Journal of Biomedical Informatics 111 (2020) 103570

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

# Using hidden Markov model to predict recurrence of breast cancer based on sequential patterns in gene expression profiles

Mohammadreza Momenzadeh [a], Mohammadreza Sehhati [a,b,c,*], Hossein Rabbani [a,b]

[a] Department of Biomedical Engineering, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran
[b] Medical Image and Signal Processing Research Center, Isfahan University of Medical Sciences, Isfahan, Iran
[c] Department of Bioinformatics, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran

ARTICLE INFO

*Keywords:*
Breast cancer recurrence
DNA microarray
Classification
Hidden Markov model (HMM)
Gene set enrichment

ABSTRACT

A new approach is presented to predict breast cancer recurrence through gene expression profiles using hidden Markov models (HMM). In this regard, 322 genes were selected from 44 published gene lists related to breast cancer prognosis. Afterwards, using gene set enrichment analysis, 922 gene sets were found from subsets of genes with the same biological meaning. In order to extract the sequential patterns from gene expression data, we ranked the gene sets using appropriate criteria and used HMM in which the ranked gene sets considered as observation sequences and hidden states represented priority of gene sets for discriminating between expression profiles. In this experiment, seven publicly available microarray datasets, including 1271 breast tumor samples, were used to classify cancer patients into two groups according to risk of recurrence. Our experiments indicated the greater performance and more robustness of the proposed model compared with other widely used classification methods.

## 1. Introduction

Breast cancer recurrence is a complex biological process regulated by several important genes. In order to discover this regulation process, DNA microarray technology has been introduced and widely used for simultaneous analysis of expression levels in thousands of genes. Given the fact that differentially expressed genes in different tumor cells from patients determine the recurrence potential among patients, there has been a tendency for breast cancer recurrence study based on the analysis of high dimensional gene expression data. Due to important information that can be provided for treatment of breast cancer patients, there is a strong motivation to propose new approaches to efficiently identify a small group of genes for predicting late recurrence after 5 years of follow up. In summary, all related studies in the field are trying to answer several main questions including: (1) Which genes are responsible for breast cancer recurrence? (2) How the structure and training procedure of classification models should be modified to improve their predictive power for breast cancer recurrence prediction? (3) How the interaction between genes can be used to improve the classification performance? (4) Does integrating the gene expression data with other resources such as GO and pathways can be helpful? (5) Can we improve the stability and generality of the predictive models for confidently applying them on different independent datasets? Meanwhile, another new question that promoted us to perform this research is that how sequential patterns in the gene expression profiles can be found and used for breast cancer recurrence prediction?

In the first major study van't Veer et al. [1,2] effectively predicted the 5-year recurrence status in a group of breast cancer patients. They found a list of 70 genes (NKI70), which is available as a breast cancer prognostic test and were 60%–70% accurate in predicting recurrence of breast cancer in a limited group of patients. Later Wang et al. [3] identified a list of 76 genes which exhibited 93% sensitivity and 48% specificity in a testing set of 171 samples.

In 2017, Choi et al. [4] improved prediction of breast cancer by identifying heterogeneous prognostic genes. They clustered data samples of several microarray datasets by K-means algorithm and applied modified PageRank algorithm to functional interaction (FI) networks using levels of gene expression samples in each cluster and receive better outcome prediction.

In some other studies, network based classification methods used for improving the prediction power of breast cancer metastasis models [5–8]. Tian et al. [9] had integrated protein–protein interaction (PPI)

* Corresponding author at: Department of Bioinformatics, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Hezar Jereeb Street, 81746 Isfahan, Iran.
*E-mail address:* mr.sehhati@amt.mui.ac.ir (M. Sehhati).

https://doi.org/10.1016/j.jbi.2020.103570
Received 6 April 2020; Received in revised form 6 September 2020; Accepted 10 September 2020
Available online 19 September 2020

information with expression data to identify the genes involved in breast cancer metastasis using random forest for classification. Recently Zhibo Wang et al. [10] proposed a deep neural network based method which called Network-based multi-task learning models for biomarker selection and cancer prediction. They tested two breast cancer microarray gene expression datasets by running network-based algorithm on several gene graphs. However, different types of gene chip and test environment lead to different performances and as a result making it difficult to have a general method based on the co-expression networks.

Generally, curse of high dimensions, small number of samples and different technological platforms [11–13] are the most important limitations of microarray data that lead to various analytical methods such as: dimensionality reduction techniques [14], gene selection methods [1,15,16], applying PPI [5], analyses of expression data based on gene sets obtained from the Gene Ontology (GO) [17] and considering the interactions between genes [18]. Haibe-Kains et al. [19] reported that classification models applying a single gene or multiple selected genes obtained as accurate or even better efficiency than classifiers using expression of the whole genome for breast cancer prediction. Vliet et al. [20] utilized several module-based classifiers as breast cancer predictors and compared with gene based classifiers. They found that classifiers rely on gene sets achieve better performance and significantly lower variance on the validation data compared to individual gene-based classifiers. Abraham G et al. [21] applied centroid classifier on five well-known microarray datasets to classify the risk of recurrence and observed that prognostic gene signatures obtained from gene sets are more stable than individual genes. They demonstrated that using gene set can reduce the noise of dataset and increase performance of classification. Gene set enrichment analysis (GSEA) is an analytical method for interpreting and evaluating gene expression microarray data at the level of gene sets. Recently successful researches have been done on the prognosis of breast cancer recurrence using GSEA [22–24].

Most recently, Rueda et al. [25] proposed a multistate Markov model for breast cancer recurrence that recognized different recurrence patterns across different molecular subgroups. This model approximates the risk of breast cancer recurrence by evaluating the transition rates via four visible states. However, the worth of this model is in doubt, as most patients do not have a distinct transition through all of the model states and, in particular, specific recurrence states are not preceded by a detectable preface state [26].

In bioinformatics, the HMM is well-known for its application in modeling the relation between biological sequences specially in sequence alignment [27], gene prediction [28] and so on. In sequence alignment, HMM-based methods are used to find relationships and similarity between sequences of DNA, RNA, or Amino-acids. In this application the conserved positions in the aligned sequences will construct the main states in the HMM topology. Moreover, in gene prediction, HMM is used for finding the location of protein coding regions or prediction of functional elements of genomes such as regulatory regions. Our knowledge about the gene sequence characteristics are crucial in order to design a suitable topology for HMM in this application. In general, HMM is a probabilistic model for describing data with sequential process in adjacent samples like the time series data [29]. Similarly, cancer metastasis is a sequential process [30] that starts with the spread of primary tumor cells and take place through different sequential pathways. As a pioneer, Nguyen et al. [31] applied HMM on gene expression profiles for cancer classification and compared the results with a range of prevalent classifiers such as k-nearest neighbors (KNN) and support vector machine (SVM) in a way that HMM yielded the best performance. However, they used a simple HMM with two states for classifying samples into normal and tumor for different cancer types.

In our study, we aim to redefine the states and observations in a hidden Markov model (HMM) structure for predicting the risk of late recurrence based on sequential patterns in gene expression profiles. In our HMM-based approach, we also used GSEA to obtain appropriate gene sets that ranked to represent the observation sequences in our recurrence predictor. Our methodology is similar to the well-known text classification applications of HMM, with this difference that words are replaced by gene sets [32–34]. We organized this paper as follow: In Section 2 the design of the proposed method is introduced. Experimental results are expressed and discussed in Section 3 and finally, Section 4 concludes the paper.

## 2. Methods

### 2.1. Datasets

Seven breast cancer microarray datasets were used in our experiments include: GSE2034 [3], GSE7390 [35], GSE6532 [36], GSE4922 [36], GSE3494 [37], GSE2990 [38] and GSE11121 [39] which are freely available from NCBI GEO. All seven datasets are Affymetrix HG-U133A microarray platforms. Quality control probes and probes with variance values close to zero were removed. Moreover, probes which have more than 15% missing expression levels were eliminated. These datasets include samples with both lymph-node-negative and nod-positive breast cancer. Moreover, the datasets include samples with both estrogen receptor positive and negative breast cancer. According to the time of distant metastasis and 5 years cut of point, samples were classified into two categories: low and high risk. We combined all datasets together to make a larger population and removed samples that were treated with hormone therapy drugs after surgery. Because hormone therapy will bias the expression data and change the outcome. Therefore, we should remove the corresponding samples, which affected by this confounding variable, to predict the outcome based on the primary expression data. In this regard our experimental dataset consists of 1271 samples (892 low risk and 379 high risk samples) in a way that all the remaining probes mapped to 12,172 gene expression levels [40]. Subsequently, log2 function was applied on the all gene expression of each dataset independently and then normalized through sample vectors and module vectors simultaneously. In normalization process, the mean value was subtracted from all expression levels and the corresponding results divided by its standard deviation. Afterwards arctan function was used to limit the range of all expression levels uniformly through the following:

$$arctan\left(\frac{X-\mu}{\sigma}\right) \tag{1}$$

where X, μ, and σ, are the value, mean and standard deviation of expression levels after applying log2 function, respectively.

### 2.2. Basics of hidden Markov models

HMM firstly introduced by Baum et al. in a series of paper in the late 1960s and early 1970s [41–44]. In probability theory the (first-order) Markov property refers to stochastic process in which the future state of system relates only to the current state [45]. HMM is a statistical tool that can be used for modeling generative sequences described by a set of observable events (say symbols) that depend on invisible sequence of factors (say states). An HMM contains two stochastic processes, and thus it is also named a doubly-embedded Markov process. The first process describes invisible route of hidden states and the second process represent visible process of visible symbols. The hidden states being modeled as Markov chain [41,43], and the occurrence of the observation symbol depends on the underlying state. Accordingly, each HMM is described by five elements as follow:

$$\lambda = \{N, V, A, E, \pi\} \tag{2}$$

1. The $N$ states of the model is denoted by $s$ as:

$$S = \{S_1, \cdots, S_N\} \tag{3}$$

2. The $M$ observation symbols of each state are denoted by $V$ as:

$$V = \{\nu_1, \cdots, \nu_M\} \tag{4}$$

3. $A = \{a_{ij}\}c$ is a $N \times N$ matrix, which called the transition matrix and represents state transition probability distribution where $a_{ij}$ cell is the probability of moving from state $S_i$ to state $S_j$:

$$a_{ij} = P\{q_{t+1} = S_j | q_t = S_i\}, \quad 1 \le i,j \le N \tag{5}$$

and $q_t$ indicates the current state.

The transition probability distribution must fulfil the normal stochastic constraints:

$$a_{ij} \ge 0, 1 \le i,j \le N, \quad \sum_{j=1}^{N} a_{ij} = 1, 1 \le i \le N \tag{6}$$

4. $E = \{e_j(k)\}$ is the emission matrix of dimension $N \times M$, that indicates probability distribution of observation symbol in each state and the $e_j(k)$ cell is the probability that symbol $v_k$ is produced in state $S_j$.

$$e_j(k) = P\{o_t = \nu_k | q_t = j\}, \quad 1 \le j \le N, 1 \le k \le M \tag{7}$$

where $v_k$ represents the $k^{th}$ observation symbol, and $O_t$ characterizes the current parameter vector. In emission matrix construction the following constraints should be satisfied:

$$e_j(k) \ge 0, \quad 1 \le j \le N, \quad 1 \le k \le M \quad \text{and} \quad \sum_{j=1}^{N} e_j(k) = 1, \quad 1 \le j \le N \tag{8}$$

5. $\pi = \{\pi_i\}$ is the initial state probability vector that determines the probabilities of states in time = 0.

$$\pi_i = p\{q_1 = i\}, \quad 1 \le i \le N \tag{9}$$

Accordingly, we describe an HMM with parameter set $\lambda = (A; E; \pi)$.

### 2.3. Three basic problems in HMM

There are three well known basic problems which are solved by HMM for real world applications [46,47].

Evaluation: In the evaluation problem we compute $P\{O|\lambda\}$, i.e. the probability that the given observation sequence $O=\{o_1, o_2, \ldots, o_T\}$ is produced by the model $\lambda$. This problem can be solved using forward or backward algorithm.
Decoding: In the decoding problem we discover most possible hidden state sequence related to the given observation sequence $O$ and the model $\lambda$. The Viterbi algorithm is used to solve this problem.
Learning: In the learning problem, the parameters of the model will be adjusted to maximize $P\{O|\lambda\}$ whereat the model $\lambda$ and an observation sequence O are given. The learning problem can be solved by Baum-Welch algorithm.

In the proposed method the evaluation problem is used as our solution to microarray data classification problem.

### 2.4. Microarray data classification using HMM

In our experiment, the problem of microarray data classification was resolved by automatically assigning a binary recurrence risk label (low/high) to new instances (unobserved tumor samples).

In the proposed approach we built two HMM classifiers, which trained on sequential patterns in gene expression profiles of two patient groups. This task is similar to the approach proposed by Kwan Yi et al. [34] for content classification in medical documents. Fig. 1 illustrates the framework of the proposed method. Accordingly, in order to classify a new sample, the probability of generating this sample by each of the two trained HMMs will be evaluated. Consequently, the model that showed the maximum probability value determines appropriate class label for the input sample.

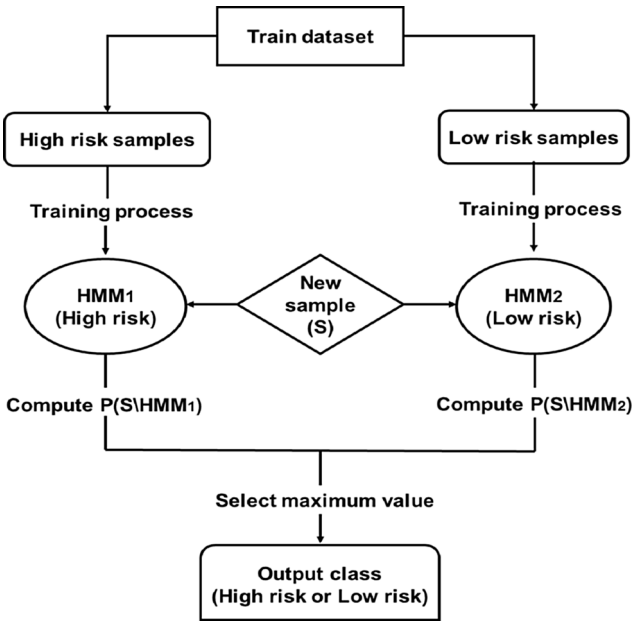



**Fig. 1.** Classification process with HMM.

### 2.5. Gene set modules construction

The proposed approach aims to find sequential patterns in gene expression profiles to classify samples. For this purpose, microarray gene expression data needs to be converted into symbols that HMM can handle. As previously mentioned our HMM classifier is inspired from document classification task using bag-of-words method [48]. This method characterizes each document by a vector in a way that vector elements represent the number of occurrences of keywords in each document. Words with more appearance are more relevant, because they are represented the best symbol for the document content. In our proposed method the words were replaced by gene set modules. As alphabets come together to form the words with different semantics, genes come together to make the gene modules with similar biological meaning which named gene sets. GSEA considers microarray data as gene sets based on previous biological information about biochemical pathways or gene co-expression network (GCN) [49,50].

From a statistical perspective, the analysis of group of genes instead of individual genes has some advantages include: increasing processing power, reducing noise and decreasing dimensionality problem such as model complexity, computational cost and processing time of individual genes analysis. From the biological point of view, gene set enrichment analysis provides some biological information about underlying pathways. For example, consideration of biological mechanism that related to breast cancer prognosis or finding functional mechanism in a cell such as activated certain pathway in a special tissue that underlying some treatments can be mentioned as provided biological information form gene sets [51,52].

For constructing desired gene sets, we need genes relevant to breast cancer as the seed points. For this purpose, we extracted 322 genes from 44 published gene lists in previous experiments which related to breast cancer prognosis [53]. The significance of this gene list relies on that these genes are confirmed at least two times in other publications besides several quality criteria. In the second step, in order to make gene sets, we used GeneCodis software tool [54]. GeneCodis incorporates information of different nature for singular and modular GSEA (e.g. functional, regulatory or structural) by search for finding frequent patterns in the annotation space and calculating their statistical

relationship. Different annotations containing the three GO categories (biological process, molecular function and cellular component), KEGG pathways, InterPro Motifs, and Swiss-Prot keywords can be analyzed using GeneCodis [55,56]. Therefore, it utilizes the integrative capacity to clarify different aspects of a data. We entered 322 gene IDs to GeneCodis while GO biological process was selected. After GeneCodis process finished we obtained 922 gene modules.

### 2.6. Gene set selection

922 gene sets obtained from the GeneCodis tool should be filtered to select the best modules. In other words, there was a need for a gene selection method to select the best gene modules. Applying the appropriate ranking criterion in GSEA has serious effect on the final results of pathway enrichment analysis. In 2017, Zyla J et al. [57] used 28 benchmark datasets to estimate the false positive rate and sensitivity of GSEA for 16 different ranking criteria which contain some new approaches. In their study, the Moderated Welch Test (MWT) showed the best overall sensitivity. We ranked gene sets by MWT statistic which in our experiments has the best results too. After ranking gene set by MWT, top-ranked gene sets were selected as the observations.

### 2.7. Gene set representative

In order to use gene sets as observation sequences in the proposed HMM classifier, we need to choose a representative for each module. In this step, we searched a suitable statistic criterion as the representative of expression levels of all genes in a module. In a comprehensive study, Abraham G et al. [21] compared statistical mean, median, t-statistic and raw data as representative of genes. All of utilized statistics are unsupervised, in the other words they do not take into account recurrence class. The one-sample t-statistic which denoted by following relation compares the mean of a sample to a determined value and tests for a deviation from that value [52,58].

$$t_{ij} = \frac{\sqrt{n_i} \times mean_{ij}}{std_{ij}} \tag{10}$$

where $mean_{ij}$ is the statistical mean and $std_{ij}$ is the standard deviation of the genes in set $i$ in the $j$th sample. In the proposed method one-sample t-statistic is used as representative of gene module which has the best efficiency in Abraham G et al. experiment and also in our experiment.

### 2.8. Proposed HMM classifier structure

After demonstrating the modules by statistical representative, sort representatives of each sample in descending order. In the proposed model, hidden states represent the difference in relevance levels or ranking of modules in a microarray sample. Each state indicates a relevance level for gene module occurring in the samples. According to this, the first state represents the most relevant modules for discriminating a group of patients in the microarray dataset. The most relevant observations for the second state are the modules getting the second level of relevance in the samples, and so on. The $N$ states of the model is an adjustable parameter that depends on the training dataset and the amount of flexibility that we need in our model. According to these explanations, each sample is finally represented by a vector or a module list ordered decreasingly by their ranking. The proposed HMM architecture that used to indicate a predefined category is structured by implementing the following instructions:

1. The total of samples is taken as descending sorted module representatives from the training dataset and create set of observation symbols $V$. Therefore, there is a symbol for each module in a way that for each HMM, observation symbols are similar.

2. As previously mentioned, states represent ranking of modules. In this regard, states are arranged from the first rank to the last rank. Thus, the state transitions creating a left-right HMM [59] without self-state loops, where only transition from state $i$ to state $i + 1$ is allowed. The transition matrix of this topology is defined as:

$$a_{ij} = \begin{cases} 1 & \text{if } j = i + 1 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

3. The probability distribution of output observations (modules) depends on the training dataset and the related category c. A module symbol $v_j$ will have a higher probability of observation than module symbol $v_k$ at a given state $s_i$ if the number of occurrences of the symbol $v_j$ is greater than $v_k$. Considering a category c and a dataset $D_c$ for samples who are at the risk of breast cancer recurrence, the emission matrix for an HMM that represents the category c (high risk or low risk) is defined as follows:

$$e_i(v_k) = \frac{\sum_{d \in D_c} F_d(v_k, i)}{NumberofallModules} \tag{12}$$

where

$$F_d(v_k, i) = \begin{cases} 1 & \text{if module } v_k \text{ appears at } i\text{th rank in sample d} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

$e_i(v_k)$ denotes the probability of the module/symbol $v_k$ emitted at the state $s_i$.

4. The initial state probability distribution $\pi$ in the proposed HMM method is 1 for the first state.

Fig. 2 illustrates the training process and topology of the proposed HMM method with two categories: high risk and low risk. Fig. 2a shows how the modules are sorted in descending order based on their representative value for each sample. Therefore, each sample is represented by a vector whose elements represent ordered modules.

Fig. 2b shows HMM topology of the proposed method which includes state transition structure and emission probabilities for high risk samples. It can be observed that states represent relevance level of gene modules occurring in the samples in a way that state transitions topology is a left–right HMM without self-state loops. It is important to note that only three states and six ordered modules are considered in this example. However, we considered the equal number of modules and states in the proposed model.

### 2.9. Classification of new sample

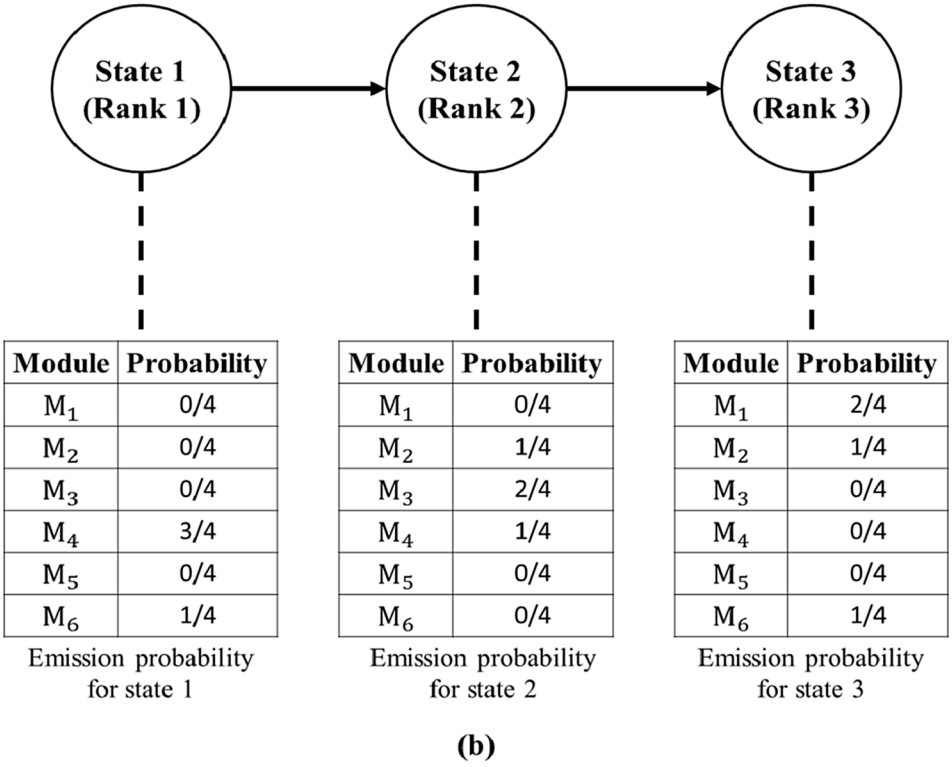
In order to classify a new sample, when two HMMs are separately trained for high risk and low risk categories, we should arrange the new sample similar to the ordered module list in the training process. Then, evaluate the probability of that module sequence generated by the two HMMs as illustrated in Fig. 1. Afterward, compare the probability value of the module sequence generated by each HMM and the category that represents the highest probability value is chosen and considered as the class label of input sample. As previously mentioned this is the evaluation problem of HMM and calculation of these probabilities is performed by using the forward algorithm.

### 2.10. Experimental settings

Since in gene expression microarray datasets there are often a small number of samples, the k-fold cross validation with k = 10 or higher is used generally [60,61]. In k-fold cross-validation, first the samples are randomly divided into k equal subsamples. Then every time one of the k subsamples and k-1 subsamples are used as validation and training data, respectively. There are several metrics to measure the effectiveness of a classification task. The AUC which defined as the area under the receiver operating characteristic curve is generally used to measure the performance of classification methods [62]. The AUC is an important evaluation criterion in medical applications that used to estimate effectiveness of diagnosis system with binary classification. The

| | 1st | 2nd | 3rd | 4th | 5th | 6th | Class |
|---|---|---|---|---|---|---|---|
| $S_1$ | $M_4$ | $M_2$ | $M_6$ | $M_3$ | $M_5$ | $M_1$ | H |
| $S_2$ | $M_6$ | $M_4$ | $M_2$ | $M_1$ | $M_3$ | $M_5$ | H |
| $S_3$ | $M_2$ | $M_5$ | $M_1$ | $M_6$ | $M_3$ | $M_4$ | L |
| $S_4$ | $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_6$ | $M_5$ | L |
| $S_5$ | $M_4$ | $M_3$ | $M_1$ | $M_2$ | $M_5$ | $M_6$ | H |
| $S_6$ | $M_4$ | $M_3$ | $M_1$ | $M_6$ | $M_5$ | $M_2$ | H |

(a)




| Module | Probability |
|---|---|
| $M_1$ | 0/4 |
| $M_2$ | 0/4 |
| $M_3$ | 0/4 |
| $M_4$ | 3/4 |
| $M_5$ | 0/4 |
| $M_6$ | 1/4 |

Emission probability for state 1

| Module | Probability |
|---|---|
| $M_1$ | 0/4 |
| $M_2$ | 1/4 |
| $M_3$ | 2/4 |
| $M_4$ | 1/4 |
| $M_5$ | 0/4 |
| $M_6$ | 0/4 |

Emission probability for state 2

| Module | Probability |
|---|---|
| $M_1$ | 2/4 |
| $M_2$ | 1/4 |
| $M_3$ | 0/4 |
| $M_4$ | 0/4 |
| $M_5$ | 0/4 |
| $M_6$ | 1/4 |

Emission probability for state 3

(b)

**Fig. 2.** Example of HMM training process for high risk samples. (a) Sample vectors with their module ordered for high risk and low risk categories. (b) Transition matrix structure and emission probabilities for high risk samples.

Matthews correlation coefficient (MCC) [63] is another useful performance criterion that used to measure efficiency of binary classification specially for imbalanced datasets. The MCC has a range of $-1$ to $+1$ where a coefficient of $-1$ indicates an absolutely wrong label is assigned to all samples, 0 indicates random prediction among different samples and $+1$ demonstrates a truly correct classification in the whole dataset. The MCC can be calculated form the confusion matrix by the following:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{14}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

In order to make reliable evaluation and create robust comparison between classification methods we increased the number of estimation and repeated 10-fold cross-validation procedure 20 times.

## 3. Experimental results

In this section, we have demonstrated the results obtained from the whole procedure for parametrization of the proposed model, validation of model against randomization, evaluation of classification results and computational cost of the proposed model in comparison with other approaches. The experiments implemented on seven well-known breast cancer datasets according to risk of breast cancer recurrence. In order to make a larger population we combined all datasets together and for better evaluation used 10-fold cross-validation which repeated 20 times. To perform a better comparison, we compared our proposed HMM with maximum likelihood and Baum-Welch as two routine training algorithms of HMM, and KNN and SVM as two prevalent classifiers.

### 3.1. Module representative

In the first experiment, we compared statistical mean, median, one sample t-statistic and individual genes as potential candidates for representing the selected gene set modules. Fig. 3 illustrates the classification results of the proposed HMM method on the training dataset

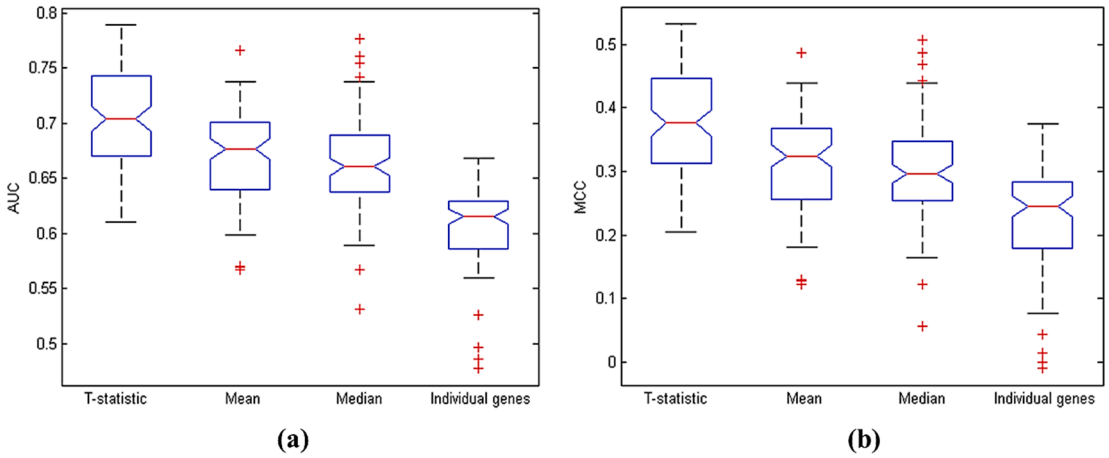



**Fig. 3.** Module representative comparison: (a) using t-statistic, mean, median and individual genes as module representative for proposed HMM classification in terms of AUC, (b) in terms of MCC.

using the mentioned representatives. According to Fig. 3a and b, which respectively show boxplots comparisons of different representatives based on AUC and MCC, t-statistic indicated a higher performance in comparison to other representatives. Therefore, we chose one-sample t-statistic as the best representative for extracting gene set modules. Furthermore, smaller standard deviation of t-statistic box plots against individual genes indicates that using gene set module instead of level of expression in individual genes reduces noise and decreases dimensionality problem. Therefore, we used one sample t-statistic as representative of each gene module in our experiments.

### 3.2. Randomization trial

In the second experiment, we examined the validity and reliability of the proposed model by performing a randomization trial. In this regard, we performed the tests by four different procedures and compared the results based on AUC and MCC metric. In the first test, relative gene modules in test samples were arranged similar to the ordered module list prepared in the training stage. In the second test, which called random train, high risk and low risk classes are created in random order of modules for the training stage and similarly, they are arranged randomly in the testing stage. In the third test, which called random test1, training stage was performed based on proposed HMM and module symbols are randomly arranged for test. Finally, in the random test2, training was performed based on proposed method but for test one sample arranged similar to training module list and others are randomly arranged. Fig. 4a and b shows comparison of these four validation tests in terms of AUC and MCC, respectively. According to the obtained results, random train and random test1 indicates random classification (MCC around 0). Moreover, by comparing random test2 with random test1 boxplots, it can be observed that using proposed procedure have positive effect in classification. We conclude from Fig. 4 that our model is sensitive to the sequential patterns in data which depending on Markov property and that leads to improved predictive power of the proposed model.

### 3.3. Classification experiment

In the third experiment, we trained our HMM classification model in a specific way, which illustrated in Fig. 2, for prediction of breast cancer recurrence. In this regard, we compared our method with other prevalent training method of HMM such as maximum likelihood estimation and Baum-Welch algorithm [64]. Moreover, we compared our results with KNN, SVM and decision tree (DT) classifiers because KNN and SVM are prevalent in studies on gene expression data classification [47] while DT is a rule-based classification method. To adjust classifiers parameters, we used $k = 5$ for KNN, Gaussian kernel for the SVM and binary tree structure for DT. which lead to best results in our datasets.

In the first step, to determine the best representative of selected gene modules we used different representatives including t-statistic, statistical mean, median and individual genes for evaluation and comparison of the proposed method, KNN, SVM and DT. Table 1 illustrates comparison results of the mentioned classification methods across different module representatives based on AUC (percentage). The numerical results of Table 1 show that t-statistic is the best representative of gene modules

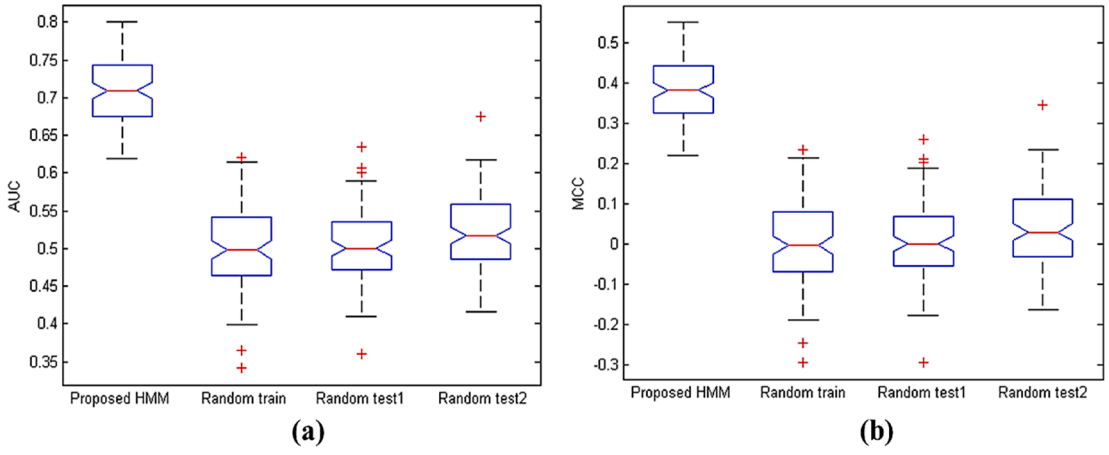



**Fig. 4.** Proposed model validation test: (a) proposed HMM classification comparison with random tests and fix modules in terms of AUC, (b) in terms of MCC.

**Table 1**
Comparison results of classification methods using different representatives.

| | T-statistic | Mean | Median | Individual genes |
|---|---|---|---|---|
| Proposed Method | 71 ± 4.48 | 67 ± 4.97 | 65 ± 6.32 | 65 ± 5.71 |
| SVM | 66 ± 5.17 | 64 ± 5.31 | 62 ± 4.37 | 64 ± 5.93 |
| KNN | 63 ± 4.94 | 62 ± 4.74 | 61 ± 4.33 | 62 ± 5.06 |
| DT | 64 ± 4.85 | 63 ± 4.41 | 61 ± 5.48 | 63 ± 5.34 |

over all of classification method and therefore we used t-statistic in downstream analysis in our experiments.

In order to compare the performance of each classification method across different number of features, we selected different number of modules (15, 20, 25 and 30) for applying to each classifier. Fig. 5 shows comparison results of different classification methods over different number of selected modules. Fig. 5a demonstrates that proposed model, represents better AUC for various number of selected modules. Also according to Fig. 5b, our method indicates better results in terms of MCC. Baum-Welch method has close results to our method and in the second place among all methods. Maximum likelihood is another training method of HMM is in the third place. Therefore, HMM, regardless of utilized training method, showed better results than KNN, SVM and DT. In Fig. 5 it can be observed that maximum classification results obtained by supplying 20 modules in the proposed method. Classification results of breast cancer recurrence for 20 modules (features) are illustrated in Fig. 6 and proposed method shows best performance. Fig. 6a illustrates AUC boxplot of comparison between proposed HMM method, maximum likelihood, Baum-Welch, KNN, SVM and DT. Therefore, proposed HMM not only lead to greater efficiency than other HMM training but also lead to better performance than KNN, SVM and DT.

Fig. 6b also confirms the greater median values of the proposed method in terms of MCC. Moreover, the relatively smaller interquartile changes of the proposed method box plots against other methods demonstrate the stability and robustness of the proposed method.

Furthermore, to make a fair comparison between related works, classification performance was evaluated over each independent dataset. Because most of the related works using k-fold cross-validation and evaluating results based on AUC, we repeated 10-fold cross-validation 20 times and reported the results based on the mean of the AUC. Table 2 illustrates comparisons of predicting recurrence between the proposed method and related works across independent breast cancer datasets. The numerical results demonstrate that our method reached the maximum AUC which shows the best performance of our model among related works of predicting recurrence of breast cancer.

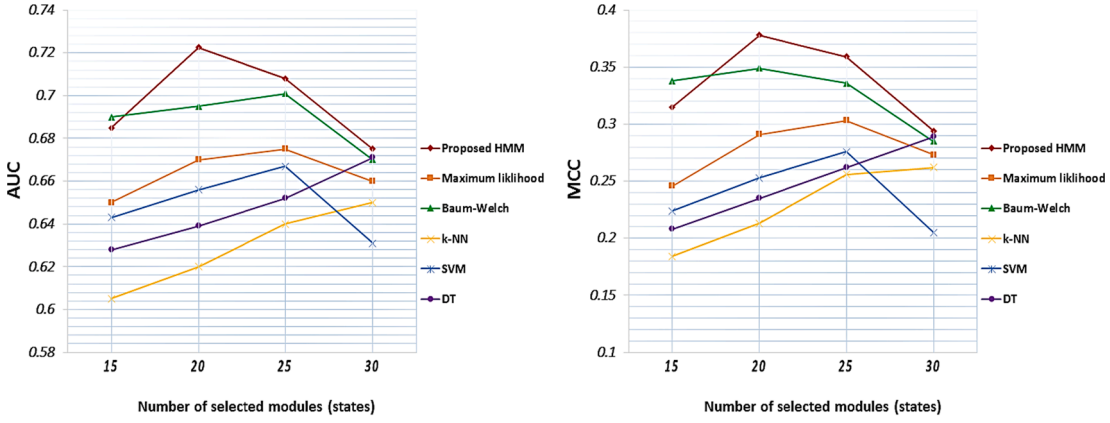
In the final classification test we have used van't Veer et al.'s dataset [1] for independent validation. In this regard, we have trained the proposed model by the seven datasets and tested van't Veer et al. dataset as independent dataset to evaluate the performance of our model. According to the results of this external validation test the proposed method reached the AUC of 73 ± 4.56 which is maximum value among different utilized methods. The measured AUC values for maximum likelihood, Baum-Welch, KNN, SVM and DT are 67 ± 5.86, 70 ± 5.43, 68 ± 5.31, 62 ± 6.17 and 65 ± 4.72 respectively.

### 3.4. Computational cost

In the fourth experiment, we evaluated the computational cost of our proposed HMM method against maximum likelihood estimation and Baum-Welch algorithm. The experiments are implemented by MATLAB installed on a PC that has the Intel(R) Core(TM) i5-2300 CPU @ 2.80 GHz with 16 GB RAM running on the 64-bit Windows7 Operating System. In order to report results of computational cost, we computed average processing time of 10-fold cross-validation for each HMM classification method. Results of computational cost are calculated in second and reported in Table 3 for different number of states (15, 20, 25 and 30). The numerical results of Table 3 demonstrate a significantly lesser processing time of proposed HMM classification compared to other HMM classifiers. Moreover, Baum-Welch exhibits very large processing time for different number of states so it is the most computationally expensive method. Thus, the proposed HMM topology for prediction of breast cancer recurrence has significantly lower computational cost, besides the higher classification results, in contrast with maximum likelihood and Baum-Welch which are prevalent training algorithms for HMM.

## 4. Conclusions

This study presented a novel HMM classification topology for prediction of breast cancer recurrence using microarray gene expression data. In this topology gene set modules represented observation sequences. Modules obtained from pathway analysis on important genes associated with breast cancer prognosis. The MWT statistic was applied for selection of best gene modules and one-sample t-statistic was used as module representatives. The proposed model considered sequential patterns in gene expression data in which gene module orders reflected Markov property and hidden states indicated the priority in the relevance levels between modules. The greater performance of proposed HMM demonstrated in terms of AUC and MCC over different number of selected modules. Also, smaller standard deviation results of the proposed method confirmed that our model was the most robust method against other classification methods. Moreover, proposed method




(a) (b)

**Fig. 5.** Results of breast cancer recurrence prediction over different number of selected modules: (a) in terms of AUC, (b) in terms of MCC.

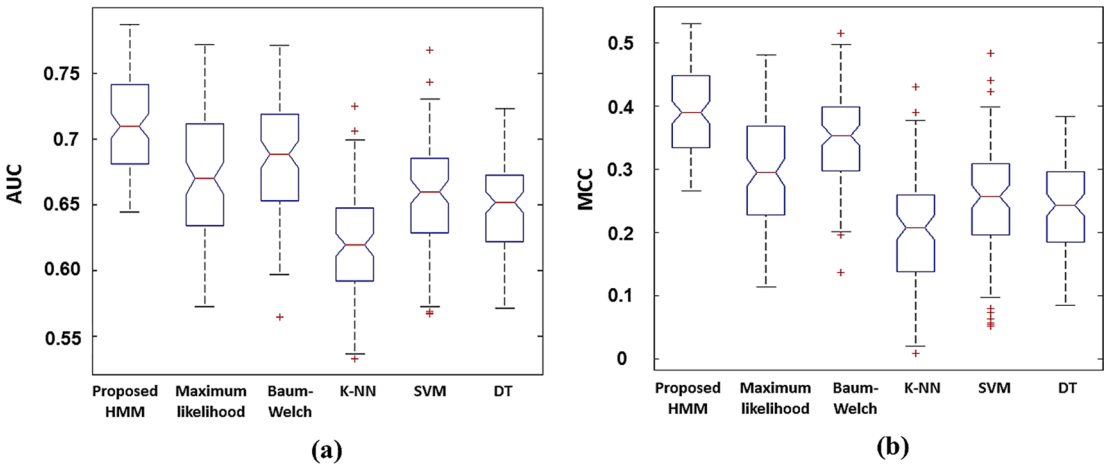



**Fig. 6.** Comparison of breast cancer recurrence prediction using KNN, SVM, DT and HMM trained by proposed method, maximum likelihood, and Baum-Welch based on (a) AUC, and (b) MCC metric.

**Table 2**
The mean AUC scores of predicting recurrence between proposed methods and related work among independent breast cancer datasets.

| | GSE2034 | GSE7390 | GSE6532 | GSE4922 | GSE3494 | GSE2990 | GSE11121 |
|---|---|---|---|---|---|---|---|
| Proposed | 72.40 | 74.50 | 76.58 | 68.23 | 74.26 | 76.48 | 72.26 |
| [9] | 71.09 | 59.18 | 67.23 | – | – | – | 68.08 |
| [4] | – | 69.9 | – | 65.4 | 72.3 | – | – |
| [10] | – | 67.1 | 70.7 | – | – | – | – |
| [8] | – | – | – | – | – | 72.5 | – |

**Table 3**
Processing time of HMM classifiers.

| | 15 states | 20 states | 25 states | 30 states |
|---|---|---|---|---|
| Proposed HMM | 0.65 s | 0.86 s | 1.141 s | 1.55 s |
| Maximum likelihood | 1.584 s | 2.356 s | 3.357 s | 4.296 s |
| Baum-Welch | 789.058 s | 1577.162 s | 2986.04 s | 4850.104 s |

represented significantly less processing time than other HMM classifiers. Furthermore, most studies on microarray gene expression data have devoted to analysis of individual gene levels. We demonstrated that using gene set modules as observation symbols of HMM have better performance than individual gene levels, and can reduce noise, decrease dimensionality problem and improve understanding about underlying biological pathways. In the future work, we will improve our gene set modules by considering other information sources such as protein interaction networks and pathway databases.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2020.103570.

## References

[1] L.J. Van't Veer, H. Dai, M.J. Van De Vijver, Y.D. He, A.A. Hart, M. Mao, et al., Gene expression profiling predicts clinical outcome of breast cancer, Nature 415 (2002) 530–536.
[2] M.J. Van De Vijver, Y.D. He, L.J. Van't Veer, H. Dai, A.A. Hart, D.W. Voskuil, et al., A gene-expression signature as a predictor of survival in breast cancer, N. Engl. J. Med. 347 (2002) 1999–2009.
[3] Y. Wang, J.G. Klijn, Y. Zhang, A.M. Sieuwerts, M.P. Look, F. Yang, et al., Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer, The Lancet. 365 (2005) 671–679.
[4] J. Choi, S. Park, Y. Yoon, J. Ahn, Improved prediction of breast cancer outcome by identifying heterogeneous biomarkers, Bioinformatics 33 (2017) 3619–3626.
[5] H.Y. Chuang, E. Lee, Y.T. Liu, D. Lee, T. Ideker, Network-based classification of breast cancer metastasis, Mol. Syst. Biol. 3 (2007) 140.
[6] J. Ruan, A.K. Dean, W. Zhang, A general co-expression network-based approach to gene expression analysis: comparison and applications, BMC Syst. Biol. 4 (2010) 8.
[7] M. Chen, M.W. Deem, Hierarchy of gene expression data is predictive of future breast cancer outcome, Phys. Biol. 10 (2013), 056006.
[8] C. Park, J. Ahn, H. Kim, S. Park, Integrative gene network construction to analyze cancer recurrence using semi-supervised learning, PLoS ONE 9 (2014), e86309.
[9] X. Tian, M. Xin, J. Luo, M. Liu, Z. Jiang, Identification of genes involved in breast cancer metastasis by integrating protein–protein interaction information with expression data, J. Comput. Biol. 24 (2017) 172–182.
[10] Z. Wang, Z. He, M. Shah, T. Zhang, D. Fan, W. Zhang, Network-based multi-task learning models for biomarker selection and cancer outcome prediction, Bioinformatics 36 (2020) 1814–1822.
[11] M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C.M. Perou, et al., Adjustment of systematic microarray data biases, Bioinformatics 20 (2004) 105–114.
[12] L. Ein-Dor, O. Zuk, E. Domany, Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer, Proc. Natl. Acad. Sci. 103 (2006) 5923–5928.
[13] A.-C. Haury, P. Gestraud, J.-P. Vert, The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures, PLoS ONE 6 (2011), e28210.
[14] H. Xie, J. Li, Q. Zhang, Y. Wang, Comparison among dimensionality reduction techniques based on Random Projection for cancer classification, Comput Biol Chem. 65 (2016) 165–172.
[15] M. Sehhati, A. Mehridehnavi, H. Rabbani, M. Pourhossein, Stable gene signature selection for prediction of breast cancer recurrence using joint mutual information, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 12 (2015) 1440–1448.
[16] T. Nguyen, A. Khosravi, D. Creighton, S. Nahavandi, A novel aggregate gene selection method for microarray data classification, Pattern Recogn. Lett. 60 (2015) 16–23.
[17] J. Li, A.E. Lenferink, Y. Deng, C. Collins, Q. Cui, E.O. Purisima, et al., Identification of high-quality cancer prognostic markers and metastasis network modules, Nat. Commun. 1 (2010) 34.
[18] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learn. 46 (2002) 389–422.

[19] B. Haibe-Kains, C. Desmedt, C. Sotiriou, G. Bontempi, A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? Bioinformatics 24 (2008) 2200–2208.
[20] M.H. van Vliet, C.N. Klijn, L.F. Wessels, M.J. Reinders, Module-based outcome prediction using breast cancer compendia, PLoS ONE 2 (2007), e1047.
[21] G. Abraham, A. Kowalczyk, S. Loi, I. Haviv, J. Zobel, Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context, BMC Bioinf. 11 (2010) 277.
[22] F.S. Varn, M.H. Ung, S.K. Lou, C. Cheng, Integrative analysis of survival-associated gene sets in breast cancer, BMC Med Genomics. 8 (2015) 11.
[23] A.P. Kumar, A.J. Kovatich, A. Biancotto, F. Cheung, J.K. Davidson-Moncada, L. Kvecher, et al., Abstract P4–09-14: Analysis of breast cancer recurrence using gene set enrichment analysis, AACR (2018).
[24] S.R. Sirkisoon, R.L. Carpenter, T. Rimkus, A. Anderson, A. Harrison, A.M. Lange, et al., Interaction between STAT3 and GLI1/tGLI1 oncogenic transcription factors promotes the aggressiveness of triple-negative breast cancers and HER2-enriched breast cancer, Oncogene 1 (2018).
[25] O.M. Rueda, S.-J. Sammut, J.A. Seoane, S.-F. Chin, J.L. Caswell-Jin, M. Callari, et al., Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups, Nature 567 (2019) 399.
[26] J. Cuzick, Predicting late recurrence in ER-positive breast cancer. Nature Reviews, Clinical Oncol. 1 (2019).
[27] S.R. Eddy, Multiple alignment using hidden Markov models, Ismb (1995) 114–120.
[28] M. Stanke, S. Waack, Gene prediction with a hidden Markov model and a new intron submodel, Bioinformatics 19; 2003:ii215–ii25.
[29] J. Alon, S. Sclaroff, G. Kollios, V. Pavlovic, Discovering clusters in motion time-series data. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003 Proceedings: IEEE; 2003. p. I-I.
[30] G.N. Armaiz-Pena, S.W. Cole, S.K. Lutgendorf, A.K. Sood, Neuroendocrine influences on cancer progression, Brain Behav. Immun. 30 (2013) S19–S25.
[31] T. Nguyen, A. Khosravi, D. Creighton, S. Nahavandi, Hidden Markov models for cancer classification using gene expression profiles, Inf. Sci. 316 (2015) 293–307.
[32] L. Borrajo, A.S. Vieira, E.L. Iglesias, TCBR-HMM: an HMM-based text classifier with a CBR system, Appl. Soft Comput. 26 (2015) 463–473.
[33] A.S. Vieira, L. Borrajo, E.L. Iglesias, Improving the text classification using clustering and a novel HMM to reduce the dimensionality, Comput Methods Programs Biomed. 136 (2016) 119–130.
[34] K. Yi, J. Beheshti, A hidden Markov model-based text classification of medical documents, J. Inform. Sci. 35 (2009) 67–81.
[35] C. Desmedt, F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, et al., Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series, Clin. Can. Res. 13 (2007) 3207–3214.
[36] S. Loi, B. Haibe-Kains, C. Desmedt, F. Lallemand, A.M. Tutt, C. Gillet, et al., Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade, J. Clin. Oncol. 25 (2007) 1239.
[37] L.D. Miller, J. Smeds, J. George, V.B. Vega, L. Vergara, A. Ploner, et al., An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival, Proc. Natl. Acad. Sci. 102 (2005) 13550–13555.
[38] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, et al., Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis, J. Natl Can. Inst. 98 (2006) 262–272.
[39] M. Schmidt, D. Böhm, C. Von Törne, E. Steiner, A. Puhl, H. Pilch, et al., The humoral immune system has a key prognostic impact in node-negative breast cancer, Can. Res. 68 (2008) 5405–5413.
[40] Q. Li, N.J. Birkbak, B. Gyorffy, Z. Szallasi, A.C. Eklund, Jetset: selecting the optimal microarray probe set to represent a gene, BMC Bioinf. 12 (2011) 474.
[41] L.E. Baum, T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, Ann. Math. Stat. 37 (1966) 1554–1563.
[42] L.E. Baum, J.A. Eagon, An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology, Bull. Am. Math. Soc. 73 (1967) 360–363.
[43] L.E. Baum, T. Petrie, G. Soules, N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, Ann. Math. Stat. 41 (1970) 164–171.
[44] L. Baum, An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process, Inequalities. 3 (1972) 1–8.
[45] M.H. Davis, Markov models & optimization, Routledge, 2018.
[46] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (1989) 257–286.
[47] M. Momenzadeh, M. Sehhati, H. Rabbani, A novel feature selection method for microarray data classification based on hidden Markov model, J. Biomed. Inform. 103213 (2019).
[48] T. Nikolaos, T. George, Document classification system based on HMM word map, in: Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, ACM; 2008. p. 7–12.
[49] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. 102 (2005) 15545–15550.
[50] B. Efron, R. Tibshirani, On testing the significance of sets of genes, Ann. Appl. Statist. 1 (2007) 107–129.
[51] S.-Y. Kim, Y.S. Kim, A gene sets approach for identifying prognostic gene signatures for outcome prediction, BMC Genomics 9 (2008) 177.
[52] M. Ackermann, K. Strimmer, A general modular framework for gene set enrichment analysis, BMC Bioinf. 10 (2009) 47.
[53] M. Lauss, A. Kriegner, K. Vierlinger, I. Visne, A. Yildiz, E. Dilaveroglu, et al., Consensus genes of the literature to predict breast cancer recurrence, Breast Can. Res. Treat. 110 (2008) 235–244.
[54] P. Carmona-Saez, M. Chagoyen, F. Tirado, J.M. Carazo, A. Pascual-Montano, GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists, Genome Biol. 8 (2007) R3.
[55] R. Nogales-Cadenas, P. Carmona-Saez, M. Vazquez, C. Vicente, X. Yang, F. Tirado, et al., GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information, Nucleic Acids Res. 37 (2009) W317–W322.
[56] D. Tabas-Madrid, R. Nogales-Cadenas, A. Pascual-Montano, GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics, Nucleic Acids Res. 40 (2012) W478–W483.
[57] J. Zyla, M. Marczyk, J. Weiner, J. Polanska, Ranking metrics in gene set enrichment analysis: do they matter? BMC Bioinf. 18 (2017) 256.
[58] T.K. Kim, T test as a parametric statistic, Kor. J Anesthesiol. 68 (2015) 540.
[59] Y. Zhu, L.C. De Silva, C.C. Ko, Using moment invariants and HMM in facial expression recognition, Pattern Recogn. Lett. 23 (2002) 83–91.
[60] U.M. Braga-Neto, E.R. Dougherty, Is cross-validation valid for small-sample microarray classification? Bioinformatics 20 (2004) 374–380.
[61] M. Momenzadeh, M. Sehhati, A. Mehri Dehnavi, A. Talebi, H. Rabbani, Automatic diagnosis of vulvovaginal candidiasis from Pap smear images, J. Microsc. 267 (2017) 299–308.
[62] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recogn. 30 (1997) 1145–1159.
[63] S. Boughorbel, F. Jarray, M. El-Anbari, Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric, PLoS ONE 12 (2017), e0177678.
[64] R. Durbin, S.R. Eddy, A. Krogh, G. Mitchison, Biological sequence analysis: probabilistic models of proteins and nucleic acids, Cambridge University Press, 1998.