Data Mining for DNA Viruses with Breast Cancer and its Limitation

Ju-Hsin Tsai
Central Taiwan University of Science and Technology, and Surgical Department,
Shian-De General Hospital, No. 420. Yichang Rd.,
Taiping City, Taichung County 411,
Taiwan

1. Introduction

Breast cancer is very common worldwide, with 800,000 new cases diagnosed each year [Parkin et al,1999]. Among Taiwanese women, breast cancer is the second most common form of cancer (Cancer Registry Annual Report in Taiwan, 1998-2002) and the fourth leading cause of cancer-related death (Public Health Annual Report in Taiwan, 2002). The risk factors for development of breast cancer in Taiwan, a low incidence area, are similar to those in a moderate-to-high risk area [Yang et al,1997]. Although there are recognized factors that increase the risk of breast cancer, its causes are still unknown and thus there is no way of preventing it. This paper explores the possibility that viruses play a role in development of breast tumors. DNA viruses have been recognized as oncogenic in humans: Examples include EBV, which is associated with Burkitt lymphoma and nasopharyngeal carcinoma, HPV, which is associated with cervical cancer, and hepatitis B virus, which is associated with hepatocellular cancer. Viral DNA sequences have been found in breast tumors [Tsai et al,2005, ue et al,2003. Fina et al,2001. Labecque et al,1995. Horiuchi et al,1994. Kleer et al,2002], but it is not certain that the presence of the virus is related to development of breast tumors because similar sequences have been found in normal mammary tissue Tsai et al,2005. Pogo et al,1997].

Breast cancer is a multistep disease, and infection with a DNA virus could play a role in one or more of the steps in this pathogenic process [Labecque et al,1995]. In addition, it has been hypothesized that familial breast cancer and sporadic breast cancer are caused by different mechanisms. More recently, the differences between familial and sporadic breast cancers have be shown to be compatible with Knudson's 'two - hit' hypothesis [Knudson,1971. Richardson,1997],which suggests that at least two mutations are required before a cell becomes malignant. The reason thatwomen with familial predisposition to breast cancer are likely to develop it at a younger age and are also morelikely to develop bilateral disease is because they have inherited one of the two genetic defects (such as mutation of p53) that are required for breast cancer. These women require only one 'hit' to get breast cancer, whereas women with non-familial breast cancer start with no major mutations and thus need two'hit' Thisobservation is consistent with the hypothesis that breast cancer is caused in part by a virus. One of the 'hit'required for development of breast cancer may be infection with a

Source: Data Mining in Medical and Biological Research, Book edited by: Eugenia G. Giannopoulou,
ISBN 978-953-7619-30-5, pp. 320, December 2008, I-Tech, Vienna, Austria

breast cancer virus. There are a few reports about the relationship between fibroadenoma and virus infection [Kleer et al,2002. Lau et al,2003], but they only investigated EBV andfibroadenoma. As far as we know, we are the first to report about relationships among multiple viruses and fibroadenoma [Tsai et al,2005]. Thus, the question arises whether or not breast tumors (either benign – fibroadenoma – or malignant – breast cancer) are influenced by infection with oncogenic viruses.

2. Materials and methods

In the current study, we explored possible relationships among DNA viruses with breast cancer, fibroadenoma, and normal mammary tissue with 106 data points (tissue samples), including 62 specimens of non- familial invasive ductal breast cancer from women and 32 mammary fibroadenomas and 12 normal mammary tissues from women cared for at Chung-Shan Medical University Hospital, with tissues collected as previously described [3]. DNA extraction, Polymerase Chain Reaction (PCR) assay, and Southern Hybridization as described previously [Tsai et al , 2005] were applied here. Genome regions, primers, and thermal cycler programs for iden- tification of each virus (e.g., HSV-1, EBV, CMV, HPV, and HHV-8) were listed in the previous report [Tsai et al , 2005].

Using PCR and Southern hybridization, 69 breast cancer and 44 non-breast cancer specimens were screened for the presence of, β-globin, the internal control. However, five breast cancer tissue samples had negative results for β-globin, and the two breast cancer specimens from patients with familial histories of breast cancer were excluded from the study. All 44 non-breast cancer specimens were positive for. β-globin. Among the 62 breast cancer samples, 8 (12.90%) were positive for HSV-1, 28 (45.16%) for EBV, 47 (75.81%) for CMV, 8 (12.90%) for HPV, and 28 (45.16%) for HHV-8. In the non-breast cancer control groups, 8/12 (66.67%) were CMV-positive normal samples, whereas the results for fibroadenoma samples (total, 32) were 20 (62.50%)HSV-1-positive, 16 (50.00%) EBV-positive, 20 (62.50%) CMV-positive, 2 (6.25%) HPV-positive, and 28 (87.50%) HHV-8-positive.

We submitted a data mining approach to the current research that included artificial neural networks (ANNs) and agglomerative hierarchical clustering techniques (AHCTs). With the proposed data mining approach (named ANN-AHCT hereafter), the different combinations of DNA viruses possible in breast cancer, fibroadenoma, and normal mammary tissue were classified by ANNs; then, AHCTs clustered the common characteristic during the same classification.

2.1 Artificial neural networks (ANNs) model

ANNs are composed of processing elements (nodes or neurons) and their connections. The nodes are inter- connected layer-wise among themselves. Each node in each successive layer receives the inner product of synaptic weights with the outputs of the nodes in the previous layer. The operation of single node is shown in Fig. 1. ANNs have been shown to be effective for addressing complex nonlinear problems. The two types of learning networks are supervised and unsupervised. For a supervised learning network, a set of training input vectors with a corresponding set of target vectors is trained to adjust weights in the ANN. For an unsupervised learning network, a set of input vectors is proposed; however, no target vectors are specified. In this study, a supervised learning network was thought to be more suitable for the classification problem. Several well-known supervised learning ANNs are the back-propagation (BP), learning vector quantization, and counter propagation network. The BP model is used most extensively and can provide better solutions for many applications [Lippmann,1987. Dayhoff,1990]. Therefore, the BP model was selected for the current study.

A BP neural network consists of three or more layers, including an input layer, one or more hidden layers, and an output layer. Fig. 2. illustrates a basic BP neural network with three layers. BP neural network learning works on a gradient-descent algorithm [Funahashi, 1989]. The BP neural network initially receives the input vector and directly passes it into the hidden layer(s). Each element of the hidden layer(s) is used to calculate an activation value by summing up the weighted input, and the sum of the weighted input will be transformed into an activity level by using a transfer function. Each element of the output layer is then used to calculate an activation value by summing up the weighted inputs attributed to the hidden layer. Next, a transfer function is used to calculate the network output. The actual network output is then compared with the target value. The BP neural network algorithm refers to the propagation of errors of nodes from the output layer to nodes in the hidden layer(s).These errors are used to update the network weights. The amount of weights to be added to or subtracted from the previous weight is governed by the delta rule. After the knowledge representation is determined, the BP neural network will be trained to attempt the classification behavior. The number of hidden layers and the number of nodes in each hidden layer are determined during the training phase. In this study, a fully connected feedforward neural network was used, and its network parameters and stopping criterion were set.

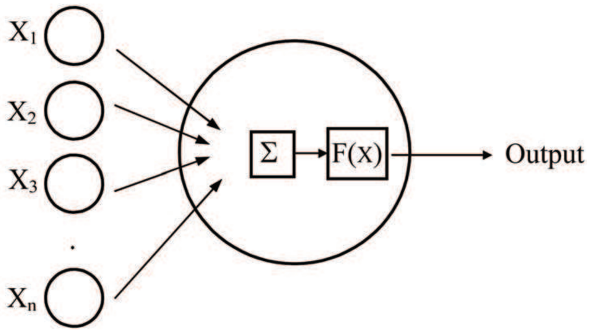

Fig. 1. Node operation

To be able to attempt the classification behavior, a learning rule needs to be used in the BP neural network. In the case of a multi-layer perception, this rule should also be able to adapt the weights of all connections in order to model a nonlinear function. The learning rule used most frequently for this purpose is the BP rule. It acts through the following two steps. First, the generalized difference D *(t) is calculated by

$$D_i^*(t) = (A_i^*(t) - A_i(t)) * A_i(t) \quad (1 - A_i(t)), \tag{1}$$

where A_i^* is the desired activation of output unit i, and A_i^* (*t*) is the generated activation of this unit. In order to obtain the generalized difference the calculated difference A_i^* (*t*) - $A_i(t)$ is multiplied by the simplified derivative of the activation function $A_i(t)$* (1 -$A_i(t)$). Second, the generalized differences of the units in the output layer are propagated back through the weighted connections to the units of the hidden layer(s). The generalized difference collected from a hidden unit is multiplied by the simplified derivative of the unit's activation function in order to obtain the generalized difference of the hidden unit

$$D_j^*(t) = \sum_{i=1}^{n} (W_{ij}(t) * D_i^*(t)) * A_j(t) * (1 - A_j(t)). \tag{2}$$

Using the generalized difference D*(t), the weights are adjusted by

$$W_{ij}(t+1) = W_{ij}(t) + C * D_i^*(t) * A_j(t). \tag{3}$$

The adaptation size of the weight Wij(t) of the connection used to send information from unit j to unit i is influenced by the existing weight Wij(t), the learning rate C, the generalized difference D_i^* (*t*), and the actual activation Aj(t) of unit j. To reduce the probability of weight change oscillation, a weight momentum term is added to adjust the weight. The weight momentum term is constructed by previous adjustment of the weight $D^*W_{ij}(t)$ and a constant value B, so

$$W_{ij}(t+1) = W_{ij}(t) + C * D_i^*(t) * A_j(t) + B * D * W_{ij}(t). \tag{4}$$

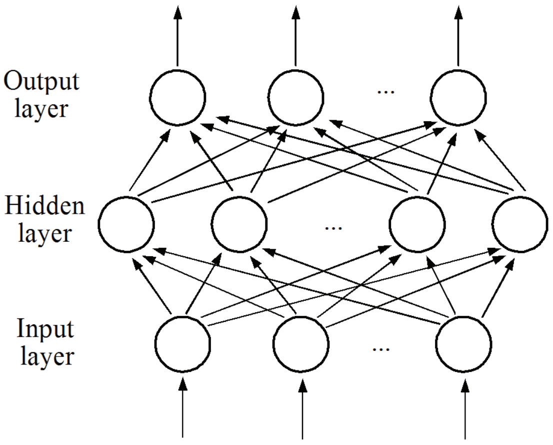

Fig. 2. A back-propagation (BP) neural network

If more hidden layers are implemented, the BP rule will use the generalized differences of the hidden units of the BP neural network to get the hidden units of the hidden layer closer to the input layer. To test the network, test set data are assigned to the networks, and then the output is evaluated. The network should be able to interpolate and, possibly, extrapolate.

In this study, through the above-mentioned principle for construction of a BP model, we collected training and testing patterns by randomly selecting data from the total number of 106 (specimens, or data points) to correlate the presence of DNA viruses with breast cancer, fibroadenoma, and normal mammary tissue to develop a BP model that could obtain underlying relationships. The BP model can be constructed without requiring any assumptions concerning the functional form of the relationship among the DNA viruses with breast cancer, fibroadenoma, or normal mammary tissue. The developed BP model can classify the behavior of all possible combinations of DNA viruses. Then, all possible DNA virus combinations in the developed classification model would be presented and the estimated probability of breast tumor (benign or malignant) would be computed.

2.2 Agglomerative hierarchical clustering techniques (AHCTs)

AHCTs are a statistical method that serially fuses n individuals into groups in order to obtain partitions. When AHCTs have placed two individuals into the same group, the two individuals cannot subsequently appear in different groups, that is, the AHCT procedure is irreversible. AHCTs' main objective is to organize data in order to form clusters that contain all individuals. In this project, AHCTs were applied to cluster DNA viruses that belong to the same classification based on the ANN's classification behavior. The decision- maker's duty is to seek the 'best' fitting number of clusters needed to decide how to organize data. The AHCT procedure is as follows [Salton,1989]:

P_n, $P_{n\,1}$, ... , P_1 represents a series of partitions of data. The first, Pn, includes n single member clusters, and the last, P1, includes a single group that contains all n individuals. The basic operation to form P_n, $P_{n\,1}$, ... , P_1 is similar.

STEP 0: Each cluster C_1, C_2, ... , C_n includes a single individual.

STEP 1: To find the nearest pair of distinct clusters, say C_i and C_j, then to merge C_i and C_j, and to delete C_j, decrease the number of clusters by one.

STEP 2: If the number of clusters equals one then stop, or else return to STEP 1.

Three inter-group measures of AHCTs differ primarily in the distances between or similarity of two clusters.

One is single linkage clustering (5), another is complete linkage clustering (Eq. 6), and the other is average linkage clustering (Eq. 7) [Bunke & Shearer,1998. Wallis et al,2001].

$$d_{AB} = \min_{\substack{i \in A \\ i \in B}} (d_{ij}), \tag{5}$$

$$d_{AB} = \max_{\substack{i \in A \\ i \in B}} (d_{ij}) \tag{6}$$

where dA_B is the distance between two clusters A and B, and d_{ij} is the distance between individuals i and j. (This could be a Euclidean distance or one of a variety of other distance measures.)

$$d_{AB} = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{i \in B} d_{ij}, \tag{7}$$

where n_A and n_B are the number of individuals in clusters A and B.

The average linkage cluster is the same as single linkage or complete linkage. However, the cluster criterion of the average linkage cluster is the average distance from all individuals in one cluster to all individuals in another. Unlike single linkage clustering and complete linkage clustering, average linkage clustering does not depend on extreme values to partition all members of the cluster. The other merit for using the average linkage approach is to form clusters with small within-cluster variation. Average linkage clustering also tends to be biased toward production of clusters with approximately the same variance. So, the authors of the current work considered the average linkage cluster's advantages and the steps of computing Eq. (7) to further cluster the DNA viruses in the ANN-AHCT approach.

2.3 The proposed ANN-AHCT approach

In order to obtain the relationship among combinations of DNA viruses with breast cancer, fibroadenoma, and normal mammary tissue, an effective and feasible data mining method, the ANN-AHCT approach, was proposed. It includes two phases that are summarized in Fig. 3, which shows the structure of the ANN-AHCT approach flow chart.

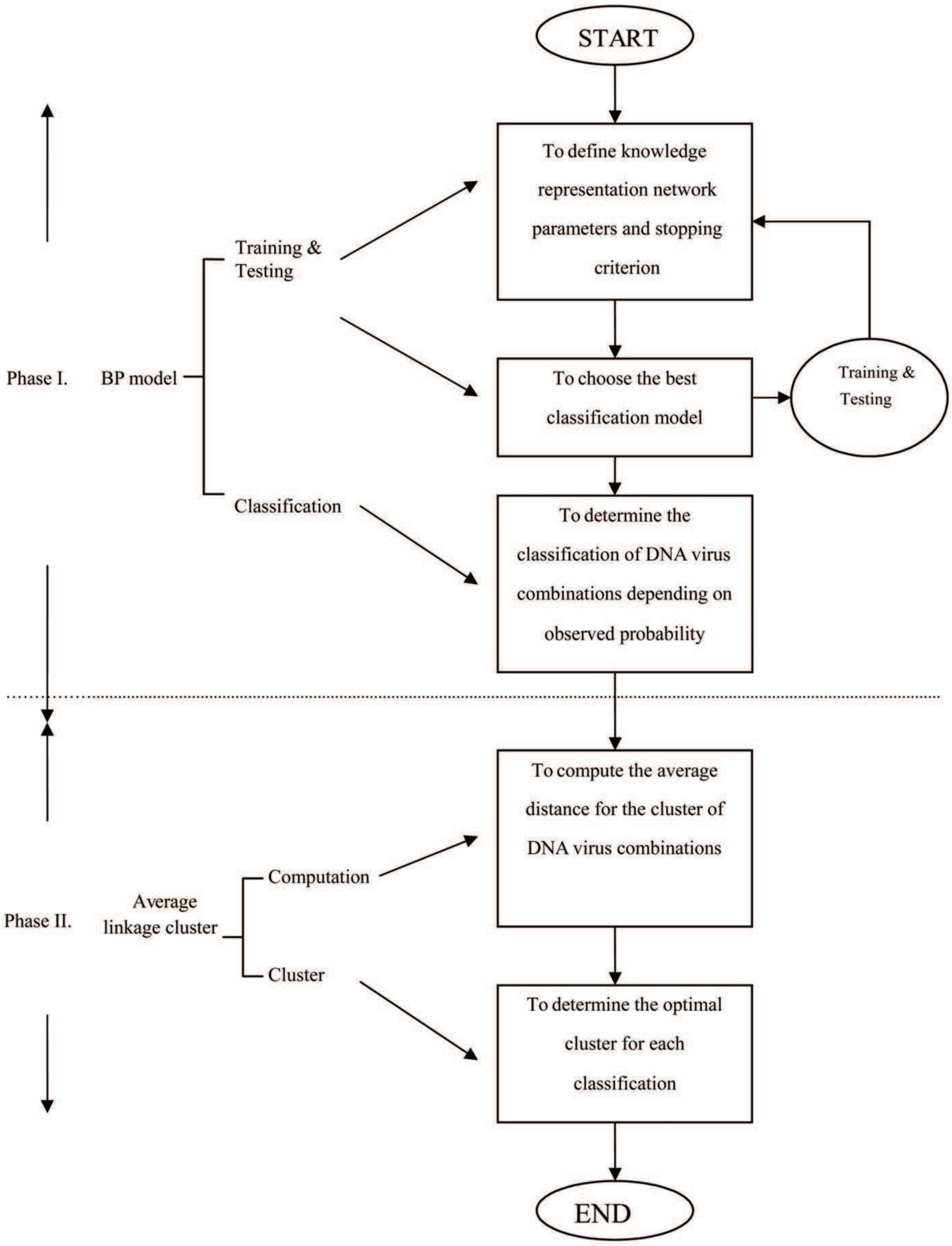
START
To define knowledge representation network parameters and stopping criterion
Training & Testing
Phase I.
BP model
To choose the best classification model
Training & Testing
Classification
To determine the classification of DNA virus combinations depending on observed probability
To compute the average distance for the cluster of DNA virus combinations
Computation
Phase II.
Average linkage cluster
Cluster
To determine the optimal cluster for each classification
END

Fig. 3. The flow chart ANN-AHCT approach.

Phase I. Using the BP model to classify all combinations of DNA viruses.

To define the knowledge representation, network parameters and stopping criterion is most important to obtain the best classification BP model. The knowledge representation defines the relationship among the DNA viruses with breast cancer, fibroadenoma, and normal mammary tissue. The number of input nodes is equal to the number of DNA viruses (five, HSV-1, EBV, CMV, HPV, HHV-8); the input values are the value of (positive, negative) DNA viruses. In addition, the number of output nodes is three (breast cancer, fibroadenoma, and normal mammary tissue); the output values are the code. This means that if breast cancer is present, the code is 1, otherwise, the code is 0; similarly, if fibroadenoma is present, the code is 1, otherwise, the code is 0; if normal mammary tissue is present, the code is 1, otherwise, the code is 0. In addition, the network parameter-learning rate and moment will be set to assist the trained network to attempt convergence andstabilization in classification behavior.

The stopping criterion is set to lower the root mean square error (RMSE) in the training and testing processes. In this study, in order to obtain the appropriate BP model, the iteration was set to 10,000 and the learning rate was set to dynamically auto-adjust from 0.01 to 0.3 for rapid effective learning and stable behavior as observed by mildly varying values of RMSE. Table 1 lists these appropriate BP architectures and their momentums.

In order to obtain the best BP model from Table 1, selection of the best classification model is done through selecting the lowest RMSE of training and testing or the highest classification correction rate. The architecture (input nodes-hidden nodes-output nodes) 5-2-3 was selected to obtain a better performance. In order to obtain the relationship among each different combination of DNA viruses with breast cancer, fibroadenoma, and normal mammary tissue, the total number 32 (2 * 2 * 2 * 2 * 2, the combination of DNA viruses) was inputted into the architecture 5-2-3, and then the classification of breast tumor and occurrence probability could be obtained. Placement of all DNA virus combinations into Tables 2–5 depended on the possible occurrence probability of breast cancer, fibroadenoma, and normal mammary tissue.

In Table 2, 10 DNA virus combinations would result in the highest probability of fibroadenoma; in Table 3, 15 DNA virus combinations would result in the highest probability of breast cancer; and in Table 4, 4 DNA virus combinations would result in a higher probability of breast cancer than of normal mammary tissue, a value that was also higher than the probability of fibroadenoma. In Table 5, 3 DNA virus combinations resulted in a higher probability of breast cancer or fibroadenoma, that is, a breast tumor.

Phase II. Using average linkage clustering to obtain different clusters during each classification.

In order to obtain the common combination for each classification, further clustering of each classification was necessary. This study used average linkage clustering to obtain different clusters in each classification. In Tables 2 and 3, the results of average linkage clustering were computed and depicted the possible cluster results shown in Tables 6 and 7. To decide the best number of clusters in Table 6, a medical expert in breast tumors at the Chung-Shan Medical University Hospital suggested that three clusters would be best. The first cluster is labeled 10 and 12. The second cluster is labeled 26, 28, 30, and 32. The third cluster is labeled 18, 20, and 24. The label 27 was clustered in the three clusters with difficulty. In addition, to decide the best number of clusters in Table 7, the medical expert suggested that four clusters would be best. The first cluster is labeled 2,4, 6, and 8. The second cluster is labeled 9 and 11. The third cluster is labeled 13, 15, 29, and 31. The fourthcluster is labeled 17, 19, 21, and 23.

The label 14 was clustered in the four clusters with difficulty. In Table 3, where the combination included only four situations, it was not necessary to further cluster. Similarly, in Table5, the combination included only three situations, so it was not necessary to further cluster.

Using used Kendall's tau-b test to examine the correlation between number of virus infections and survivals in breast cancer patients. The results suggest that the number of infecting viruses is related to the overall and relapse-free survivals of breast cancer patients (correlation coefficient=-0.275; P=0.021). In order to further investigate the relationship between viral factors and overall and relapse-free survivals in our sample of breast cancer patients, univariate log-rank analysis was used. Both overall and relapse-free survival rates were significantly different (P=0.001 and P=0.000, respectively; Table 9) comparing V0, V1, V2, V3 and V4 subgroups (zero, one, two, three and four virus infections, respectively).

Architecture (nodes) input-hidden-output	Momentum	RMSE		Correction rate (%)
		Training	Testing	
5-1-3	0.52	0.02358	0.02334	89
5-2-3	0.45	0.02227	0.02241	93
5-3-3	0.63	0.02587	0.02432	85
5-4-3	0.43	0.02262	0.02258	93
5-5-3	0.60	0.02428	0.02518	85
5-6-3	0.55	0.02344	0.02488	89
Iteration 10,000.				

Learning rate 0.01−0.

Table 1. Network options for current study

Label	HSV-1	EBV	CMV	HPV	HHV-8	Breast cancer	Normal mammary tissue	Fibroadenoma
10	-	+	-	-	+	0.10	0.00	0.90
12	- -	+	-	+	+	0.02	0.00	0.98
26	+	+	-	-	+	0.00	0.00	1.00
28	+	+	-	+	+	0.00	0.00	1.00
30	+	+	+		+	0.00	0.00	1.00
32	+	+	+	+	+	0.00	0.00	1.00
18	+	-	-	-	+	0.01	0.00	0.99
20	+	-	-	+	+	0.00	0.00	1.00
24	+	-	+	+	+	0.16	0.00	0.84
27	+	+	-	+	-	0.28	0.00	0.72

Table 2. DNA virus combinations for fibroadenoma

Label	HSV-1	EBV	CMV	HPV	HHV-8	Breast cancer	Normal mammary tissue	Fibroadenoma
2	-	-	-	-	+	0.96	0.00	0.04
4	-	-	-	+	+	0.90	0.00	0.10
6	-	-	+	-	+	1.00	0.00	0.00
8	-	-	+	+	+	0.99	0.00	0.01
9	-	+	-	-	-	0.99	0.00	0.01
11	-	+	-	+	-	0.99	0.00	0.01
13	-	+	+	-	-	0.99	0.00	0.01
15	-	+	+	+	-	0.99	0.00	0.01
29	+	+	+	-	-	0.97	0.00	0.03
31	+	+	+	+	-	0.92	0.00	0.08
17	+	-	-	-	-	0.98	0.00	0.02
19	+	-	-	+	-	0.98	0.00	0.02
21	+	-	+	-	-	0.99	0.00	0.01
23	+	-	+	+	-	0.99	0.00	0.01
14	-	+	+	-	+	0.79	0.00	0.21

Table 3. DNA virus combinations for breast cancer

Label	HSV-1	EBV	CMV	HPV	HHV-8	Breast cancer	Normal mammary tissue	Fibroadenoma
1	-	-	-	-	-	0.59	0.28	0.12
3	-	-	-	+	-	0.69	0.21	0.10
5	-	-	+	-	-	0.50	0.36	0.14
7	-	-	+	+	-	0.61	0.27	0.12

Table 4. DNA virus combinations and probability for breast cancer

Label	HSV-1	EBV	CMV	HPV	HHV-8	Breast cancer	Normal mammary tissue	Fibroadenoma
16	-	+	+	+	+	0.50	0.00	0.50
22	+	-	+	-	+	0.43	0.00	0.57
25	+	+	-	-	-	0.60	0.00	0.40

Table 5. DNA virus combinations and probability of breast cancer or fibroadenoma

Cluster (total)	{Label}
1	{24, 20, 18, 32, 30, 28, 26, 12, 10}
2	{24, 20, 18} {32, 30, 28, 26, 12, 10}
3	{24, 20, 18} {32, 30, 28, 26} {12, 10}

Table 6. Possible total number of clusters for fibroadenoma

Cluster (total)	{Label}
1	{23, 21, 19, 17, 31, 29, 15, 13, 11, 9, 8, 6, 4, 2}
2	{23, 21, 19, 17, 31, 29, 15, 13, 11, 9} {8, 6, 4, 2}
3	{23, 21, 19, 17} {31, 29, 15, 13, 11, 9} {8, 6, 4, 2}
4	{23, 21, 19, 17} {31, 29, 15, 13} {11, 9} {8, 6, 4, 2}

Table 7. Possible total number of clusters for breast cancer

Moreover, as shown in Table 9, significant differences were also demonstrated in comparisons of the respective survival rates for the virus infection subgroups: V(0,1), V2, V3 and V4; V(0, 1), V2, V(3, 4), V(0, 1), V(2, 3), V4, and, V(0, 1) and V(2, 3, 4) (P<0.005 or <0.001) (n, n+1.., indicates virus number). Except for the V0 vs. V1 vs. V2 vs. V3 vs. V4 variable, only when V(0,1) was grouped for comparison with other multiply virus-infected subgroups, however, were the overall and relapse-free survivals significantly different. These results suggest that the number of virus infections is related to the overall and relapse-free survivals in our sample of breast cancer patients, moreover, the overall and relapse-free survivals of multiply (more than two) virus-infected breast cancer patients group is significantly different from the no virus- or one virus-infected group.

3. Discussion

Previous studies have provided direct evidence that viruses exist in human breast tumors and suggest that viruses are one risk factor for breast tumors [Tsai et al,2005. Pogo et al,1997. Brower, 2004]. However, some of these studies are disputed [Gopalkrishna et al,1996. McCall et al,2001]. From the sample data, this study detected DNA of the five viruses HSV-1, HPV, CMV, EBV, and HHV-8 in some tissue samples from patients with breast cancer or fibroadenoma or women with normal mammary tissue. Only CMV was detected in some normal mammary tissue samples. In contrast, breast cancer and fibro- adenoma had a much higher frequency for the presence of DNA belonging to two or more viruses (76.90% and 100.00%, respectively) than the presence of DNA from one virus (23.10% and 0.00%, respectively). This suggests that multiple viral infection is closely associated with benign or malignant breast tumors. Lawson et al. speculated that EBV may enhance the action of the human homologue of the mouse mammary tumor virus (HHMMTV) because it is known that some viruses remain dormant unless activity is promoted by the activity of another virus [Mckeating et al,1990. Biegalke, Geballe,1991]. Therefore, different viruses have different oncogenic potencies in mammary gland tissue.

In Table 2 showing data from this study, both the HSV-1 and HHV-8 positive group (clusters 2 and 3) with or without EBV infection nearly always had the pathological diagnosis of fibroadenoma (nearly 99.00%). The first cluster of Table 2 has the same result, but with the conditions HSV-1(-), EBV(+), CMV(-), HHV-8(+). In Table 8, the individual effect of HSV-1 and HHV-8 is seen between fibroadenoma and breast cancer. The HSV-1(+) group shows OR (Odds Ratio) = 0.09, 95%CI (Confidence Interval) = 0.03–0.25, P (P-value) < 0.001. The HHV-8(+) group shows OR = 0.12, 95%CI = 0.04–0.38, P < 0.001. The individual effect of HSV-1 and HHV-8 between fibroadenoma and breast cancer shows that both HSV-1 and HHV-8 positive groups appear to show a strongly protective effect against the progression from fibroadenoma to breast cancer. When the combination

HSV-1(-), HHV-8(+) is compared with HSV-1(+), HHV-8(+) for fibroadenoma and breast cancer cases, OR = 20.83, 95%CI = 4.88–88.87, P < 0.001. There is a similar difference between cluster 1 of Table 2 and cluster 1 of Table 3: Apart from the combination HSV-1(-), HHV-8(+),the other condition found to be present was EBV(+), CMV(-), which was associated with a roughly 99.00% probability of fibroadenoma. In our previous report [Tsai et al , 2005], when we took the comparative results of the mammary fibroadenoma group with normal tissues into account (data not shown), EBV was closely related to fibroadenoma (P < 0.01). Kleer et al. suggested that EBV was associated with fibroadenoma in an immunosuppressed population, with infection localized specifically to epithelial cells. Gandhi et al. suggested that the replication of CMV in the absence of an effective immune response is central to pathogenesis of disease. Therefore, complications such as tumor formation are primarily seen in individuals whose immune systems are immature or are suppressed by drug treatment or coinfection with other pathogens. In this study,EBV appears to be related closely to fibroadenomas in non-immunosuppressed women. In this classification,the prerequisite was CMV(+) or CMV(-). When HSV-1(-), HHV-8(+) accompanied EBV(-) (cluster 1 of Table 3), there was a greater than 96.00% probability of breast cancer.

DNA viruses have been recognized as oncogenic in humans, and viral DNA sequences have been found in breast tumors [Tsai et al,2005. Kleer et al,2002]. Different viruses have different carcinogenicity, as has been shown in models of mouse mammary cancer induced by 7.12-dimethylbenz(a)anthrance (DMBA). Qing et al. gave the classical dosage (1mg DMBA give once a week for six weeks) intragastrically to female SENCAR mice, with resulting high toxicity; the major tumor type was lymphoma. Lowering the dose to 60 mcg/day produced less toxicity; in terms of tumor type, there was a 75% incidence of lymphoma and a 30% incidence of mammary carcinoma. However, 20mcg DMBA given five times per week for six weeks resulted in a 65–70% incidence of mammary carcinoma. It is known that some viruses remain dormant unless activity is promoted by the presence of another virus [Mckeating et al,1990. Biegalke et al,1991]. In contrast, some virus activity may be diminished by the presence of another virus. This may be the basis for the different result between HSV-1(-), HHV-8(+) with EBV(+) or EBV(-), CMV(-).

In Table 3, cluster 2 shows HSV-1(-), EBV(+), CMV(-), HHV-8(-). In Table 8, comparing HSV-1(-), HHV-8(-) to HSV-1(+), HHV-8(+) yields OR = 48.33, 95%CI = 9.75–239.65, P < 0.001, showing that the combined effect of HSV-1 and HHV-8 is more likely to be breast cancer than fibroadenoma. In this cluster, other prerequisites were EBV(+), CMV(-). Labecque et al. and Bonnet et al. , who detected the presence of EBV by PCR, found the virus more frequently in malignant tumors than in non-malignant tumors. Disputes surrounding some of these studies [Gopalkrishna et al,1996. Chu et al,1998] may be due to the focus on the relationship of a single virus with breast cancer. In Table 3, cluster 3 shows EBV(+), CMV(+), HHV-8(-) events. Late expo-sure to a common virus, such as human CMV [Richardson,1997], or delayed exposure to EBV [Yasui et al,2001] has been suggested as a risk factor of breast cancer. In Table 6, the individual effect of HHV-8(+) suggests a strongly protective effect on progression from fibroadenoma to breast cancer (OR = 0.12, 95%CI = 0.04–0.38, P < 0.001). These pre- requisites resulted in a nearly 97.00% probability of breast cancer. Cluster 4 of Table 3 shows that HSV-1(+), EBV(-), HHV-8(-) was a prerequisite. In Table 8, which compares HSV-1(+), HHV-8(-) to HSV-1(+), HHV-8(+), there is OR = 25.00, 95%CI = 5.92–105.58, P < 0.001 for the combined effect of HSV-1 and HHV-8 with the greater likelihood of breast cancer than fibroadenoma. Beside these factors, EBV as a negative condition showed a nearly 98.00% possibility of breast cancer. Bonnet et al., who detected EBV by PCR, found EBV was detected more frequently in breast tumors that were hormone-receptor negative (P = 0.01). However, it may still be necessary in the study of breast tumors to investigate hormonal influences.Moore et al. indicated that a virus was oncogenic in the estrogen milieu of female mice of a strain with genetic susceptibility to mammary tumors.

There is experimental evidence that insulin, glucocorticoids, estrogen, and progestins synergize with MMTV in genetically susceptible female mice to cause mammary cancer [McGrath et al,1978]. However, insulin and glucocor-ticoids are physiologically required in certain amounts and consistently over time. Therefore, there is particular importance in estrogen and progesterone and their correlation to breast cancer [McGrath et al,1978]. Mastopathia cystica (fibrocystic disease, mammary dysplasia) was induced by DMBA in neonatally androgenized female Spraque-Dawley rats [Yoshida, 1994]. In these androgenized rats, no corpora lutea were found in the ovaries. The rat mastopathia cystica condition varied widely, with two kinds of macroscopically detectable tumor-forming lesions (solid and cystic). The development of rat mastopathia cystica was dependent on estrogen [Yoshida, 1994]. Although mammary car-cinoma occurred more frequently in neonatally non-androgenized rats than did fibroadenoma [Yoshida, 1994. Yoshida et al,1980], it seems even mammary tissues induced by the same carcinogen require a specific hormonal state to result in a benign or malignant condition.

In Table 4, the prerequisite was negative status for HSV-1, EBV, and HHV-8. For the three combinations either CMV or HPV, neither CMV nor HPV, or both CMV and HPV in the absence of any of the first three viruses, carcinogenic potency was not linked absolutely to a malignant or benign tumor. Different studies of HPV in breast cancer present conflicting results [Gopalkrishna et al,1996. Morimoto et al,1999]. Yu et al. [Morimoto et al,1999], who detected HPV-33 by PCR, suggested that it may be involved in human breast cancer. However, the positive rate of HPV-33 in China was 43.75% (14/32), while in Japan it was only 8.33% (1/12). Richardson et al. hypothesized that some breast cancers might be caused by late exposure to a common virus such as CMV. According Richardson's hypothesis, in breast cancer exposure to the virus could be a ''hit,'' while other ''hits'' [Knudson,1971] could be genetic susceptibility (such as an inherited mutation of p53) or uninterrupted exposure to a combination of estrogen and progesterone. Therefore, absent an enhancing factor (HSV-1, EBV, HHV-8), the incidence of breast cancer in this group was only 60.00%.

In Table 5, there is no prerequisite: Random virus infection in the mammary tissue produces an incidence of either breast cancer or fibroadenoma that is almost the same (50.00%). However, despite lack of specificity in resulting pathology, viruses are clearly a tumorigenesis factor in the mammary gland. Endogenous estrogens are central to the etiology of breast cancer [Adami et al,1998] because in the absence of estrogens breast cancer does not occur. A recent prospective study of Japanese women indicated that levels of serum estrogens were positively correlated with risk of breast cancer. In humans, there are strong associations between dietary pattern and level of circulating estrogen, with energy-rich diets correlated with high circulating estrogen levels [Kabuto et al,1998].Well-conducted case–control and ecological studies in populations with a low risk of breast cancer such asthose in China, Japan, and Indonesia, have shown that the risk of breast cancer is up to seven times higher in women who consume the highest level of fats and energy within those populations [Goldin et al,1986. Hirayama,1978]. Chen and Liaw [2002], who conducted an ecological study of dietary fat intake and mortality rates from breast cancer and colorectal cancer in Taiwan, found a positive correlation between fat and both breast cancer and colorectal cancer. Lawson et al. [2001] hypothesized that viruses such as HPV and EBV act as cofactors with diet, estrogen, and other hormones in initiation and promotion of some types of breast cancer in genetically susceptible women.

Triple negative breast cancers have more aggressive clinical course than other forms of breast cancer and the incidence was 10-15% [Sorlie et al,200 and 2003. Iwase and Yamamoto,2008]. In the current study, the viral prereguisites for breast carcinogenesis almost showed the single virus-infected events and the incidence was two folds(27.3%) of previous studies (data not shown). In the author unpublished data showed that the overall and relapse – free survivals of multiple (more than two) virus – infected breast cancer patients group is significantly better than the no virus – or one virus infected group. It suggest that current method only detect the aggressive factors of the viral prereguisites for breast carcinogenesis.

HSV-1				
Negative (-)	12	54	1	
Positive (+)	20	8	0.09 (0.03–0.25)	P < 0.001
HHV-8				
Negative (-)	4	34	1	
Positive (+)	28	28	0.12 (0.04–0.38)	P < 0.001
HSV-1(+) HHV-8(+)	20	3	1	
HSV-1(-) HHV-8(+)	8	25	20.83 (4.88–88.87)	P < 0.001
HSV1(+) HHV-8(+)	20	3	1	
HSV1(+) HHV-8(-)	8	30	25.00 (5.92–105.58)	P < 0.001
HSV1(-) HHV-8(-)	4	29	48.33 (9.75–239.65)	P < 0.001

Table 8. Odds ratios and P-values from sample data (breast cancer and fibroadenoma)

Variable	Overall survival	Relapse-free survival
	P value	P value
V0 vs. V1 vs. V2 vs. V3 vs. V4	0.001*	＜0.001*
V0 vs. V(1, 2) vs. V(3, 4)	0.359	0.189
V0 vs. V(1, 2, 3) vs. V4	0.645	0.598
V0 vs. V(1, 2, 3, 4)	0.515	0.979
V(0,1) vs. V2 vs. V3 vs. V4	0.013*	＜0.001*
V(0, 1) vs. V2 vs. V(3, 4)	0.005*	＜0.001*
V(0, 1) vs. V(2, 3) vs. V4	0.005*	＜0.001*
V(0, 1) vs. V(2, 3, 4)	0.001*	＜0.001*
V(0, 1, 2) vs. V3 vs. V4	0.543	0.187
V(0, 1, 2) vs. V(3, 4)	0.288	0.078
V(0, 1, 2, 3) vs. V4	0.539	0.312

Table 9. Results of log-rank analysis for overall and relapse-free survival

4. Conclusion

In previous research on the relationships among DNA viruses and breast cancer, fibroadenoma, and normal mammary tissue, only statistical analysis was used. However, when using statistical methods to classify DNA viruses and predict their probability in breast cancer, fibroadenoma, and normal mammary tissue,the assumption of statistics is necessary. Thus, statistics may become inappropriate to deal with problems of prediction, classification, and cluster in this study setting. In order to overcome this difficulty, our approach used an ANN and AHCTs to achieve the research objective.

Our findings suggest that in Taiwan, at least, the viral prerequisites of HSV-1(-); EBV(-), HHV-8(+);HSV-1(-), EBV(+), CMV(-), HHV-8(-); EBV(+), CMV(+,) HHV-8(-); and HSV-1(+); EBV(-), HHV-8(-) have a role in breast carcinogenesis. HSV-1(+) and HHV-8(+) have a strongly protective effect on progression from fibroadenoma to breast cancer. ANN and AHCTs seems to be a contributory method for detecting aggressive viral factors but not the better one. However, it is likely that infection by multiple viruses is important in development of either benign or malignant breast tumors. Further investigation is required to clarify which oncogenic viruses have protective effects and to clarify correlation of viral factors and hormone status with development of benign and malignant breast tumors.

5. References

D. M. Parkin, P. Pisani, J. Ferlay, Global cancer statistics, Ca. Cancer. J. Clin. 49 (1999) 33–64.

P. S. Yang, T. L. Yang, C. L. Liu, C. W. Wu, C. Y. Shen, A case–control study of breast cancer in Taiwan – a low incidence area, Br. J. Cancer 75 (1997) 752–756.

J. H. Tsai, C. H. Tsai, M. H. Chang, S. J. Lin, F. L. Xu, C. H. Yang, Association of viral factors with non-familial breast cancer in Taiwan by comparison with non-cancerous, fibroadenoma, and thyroid tumor tissues, J. Med. Virol. 75 (2005) 276–281.

S. A. Xue, I. A. Lampert, J. S. Haldane, J. E. Bridger, B. E. Griffin, Epstein-Barr virus gene expression in human breast cancer: protagonist or passenger? Br. J. Cancer 89 (2003) 113–119.

F. Fina, S. Romain, L. H. Ouafik, J. Palmari, A. F. Ben, S. Benharkat, P. Bonnier, F. Spyratos, J. A. Foekens, C. Rose, M. Buisson, H. Gerard, M. O. Reymond, J. M. Seigneurin, P. M. Martin, Frequency and genome load of Epstein-Barr virus in 509 breast cancer from different geographical area, Br. J. Cancer 84 (2001) 783–790.

L. G. Labecque, D. M. Barnes, I. S. Fentiman, B. E. Griffin, Epstein-Barr virus in epithelial cell tumors: A breast cancer study, Cancer Res. 55 (1995) 39–45.

K. Horiuchi, K. Mishima, M. Ohasawa, K. Aozasa, Carcinoma of stomach and breast with lymphoid stroma: Localization of Epstein-Barr virus, J. Clin. Pathol. 47 (1994) 538–540.

C. G. Kleer, M. D. Tseng, D. E. Gutsch, R. A. Rochford, Z. Wu, L. K. Joynt, M. A. Helvie, T. Chang, K. L. Van Golen, S. D. Merajver, Detection of Epstein-Barr virus in rapid growing fibroadenomas of the breast in immunosuppressed hosts, Mod. Pathol. 15 (2002) 759–764.

B. G. Pogo, J. F. Holland, Possibilities of a viral etiology for human breast cancer. A review, Biol. Trace Elem. Res. 56 (1997) 131–142.

A. G. Knudson, Mutation and cancer: statistical study of retinoblastoma, Proc. Natl. Acad. Sci. USA 68 (1971) 820–823.

A. Richardson, Is breast cancer caused by late exposure to a common virus? Med. Hypotheses 48 (1997) 491–497.

S. K. Lau, Y.-Y. Chen, G. J. Berry, S. A. Yousem, Epstein-Barr virus infection is not associated with fibroadenomas of the breast in immunosuppressed patients after organ translation, Mod. Pathol. 16 (12) (2003) 1242–1247.

R. P. Lippmann, An introduction to computing with neural nets, IEEE ASSP Mag. (1987) 4–12. April.

J. E. Dayhoff, Neural Network Architecture, Van Nostrand Reinhold, New York, 1990.

K. Funahashi, On the approximate realization of continuous mappings by neural networks, Neural Networks 2 (1989) 183–192.

G. Salton, Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley, Reading, MA, 1989.

H. Bunke, K. Shearer, A graph distance metric based on the maximal common subgraph, Pattern Recognit. Lett. 19 (1998) 255–259.

M. L. Fernandez, G. Valiente, A graph distance metric combing maximum common subgraph and minimum common supergraph, Pattern Recognit. Lett. 22 (2001) 753–758.

W. D. Wallis, P. Shoubridge, M. Kraetz, D. Ray, Graph distances using graph union, Pattern Recognit. Lett. 22 (2001) 701–704.

V. Brower, Accidental passenger or perpetrators? Current virus-cancer research, J. Natl. Cancer Inst. 96 (2004) 257–258.

V. Gopalkrishna, U. R. Singh, P. Sodhani, J. K. Sharma, S. T. Hedau, A. K Mandal, B. C. Das, Absence of human papillomavius DNA in breast cancer are revealed by polymerase chain reaction, Breast Cancer Res. Treat. 39 (1996) 197–202.

J. S. Chu, C. C. Chen, K. J. Chang, In situ detection of Epstein-Barr virus in breast cancer, Cancer Lett. 124 (1998) 53–57.

S. A. McCall, J. H. Lichy, K. E. Bijwaard, N. S. Aguilera, W. S. Chu, J. K. Taubenberger, Epstein-Barr virus detection in ductal carcinoma of breast, J. Natl. Cancer Inst. 93 (2001) 48–150.

J. S. Lawson, D. Tran, W. D. Rawlinson, From Bittner to Barr: a viral, diet and hormone breast cancer aetiology hypothesis, Breast Cancer Res. 3 (2001) 81–85.

J. A. Mckeating, P. D. Griffith, R. A. Weiss, HIV susceptibility conferred to human fibroblasts by Cytomegalovirus-induced FC receptor, Nature 343 (1990) 659–661.

B. J. Biegalke, A. P. Geballe, Sequence requirements for activation of the HIV-1 LTR by human cytomegalovirus, Virology 183 (1991) 381–385.

M. K. Gandhi, R. Khanne, Human Cytomegalovirus: clinical aspects, immune regulation, and emerging treatment, Lancet Infect. Dis. 4 (2004) 725–738.

W. G. Qing, C. J. Conti, M. LaBati, D. Johnston, T. J. Slaga, M. C. MacLeod, Induction of mammary cancer and lymphoma bymultiple, low oral doses of 7.12-dimethylbenz(a)anthrance in SENCAR mice, Carcinogenesis 18 (1997) 553–559.

M. Bonnet, J. M. Guinebretiere, E. Kremmer, V. Grunewald, E. Benhamon, G. Contesso, I. Joab, Detection of Epstein-Barr virus in invasive breast cancer, J. Natl. Cancer Inst. 91 (1999) 1376–1381.

Y. Yasui, J. D. Potter, J. L. Stanford, M.A. Rossing, M. D. Winget, M. Bronner, J. Daling, Hypothesis: breast cancer risk and ‘‘delay’’ primary Epstein-Barr virus infection, Cancer Epidemiol. Biomark. Prevent. 10 (2001) 9–16.

D. H. Moore, J. Charney, B. Kramarsky, E. Y. Lasbraques, N. H. Sarkar, M.J. Brennan, J. H. Burrows, S. M. Sirsat, J. C. Paymaster, A. B. Vaidya, Search for a human breast cancer virus, Nature 229 (1971) 611–615.

C. M. McGrath, R.F. Jones, Hormonal induction of mammary tumor viruses and its implication for carcinogen, Cancer Res. 38 (1978) 4112–4125.

H. Yoshida, Experimental study of pathogenesis of mastopathia cystica, Jpn. J. Breast Cancer 9 (1994) 185–193 (Japanese withEnglish summary).

H. Yoshida, R. Fukunish, Y. Kato, K. Matsumoto, Progesterone-stimulated growth of mammary carcinomas induced by 7,12-dimethylbenz(a)anthracene in neonatally androgenized rats, JNCI 65 (1980) 823–828.

Y. Yu, T. Morimoto, M. Sasa, K. Okazaki, Y. Harada, T. Fujiwara, Y. Irie, E. Takahashi, A. Tanigami, K. Izumi, HPV33DNA in premalignant and malignant breast lesions in Chinese and Japanese population, Anticancer Res. 19 (1999) 5057–5062.

H.-O. Adami, L.B. Signorello, D. Trichopoulos, Toward an understanding of breast cancer etiology, Cancer Biol. Semin. 8 (1998) 255–262.

M. Kabuto, S. Akiba, R.G. Stevens, K. Neriishi, C.E. Land, A prospective study of estradiol and breast cancer in Japanese women, Cancer Epidemiol. Biomark. Prev. 9 (2000) 575–579.

B. R. Goldin, H. Aldercreutz, S.L. Gorbach, M. N. Woods, J. T. Dwyer, T. Conlon, E. Bohn, S. N. Gershoff, The relationship between estrogen levels and diets of Caucasian American and Oriental immigration women, Am. J. Clin. Nutr. 44 (1986) 945–953.

J. M. Yuan, O. S. Wang, R.K. Ross, B. E. Henderson, M. C. Yu, Diet and breast cancer in Shanghai and Tianjin, China, Br. J. Cancer 71 (1995) 1353–1358.

T. Hirayama, Epidemiology of breast cancer with special reference to the role of diet, Prev. Med. 7 (1978) 173–195.

K. Wakai, D. S. Dillon, Y. Ohno, J. Prihartono, S. Budiningsih, M. Ramli, I. Darwis, D. Tjindarbumi, G. Tjahjadi, E. Soestrisno, E.S. Roostini, G. Sakamoto, S. Herman, S. Cornain, Fat intake and breast cancer risk in an area where fat intake is low: a case control study in Indonesia, Int. J. Epidemiol. 29 (2000) 20–28.

K.-J. Chen, Y.-P. Liaw, An ecological study of dietary fat intake and mortality rates from breast cancer and colorectal cancer in Taiwanese women, Nutr. Sci. J. 27 (2002) 202–210 (Chinese with English summary).

J. H. Tsai, C. S. Hsu, C. H. Tsai, J. M. Su, Y. T. Liu, M. H. Cheng, J. C. C. Wei, F.L hen, C. C. Yang: Relationship between viral factors , axillary lymph node status and survival in breast cancer. J Cancer Res Clin Oncol (2007)133:13-21

T. Sorlie, C . M Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M.B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lonning, and A-L Borresen-Dale: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.: Proc Natl Acad Sci USA 98(2001)10869-10874.

T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, ANobel, S. Deng, H.Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lonning, P. O. Brown, A-L. Borresen-Dale, and D. Botstrin. Repeated observation of breast tumor subtypes in independent gene expression data sets : Proc Natl Acad Sci 100(2003)8418-8423

H. I wase and Y. Yamamoto: Biological characteristico of triple negative breast cancer. Jpn J Breast Cancer 23(2008)75-80

Data Mining in Medical and Biological Research
INTECH