Introduction

We proposed a hybrid computational method to select important high-impact genes in a high dimensional dataset. The proposed method is based on area under the curve (AUC) and hidden markov model (HMM). Here we first describe the Background of the techniques used in the proposed model and then discuss the Proposed Method in details.

Background of the techniques used

Binary Classification

In binary classification, only two classes are involved. Therefore, each instance *I* is mapped to one element of the set of positive and negative class labels i.e., {+, - } or {p, n}.

Classification model

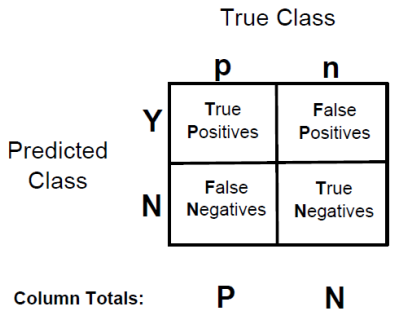
A classification model maps instances to predicted classes. In order to distinguish between the actual class and the predicted class, different labels can be used for the class predictions produced by a model like {Y,N}.

Binary Classification Outcomes

The four possible outcomes from a binary classifier and an instance are as follows.

true positive: If the instance is positive and it is correctly classified as positive
false negative: If the instance is positive and it is incorrectly classified as negative
true negative: If the instance is negative and it is correctly classified as negative
false positive: If the instance is negative and it is incorrectly classified as positive

Figure 1 shows the confusion matrix, also called contingency table, and the common performance metrics calculated from it.



$$\text{fp rate} = \frac{FP}{N} \qquad \text{tp rate} = \frac{TP}{P}$$

$$\text{precision} = \frac{TP}{TP + FP} \qquad \text{fp rate} = \frac{FP}{N}$$

$$\text{accuracy} = \frac{TP + TN}{P + N}$$

Figure 1 Confusion matrix and common performance metrics calculated from it

Receiver operating characteristic (ROC) curve

An ROC curve is a popular method used to evaluate the discriminative performance of binary classifiers in machine learning. By varying the discrimination threshold for a binary classifier, the ROC curve measure can be obtained by plotting the curve of true positive rate (sensitivity) versus false positive rate (1 - specificity). When the ROC curve matches with the upper left corner of the

ROC space, the best performance would be achieved as this would yield 100% sensitivity and 100% specificity. Moreover, the closer the ROC curve is to the upper part of the ROC space, the better the performance of the classifier.

Figure 2 shows ROC curves for 3 different predictors whereas the dotted line shows the line that denotes the average AUC.

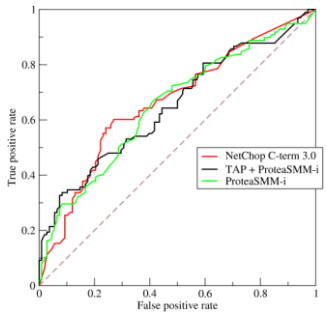


Figure 2 ROC curves for 3 different predictors

Area under the ROC curve (AUC)

AUC is a common method is to calculate the area under the ROC curve to reduce ROC performance to a single scalar value that represents the expected performance. The value of any AUC will always lie between 0 and 1 since the AUC calculates the portion of the area of the unit square. However, since random guessing produces the diagonal line between (0, 0) and (1, 1) with an area of 0.5, no realistic classifier should have an AUC less than 0.5. Hence, an AUC value close to 1 indicates better performance. Figure 3 shows the areas under two ROC curves, A and B where classifier B has greater area and, therefore, better average performance than classifier A.

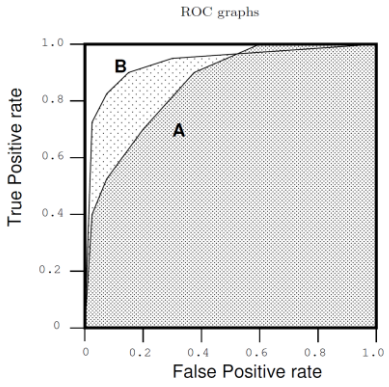


Figure 3 An ROC graph that shows the area under two ROC curves for two classifiers A and B.

Hidden Markov Model

An HMM is a model in which the system being modeled is assumed to be a Markov process with unknown parameters. HMM builds a causal model for observation sequence O = (o_1, o_2,…, o_n) by introducing corresponding 'hidden states' q = (q_1, q_2,…, q_m). The parameters of the HMM are λ = (a; b; π) where *a* is the parameter for the transition model $P(q_t|_{qt})$ and *b* is the parameter for the observation model $P(o_t|q_t)$ and π denote the initial probability. The hidden parameters are determined from the observable parameters. The extracted model parameters can then be used to perform further analysis; for example, pattern analysis or feature subset selection.

For better understanding the HMM, let us consider an example of persons and stick model. Assume there are N persons in a closed room (see Figure 4) and an observer standing outside the room. Each person is holding a number of colored sticks. The sticks are of M distinct color. Now, each person throws the colored sticks, one after another, out of the room. At this stage, the only visible outcome, to the observer, is the sequence (order) of colored sticks received. He does not know who, out of N persons,) has thrown which colored sticks.

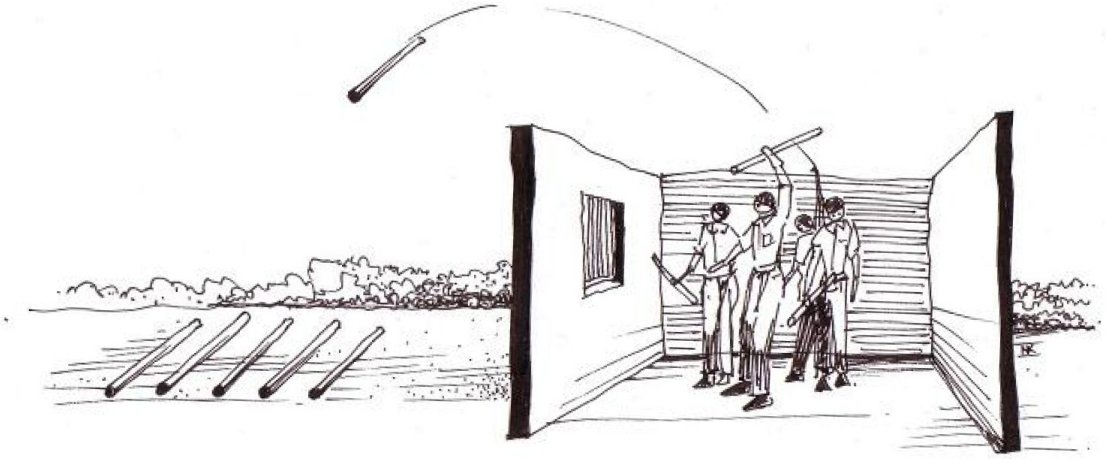


Figure 4 The person and stick model

In this example, both the selection of a person and color are completely random. The individuals are hidden from the observer, making it a hidden process. The overall process generates a sequence of colors/colored sticks, which forms the output observation sequence. For this process, we neither know the sequence in which a person is throwing the colored sticks, nor the sequence of persons throwing the sticks. The output sequence is very much dependent on the transition probabilities between the various persons/states, and the choice of initial person/state at the beginning. This example can be associated with an HMM, where the states are hidden (like the persons in the example) and the output is the sequence of observations (colored sticks).

Proposed Method

Our proposed method consists of two passes. In pass one, we rank all genes and determine the highly ranked genes one-at-a-time. In the second pass, we investigate the interrelationship amongst the genes using a Hidden Markov Model (HMM) to find the best subset of genes among the ranked genes. Details of these passes are described in the subsequent sections. Figure 5 shows the two passes with the various operations in them. The two passes are discussed in details in the subsequent sections.

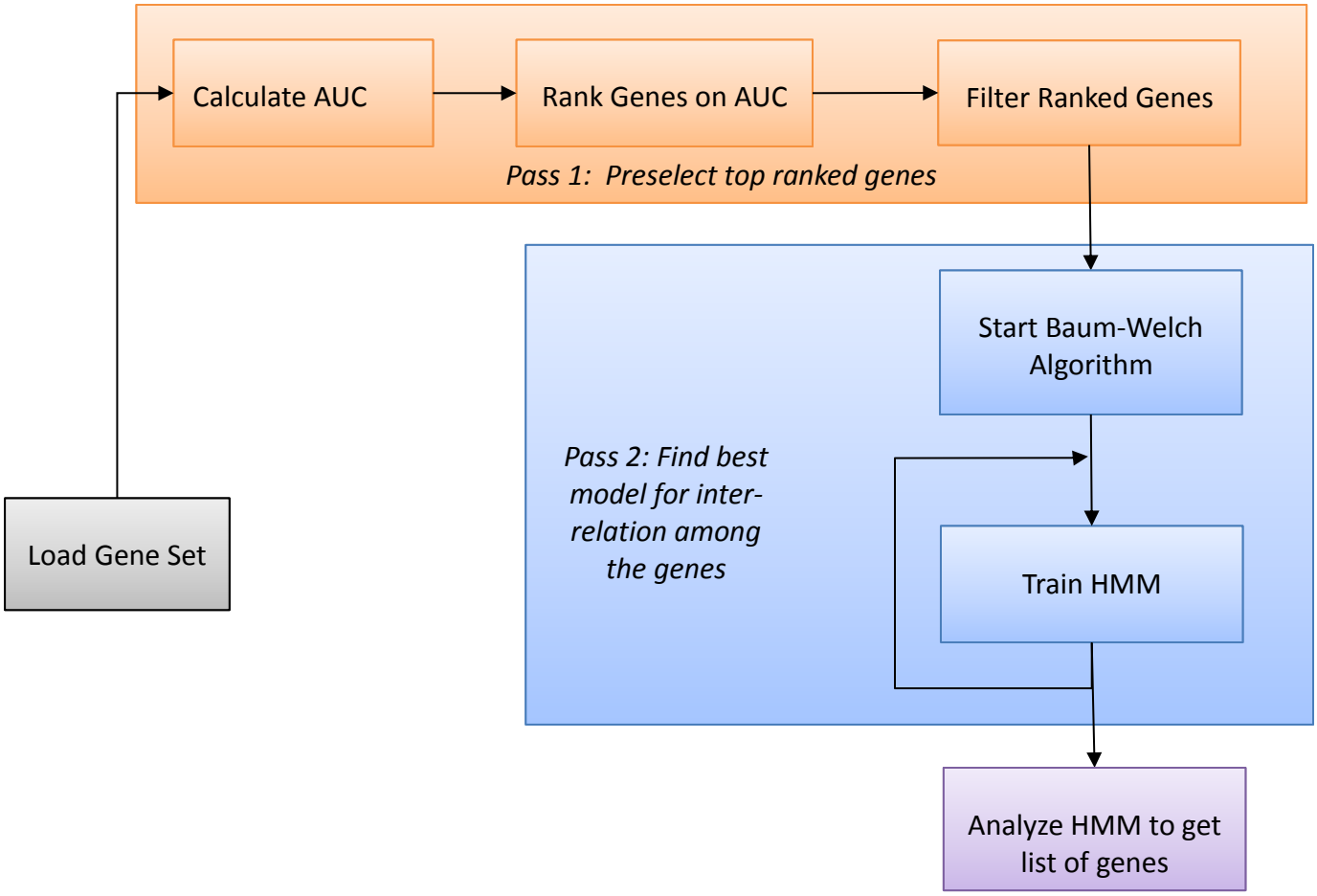


Figure 5 The proposed method

Ranking of genes using AUC

We start this process by calculating the AUC for each gene as described below:

1. Select the gene
2. Sort the gene expression with the class level. This class level is the actual value when calculating the AUC score
3. Calculate AUC
    a. For the predicted value, assign *+1* incrementally, from the first sample till the last and calculate the AUC for each. (Start by assigning a *+1* to the first sample. Then calculate the different measures like *TP*, *FP*, *Sensitivity*, *and Specificity* for the actual value and the predicted value and calculate the AUC. Next assign *+1* to the first 2 samples, calculate the measures and the AUC. Repeat until all samples have +1 as their predicted value, calculate the different measures and calculate the AUC. Finally select the maximum AUC among all the AUCs calculated for various threshold settings)
4. Goto step 1 and select the next gene.

When this procedure ends, each gene will have a maximum AUC score. Then we filter all less important genes by setting a threshold value. The threshold value that we used is $AUC \geq 0.8$, i.e. if the AUC value of any gene is less than or 0.8, we discard this gene as unimportant. Those genes that have AUC values higher than the threshold are passed to the next phase. Figure 6 shows the pictorial view of calculating the AUC

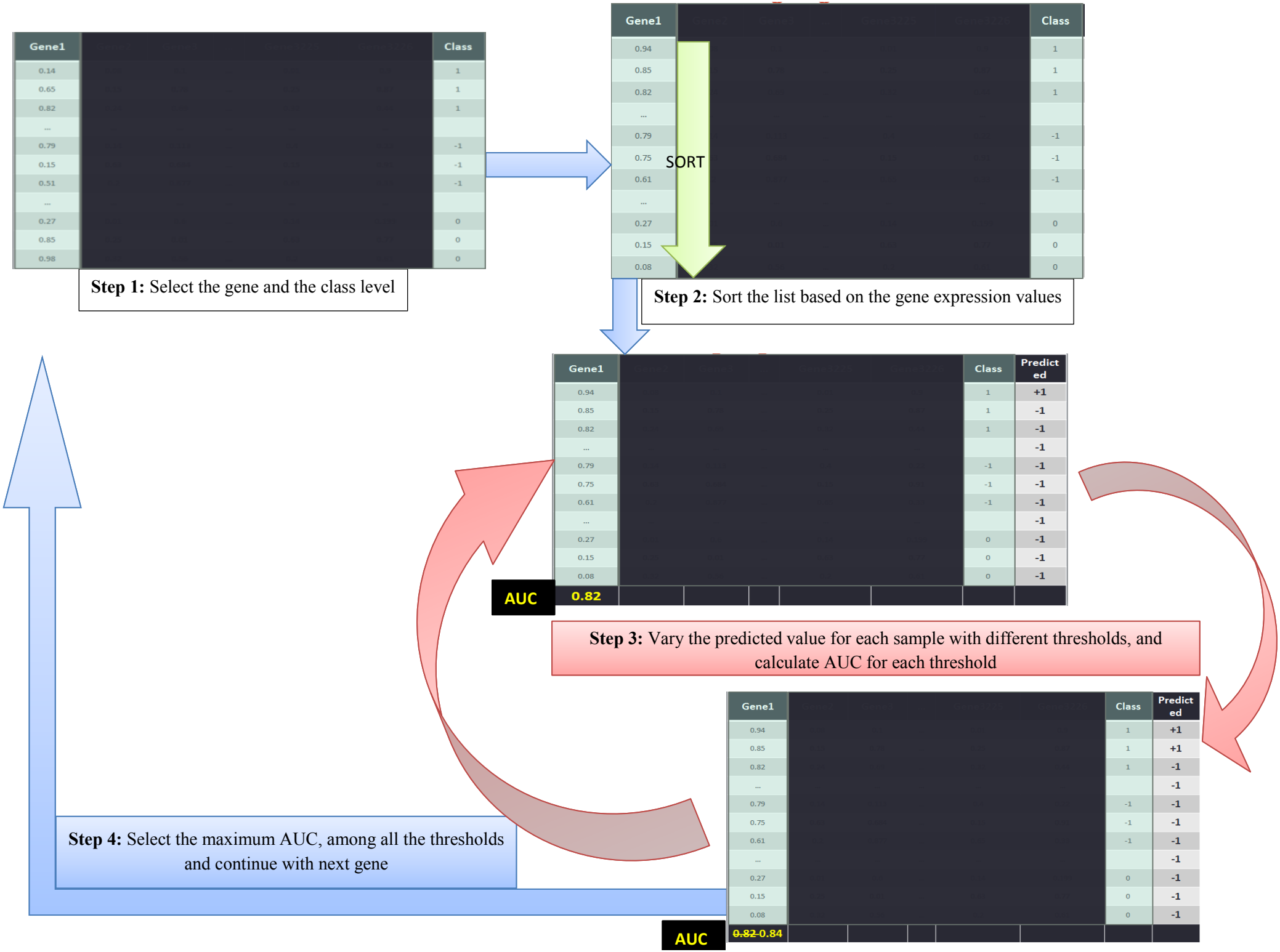


Figure 6. Procedure of calculating AUC for each gene

Gene subset selection using Hidden Markov Model

This pass begins by receiving a list of important genes that have an AUC ≥ 0.8 from the previous pass. In our HMM, each state represents a gene. We start training the hidden markov model and calculate the two-way Wilcoxon Rank-sum measurement for each state (gene) with the other state (gene). As show in in Figure 7, the emission probability matrix consists of the p-values from the rank-sum statistic. One cell in the matrix is the numerical value of the ranksum measure between two genes expression vectors (expression profiles values). Hence, if we have 20 genes, the matrix size will have 20 rows and 20 columns (20x20) where each cell holds the ranksum measure of each gene with all the other genes. The proposed null hypothesis for comparing any two genes is that "two genes are similar if they have a p value greater than 0". Since each state represents a gene, we also need to analyze the pattern of the p-values in the emission matrix.

In order to select the best possible model that fits the given data, we use the well-known Baum Welch expectation maximization algorithm. It runs in a loop either for 50 cycles or until the log likelihood value of the new model is less than the log likelihood value of the previous model (indicating no better model possible). This ensures that the best model is selected among all the possible models that fit the given data.

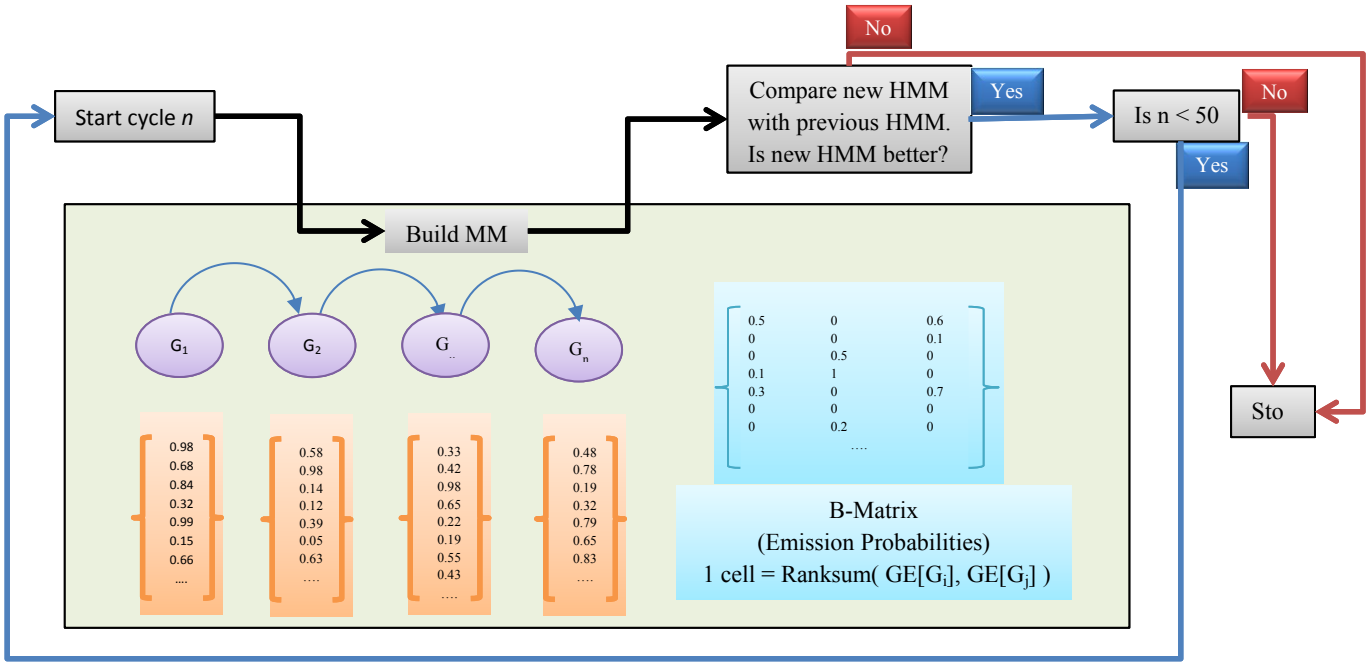


Figure 7 Building HMM

Once the best model has been selected, we study the outputs of the hidden markov model especially the emission probability matrix. We identify all rows and columns that have non-zero values and study the pattern in which the values are spread over the matrix. Then we compare the values with the null hypothesis to identify the genes that are distinct as well as those genes that are common. Figure 8 shows the emission matrix with values and one common and one unique gene. In case of common genes i.e., genes whose p-values are non-zeroes and their patterns are similar in the emission matrix, we select the gene that has the highest AUC among them.

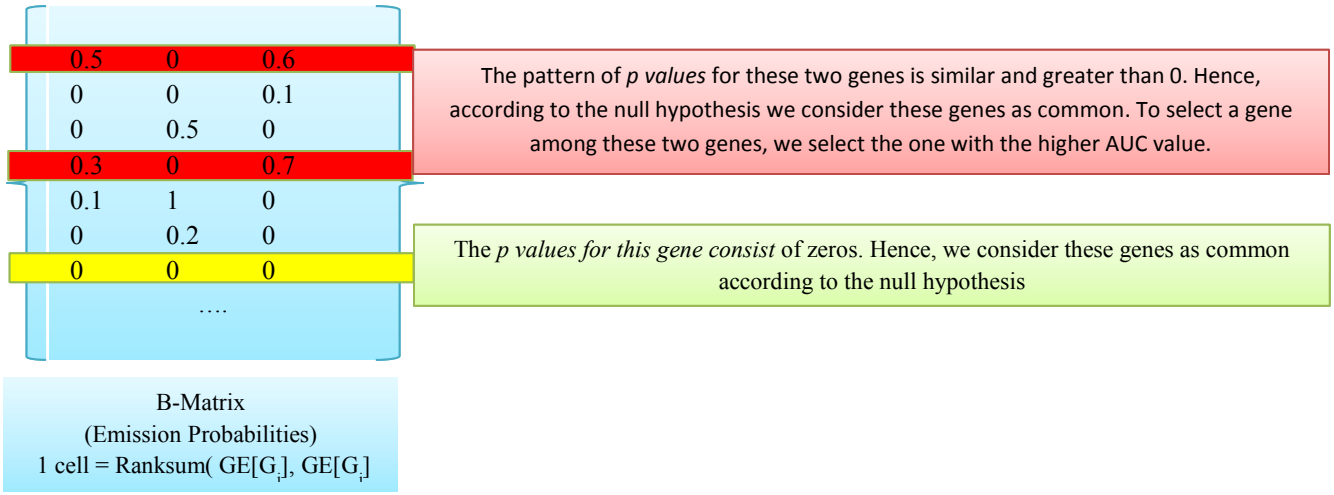


Figure 8 Analyzing the emission probability matrix

After analyzing the emission matrix, we have a list of unique genes. Initially, we used the whole dataset for modeling the HMM and got a list of genes that were important. Then, we used leave one out cross validation (LOOCV) to select only the most important genes. For each cycle, we leave one sample out and build the model with the remaining samples. Since our dataset has 15 samples for BRCA1 vs. BRCA2 analysis, the LOOCV results in 15 different genes list and the first list without LOOCV. We selected only those genes that were most recurring among the 15 LOOCV lists and whose AUC values were highest. The full list of genes, with and without LOOCV, is provided in the supplementary documents.