into a lesion and the extraction of a small sample of cellular material which is

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

smeared onto glass slides. The cells are stained and examined morphologically by cytopathologists. **[6]** The features are computed from digitized image of a fine needle aspirate to a breast mass. They describe characteristics of the cell nuclei present in the image. **[6]**

The aim of our work is to study and apply Dumpster Shafer theory of mathematical belief to fuse breast cancer data obtained from different diagnostic techniques in the management of breast disease. Input data, consisting of feature vectors ported into three different classifiers as input. The classifiers we used in this study are SVM with linear, Polynomial and RBF kernel. Each classifier provides beliefs of two classes benign and malignant. These beliefs are then combined to reach a final diagnosis using Dempster's combination formula. The experiments are carried on two types of breast cancer data. One is Fine Needle Aspirates Cytology (FNAc) data, other is obtained from gene expression pattern in peripheral blood cells. FNAc breast cancer data collected by physician W.O. Wolberg, University of Wisconsin Hospitals, contained 699 samples, 458 of which were benign and 241 of which were malignant. **[1]** FNAc data set is publicly available on UCI machine learning repository. Gene expression data consist of 60 blood samples obtained from 56 different women of which 24 were malignant and 36 were benign. **[6]** We have used leave one out cross validation. To implement this method, the available data was divided into k disjoint sets; k models were trained using different combination of k−1 partitions and were tested on the remaining partition. Cross-validation makes good use of the available data as each sample is used both as training and test data. Cross-validation is therefore especially useful where the amount of available data is insufficient to form the usual training, validation and test partitions required for split-sample training. **[12]**

Methodology:

To describe the methodology in figure 1 we start with the visualization of FNAc and microarray data. Each element of FNA cytology pattern sets consisted of 9 cytological characteristics. Each of 9 cytological characteristics of breast Fine Needle Aspirates (FNA) differs between benign and malignant samples.

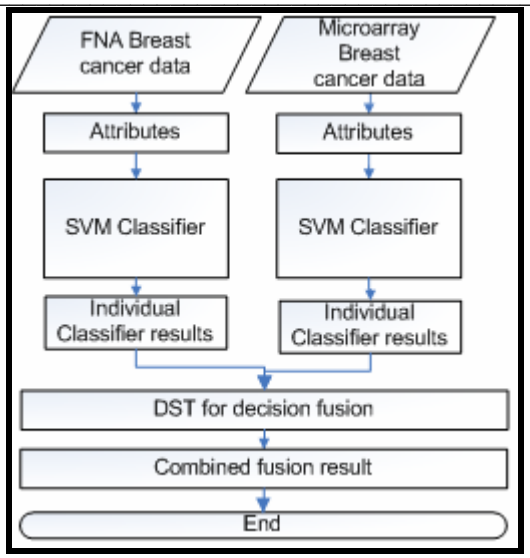

Figure 1: FNA-Cytology and gene-expression data fusion methodology using Dempster-shafer theory of evidence

The nine independent parameters of FNAc data are: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. **[18]** Each of these characteristics is assigned a number between 1 and 10, with the largerest numbers generally indicating a greater likelihood of malignancy. However, not a single measurement by itself can be used to determine whether the sample is benign or malignant.

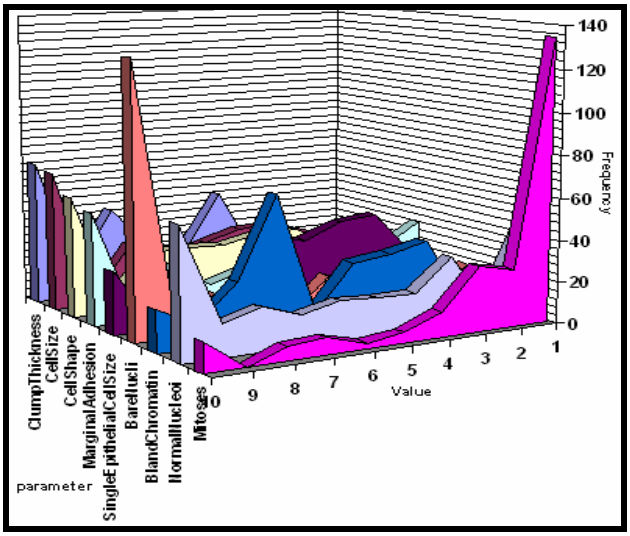

Figure 2: Visualization of FNAc malignant data set

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

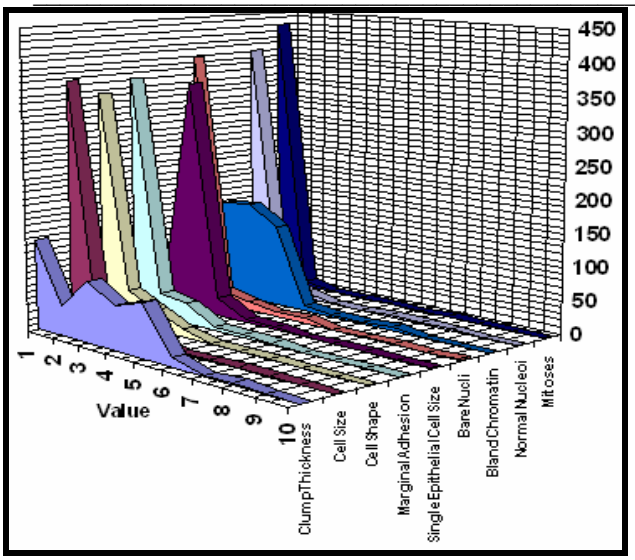

Figure 3: Visualization of FNAc benign data set

It was noted in figure 3 that the benign samples had lower parameter values than the malignant samples shown in figure 2. It was apparent that simultaneous simple frequency distribution histograms all nine parameters for each class, would graphically illustrate any differences between the two classifications, which is highlighted in figure 2 and 3.

The second data is from Sharma et al. that consist of microarray gene-expression pattern of 1368 genes in peripheral blood cells of 24 women with malignant breast cancer and 36 women with benign cancer. **[6]** Out of 1368 genes a panel of 37 genes had been identified with distinct expression patterns in malignant versus benign samples. We have used data matrix of 60 samples and 37 genes with two classes benign and malignant.

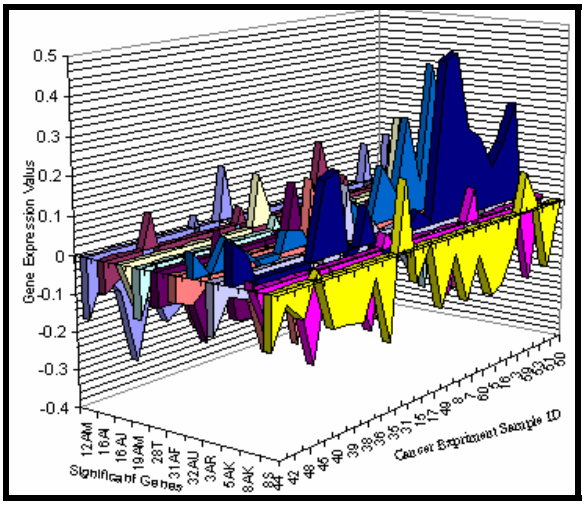

Figure 4: Visualization of microarray malignant data set

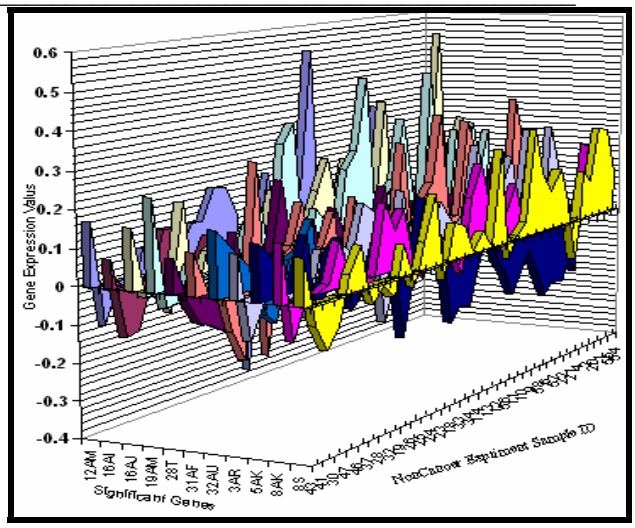

Figure 5: Visualization of microarray benign data set

The relative expressions of 11 features of selected genes are presented in figure 4 and 5. The expression level of cancer genes are shown in figure 4 with 36 samples of women with malignant cancer. The expression level of 11 genes with benign cancer of 24 samples is shown in figure 5.

The classifiers provide the category of the cancer class: benign and malignant. Each classifier provides belief for classes. These beliefs are then combined using Dempster's rule of combination. Dempster's combination involves evaluation of beliefs and uncertainties from individual classifiers along with classifier decision. The SVM believe is calculated using decision function in equation (6). Uncertainty is calculated using equation (13). Dempster's rules of combining belief and uncertainty are described in equation (12). Dempster's rules of combination involve evaluation of beliefs and uncertainties from individual data source along with classifier decision. We perform classification on all three kernels and chose the best classification result out of three. We then combined the beliefs of SVM classifier of microarray (S_{mic}), and SVM classifier of FNA (S_{fna}). Let's assume that Bel_S_{mic}(B) is the classifier beliefs of microarray and Bel_S_{fna}(B) is the beliefs of FNA of class benign. Uncertainties for two data sources are U_S_{mic} and U_S_{fna}. Combined belief for benign class is as follows:

$$Bel\_comb(B)=[Bel\_S_{mic}(B)*Bel\_S_{fna}(B)]+[U\_S_{fna}*Bel\_S_{n}\ B)]+[Bel\_S_{mic}(B)*U\_S_{fna}]] \quad (1)$$

The combined belief for malignant class is as follows:

$$Bel\_comb(M)=[Bel\_S_{mic}(M)*Bel\_S_{fna}(M)]+[U\_S_{fna}*Bel\_S_{c}(M)]+[Bel\_S_{mic}(M)*U\_S_{fna}]] \quad (2)$$

Support Vector Machine

Support vector machines are systems based on regularization techniques which performed well in many classification

problems. **[15]** SVM converts Euclidean input vector space in to higher dimensional space **[15]** and attempts to insert a separating hyperplane. The data Z is transformed into higher dimensional space. The separating hyperplane in higher dimension space satisfies

$$W.Zi + b = 0 \quad (3)$$

maximize the margin between classes Equation (4) and Equation (5) are used.

$$Max \frac{1}{\|W\|^2} \quad (3)$$

Subject to the condition

$$y_i (W.Zi + b) \geq 1 \quad (4)$$

In our experiments, a linear, polynomial (of degree 2 to 20) and RBF functions were used. The hyper plane found by an SVM in the feature space corresponds to a decision boundary in the input space. The value of the decision function can be used to evaluate belief masses. The SVM constructs a decision function that is represented in higher dimensional space by

$$D(x) = \sum_{k}^{p} \alpha_k K(x_k, x) + b \quad (5)$$

Where: **D(x)** is the decision function. **p** is the number of training examples in the training set. α is a learned parameter associated with the k^{th} training example. **K** is the kernel function which uses the k^{th} training example and the current input x; and **b** is a learned bias which is the same across all examples. The kernel function is

$$K(x_k, x) = \Phi(x_k) \bullet \Phi(x) \quad (6)$$

Dempster Shafer Theory of Evidence

DST is a generalization of the bayesian theory of subjective probability. Bayesian theory requires probabilities for each question of interest, while belief functions allow us to base degrees of belief on the probabilities of related question. The belief and the ignorance or uncertainty concerning a question can be modeled independently. **[19]** In a Dempster-Shafer reasoning system, all the mutually exclusive context interpretations are enumerated in a "frame-of-discernment", denoted Θ. A mathematical function that translates degree of support to belief is known as a Belief Function. **[19]** Basic belief m(*X*), which represents the strength or belief mass of some evidence for event *X* provided by the source of information under consideration, has the following properties

$$m(\phi) = 0 \text{ and } \sum_{x \subseteq \Theta} m(X) = 1 \quad (7)$$

Here ϕ is empty, indicates belief of empty set is always zero and Θ represents the total event space. The belief function for an event A is given by

$$Bel(A) = \sum m(X) \text{ Where } X \subseteq A \text{ and } A \subseteq \Theta \quad (8)$$

To understand how DST is related to our work, let's consider, there is a cancer patient, and from the reality constraints person have "Malignant" or "Benign" cancer. Now our task is to specify the cancer as one of the four possibilities described as: $\Theta = \{M, B, \{M, B\}, \phi\}$

Meaning person is "Malignant", "Benign", "either Malignant or Benign" (which is actually an indication of ignorance or uncertainty), or ϕ "neither Malignant nor Benign" (which is an indication of exceptional situation). With the frame of discernment Θ defined, each data source Di would contribute its outcome by assigning its beliefs over Θ. This assignment function is called the "probability mass function" of Di, denoted mi. So, according to Di's outcome, the probability that "the cancer is Malignant" is indicated by a "confidence interval" whose lower bound is a "belief" and whose upper bound is an "Uncertainty".

[$Belief_i$(M), $Uncertainty_i$(M)]

$Belief_i$(M) is quantified by all pieces of evidence E_k that support proposition "Malignant"

$$Belief_i(M) = \sum_{E_k \subseteq M} m_i(E_k) \quad (9)$$

Uncertainty (M) is quantified by all pieces of evidence E_k that do not rule out proposition "Malignant":

$$Uncertainty_i(M) = 1 - \sum_{E_k \cap M = \phi} m_i(E_k) \quad (10)$$

For each proposition in Θ, e.g., "Malignant", Dempster-Shafer theory gives a rule of combining data source Di's outcome mi and data source Dj's outcome mj

$$(m_i \oplus m_j)(M) = \frac{\sum_{E_k \cap E_{k'} = M} m_i(E_k) m_j(E_{k'})}{1 - \sum_{E_k \cap E_{k'} = \phi} m_i(E_k) m_j(E_{k'})} \quad (11)$$

The evaluation of uncertainty

We used Decision function in SVM to evaluate the belief masses '$m_{(i)}$'. Now we evaluate the uncertainty according to belief masses. If the value of beliefs for K classes is close to each other, then the classifier is more uncertain about its decision. **[19]** This mean as the beliefs start spreading apart uncertainty starts decreasing. Let uncertainty be denoted as S(U)

$$S(U) = -1 \sum_{i=1}^{k} (m(i) - \frac{1}{K})^2 \quad (12)$$

Results & Discussion:

This section provides the results of individual classifiers as well as the combination of classifiers using DST for breast cancer data. The performance of SVM based classifiers with linear, polynomial and RBF kernel has been evaluated using sensitivity, specificity, positive predicted value (PPV),

negative predicted value (NPV) and accuracy. To highlight these parameters, let's for some class A, the results would be True Positive (TP) if samples of class A are predicted as A and the results would be False Negative (FN) if samples of class A are predicted as non-A. The result would be False Positive (FP) if samples of non-A predicted as A, and True Negative (TN) if samples of non-A predicted as non-A. The following parameters are used to characterize performance of classifier and are given below

Sensitivity=TP/(TP+FN)	(13)
Specificity=TN/(FP+TN)	(14)
PPV=TP/(TP+FP)	(15)
NPV=TN/(FN+TN)	(16)

Sensitivity is the probability for a class A sample to be correctly predicted as class A, Specificity is the probability for a non class A sample to be correctly predicted as non-A, PPV is the probability that a sample predicted as class A actually belongs to class A, NPV is the probability that a sample predicted as non class A actually does not belong to class A. For each classification method and each class, these parameters are listed in the tables below. Table 1 shows the results of SVM classifiers on FNA data. The overall accuracy is 90.25% with sensitivity of malignant is 91.6% and benign is 88.8%. Table 2 shows the performance of the microarray data the overall accuracy is 81.94% with sensitivity of malignant is 80.50% and benign is 83.30%. Table 3 shows the result of application of DST to fuse the classifiers. The overall accuracy shown in Table 3 is 94.4% with sensitivity of malignant is 97.1% and benign is 94.4%. Table 3 shows improved accuracy using information fusion with DST.

Class	Sensitivity	Specificity	PPV	NPV
Malignant	0.916	0.888	0.891	0.923
Benign	0.888	0.916	0.923	0.891

Table 1: Performance of the Support Vector Machine Classifier on FNA data

Class	Sensitivity	Specificity	PPV	NPV
Malignant	0.805	0.833	0.828	0.878
Benign	0.833	0.805	0.878	0.828

Table 2: Performance of the Support Vector Machine Classifier on gene expression data

Class	Sensitivity	Specificity	PPV	NPV
Malignant	0.971	0.944	0.921	0.972
Benign	0.944	0.971	0.972	0.921

Table 3: Performance of the combined result of fusion using DST

SVM-Micro			SVM-FNA			Combined-Fusion (DST)		
	M	**B**		**M**	**B**		**M**	**B**
M	29	7	**M**	33	3	**M**	35	1
B	6	30	**B**	4	32	**B**	3	33

Table 4: Confusion matrices of individual classifiers and the combined result of fusion using Dempster Shafer Theory

A confusion matrix in Table 4 shows the classification results of classes Malignant (M) and Benign (B) classes. Fusion with DST shows the maximum accuracy where 35 malignant classes were correctly identified while 1 was classified as benign. The use of FNA data with the SVM classifiers identified 33 benign classes and 3 were incorrectly classified as malignant. Table 5 shows that when two single sources of data: gene expression and Cytology FNA data were fused using Dempster Shafer it showed higher accuracy.

	SVM-Micro	SVM-FNA	Combined Fusion (DST)
Overall Accuracy	82.00	90.27	94.44

Table 5: Accuracy of classifiers for test cases on malignant and benign

Overall accuracies of individual and DST classifiers in Table 5 show that fusion by DST has improved the breast cancer prediction as compared to individual classifiers.

Conclusion:

We have looked at the fusion of data from disparate sources for the prediction of breast cancer tumors. We have demonstrated our methodologies for fusing data form FNAc and microarray data set to achieve a batter overall prediction of breast cancer tumors. The paper has presented a method for fusing medical data using multiple classifiers where uncertainty and unequal costs of errors are present. The fusion framework has been presented for the computation of belief functions and uncertainty values from individual classifiers and data fusion through the Dempster-Shafer theory, in which class differentiation quality is used for the computation of uncertainties. The fusion approach has shown the best classification accuracy of breast tumor classification. The fusion approach remained robust in the presence of fairly different classifier performances. The ability to handle such situations robustly and the ability to classify samples in the presence of classifier uncertainty, makes this approach attractive for healthcare applications.

References:

[01] J. Russo & I. H. Russo, *Oncology Research,* 11:169 (1999) [PMID:10566615]

[02] A. S. Ketcham & W. F. Sindelar, *Progress in Clinical Cancer,* 6:99 (1975) [PMID:1105679]

[03] Antoniou, *et al., Am J Hum Genet.*, 72:1117 (2003) [PMID:12677558]

[04] G. Mitchell, *et al., Breast Cancer Research,* 7:R1122 (2005) [PMID:16457692]

[05] D. L. Page, *et al., Cancer,* 55:2698 (1985) [PMID:2986821]
[06] P. Sharma, *et al., Breast Cancer Research,* 7:R634 (2005) [PMID: 16168108]
[07] J. Torhorst, *et al., American Journal of Pathology,* 159:2249 (2001) [PMID:11733374]
[08] Y. Miki, *et al., Science,* 266:66 (1994) [PMID:7545954]
[09] A. Francisco, *et al., IEEE Trans. On Biomedical Eng.,* 46 (1999) [PMID:10513121]
[10] S. Paquerault, *et al., Medical Physics,* 29:238 (2002) [PMID:11865995]
[11] S. Dudoit, *et al., Journal of the American Statistical Association*, 77 (2002)
[12] M. Stone, *Journal of the Royal Statistical Society*, 36:111 (1974)
[13] P. S. Brown, *et al., Proc. Natl. Acad. Sci.,* 97:262 (1997) [PMID:10618406]
[14] T. S. Furey, *et al., Bioinformatics,* 10:906 (2000) [PMID:11120680]
[15] O. Chapelle, *et al., Machine Learning,* 46:131 (2002)
[16] Y. Freund, *et al., Journal of Machine Learning Research,* 4:933 (2003)
[17] Y. Freund, *et al., Journal of Computer and System Sciences,* 55:119 (1997)
[18] W. H. Wolberg & O. L. Mangasarian, *Proc Natl Acad Sci.,* 87:9193 (1990) [PMID:2251264]

Edited by S. Krishnaswamy
Citation: Raza *et al.*, Bioinformation 1(5): 170-175 (2006)