A neural network-based biomarker association information extraction approach for cancer classification

Hong-Qiang Wang a,*, Hau-San Wong a, Hailong Zhu b, Timothy T.C. Yip c

a Department of Computer Science, City University of Hong Kong, Hong Kong, China
b Department of RIPT, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China
c Department of Clinical Oncology, Queen Elizabeth Hospital, Hong Kong, China

ARTICLE INFO

Article history:
Received 18 August 2008
Available online 6 January 2009

Keywords:
Cancer classification
High-throughput technology
Neural network
Biomarker Association Network

ABSTRACT

A number of different approaches based on high-throughput data have been developed for cancer classification. However, these methods often ignore the underlying correlation between the expression levels of different biomarkers which are related to cancer. From a biological viewpoint, the modeling of these abnormal associations between biomarkers will play an important role in cancer classification. In this paper, we propose an approach based on the concept of Biomarker Association Networks (BAN) for cancer classification. The BAN is modeled as a neural network, which can capture the associations between the biomarkers by minimizing an energy function. Based on the BAN, a new cancer classification approach is developed. We validate the proposed approach on four publicly available biomarker expression datasets. The derived Biomarker Association Networks are observed to be significantly different for different cancer types. Extensive comparisons show the superior performance of the BAN-based classification approach over several conventional classification methods.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

High-throughput biological technology can simultaneously assess the levels of expression of tens of thousands of biomarkers in tumors, and it was demonstrated that it is capable of providing more reliable cancer classification owing to the resulting more complete molecular understanding of tumors [1–5]. Currently, developing efficient computational methods based on high-throughput technology for cancer classification is an important and challenging task in pattern recognition and bioinformatics [6–8].

There have been various methods developed to address this issue [9–11]. These methods are in general data-driven and depend on a linear or non-linear discriminant function. Several representative examples of linear models include the compound covariate method proposed by Hedenfalk et al. [12], the shrunken centroid method by Tibshirani et al. [13], and the optimized linear model using the partial least square (PLS) technique [11]. Pochet et al. assessed the effectiveness of non-linearity in cancer classification, and the resulting conclusion is that the non-linear techniques tend to outperform the linear techniques [14]. The applications of a number of non-linear methods including k-nearest neighborhood [6], logistic models [15,16], and factor mixture models [17] to cancer classification further confirmed the conclusion. A number of neural network techniques have also been developed and applied to cancer classification [10,18–20]. However, due to the non-typical “high-dimension and small sample” property of micro-array data, many neural network classifiers exhibit poor generalization on unknown data in cancer classification problems [21,22]. In general, the overfitting problem tends to occur in the neural network-based approaches, which were previously designed for the case where there are a large number of training samples available in a low-dimensional space. In contrast, the support vector machine (SVM) technique motivated by statistical learning theory can achieve a better performance for data of high dimensionalities and small sample sizes [23–26]. A number of investigators have developed and applied various kinds of SVMs to the analysis of gene expression data [9,27–30]. However, as a data-driven model, SVM cannot interpret the differences between the various cancer types from a biological viewpoint, and cannot help to retrieve the related biological information. In recent years, as an alternative method, model-driven cancer classification methods have received more and more attention [31–33]. For example, a number of investigators have proposed to model the association relationships of genes for cancer classification [31,34,35]. Antonov et al. assumed that expression levels of a set of pre-selected genes are subject to multiple positive and negative correlational relationships with each other [31]. As a result, by adopting a weighted sum
of the logarithms of the expression levels, the authors proposed a classification model based on biomarker association information. A main disadvantage of the model is that the same model parameters are used for all cancer classes, such that biomarker association patterns characteristic to each class cannot be identified in an efficient way [36].

From a biological viewpoint, there may exist underlying abnormal biomarker association patterns which are responsible for a cancer or cancer subtype [37–39]. In this paper, we propose to construct a network model, referred to as Biomarker Association Network (BAN), and apply it to cancer classification. The BAN is different from the concept of a general regulatory network. The regulatory network aims to understand the cell development process of organisms, and is modeled using a time series of biomarker expression data. For these networks, the following modeling tools are often used: Boolean networks [40], Bayesian networks and dynamic Bayesian networks [39,41], or the differential equation technique [42,43]. We employ neural network theory to model associations between biomarkers. A large number of methods have been developed to detect non-trivial changes related to biological responses in biomarker expression, but most of them have limited effectiveness due to little or no consideration of biomarker association information. Cellular processes often affect sets of biomarkers acting in concert. The main focus of our proposed approach is to detect the associations between the biomarkers, which exhibit more subtle but consistent changes related to particular cancer types. To our knowledge, few works have been done on using the network concept for cancer classification because it is not easy to define the relationship between a network and a single cancer sample, and determine the class of a sample by the network.

We model BAN as a fully connected neural network to capture the biomarker association patterns related to a particular type of cancer: network nodes represent biomarkers and the connection weights represent association coefficients between biomarkers. The input of the BAN denotes the observed expression levels of biomarkers, while the output denotes the estimated levels by the association patterns. In view of this objective, we define an energy function as a measure of the disagreement between the input and output of the BAN. The definition of the energy function enables the BAN to efficiently capture the associations between biomarkers, and a classification criterion can be easily designed for a network-based classifier. A novel cancer classification framework is then formulated by integrating the BANs of different cancer classes by a competition mechanism, in which a test sample is classified into the class whose BAN has the lowest energy.

We validate our proposed approach on four real-world datasets including 1 protein expression dataset, nasopharyngeal carcinoma (NPC) [44], and three gene expression datasets, leukemia [3], colon [45] and breast [12]. Experimental results show the excellent discriminative power of the BAN-based classifiers as well as the effectiveness and efficiency of the BANs in capturing the biomarker association patterns responsible for a cancer type or subtype. Several conventional classification methods including Fisher discriminant analysis (FDA), k-nearest neighborhood (KNN), Bayesian network (BN) classifier, support vector machines with linear kernel (linear-SVM) and radial basis function kernel (rbf-SVM), are compared with our approach, and the comparison results show that the BAN classifiers have better performance than the conventional methods.

2. Methods

2.1. The framework of the algorithm

The aim of our algorithm is to identify the association patterns between a small set of biomarkers to facilitate cancer classification. Although there are a large number of dynamic variables coming into play in a biological system, not all take part in a particular biological process [46–48]. We need to restrict the analysis to a subset of core biomarkers for a cancer class. In view of this, we determine key biomarkers from tens of thousands of biomarkers in high-throughput expression data by considering two factors: single biomarker correlation degree and combinative performance. Following the former, we first choose a subset of candidate biomarkers based on two single-variable selection rules, signal-to-noise-ratio (SNR) [3] and regulation probability (RP) [49], and then determine a best combination from the subset according to their cross-validation performances for constructing Biomarker Association Networks.

For a biomarker, the SNR and RP criteria are, respectively, determined as follows:

\[
R_{\text{SNR}} = \frac{\left| \mu_1 - \mu_2 \right|}{\sigma_1 + \sigma_2}
\]

\[
R_{\text{RP}} = \frac{1}{l} \left| (l_d - l_u) \right|
\]

where \(l_d\) and \(l_u\) represent the number of training samples in which the biomarker is up-regulated, and \(l_u\) represents the number of training samples in which the biomarker is down-regulated. We estimate \(l_d\) and \(l_u\) using a maximum likelihood estimation approach [49]. The two criteria are integrated to calculate a combined rank for each biomarker to select the subset of biomarkers. Specifically, the combined ranks are calculated by averaging the two ranks associated with the two criteria. The SNR and RP criteria focus on different aspects: the former provides a quantitative measure of the difference between cancer classes based on their mean expression levels, while the latter assesses the probability of differential expression between the classes. The integration of the two criteria can more efficiently filter out irrelevant biomarkers which are not essential to cancer classification.

To overcome the limitations of the above single-variable selection, we further employ biomarker combinative performances to select the best combination from the subset obtained above. The combinative performance is estimated based on performing stratified k-fold cross-validation on the data sets, and the following error estimation is used:

\[
\text{Accu} = 1 - \frac{1}{G} \sum_{g=1}^{G} \frac{t_g}{n_g}
\]

where \(t_g\) is the number of correctly classified samples of the \(g\)th class, \(n_g\) is the total number of samples in the \(g\)th class, and \(G\) is the number of classes. This error criterion can overcome the problems of small sample and sample imbalance in the biomarker expression data. In the stratified cross-validation process, samples of each class are randomly divided into \(k\) equal subsets. One of the subsets from each class is chosen, and these are combined to construct a holdout test set, while the remaining data are used for training. This process is repeated, and the mean error is used to quantify the classification performance. Finally, the best combination with the lowest mean error is selected, and based on this combination, the proposed algorithm constructs BAN for each cancer class to perform cancer classification. Fig. 1 shows the framework of the proposed algorithm.

2.2. Linear Biomarker Association Network (LN)

We model the linear Biomarker Association Network (LN) as a fully connected neural network, as shown in Fig. 2. The underlying
assumption is that the expression level of a biomarker is associated with some other biomarkers and can be estimated using a linear combination. Consider a set of \( p \) biomarkers. The input of the LN represents the observed levels of the \( p \) biomarkers, denoted as \( \mathbf{x} = [x_1, x_2, ..., x_p]^T \), and the output represents the corresponding estimated results, denoted as \( \mathbf{y} = [y_1, y_2, ..., y_p]^T \). Let \( A = \{a_{ij} \in R, a_{ij} = 0: i, j = 1, 2, ..., p\} \) denotes the connection matrix of the network, the operation of the LN can be represented as follows:

\[
y = Ax + B, \quad B = [b_1, b_2, ..., b_p]^T
\]

where \( b, i = 1, 2, ..., p \), represents the expression baseline of the \( i \)th biomarker. In addition, \( A \) is referred to as the association matrix, of which each element \( a_{ij} \), referred to as association coefficient, represents the expression effect of biomarker \( i \) on biomarker \( j \). The coefficient can be positive or negative: a positive value indicates expression promotion, and a negative value indicates expression repression. The constraint \( a_{ii} = 0 \) is imposed based on the assumption that there is no self-association. The expression baselines can be viewed as background expression levels that compensate for the absence of self-association. The structure of the LN is thus related to that of the auto-associative neural network (refer to [50]).

To analyze a BAN, an energy function, denoted by \( E \), is defined as a measure of the disagreement between the input and output of the network, i.e.,

\[
E = \frac{1}{2}(\mathbf{y} - \mathbf{x})^T(\mathbf{y} - \mathbf{x}) \tag{5}
\]

which characterizes the dissimilarity between the observed levels and the estimated levels. Further, substituting Eq. (4) into Eq. (5) the energy function can be rewritten as

\[
E = \frac{1}{2}(\mathbf{Ax} + \mathbf{B})^T(\mathbf{Ax} + \mathbf{B}) \tag{6}
\]

From Eq. (6) the energy function only depends on the input of the BAN and can be used to characterize biomarker association patterns. If a low energy state is maintained for a specific cancer type, it is hypothesized that the BAN characterizes an association pattern responsible for that cancer type and encapsulates it in the parameters \( A \) and \( B \). In what follows, an algorithm is developed to obtain a BAN associated with a particular cancer group.

Let \( X = [x_1, x_2, ..., x_l] \) denote the set of \( l \) known samples of a cancer class, the association coefficients can be determined through the minimization of the total energy of the BAN for all \( l \) samples, i.e., by solving the following optimization problem:

\[
\text{Minimize } f = \frac{1}{2} \sum_{j=1}^{l} (\mathbf{A} - \mathbf{I})x_j^T (\mathbf{A} - \mathbf{I})x_j + \mathbf{B}^T (\mathbf{A} - \mathbf{I})x_j + B
\]

s.t. \( a_{ii} = 0, i = 1, 2, ..., p \)

where \( l \) is the identity matrix. Let \( \mathbf{A} = (\mathbf{A} - \mathbf{I}) \), the optimization problem is reduced to

---

**Fig. 1.** Framework of the algorithm.

**Fig. 2.** Biomarker Association Network (BAN).
Minimize \( f = \frac{1}{2} \sum_{x} (\tilde{A}x + B)^T (\tilde{A}x + B) \)

s.t. \( \tilde{a}_i = -1, i = 1, \ldots, p \)

By rewriting \( \tilde{A} \) as \( [\tilde{A}_1; \tilde{A}_2; \ldots; \tilde{A}_p] \), we expand the objective function as follows:

\[
f = \frac{1}{2}(\tilde{A}_1X + b_1e)(\tilde{A}_1X + b_1e)^T + \frac{1}{2}(\tilde{A}_2X + b_2e)(\tilde{A}_2X + b_2e)^T + \cdots + \frac{1}{2}(\tilde{A}_pX + b_pe)(\tilde{A}_pX + b_pe)^T
\]

where \( e \) is an \( I \)-dimensional row vector consisting of \( I \) 1s. As far as the relationships between biomarkers are mainly concerned, \( B \) is not crucial to the BAN. In fact, by normalizing the expression data to a mean of zero and a variance of one, \( B \) can be reduced to zero. As a result, for computational convenience, we set \( B = 0 \). The objective function thus reduces to

\[
f = \frac{1}{2} (\tilde{A}_1X)(\tilde{A}_1X)^T + \frac{1}{2} (\tilde{A}_2X)(\tilde{A}_2X)^T + \cdots + \frac{1}{2} (\tilde{A}_pX)(\tilde{A}_pX)^T
\]

(10)

Further, the function can be rewritten as

\[
f = \frac{1}{2}(U_1 - Z_1\phi_1)^T(U_1 - Z_1\phi_1) + \frac{1}{2}(U_2 - Z_2\phi_2)^T(U_2 - Z_2\phi_2) + \cdots + \frac{1}{2}(U_p - Z_p\phi_p)^T(U_p - Z_p\phi_p)
\]

(11)

where \( Z_i = [x_{i1}; i = 1, \ldots, p; j = 1, \ldots, n] \), \( U_i = [x_{i1}; x_{i2}; \ldots; x_{in}] \), and \( \phi_i = [a_{i1}, a_{i2}, \ldots, a_{i(n+1)}, \ldots, a_{ip}] \). By taking the derivative with respect to \( \phi_i \), we obtain

\[
\frac{df}{d\phi_i} = -Z_i^T U_i + Z_i^T Z_i \phi_i = 0, i = 1, 2, \ldots, p
\]

(12)

and the association coefficients can be obtained as

\[
\phi_i = (Z_i^T Z_i)^{-1}Z_i^T U_i, i = 1, 2, \ldots, p
\]

(13)

If the inverse of \( Z_i^T Z_i \) does not exist, the pseudo-inverse can be used to compute the solution as in [51].

2.3. Non-linear Biomarker Association Network (NLN)

Non-linearity has been shown to provide a better capability to explore the complex relationships between variables as well as to efficiently reduce noise [52, 51]. In view of this, a non-linear BAN (NLN) is developed to characterize biomarker association patterns. The NLN can be constructed by adding a sigmoid transformation unit to the LN. In general, other non-linear transformations can also be used. Let \( x = [x_1, x_2, \ldots, x_n]^T \) denotes the input sample, the corresponding output \( v = [v_1, v_2, \ldots, v_p]^T \) is determined as follows:

\[
v_i = \text{Sigmoid}(x_i) = \left(1 + e^{-\beta x_i}\right)^{-1}, i = 1, 2, \ldots, p
\]

(14)

where \( \mu_i \) and \( \sigma_i \) represent the mean and standard deviation of the expression level of biomarker \( i \), respectively, and can be estimated by training samples. In addition, \( \beta \in [0, 1] \) is a tunable parameter, referred to as the sigmoid coefficient. The transformed vector \( v \) is then used as the input to a LN to form a NLN. The energy function of the NLN can be written as

\[
E = \frac{1}{2} (Av + v + B)(Av + v + B)
\]

(15)

The NLN can be optimized by choosing a proper sigmoid coefficient.

2.4. Cancer classification based on BANs

Considering a \( C \)-class cancer classification problem, the BAN-based classifier can be designed by combining the \( C \) BANs, as shown in Fig. 3. Given an unknown sample \( x \), the system first computes the energy of each BAN \( E_c, c = 1, \ldots, C \), and then a winner-takes-all competition mechanism is applied to determine the BAN with the minimum energy as the winner. Finally, the system outputs a \( C \)-dimensional label vector \( L \) consisting of \( (C - 1) \) 0-elements and one 1-element indicating the predicted class. The idea behind the cancer classification system is that distinct types of cancer may be due to different biomarker association patterns.

Fig. 4 provides a geometric interpretation of the BAN-based system based on real gene expression data (the leukemia data). Fig. 4(A) shows the distributions of samples of the two classes in the expression space based on three optimally selected genes. From this figure, it can be seen that the two classes are not well separated. However, they can be completely separated in the error space, in which each sample is represented in terms of the sum of its fitting errors with respect to the association patterns of different classes, and the energy space, in which the samples are represented in terms of the squared sum of the fitting errors, as shown in Fig. 4(B) and (C). This result is due to that the class difference is mainly characterized by hidden association patterns, instead of the original biomarker expression levels. In particular, the margin between the two classes is further enlarged in the energy space when compared to that in the error space, as energy is calculated as a square function of the fitting errors. Compared with the error space, the energy space is introduced to highlight the hypothesis that, rather than performing a simple additive combination of multiple association patterns, the introduction of non-linear operations will facilitate the characterization of the interaction between the different biomarkers. In addition to the optimal genes, we also illustrate the discriminatory power of the energy function based on sub-optimal genes, as shown in Fig. 5. Fig. 5(A) indicates that the three genes do not lead to a good separation of the two classes in the expression space. However, from Fig. 5(B) in which the samples are represented in the energy space, it can be seen that the classes can be distinguished to a certain extent. This figure provides further support to the capability of our proposed energy function to detect differential biomarker association patterns between various cancer types to facilitate classification. The smaller degree of improvement in the classification performance is due to the sub-optimality of the three selected genes in the current case, which highlights the necessity of the biomarker selection process.

To evaluate the classification system, we propose a prediction strength (PS) index. Consider a binary case. According to the above description, given a sample, the system will compute two energy values, \( E_{\text{min}} \) and \( E_{\text{iso}} \), which correspond to the values of the winning faction.
Fig. 4. Geometric interpretation of the BAN-based network classifier. (A) The sample distribution in the expression space based on three optimally selected genes. (B) The sample distribution in the error space. (C) The sample distributions in the energy space. Note that subfigures (B) and (C) use the same ranges and scales for the coordinate axes.

Fig. 5. The discrimination power of the energy function based on sub-optimal biomarkers. (A) The sample distribution in the expression space based on gene expression levels. (B) The sample distribution in the energy space based on the energy function. The pink line in (B) indicates a possible class boundary. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
and losing BANs, respectively. Based on the two energy values, the PS index for the predicted sample is computed as follows:

$$ps = \frac{E_{\text{line}} - E_{\text{win}}}{E_{\text{win}} + E_{\text{line}}}$$

(16)

The PS index ranges from 0 to 1 and reflects the relative classification margin. Ideally, $E_{\text{line}}$ is equal to zero, and the maximum value of 1 for the prediction strength can be attained.

3. Experimental results

We evaluate our approach on one protein expression dataset, Nasopharyngeal carcinoma (NPC) [44], and three gene expression datasets, leukemia[3], colon [45] and breast [12]. In the NPC dataset, each patient sample consists of 530 spectra peaks and is labeled as chemo-responders (RS) or non-responders (NR). The total 54 samples (44 chemo-responders and 10 non-responders) are allocated to the training set and testing set, with each having 27 samples [44]. The leukemia dataset has 72 samples, of which 47 correspond to acute lymphoblastic leukemia (ALL) and 25 to acute myeloid leukemia (AML), each consisting of the expression levels of 7129 genes. The leukemia dataset is split into a training set of 38 samples and a test set of 34 samples as done in [3,31]. The colon dataset consists of 62 samples, of which 22 are normal and 40 tumor tissue, and each sample contains the expression levels of 2000 genes [45]. The breast dataset consists of multiple classes, and is used to test the multi-class classification performance of our approach. The total 22 samples in the data set are categorized into three classes: 7 with BRCA1 mutation, 8 with BRCA2 mutation and 7 sporadic breast cancers, and each sample is represented by the expression levels of 3226 genes [12]. The adopted training and testing methodology can be described as follows: for the datasets with pre-specified training/test split, we use the training set to select the best combination of biomarkers to construct BAN-based cancer classifiers, and the test set to evaluate the performance of the trained classifiers. Specifically, in the biomarker combination selection, the SNR-PR integration method is first applied to the whole training set to choose a subset of biomarkers, and then the training set is divided into $k$ folds to evaluate the cross-validation performance of each combination. For the datasets without pre-specified training/test split, we evaluate BAN-based classifiers using a leave-one-out cross-validation approach on the whole dataset. In each step, one sample is selected for testing, and the rest are used for training.

3.1. Application to the NPC data

For this data set, we first choose a panel of 30 protein biomarkers and search for the best biomarker combination from this panel. The search is performed by initially setting the number of biomarkers $p$ to 2, and then gradually increasing this number. The classification error of each combination is estimated in the following manner: we apply stratified 3-fold cross-validation, and repeat the process 10 times with different 3-fold splits of samples to obtain the mean errors. The stratified cross-validation guarantees that the original proportion of each class is consistently represented in both the training and test sets. In the cross-validation process, samples of each class are divided into three equal subsets. One subset from each class is chosen, and these are combined to construct a holdout test set, while the remaining are used for training. This operation ensures that both training and testing sets have the same ratio of samples between the different classes as that in the original data set. Considering the small number of samples in the expression data set, we set the number of folds to 3. The 10-time repetition of the cross-validation is used to provide a further reduction of the variance of the estimated errors. The two types of BANs, LN and NLN, are, respectively, constructed for the NPC data, and for the NLN case, the optimal sigmoid parameter $\beta$ is chosen from the set $\{0.0001,0.001,0.005,0.01,0.05,0.1,0.5\}$.

Fig. 6(A) shows the evolution curves of the minimum errors with $p = 2-6$. From this figure, it can be seen that the misclassification rates of both LN and NLN classifiers exhibit similar trends: they first decrease and then reach a minimum. It is also observed that the NLN classifiers attain better training accuracies than the LN classifiers. Table 1 presents the accuracies of these classifiers on the independent test set. It can be seen that the test accuracy values are consistent with the training ones shown in Fig. 6(A), irrespective of whether the LN or the NLN classifier is used, which indicates the good generalization performances of the BAN classifiers. From Fig. 6(A) and Table 1, it is also observed that the performance of these BAN-based classifiers steadily improves with more biomarkers used, indicating the positive contributions of the additional relationships captured by the Biomarker Association Network.

From Table 1, the best test accuracy based on our approach is 92.6%, which is achieved by the five-biomarker LN classifier, and the specificity and sensitivity are 100% and 80%, respectively. We then perform further analyses: first, we compute the prediction strengths (PS) for the correctly classified samples by Eq. (16), and illustrate them in the left-half panel of Fig. 7. From this figure, almost all of the PS values are significantly higher than 0.4, and the median PS (the horizontal bar) is close to 1 (0.99). The high PS values further confirm the reliability of the classifier. Table 2 lists the related biological information of the five protein biomarkers involved in the classifier [44], indicating that all the five biomarkers are related to the discrimination of NR and RS with significance levels less than 0.05. Fig. 8 compares the expression distributions of these biomarkers in the RS and NR groups, indicating that each of them has significantly different expression distributions across the two groups.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Datasets</th>
<th>$p = 2$</th>
<th>$p = 3$</th>
<th>$p = 4$</th>
<th>$p = 5$</th>
<th>$p = 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>LN classifier</td>
<td>NPC</td>
<td>0.741</td>
<td>0.778</td>
<td>0.815</td>
<td>0.889</td>
<td>0.815</td>
</tr>
<tr>
<td>Leukemia</td>
<td>0.735</td>
<td>0.735</td>
<td>0.794</td>
<td>0.853</td>
<td>0.941</td>
<td></td>
</tr>
<tr>
<td>NLN classifier</td>
<td>NPC</td>
<td>0.889</td>
<td>0.852</td>
<td>0.889</td>
<td>0.926</td>
<td>0.926</td>
</tr>
<tr>
<td>Leukemia</td>
<td>0.853</td>
<td>0.971</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Table 1
Classification performances of the LN and NLN classifiers on the NPC and leukemia data.
Next, the BANs in the classifier are analyzed to discover the difference of the captured association patterns of the two classes. Fig. 9 visually compares the two resulting NLNs. In Fig. 9(A) and (B), red and green lines between nodes represent positive and negative associations, respectively, and the widths of the lines indicate the association strength. From the two figures, it can be seen that the two NLNs have significantly different connection patterns, suggesting that the two groups exhibit different association patterns. For example, in the RS group, association coefficients between biomarkers M7993_48 and M2193_50 are positive (0.26 and 0.32), indicating mutual expression enhancement. However, in the NR group, the corresponding coefficients become negative (-0.89 and -0.34), which indicate mutual repression. Another significant difference is observed when comparing the biomarkers M7993_48 and M9209_62: their coefficients are very small (0.035, 0.025) in the RS group, suggesting that there is little or no association, while in the NR group, M7993_48 is positively correlated with M9209_62 by a very large coefficient (12.15). Furthermore, the two NLNs have different dominant biomarkers, as marked in the hexagon nodes in Fig. 9(A) and (B). Such dominant biomarkers are significantly correlated with all the other biomarkers, which, as a hub biomarker of the network, play a crucial role in cancer development [53]. The biomarkers having smaller association coefficients than any others are marked in the smaller circle nodes in Fig. 9(A) and (B), which may play only a minor role in the Association Networks. Similar to the expression profile, we can view an association coefficient as an association channel, and form an association spectrum for a cancer class by combining all the association channels to gain a more in-depth understanding of the association patterns associated with cancer. Fig. 9(C) compares the association levels on each association channel between the two classes in a bar diagram form, and Fig. 9(D) shows the association spectra of the two classes across the whole channel range. The two figures more clearly indicate the remarkable difference of the association patterns between the two classes.

We then compare the classification performance of the NLN classifier with those of conventional methods including Fisher discriminant analysis (FDA), k-nearest neighborhood (KNN), Bayesian network classifier (BN) and support vector machines (SVMs) with both linear kernels (linear-SVM) and radial basis function kernels (rbf-SVM). Among these conventional methods, the BN classifier, linear-SVM and rbf-SVM require their parameters to be tuned for best performance. For the BN classifier, we varied the bin number in the range {2, 3, 4, 5, 6}. For the linear-SVM, we varied the regularization parameter in the range $2^{-2}$, $2^{-1}$, ..., $2^{-12}$. For the rbf-SVM, a two-dimensional grid search technique was employed to optimize the two parameters, regularization factor and kernel width, with the search ranges of $2^{-2}$, $2^{-1}$, ..., $2^{-12}$ and $2^{-10}$, respectively. These methods were implemented based on the public-domain software WEKA [54].

### Table 2

The five protein biomarkers for the NPC data.

<table>
<thead>
<tr>
<th>Spectrum name</th>
<th>m/z value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>M9177_96</td>
<td>9177.96</td>
<td>2.3E-2</td>
</tr>
<tr>
<td>M9209_62</td>
<td>9209.62</td>
<td>3.1E-2</td>
</tr>
<tr>
<td>M9548_54</td>
<td>9548.54</td>
<td>2.5E-2</td>
</tr>
<tr>
<td>M7993_48</td>
<td>7993.48</td>
<td>4.8E-2</td>
</tr>
<tr>
<td>M2193_50</td>
<td>2193.50</td>
<td>3.0E-2</td>
</tr>
</tbody>
</table>

Fig. 7. Prediction strengths for the NPC and leukemia datasets. Median PS is denoted by a horizontal line for each dataset. Almost all of the samples have significantly higher PS values than 0.4.

Fig. 8. The expression levels of the five proteins in chemo-responders (RS) and non-responders (NR). The bars indicate mean values.
lists the classification results of these conventional classifiers on the independent test set with different numbers (p) of protein markers used. From the table, it can be seen that the non-linear methods tend to have better classification performances. When \( p = 2 \) or 3, the NLN classifier, the KNN and the rbf-SVM attain the same accuracy that is the highest among the results of all the methods. With \( p \) increasing, the performance of our NLN classifier further improves, while those of the KNN and rbf-SVM do not. When \( p = 5 \), the NLN classifier attains an accuracy of 92.6\%, which is significantly higher than the best accuracy of the rbf-SVM (88.9\%). In general, our NLN classifiers attain higher accuracies than the previous approaches, irrespective of the number of proteins used. The conventional methods are either linear, non-linear, or joint probabilistic distribution-based, and are widely used in bioinformatics and pattern recognition [9,32,25]. Compared with them, our approach has the unique capability of extracting the association patterns hidden in the expression data, which accounts for its better performance.

It is well-known that SVM tends to exhibit better performance for high-dimensional data sets. To explore the potential classification performance limit of the NPC dataset, we selected 10, 20, 50, 100 and 200 proteins by the RP and SNR criteria, and re-applied SVM to the NPC data. Table 4 presents the resulting test accuracies, indicating that the highest accuracy by SVM is lower than that (92.6\%) of our NLN classifier. The better discrimination power of our approach is due to the inclusion of the mutual association information that is not available in SVM. As for the BN classifier, although it is capable of characterizing the relationships of the diff-

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Methods</th>
<th>( p = 2 )</th>
<th>( p = 3 )</th>
<th>( p = 4 )</th>
<th>( p = 5 )</th>
<th>( p = 6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPC</td>
<td>NLN</td>
<td>0.889</td>
<td>0.852</td>
<td>0.889</td>
<td>0.926</td>
<td>0.926</td>
</tr>
<tr>
<td></td>
<td>FDA</td>
<td>0.889</td>
<td>0.815</td>
<td>0.741</td>
<td>0.704</td>
<td>0.815</td>
</tr>
<tr>
<td></td>
<td>KNN (k = 3)</td>
<td>0.889</td>
<td>0.852</td>
<td>0.852</td>
<td>0.778</td>
<td>0.889</td>
</tr>
<tr>
<td></td>
<td>Linear-SVM</td>
<td>0.815</td>
<td>0.815</td>
<td>0.852</td>
<td>0.852</td>
<td>0.852</td>
</tr>
<tr>
<td></td>
<td>rbf-SVM</td>
<td>0.889</td>
<td>0.852</td>
<td>0.852</td>
<td>0.889</td>
<td>0.889</td>
</tr>
<tr>
<td></td>
<td>BN</td>
<td>0.852</td>
<td>0.852</td>
<td>0.852</td>
<td>0.852</td>
<td>0.889</td>
</tr>
<tr>
<td>Leukemia</td>
<td>NLN</td>
<td>0.853</td>
<td>0.971</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>FDA</td>
<td>0.912</td>
<td>0.853</td>
<td>0.971</td>
<td>0.941</td>
<td>0.941</td>
</tr>
<tr>
<td></td>
<td>KNN (k = 3)</td>
<td>0.912</td>
<td>0.882</td>
<td>0.971</td>
<td>0.971</td>
<td>0.971</td>
</tr>
<tr>
<td></td>
<td>Linear-SVM</td>
<td>0.882</td>
<td>0.882</td>
<td>0.971</td>
<td>0.971</td>
<td>0.971</td>
</tr>
<tr>
<td></td>
<td>rbf-SVM</td>
<td>0.912</td>
<td>0.941</td>
<td>0.971</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>BN</td>
<td>0.852</td>
<td>0.941</td>
<td>0.971</td>
<td>0.852</td>
<td>0.839</td>
</tr>
<tr>
<td>Colon</td>
<td>NLN</td>
<td>0.903</td>
<td>0.919</td>
<td>0.936</td>
<td>0.936</td>
<td>0.936</td>
</tr>
<tr>
<td></td>
<td>FDA</td>
<td>0.839</td>
<td>0.839</td>
<td>0.839</td>
<td>0.839</td>
<td>0.839</td>
</tr>
<tr>
<td></td>
<td>KNN (k = 3)</td>
<td>0.806</td>
<td>0.806</td>
<td>0.855</td>
<td>0.855</td>
<td>0.855</td>
</tr>
<tr>
<td></td>
<td>Linear-SVM</td>
<td>0.8395</td>
<td>0.855</td>
<td>0.855</td>
<td>0.855</td>
<td>0.871</td>
</tr>
<tr>
<td></td>
<td>rbf-SVM</td>
<td>0.855</td>
<td>0.855</td>
<td>0.887</td>
<td>0.887</td>
<td>0.887</td>
</tr>
<tr>
<td></td>
<td>BN</td>
<td>0.806</td>
<td>0.855</td>
<td>0.823</td>
<td>0.806</td>
<td>0.839</td>
</tr>
</tbody>
</table>
The obtained results are compared with those reported by other researchers, indicating the remarkable difference of the association patterns between the classes.

### 4. Application to the breast data

The breast data set is about a three-class cancer classification problem, and is used here to evaluate the multi-class classification performance of our approach. Considering the small sample size (22) of the data set, we adopt LOOCV evaluation in this experiment. From this figure, it can be seen that the NLN classifiers using four or six gene biomarkers attained an accuracy of 100%, i.e., correctly classifying all the samples, which is better than the results of the other approaches. Fig. 12 also suggests that the accuracy of the NLN classifier steadily increases when more biomarkers are used, similar to the cases of the above three data sets. The successful classification of the breast dataset confirms the multi-class cancer classification capability of our BAN-based approach, and lends more support to the effectiveness of the association patterns captured by the BAN models. Based on the six-gene model, Fig. 11(B) shows the mean association levels and the mean association spectra for each class, and show them in Fig. 11(A). This figure indicates the remarkable difference of the association patterns of the two classes.

### 5. Influence of the non-linear transformation

From the above four applications, it can be seen that the NLN classifiers perform better than the LN ones. We hypothesize that the non-linear transformation in the NLN classifiers plays a crucial role in the improvement of performance. The sigmoid coefficient $\beta$ is an important parameter of the sigmoid function, which controls the non-linear transformation and can improve the separability of the original data by reducing noise [52,51]. To investigate the influence of this coefficient and possibly obtain some guidelines for its choice, we investigate the classification performance of the NLN classifier under different values of $\beta$. From Fig. 13, it is seen that a suitable range of $\beta$ is $[0.001, 0.1]$. 

### Table 4

<table>
<thead>
<tr>
<th>No. of proteins</th>
<th>10</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.852</td>
<td>0.852</td>
<td>0.889</td>
<td>0.889</td>
<td>0.889</td>
</tr>
</tbody>
</table>

### Table 5

The four genes in the four-biomarker NLN classifier for the leukemia data.

<table>
<thead>
<tr>
<th>Access no.</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>U22376</td>
<td>c-myb gene extracted from human (c-myb)</td>
</tr>
<tr>
<td>D88422</td>
<td>CYSTATIN A</td>
</tr>
<tr>
<td>M23197</td>
<td>CD33 antigen (differentiation antigen)</td>
</tr>
<tr>
<td>M12959/X02592</td>
<td>TCRA T cell receptor alpha-chain</td>
</tr>
</tbody>
</table>

Different from the above two datasets, the colon dataset does not have a standard training/test split, and most researchers adopted LOOCV to evaluate their algorithms when using this data set [11,14]. For convenient comparison, in this study our BAN-based approach is also evaluated using LOOCV. Table 3 reports the classification results of our NLN classifier and compares them with those of the previous approaches including FDA, KNN, BN and SVMs. From this table, it can be seen that the NLN classifier significantly outperforms the other classifiers in classification accuracy, irrespective of the number of genes used. In particular, our six-gene NLN classifier can attain the highest accuracy of 95.2%. Table 8 further compares our results with the previously reported results, showing that our result is significantly better than all but one of the results. Table 9 reports the previous 3- or 10-fold cross-validation results for the colon data, showing that most of the classification accuracy values are lower than 0.90. Based on the six-gene model, we computed the mean association levels and the mean association spectra for each class, and show them in Fig. 11(A). This figure indicates the remarkable difference of the association patterns of the two classes.

#### 3. Application to the leukemia data

Next, the leukemia data set is used to evaluate our approach. Fig. 6(b) shows the training errors. Similar to the case of the NPC dataset, the training errors first decrease when more biomarkers are used, and then reach a minimum. The accuracies of these BAN classifiers on the independent test set are reported in Table 1. From this table, it can be seen that the NLN classifiers attain better accuracies compared with the LN classifiers, similar to the case of the NPC dataset. More interestingly, three of the five NLN classifiers correctly classify all the independent test samples, and the least number of genes used is only 4. In the right half panel of Fig. 6, we show the prediction strengths of the correctly classified samples by the four-biomarker NLN classifier. It is clear that all of the PS values are significantly high (> 0.4) and the median PS (the horizontal bar) is close to 1 (0.999). Table 5 lists the four genes involved in the NLN classifier. In previous works, the four genes have been shown to be related to leukemia [3,5,5,6]. For example, Gene "M23197" (CD33 anti-gen) is observed to be over-expressed in more than 90% of patients with AML, and an anti-CD33 monoclonal antibody has been introduced clinically for the treatment of AML patients [55]. Gene "D88422" is ranked no. 2 with respect to the degree of correlation with AML in [56]. Fig. 10(A) and (B) shows the four-gene NLNs of the AML and ALL classes, respectively, and Fig. 10(C) and (D) compare the association levels and association spectra of the ALL and AML classes, showing the remarkable differences of the association patterns between the two classes.

### Table 3

The four-gene NLNs of the AML and ALL classes, respectively, and Fig. 10(C) and (D) compare the association levels and association spectra for each class, and show them in Fig. 11(A). This figure indicates the remarkable difference of the association patterns of the two classes.

#### 3.3. Application to the colon data

Different from the above two datasets, the colon dataset does not have a standard training/test split, and most researchers also performed the leave-one-out cross-validation (LOOCV) on all the samples for the leukemia dataset, and the obtained results are compared with those reported by other researchers in Table 7. From Table 7, it can be seen that our NLN classifier can correctly classify all the 72 samples, which is significantly better than all but one of the results.
Fig. 10. The association patterns for the two classes of the leukemia data. (A) BAN of the ALL class and (B) BAN of the AML class, where red and green lines represent positive and negative associations, respectively, the widths of the lines indicate the strength of associations, and the big hexagon nodes indicate the dominant biomarkers of the networks; (C) comparison of the association levels in each association channel between the two classes; (D) association spectra of the two classes.

Table 6
Previous results on the test set for the leukemia data.

<table>
<thead>
<tr>
<th>Methods</th>
<th>No. of genes</th>
<th>Accuracies</th>
</tr>
</thead>
<tbody>
<tr>
<td>RPLS [15]</td>
<td>50</td>
<td>0.971</td>
</tr>
<tr>
<td>RPCR [15]</td>
<td>50</td>
<td>0.971</td>
</tr>
<tr>
<td>Factor mixture models [17]</td>
<td>242</td>
<td>0.94</td>
</tr>
<tr>
<td>PCA/SFFS/SVM [63]</td>
<td>–</td>
<td>0.94</td>
</tr>
<tr>
<td>ICA/SFFS/SVM [63]</td>
<td>–</td>
<td>1</td>
</tr>
<tr>
<td>ZSN correlation/SVM [64]</td>
<td>Not more than 64</td>
<td>0.971</td>
</tr>
<tr>
<td>FC correlation/SVM [64]</td>
<td>Not more than 64</td>
<td>0.912</td>
</tr>
<tr>
<td>“Expected” SVM-RFE/SVM [64]</td>
<td>Not more than 64</td>
<td>0.962</td>
</tr>
<tr>
<td>Two-stage SVM-RFE/SVM [64]</td>
<td>Not more than 64</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 7
Comparison of the LOOCV results of our NLN classifier with several previously reported ones on the colon data.

<table>
<thead>
<tr>
<th>Methods</th>
<th>No. of genes</th>
<th>Accuracies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our NLN classifier</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Logistic regression method [16]</td>
<td>20</td>
<td>0.972</td>
</tr>
<tr>
<td>RPLS [15]</td>
<td>50</td>
<td>1</td>
</tr>
<tr>
<td>RPCR [15]</td>
<td>50</td>
<td>0.972</td>
</tr>
<tr>
<td>Neuro-fuzzy ensemble machine [65]</td>
<td>20</td>
<td>0.958</td>
</tr>
<tr>
<td>FREM/SVM [66]</td>
<td>25</td>
<td>0.986</td>
</tr>
<tr>
<td>Bayesian variable selection approach [56]</td>
<td>5</td>
<td>0.972</td>
</tr>
</tbody>
</table>

Table 8
Comparison of the LOOCV results of the NLN classifier with previously reported results on the colon data.

<table>
<thead>
<tr>
<th>Methods</th>
<th>No. of genes</th>
<th>Accuracies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our NLN classifier</td>
<td>6</td>
<td>0.952</td>
</tr>
<tr>
<td>QDA/t-score/PLC [11]</td>
<td>50</td>
<td>0.919</td>
</tr>
<tr>
<td>RPLS [15]</td>
<td>100</td>
<td>0.887</td>
</tr>
<tr>
<td>RPCR [15]</td>
<td>1000</td>
<td>0.887</td>
</tr>
<tr>
<td>MFW [67]</td>
<td>6</td>
<td>0.952</td>
</tr>
<tr>
<td>SFSW [67]</td>
<td>10</td>
<td>0.903</td>
</tr>
<tr>
<td>Entropy-based method [68]</td>
<td>31</td>
<td>0.903</td>
</tr>
</tbody>
</table>

Table 9
Previous 3-/10-fold cross-validation results for colon data.

<table>
<thead>
<tr>
<th>Methods</th>
<th>k-fold</th>
<th>Accuracies</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACA/LVM [69]</td>
<td>10-fold</td>
<td>0.933</td>
</tr>
<tr>
<td>REDISC/PLS/SVM [70]</td>
<td>10-fold</td>
<td>0.866</td>
</tr>
<tr>
<td>RELIC/PCA/SVM [70]</td>
<td>10-fold</td>
<td>0.863</td>
</tr>
<tr>
<td>Rule groups [71]</td>
<td>10-fold</td>
<td>0.914</td>
</tr>
<tr>
<td>GA-MTL [72]</td>
<td>3-fold</td>
<td>0.857</td>
</tr>
<tr>
<td>MPE/SVM [57]</td>
<td>3-fold</td>
<td>0.879</td>
</tr>
</tbody>
</table>
4. Biological interpretations

In the proposed BAN model, the association patterns are encapsulated in the connection weights, and the expression level of a biomarker is estimated based on these patterns. The estimated values should be compared with the observed one to determine whether the patterns can efficiently explain a given biomarker expression profile. As a result, we define an energy function as a measure of the discrepancy between the estimated and observed values. More and more studies indicate that abnormal epigenetic change plays a critical role in cancer development, and biomarker expression information serves as an important indicator of neoplastic initiation and progression [59]. The biomarker association patterns detected using our approach may more consistently account for different cancer types, and are more stable than expression patterns across different samples. In particular, the non-linearity of the model can further facilitate the modeling of the complex epigenetic change, and the effectiveness has been confirmed through our experiments.

Since our model implies that different types or subtypes of cancer correspond to different BANs, each network has a high specificity for a particular cancer type. Compared with conventional

![Fig. 11. The association patterns for the different classes of (A) the colon data and (B) the breast data. The left panel compares the association levels in each channel between the different classes; the right panel compares the association spectra of the different classes.](image)

![Fig. 12. Performance comparison of our NLN classifier with several previous methods on the breast data.](image)
network modeling techniques such as Boolean network or Bayesian network, the BAN model can quantify an association between biomarkers in the form of a real coefficient. The BAN model highlights the difference of biomarker association patterns between different cancer types or subtypes, and provides an association spectrum representation for a better understanding of cancer from a biomarker association angle. Different manifestations of the same type of cancer, while corresponding to highly variable ranges of expression values for the different biomarkers, can have similar association spectra. Feedback serves as an important mechanism to maintain stability, and plays a prominent role in biomarker activity, in particular for keeping the dynamic balance of the cellular system and endowing cells with the self-repairing ability against unexpected changes in the outer environment [60]. The bi-directional association coefficients reflect this kind of feedback. In addition, the BAN model can be incrementally expanded through a step-by-step judicious inclusion of related biomarkers, thus facilitating the construction of a large-scale network.

5. Conclusion and discussion

We have proposed a new cancer classification approach based on the concept of a Biomarker Association Network (BAN). In this approach, a neural network structure is proposed to model the biomarker association patterns responsible for cancer. In particular, an energy function is designed as a measure of the disagreement between the observed levels of biomarker expression and the estimated levels so that the association coefficients can be determined by minimizing the energy function. We develop two types of BANs, LN and NLN, for cancer classification. The proposed approaches are evaluated on four publicly available data sets including one protein expression data, the NPC dataset, and three gene expression data sets, the leukemia, the colon and the breast cancer data, and the experimental results show that the proposed BAN-based classification approaches, in particular, the NLN classifier, achieve excellent classification performances on the four datasets. Extensive comparisons with five previous classification methods, FDA, KNNS, BN, SVMs with linear and non-linear kernels, are performed, which further confirm the superior performance of the NLN classifier. As well as providing high accuracy of cancer classification, the proposed approaches can also allow us to gain an understanding of the complex relationships between the various biomarkers from a biological viewpoint. Based on the real-world datasets, remarkably large differences of the association coefficients can be identified between different cancer types, reflecting the significantly different underlying biomarker association patterns.

In this study, the proposed BAN model is applied to characterize biomarker expression profile data associated with cancer, which may not include a temporal element, for the purpose of classification. In particular, the connections of the network represent specific association relationships between biomarkers which are related to cancer, and the associated weights of these connections are estimated based on empirical data. This Association Network differs from a general regulatory network in that, for the latter, the expression of a biomarker is controlled by its regulatory elements, and the regulated result will be observed only after a certain delay. In other words, while temporal data are necessary to determine the regulatory relationships between biomarkers in the case of a regulatory network [41, 61, 62], the inclusion of these type of data is not mandatory for the approximate characterization of the relationships between biomarkers in the case of an Association Network.

A limitation of the proposed approach is the expensive computation cost for the case when too many biomarkers are considered. Fortunately, we only focus on a few association patterns associated with cancer classification, and there are a number of approaches for seeking a small set of involved biomarkers. More importantly, with the increasing amount of information available in data repositories such as Gene Ontology, KEGG and GenMAPP, the search space can be efficiently reduced based on these prior knowledge. In addition, distributed computing technique can serve as an effective solution to improve the computational efficiency. Our future works will focus on exploring more efficient non-linear forms of BANs and developing large-scale BAN structures for the understanding of biomarker association patterns of cancer.

Acknowledgment

The work described in this paper was supported by a grant from the City University of Hong Kong [Project No. 7001965].

References