How to Cite:
Behera, N., Sinha, S., Srivastava, A. K., Hairat, S., Neha, N., & Prasanta, P. (2022). Analysis of microarray data by genetic algorithm . *International Journal of Health Sciences*, *6*(S6), 7721–7743. https://doi.org/10.53730/ijhs.v6nS6.11068

Analysis of microarray data by genetic algorithm

Narayan Behera
Institute of Bioinformatics and Applied Biotechnology, Electronics City, Phase I,Bengaluru – 560100, India | Department of Applied Physics, School of Natural Sciences, Adama Science and Technology University, Adama, P O Box 1888, Ethiopia | SVYASA University, Eknath Bhavan, Kempegowda Nagar, Bengaluru 560019, India
*Corresponding author email: narayanbehera@svyasa.edu.in

Shruti Sinha
Institute of Bioinformatics and Applied Biotechnology, Electronics City, Phase I, Bengaluru – 560100, India

Ankit K. Srivastava
Department of Applied Physics, School of Natural Sciences, Adama Science and Technology University, Adama, P O Box 1888, Ethiopia | School of Science, Indrashil University, Mehsana 382740, India

Suboot Hairat
Department of Biotechnology, Wachemo University, Hossana, Ethiopia

Neha
Department of Biotechnology, Deenbandhu Chhoturam University of Science and Technology, Murthal, Sonipat, India

Prasanta
Department of Biotechnology, Manav Rachna International Institute of Research and Studies, Surajkund, Faridabad, Haryana (India) -121004

Abstract---Microarray gene expression data is used to understand the actions of thousands of genes. Just a few genes out of thousands have a significant impact in any cancer process. Finding these defective genes using experimental data is impractical. To locate the relevant genes, computational techniques are required. A method to identifying cancer candidate genes from microarray data is created. Clustering of similar genes is necessary to find co-expressed genes in different biological conditions. It is important to develop methods to find the few candidate genes for cancer. An optimization process is used for such purpose. A genetic algorithm employs the principles of evolution:

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
Manuscript submitted: 9 April 2022, Manuscript revised: 27 June 2022, Accepted for publication: 18 July 2022

selection, recombination, and mutation to solve an optimization problem. Mutual information is used to find the dependency between genes. The two genes are similar if their expression levels are comparable. The similarity as well as positive and negative correlations between genes is considered while clustering them. Interdependence measure tells how the genes are correlated. The genes responsible for a sick state have higher interdependence measures. These genes are defective genes having cancer diagnostic information. Here microarray gene expression datasets from gastric cancer and colon cancer from the public domain are considered. The candidate genes found by this genetic algorithm model can identify known microarray samples as malignant or normal with high accuracy. The new algorithm creates even distribution of genes in the clusters. Furthermore, the computational tool can group genes having higher differences in gene expressions. This algorithm can assist in computational drug discovery process.

Keywords---genetic algorithm, grouping genes, classification of microarray samples, candidate genes for disease, microarray data, entropy, mutual information, optimization.

Introduction

Microarray technology measures gene expression levels across the genome of an organism. The DNA microarray technique monitors the interactions of thousands of genes simultaneously through their expression levels. Genes with varied expression levels under different experimental settings are subjected to differential gene expression analysis. Different tissue types or developmental stages of the organism might be used as experimental levels. Normal-versus-diseased state studies are a common type of analysis. Differential gene expression is analogous to gene co-regulation. It's necessary to find the genes whose expression levels are associated across the different experimental circumstances or samples being analyzed. Two genes are said to have positive correlation if the expression level of both genes rises simultaneously. On the other hand, they have negative correlation when one gene's expression level rises while that of the other falls at the same time. Gene regulation analysis requires looking into many gene profiles. Knowing the activities of genes in many biological systems has become a research area after the human genome project. Functional genomics has been transformed by the introduction of microarray technology. Scientists may now analyse hundreds of genes and their expression patterns in parallel under a variety of experimental settings. On a genome-scale, this sheds light on gene products behavior in normal and pathological conditions. The cellular activities of genes and the regulatory mechanisms in disease processes can be understood using this mechanism.

Gene function is revealed by microarray research. The procedure entails comparing the expression patterns of the desired genes. A gene profile is a set of gene expression data of a fixed gene across several samples. An array profile shows the gene expression levels of many genes in a single experimental

condition. The function of an unknown gene can be effectively inferred if the expression profile of a known gene matches the expression profile of an unknown gene. The functionalities of genes with very comparable expression patterns are used to predict the function of new genes. Data on gene expression is very useful in clinical diagnosis. They can identify expression patterns that are associated with a certain illness. Computational and statistics approaches are used to analyze microarray data. Microarray data analysis consists of numerous processes. To minimize the data volume, the raw data is first processed. A gene data matrix is a set of values that indicate the expression levels of a gene in multiple samples. The measured value deviates from the real expression level due to the measurement error caused by the faulty equipment. To cope with measurement inaccuracies, normalization is necessary.

Microarray gene expression levels for gastric cancer and colon cancer are investigated here. The literature on experimental and computational microarray data analysis can be found here (Berrar et al 2003). Razak explains the data mining process in depth (2013). In cancer bioinformatics, Liu (2009) presented computational data mining. Solmaz et al. created SEMA, an online platform for generating models of cancer processes using massive cancer genomic data sets (2019). The fundamentals of data mining techniques and their applications in genomics, proteomics, and medical analysis were described by Herbola et al (2022). In recent years, the field of microarray data analysis has exploded (Zhang et al 2009; Selvaraj and Natarajan 2011; Koschmieder et al 2012; Amaral et al 2018). Cancer identification studies involve greatly microarray data in recent days. Different data mining classification tools for microarray data with information on speed and accuracy of algorithms were systematized (Aydadenta and Adiwijaya 2017). The Deep Gene Selection (DGS) algorithm was investigated on several public microarray datasets and it provided better results in classifying different types of cancers Alanni et al 2019). The selected genes of the algorithm showed high degree of sensitiveness to the samples' classes.

Algorithms for genes cluster

Computational tools to group genes together are necessary. Microarray data clustering is a fertile area of research in bioinformatics. Under specific situations, they can show the nature of co-regulation. Genes must be chosen that can reliably categorize test samples into sick and non-diseased categories. In general, the number of genes in microarray data far exceeds the sample size. This makes sample categorization and gene selection challenging. It's just essential to have a limited number of genes containing diagnostic information. All genes taken together may increase noise in the system and makes analysis difficult. A small number of important genes can lower the data's dimensionality. Then these genes can be exploited easily for diagnostic research.

In the recent past, several clustering methods have been created. The K-means algorithm is a basic clustering technique (Lloyd 1982). But a priori determination of the number of clusters, k, is necessary in this process. The goal is only to cluster the data points. There is no need to compare them to any benchmark performance. It divides the genes into k separate partitions depending on particular attributes/features (e.g. gene expression levels). Each gene is only part

of one cluster. The microarray data is used to obtain a range of gene expression levels.

The algorithm's stages are as follows. The k cluster centers are first taken at random. Each data point is assigned to the centre that is nearest to it. The cluster centre is then determined by taking the mean values of data points from a particular cluster centre. Now the new cluster center is changed. Then all the new cluster centers are determined. The iteration process is repeated until the values of the cluster centers remain unchanged. As a result, k clusters are created to group comparable genes. The expression levels of the genes inside each cluster are comparable. Between the clusters, the genes are more different. In general, related genes in an organism perform the same activities. Other clustering techniques are mentioned below but not in depth. They don't immediately relate to the topic at hand, but they illumine clustering tools used in the literature.

The Self-Organizing Map (SOM) is an unsupervised learning method of clustering (Kohonen 1990). A neural-network type of array represents the SOM. The cells are adjusted to various input signal patterns in a systematic manner. The process of learning is unsupervised. Hence, the SOM converts high-dimensional data into low dimensional representation. The technique for hierarchical clustering features a simple visualization tool (Johnson 1967). A tree with individual members at one end is the standard depiction of this structure. Agglomerative approaches are differentiated from hierarchical clustering. A succession of gene fusions into groups is carried out. The genes are then grouped into ever finer groups. Bi-clustering is another tool relies on data mining method. The rows and columns of a matrix of data are clustered simultaneously (Cheng and Church 2000). The technique creates bi-clusters from a collection of m rows and n columns. They are subsets of rows that show similar types of behavior on a subset of columns, or the vice versa. For clustering gene data from microarray datasets, all of these clustering algorithms are routinely utilized (Eisen et al 1998; Heyer et al 1999; Golub et al 1999). The expression levels of a gene under a certain experimental setting can be a time series data. The Euclidian distance and the Pearson's correlation coefficient are two extensively used distance metrics to cluster data. The Euclidian distance (s) between two genes, x and y, with n characteristics may be calculated as follows:

$$s = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (1)$$

Here, x_i is the value of the expression level i of gene x , and y_i is the value of expression level i of gene y. The Pearson correlation coefficient (p) between two genes, x and y, with n characteristics is calculated as follows:

$$p = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})}} \qquad (2)$$

where $\bar{x}, \bar{y}$ are the average of expression values of genes x and y respectively, x_i is the value of ith expression level of gene x and y_i is the value of the ith expression level of gene y. For negative p, a negative linear relationship between x and y exists. A positive value of p implies a positive linear relationship. The genes are independent if p is equal to zero.

To calculate the similarity between two genes, traditional clustering techniques employ a Eucledian or Pearson's correlation as distance measurements. There are various restrictions to distance measurements. Genes with similar profile shapes are usually functionally linked. The Euclidian distance compares two expressions based on their expression values rather than their profile form. As a result, genes with similar expression profile shapes but big differences in gene expression are less likely to be grouped together. Genes with slight differences in expression profile shape and expression profiles that differ by a small magnitude may be grouped together. In terms of defining relationships, the Pearson's correlation outperforms the Euclidian distance. It is, nevertheless, sensitive to outliers. A new distance measure through mutual information has been suggested recently. This is the Attribute Clustering Algorithm (ACA) by Au et al 2005. It overcomes some of the drawbacks of traditional distance measurement methods. The ACA is based on the k-means idea. The genetic distance here is interdependent redundancy measure. It considers the interrelationships between any two genes.

Two genes, x and y, each with n expression levels has Interdependency Redundancy measure defined as:

$$IR(x:y) = \frac{M(x:y)}{E(x:y)} \tag{3}$$

where $M(x:y) = \sum_{k=1}^{g}\sum_{l=1}^{h} P(v_k \Lambda v_l) \log \frac{P(v_k \Lambda v_l)}{P(v_k)P(v_l)}$, (4)

and

(5)

where M (x: y) is the mutual information, E(x: y) stands for the joint entropy, g is the number of x intervals, h denotes the number of y intervals, v_k is the k th interval of x, v_l is the l th interval of y, and $P(v_k\ v_l)$ is the probability of a gene value appearing in the interval v_k and v_l, . $P(v_k)$ is the mutual information that deals with two genes' dependency, but it increases as the number of possible gene expression levels grows. Thus it is normalized by dividing it with entropy value.

The idea of Clustering by Mutual Information (CMI) is defined as follows. It is similar to the one developed by Au et al (2005). But there is an exception of the fact that they employ a different smoothing method for the dataset (Behera et al 2018). However, the gene distributions in the clusters have a limitation resulting

in narrow search space. Finding the optimum gene cluster can be biased by this limitation. The best answer in this case may not be the global optimum. Therefore, a hybrid of stochastic computation and clustering algorithm is created to solve this drawback (Behera et al 2018). Evolutionary computations are search methods to find the best solution to a research problem. They can be used to investigate a large variety of initial gene distributions in clusters. The Evolutionary Clustering Algorithm (ECA) is created keeping this idea in mind (Behera et al 2018). The ECA is an extension of CMI employing Darwinian principles of evolution by natural selection. It improves gene grouping and selection by eliminating the limits imposed by the clusters' skewed initial distribution of genes.

This technique is applied to two datasets of taken from the public domain available on the internet: gastric cancer (Tsutsumi et al 2002), and colon cancer (Alon et al 1999). The findings are compared to those of other well-known clustering techniques. A decision tree algorithm, C4.5 is used to create classifiers based on a subset of candidate genes (Quinlan 1993). The C4.5 program is a standard method used in learning (Elomaa 1994). From a set of data, a decision tree is generated. It employs certain concepts from information theory. The decision tree classifier has been used in data mining research.

Methodology

Partitioning of data

A vast majority of data is generally continuous. For any quantitative study, they must be discretized. Noise is frequently present as a result of measurement mistakes or inaccurate value entering. In the discretization process, this noise might result in a huge number of intervals. The greater number of intervals tends to increase the amount of information lost. A good method must address this problem. The gene expression levels in microarray data are generally continuous for all practical purposes. To facilitate computation of an information measure, they must be partitioned into appropriate intervals. To discretize a continuous data, the Optimal Class Dependent Discretization (OCDD) method is used (Wong et al 2004). It creates a near-global solution. The OCDD takes into account the interaction of the classes and the values of gene expressions. Then it reduces the amount of data lost (Au et al 2005). Each sample belongs to a fixed category. In this case, there are two types of samples: normal and diseased. Smoothing of the data and a chi-square test are used in the model to offset information loss due to large intervals in data. Before doing partitioning, smoothing is performed to reduce noise. The number of intervals is decreased when using the Chi-square test. The ECA method uses somewhat different smoothing and chi-square test than specified in the OCDD algorithm.

Genetic Algorithm

A genetic algorithm is a procedure to finding true or approximate solution of an optimization problem via an evolutionary search. This method can solve complex optimization issues in science and engineering (Holland 1975; Goldberg 2008). The algorithms use mutation, selection, and crossover concepts of evolutionary

biology. It is an evolutionary algorithm. It mimics Darwin's idea that Nature is the best optimizer. A population of approximate solutions undergoes Darwinian selection over several generations. Mutation (new genes introduced spontaneously as a result of random fluctuations) is the raw material for the Darwinian evolution to act. Mutation is caused by a change in the sequence of certain nucleotides in the gene.

Recombination of genes in the chromosomes occurs due to the reproductive process. As a result, new chromosomes are generated. A haploid gene system is used in this example for illustrative purposes. A chromosome's crossing site is picked up at random. The genes in the parent 1 from the beginning of the chromosome up to the crossover point and the genes in parent 2 from the crossover point to the end of the chromosome are combined. Therefore, it produces a new chromosome. When parent 1 and parent 2 are swapped a new offspring is produced. They introduce new individuals into the population that did not exist in the previous generation.

Individuals who adapt well are more likely to survive in an environment and reproduce. Their genes are passed on to future generations. So the genes that confer greater fitness to for survival of a population become more common in following generations. The fitness is measured by the net offspring an individual passes to the next generation. The chief problem is to design a fitness function appropriate for the particular research topic. This genetic algorithm has been successfully applied to numerous problems in diverse areas: evolutionary biology (Behera and Nanjindiah 1995, 1996, 1997, 2004), multiple protein sequence alignment optimization (Behera et al 2017), and analysis of microarray gene expression data (Behera et al 2018). A genetic algorithm will be utilized to extract the essential genes responsible for cancer from the microarray gene expression data. A string (individual) is represented by a series of 1s and 0s. It represents a plausible solution to the optimization issue in question. The genetic algorithm begins with a set of solutions created at random (possible solutions) by guess work. As the solutions evolve, they become better in subsequent generations. Every string in the population gets changed (recombined and maybe mutated) probabilistically. As a result, new individuals are created in each generation. With each consecutive generation, the solution improves. The fitness of each individual is assessed. To generate a new population in next generation, multiple individuals are randomly taken from the present population based on certain fitness criteria. The new population is subsequently used in the algorithm following the iteration. The method ends when the population has reached a predestined level of fitness, or when the maximum number of generations (given as input parameter) has been reached. Finally, the optimal solution corresponds to the individual having the highest fitness.

Evolutionary Clustering Algorithm

Mutual information idea of the CMI algorithm is used. The ECA implements the interdependence redundancy measure. The modes of the clusters are taken into account to find the distance metric. The gene with the largest multiple interdependent redundancy (MIR) is treated as the cluster's mode. The MIR of a

mode is defined as the total of its IR measures with all other genes in a fixed cluster. Figure 1a depicts the ECA's flow diagram.

Creation of Individuals

A two-dimensional array of integers is used to represent each individual. The gene number from the microarray data is represented by one index. The cluster number is shown by the other. It indicates that the genes relate to specific clusters in different ways. A set of gene expression values equal to the number of clusters is selected in a random manner. They become the clusters' modes. The remaining genes are assigned to the clusters. It's done by comparing the greatest IR values of the genes to the modes of the required clusters. The members of the population are created in a similar manner. An individual is a set of clusters. Its overall number of genes remains unchanged. This holds true for the population as a whole. A heterogeneous population of individuals with different gene distributions is created. Changing the seed values in the algorithm can generate different initial gene distribution. The population size is set at 300 in this model. This was discovered after analyzing the effect of population size on the optimized result (Behera et al 2018).

Mutation Operations

- Cluster Assignment Operator
  The Genetic K-Means tool was created by combining the K-means model with the genetic algorithm concept (Krishna and Murty 1999). This model gives the best outcome when compared to certain convergence criteria. In this work, a fusion of the K-mode clustering technique with an evolutionary algorithm is made. It determines each cluster's mode. It's the number of genes in that cluster with the highest MIR (at a given expression level). Other genes are allocated to the clusters with greater IR when using a specific cluster mode. The fitness of the individual and the new modes of the clusters are then determined.
- Probabilistic Mutation Operator
  The mutation rate is a parameter that can be chosen. To avoid losing individuals of higher fitness, only the top 5% of the population having high fitness is kept. A genetic algorithm's usefulness is improved in this procedure. The remaining 95% of people are randomly picked for mutation. A random number is picked to choose an individual for mutation. The random number lies between 0 and 1. If the random number is smaller than the mutation rate, an individual from the whole population is chosen randomly for the implementing mutation. A mutation rate is usually very low. An individual here is made up of a number of clusters, each of which contains a number of genes. A cluster with at least five genes is taken into account. The gene with the lowest MIR value is moved to another cluster of the same individual. A roulette wheel is built by combining the values of the IRs of the genes with the modes of the clusters in an individual. For more information, see the section below. The roulette wheel selection is used to choose the cluster to which a gene is transported. The fitness of the individual and the new modes of the clusters are next assessed. The mutation process affects all clusters. The result is a mutated individual. The

new population is made up of the top 5% of the better fit population as well as mutant and normal ones. This is done to eliminate bias caused by the top 5% of the population being chosen, to avoid early convergence and to increase variety. After examining the effects of the mutation rates, the probabilistic mutation rate is set to 0.1 to get the optimum results (Behera et al 2018). Figure 1b describes a flow chart of the operator steps.

Fitness Function

The best individual in the population is the one with the greatest fitness value. The sum of the fitness values of the total number of clusters determines an individual's fitness. It is defined as follows:

$$F = \sum_{i=1}^{k} R_i \qquad (7)$$

where F represents individual fitness, i signifies cluster number, and R_i is the cluster i's multiple interdependence redundancy measure.

Selection

To choose all of the individuals for the next generation, the roulette wheel technique is used. In a genetic algorithm, it is a common selection strategy. The relative fitness of each individual is used to create a roulette wheel. It's shown as a pie chart with the space filled by each player on the roulette wheel according to their relative fitness. The relative fitness sum (S) is computed. A random number is generated between 0 and S. The relative fitness values are added together. The corresponding individual is chosen when the total relative fitness value exceeds the random number created. An individual with a higher relative fitness value will take up more space in the pie chart. Therefore, the likelihood of selecting this individual for the next generation will be increased. Figure 1c illustrates a visual representation of the roulette wheel selection.

Termination

The change in average fitness for subsequent generations is found. The change in the mean fitness for two successive generations is calculated. This procedure is done for ten generations. When this change becomes less than 2% for ten continuous generations, the computer program is stopped. Otherwise the population moves on to the next generation.

Results

The program has run on a 3.0 GHz dual-core Intel Xeon CPU with 2 GB RAM. The research project is developed in Java 1.6.0. It is has the following sections: discretization, redundancy computation, and the ECA development. For a collection of 7129 genes and 30 samples, the discretization time was 0.6469 minutes. There were a total of 55337 intervals created. The computation of redundancy for the same dataset took 1363.3997 minutes. ECA took 48.6153

minutes to simulate. The equilibrium generation was between 12 and 14 kilowatts. One simulation of CMI took 0.7068 minutes and had 2 to 4 iterations for convergence for the same set of data. It took 70.56 minutes to run a k-means simulation with 10,000 iterations. For a larger number of iterations, k-means revealed that the answer recurred more than once. This indicated that the answer was likely to be statistically optimum.

Real data

Three gene expression data sets are utilised to assess the ECA's performance. They are of the Tsutsumi et al 2002 for gastric cancer dataset, Alon et al 1999 for colon cancer dataset, and MacDonald et al 2001 for brain cancer dataset. Table 1 lists all of the datasets and their descriptions. The CMI is simulated 50 times for each dataset. Each simulation has a different number of clusters, ranging from 2 to 20. For that simulation, the cluster with the highest individual fitness becomes the best cluster in the dataset. Because data for big values might be dispersed, the lowest value is used. The ideal cluster number for the dataset is determined by the minimum value of the optimal cluster number among the 50 simulations. The same cluster number is utilized in all simulations for all methods. The ECA is compared to the CMI and the k-means. For the sake of comparison, 10 simulations of each method are taken into account. A collection of clusters is generated for each simulation. For classification investigations, a selection of genes from each cluster is chosen. Candidate genes are what they're termed. They are made up of the most fit genes at the top of the list. The top-ranking genes in the ECA and CMI are defined as the genes with the highest multiple redundancy measure in clusters. They are the genes with the shortest Euclidian distance from the cluster's mean value for the k-means.

The leave-one-out cross-validation (LOOCV) method is used to calculate classification accuracy. The first sample is used as the test set, while the following samples are used as the training set for LOOCV. To predict whether the test sample is sick or normal, the top-ranking genes from the training set are employed. The test sample is then taken from the second sample. The training set is made up of the rest of the sets. Each individual sample is used as the test sample, and the process is repeated. The percentage of test samples accurately identified as sick or normal is used to assess classification accuracy. A greater classification accuracy number indicates that the algorithm is more effective at pinpointing the genes with the most diagnostic information.

Dataset of studies on stomach cancer (GDS1210)
The algorithms are compared based on sample classification accuracy

For each of the methods, the average classification accuracy of the top-ranking genes is computed. One example of a stomach cancer dataset is shown in Table 2. The ECA discovered the percentage disparities in classification accuracy. The CMI and the k-means are used to compare them. It clearly demonstrates the ECA's superiority to the other algorithms. Table 3 shows the results for the stomach cancer dataset. According to the research, the ECA surpasses the CMI and k-means. When the seed for producing the random number is changed for a fresh simulation in any stochastic computation, the initial gene distributions in the

clusters likewise change. For each simulation, this can result in distinct evolutionary dynamics, such as the equilibrium generation number and mean fitness. As a result, the average impacts of many simulations with various beginning gene distributions in the clusters are determined.

The following are the findings of a complete investigation of the algorithms on the stomach cancer dataset. In the ECA, there are more simulations that yield greater classification accuracy of test samples than in the CMI and k-means. This demonstrates that the ECA has a better likelihood of locating the proper candidate genes than the other methods. Table 4 shows the performance of classification accuracies for the 10 simulations of the ECA, CMI, and k-means. The table displays the number of simulations required for various numbers of top-ranking genes, as well as the classification accuracy obtained.

The investigation of classificatory genes

On the genes subset, the decision tree produced by C4.5 is examined. Only one gene, TGIF1 or, in some cases, D26129, had 96.67 percent classification accuracy in the cases. According to a review of the literature, reduced TGIF1 expression is associated with lymph node metastases. It may limit gastric cancer invasion and metastasis by down-regulating MMP9 and VEGF proteins (Liu et al 2006). D26129 is also a dominant gene that influences gastric cancer growth (Wang et al 2006). The two genes TGIF1 and D26129 are ranked extremely low by k-means and much lower by the CMI (data not shown). For three distinct simulations, the numbers beneath each algorithm's name reflect the locations of the relevant genes in the respective clusters (given in square brackets). The top three ranking genes were correctly classified in 96.67 percent of the ECA simulations. Because either TGIF1 or D26129 can only attain a 96.67 percent accuracy value, one of the genes must be among the top three. The ECA can be determined to be successful in identifying genes that carry important diagnostic information for a certain illness. Beyond a given threshold of classification accuracy, a comparison of the three methods based on the common genes detected throughout all simulations is provided. Each simulation yields a set of top-ranking genes (i.e. top 1, 2, and 3), as shown in Table 6. The common genes are the ones that appear in all of these simulations.

Algorithms based on representative genes are compared.

The degree of coherence determines the overall trend in the expression levels of a collection of co-expressed genes. Only the genetic distances from the cluster centre are taken into consideration when calculating gene similarity. The interdependence factor, on the other hand, takes into account genetic distance as well as negative and positive gene correlations. The degree of coherence of a pair of genes can be assessed in terms of similarity or positive/negative correlation. Three example genes are chosen from the top five genes to better comprehend the coherence pattern. For each method, their patterns are evaluated across distinct clusters. The most important genes in distinct clusters are chosen for the ECA and the CMI based on the size of their multiple interdependency metrics. The genes with the smallest distance from the cluster mean are chosen for the k-means. For a single simulation, the pattern of coherence of these genes for the

algorithms - ECA, CMI, and k-means - is illustrated in Figures 2, 3, and 4. The simulation is based on a completely random setting. This is true in a lot of simulations. The HMHA1 and TGIF1 genes are interdependent in the ECA because their total expression profiles have a negative connection. They are not, however, genetically related due to the huge genetic gap between them. The expression profiles of the HMHA1 and C21orf33 genes are comparable and exhibit negative relationships in some areas.

Positive and negative relationships may be seen across the expression profile shape for the TGIF1 and C21orf33 genes. The CMI also reveals interconnectedness even when genes are dissimilar owing to significant scale factors, as the HMHA1 and CUL5 genes demonstrate. The HMHA1 and COPS5 genes are comparable, with negatively and positively associated areas found across the profile. The k-means algorithm, on the other hand, only displays coherence in the sense of resemblance. It does not account for all of the genes' scaling factors or dependency. The finding is that, although the Euclidean distance clusters genes based just on similarity, the interdependent redundancy measure clusters genes based on both similarity and positive/negative correlations.

In contrast to tiny sections throughout the profile form, the ECA may group genes that demonstrate correlations in their overall expression profile shape (determined by short Euclidean distance). In addition, the ECA considers bigger scaling factors (a higher value in gene expression) than the CMI. As a consequence of these findings, it may be concluded that the ECA is more consistent in grouping genes than the CMI. This illustrates that, despite the fact that the ECA and the CMI employ the same distance metric for clustering, namely the interdependence redundancy measure, the ECA uses the interdependence measure better than the CMI. The ECA's algorithm is more powerful than the CMI's.

Study of gene distribution in clusters

The clusters formed by the ECA consist of nearly the same number of genes (Behera et al 2018). So the genes in different clusters are evenly distributed. The selected top-ranking genes provide significant information about a disease process. This is clear as the ECA classifies the microarray samples as normal or sick. Therefore, the cluster configuration obtained by the ECA is more reliable to select the candidate genes for a disease state. The k-means algorithm creates the most lopsided gene distribution in the clusters (Behera et al 2018). Compared to the k-means algorithm, the CMI shows better even-distribution in many cases. But its performance is poor in comparison to the ECA. The following trend of gene distributions in distinct clusters may be seen in a randomly selected single simulation for each of the ECA, CMI, and k-means algorithms. The ECA, for example, has divided 7129 genes into four clusters: 2025, 1695, 1720, and 1689 genes; and the CMI, on the other hand, has divided the same number of genes into 2582, 2614, 1592, and 341 genes. The k-means has clusters of 752, 89, 5148 and 1140 genes. The ECA outperforms the CMI and the k-means algorithms in terms of nearly equal gene distribution in the clusters. This holds good for a large number of simulation samples picked at random.

Colon cancer dataset
The algorithms are compared based on classification accuracy.

Mean classification accuracy for all ten simulations is determined for the total test samples using each tool. The top 4 genes per cluster are shown in Table 8 for one colon cancer dataset. The percentage difference in classification accuracy between the ECA and the CMI and the k-means algorithms is computed. Table 8 shows that the ECA outperforms the CMI and k-means algorithms when a few top-ranking genes per cluster is considered. Au et al. (2005) conducted considerable research on this colon cancer dataset. Here, analysis is done regarding the test samples' accuracy as normal or diseased.

Conclusion

The high-throughput microarray data gives the expression patterns of thousands of genes at the same time. Out of thousands of genes, the computational procedures try to select a few key genes responsible for diagnostic information of a disease. Therefore, microarray data clustering and selection for the candidate genes of cancer has become important in the recent years. A new algorithm for tackling this problem has been devised in this research. The clustering method and genetic algorithms are combined in this algorithm. The ECA developed here is a strong software tool for identifying the candidate genes for a cancer process. The ECA's Cluster Assignment Operator is powerful to speed up the algorithm's convergence without sacrificing its efficiency. The proper gene distribution in distinct clusters is found in a short amount of time. In the ECA, the Probabilistic Mutation Operator adds diversity to the population, preventing early convergence with bad clustering of genes. Other genetic algorithm operators are studied throughout the algorithm development but are abandoned due to their limitations. The typical crossover operator is tested but found to be disruptive. This ECA is a standard stochastic computation program that produces good result without the need of a crossover operator. Crossover operators, in general, help in the development better solutions in other evolutionary models. The genes with the least MIR from two random clusters are exchanged in a Swap Mutation Operator. The results suggest that it can only explore a restricted portion of the gene clustering space. For the time being, this isn't really important.

The ECA is more stable in terms of lopsided gene distribution across clusters unlike the CMA and the k-means tools. In addition, unlike the other two algorithms, the ECA clusters the same set of genes with a higher probability in multiple simulation samples. The research is carried out by obtaining the top-ranking genes from different clusters in various simulations. It is particularly effective in clustering genes based on expression profile similarity as well as positive and negative correlation between the genes. The interdependent redundancy measure is a considerably superior metric in this case. Although the CMI and the ECA both employ the interdependence measure, the ECA is superior at grouping the genes and locating the cancer causing genes due to its exposure to a much larger search space in the gene expression values. Hence the ECA is extremely good in grouping genes. Microarray datasets often contain a large number of genes. Due to the high probability of of noise in such systems, classification of test samples as normal versus diseased by machine learning

algorithms becomes problematic. It is necessary to reduce the dimensions of the data set to find the informative genes that contain the diagnostic information. For each method, a subset of genes is chosen.

The ECA surpasses the other two algorithms to classify the test samples as normal or sick by using a small number of candidate genes in most cases. The classification accuracy for gastric cancer, and colon cancer datasets is determined using the ECA's top genes. When just the top-ranking genes are utilized, some classification accuracies are higher than when the entire dataset is used. The classification accuracy for the colon cancer dataset, for example, is 93.33 percent when the entire dataset is included. But the classification accuracy is 96.67 percent when TGIF1/D26129 is used alone. This demonstrates that the ECA is capable of accurately and successfully selecting the candidate genes of a disease. In the context of drug development, their functional and diagnostic features can be investigated further. Au et al. (2005) have created the Attribute Clustering Algorithm (ACA), which primarily employs interdependent redundancy measure and the OCDD discretization procedure. They have shown that it outperforms biclustering, K-means, Self Organization Map, and other traditional tools. Their ACA is not publicly available. They don't provide the details of their discretization method, such as the smoothing procedure or the chi-square test. The CMI algorithm is identical to the ACA except that it uses a somewhat different smoothing method. The ACA and the CMI have nearly similar performances.

The ECA has been shown to be superior to the CMI and K-means algorithms. Hence it should outperform biclustering, SOM, and a few other techniques (Au et al, 2005). A fundamental problem is that most traditional algorithms only employ Euclidean genetic distance metrics in different forms. They don't employ an interdependent redundancy metric that explicitly considers positive and negative gene correlations. Therefore, the ECA is now the best algorithm for analyzing the microarray data and discover the correct cancer genes for a disease process. A three-class microarray gene expression data may be created using the current approach. Samples of normal, moderately developed, and totally sick tissues, for example, can be found in some cancer microarray data. Modifications to the discretization algorithm can be done to improve it further. To make the machine learning method more successful, the data smoothing process can be optimized. Also, other mutation operators may be investigated to improve the ECA's effectiveness.

Authors' contribution

N Behera gave the basic concept and supervised the entire research project. S Sinha developed the computer program and created the results. S Hairat and Neha helped in statistical analysis and wrote a part of the paper.

Funding

A partial funding from Philips Research Asia, Bengaluru, India is acknowledged.

Competing interest

There is not any competing interest among authors.

Ethics approval

Not applicable

References

Alanni R et al (2019). Deep gene selection method to select genes from microarray datasets for cancer classification, BMC Bioinformatics, 20, Article number 608.

Alon, U., et al. (1999) Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. *Proc. Nat'l Academy of Sciences of the United States of America* 96(12): 6745-6750

Amaral M L et al (2018) BART: bioinformatics array research tool. BMC Bioinformatics, 19, article no 296.

Au, W. et al. (2005) Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2(2): 83 – 101.

Aydadenta H and Adiwijaya (2017) On the classification techniques in data mining for microarray data classification. Journal of physics: Conference series 971, 012004.

Behera N & Nanjundiah V (1995) An Investigation into the Role of Phenotypic Plasticity in Evolution **/** *Journal of Theoretical Biology Vol.172, No. 3, 225-234.*

Behera N & Nanjundiah V (1996) The Consequence of phenotypic plasticity in cyclically varying environments: a genetic algorithm study / *Journal of Theoretical biology, Vol.178, No.2, 135-144*

Behera N & Nanjundiah V (1997) *trans*-Gene Regulation in Adaptive Evolution: a Genetic Algorithm Model / *Journal of Theoretical Biology Vol. 188, 153-162.*

Behera N & Nanjundiah V (2004) Phenotypic plasticity can potentiate rapid evolutionary change / *Journal of Theoretical Biology***,** *226, 177-184.*

Behera, N., Jeevitesh, M., Jose, J., Kant, K., Dey, A. & M Mazher (2017) / Higher accuracy

Behera, N., Sinha, S, Gupta, R, Geoncy, A., Dimitrova, N & Mazher M (2018) Analysis of gene expression data by evolutionary clustering algorithm *IEEE Explore (*DOI 10.1109/ICIT.2017.41 *in 2018)*

Behera, Narayan. 1997. "Effect of phenotypic plasticity on adaptation and evolution: a genetic algorithm analysis." *Current Science* 73:968-976

Berrar, D. P., Dubitzky, W & Granzow, M (2003) A Practical approach to Micro-array data analysis,*Kluwer Academic Publishers, London.*

Cheng, Y. and Church, G.M. (2000) Biclustering of expression data. Proc. Int. Conf. Intell. Syst. Mol.Biol,. 8, 93–103.

Cho, S.W., Kim, D H, Uhmn, S, Ko Y W, Cheong J Y and Kim, J (2007) Chronic Hepatitis and Cirrhosis Classification Using SNP Data, Decision Tree and Decision Rule. ICCSA (3): 585-596.

Cios, K.J., and Kurgan, L., (2004) Discretization Algorithm that Uses Class-Attribute Interdependence Maximization. *IEEE/ACM Transactions on Knowledge and Data Engineering* 16(2): 145 -153(2004).

discretization of continuous data. *Intelligent Data Analysis* 8(2): 151-170.
Eisen, M.B., Spellman P T, Brown P O, and Botstein D (1998) Cluster analysis and display of genome- wide expression patterns. Proc Natl Acad Sci U S A. 1998 Dec 8;95(25):14863-8
Elomaa, T.(1994) In defense of C4.5: Notes on learning one-level decision trees, Proc. of the 11th Int. Conf.on Machine Learning, Morgan Kaufmann, 62- 69.
Gandamayu, I. B. M., Antari, N. W. S., & Strisanti, I. A. S. (2022). The level of community compliance in implementing health protocols to prevent the spread of COVID-19. International Journal of Health & Medical Sciences, 5(2), 177-182. https://doi.org/10.21744/ijhms.v5n2.1897
Goldberg, D E (2008) Genetic algorithms in Search, Optimization and Machine learning, *Pearson Education, India*
Golub, T.R., et al. (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. PNAS 96, 2907-2912.
Hambali M A et al (2020) Microarray cancer feature selection: Review, challenges and research directions. International journal of cognitive computing in Engineering, Vol 1, pages 78-97.
Herbola A et al (2022) Chapter 27- Bioinformatics and biological data mining. Bioinformatics: methods and applications, Academic press, 457-471.
Heyer, L.J., Kruglyak S and Yooseph S (1999) Exploring expression data: identification and analysis of coexpressed genes. Genome Research 9, 1106-1115.
Holland, J H (1975) Adaptation in natural and artificial systems, *University of Michigan Press, Ann Arbor, MI, USA.*
Johnson S. C., (1967) Hierarchical Clustering Schemes Psychometrika, 2:241-254.
Kohonen, T., (1990) The self-organizing map Proc. IEEE 78,1464-1479.
Koschmieder A et al (2012) Tools for managing and analyzing microarray data. Briefings in Bioinformatics, 13, 46-60.
Krishna K, Murty M (1999) Genetic K-means algorithm. *IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics* 29**:**433-439.
Liu et al (2009) Computational data mining in cancer bioinformatics and cancer Epidemiology. BioMed Research international. DOI.org//1155//2009//582697.
Liu, Y., Shen M, Wen J F, and Hu Z L (2006) Expressions of TGIF, MMP9 and VEGF proteins and their clinicopathological relationship in gastric cancer. PUBMED Feb;31(1):70-4.
Lloyd, S.P., (1982) *Least Squares Quantization in PCM.* IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129-137
protein multiple sequence alignment by genetic algorithm, *Procedia Computer Science*, Vol
Quinlan, J.R. (1993) *C4.5: Programs for machine learning.* Morgan Kaufman, San Francisco.
Razak K (2013) Application in the domain of data mining. Indian Journal of Computer Science and Engineering, 1, 114-118.
Selvaraj S and Natarajan J (2011) Microarray data analysis and mining tools. Bioinformation, 6, 95-99.
Solmaz M et al (2019) Graphical data mining of cancer mechanisms with SEMA. Bioinformatics, 35, 4413-4418.

Suryasa, I. W., Rodríguez-Gámez, M., & Koldoris, T. (2021). The COVID-19 pandemic. International Journal of Health Sciences, 5(2), vi-ix. https://doi.org/10.53730/ijhs.v5n2.2937

Tsutsumi, S., Hippo Y, Taniguchi H, and Machida N (2002) Global gene expression analysis of gastric cancer by oligonucleotide microarrays. *Cancer Res* Jan 1;62(1):233-40.

Wang, L., Zhu J S, Song M Q, Chen G Q, and Chen J L(2006) Comparison of gene expression profiles between primary tumor and metastatic lesions in gastric cancer patients using laser microdissection and cDNA microarray. World J Gastroenterol November;12(43):6949-6954.

Wong, A.K.C., Liu, L L and Yang W (2004) A global optimal algorithm for class-dependent

Zhang Y et al (2009) Bioinformatics analysis of microarray data. Methods Mol. Biol. 573, 259-284

Appendix

Smoothing and Chi-square test

A segment s_i is for each gene value x_i such that it contains values ranging from x_i-w to x_i+w, where w is the segment width. The frequency of the often occurring class label and the frequency of the class label for x_i are used to determine a ratio (r) for this segment. The class label is altered to the often occurring class label if r is larger than some threshold value (t). For discretization, the default values for w and t as 5.0 and 1.3 respectively are used. The smoothing and parameter settings incorporated are almost identical to those in the OCDD method (Wong et al., 2004).The Chi-square test is used to find the statistical significance of a gene's interaction with its classification labels. It calculates if the frequency distributions of two adjacent intervals and the class label are significantly correlated. So the intervals are decided to merge. The IR between each gene and class label is calculated for this purpose. Let E be the combined entropy of a gene and a class label. A product p is determined by multiplying E by the total number of data points. The ratio of the chi value to twice the value of p is calculated. It is known as rt. The intervals are merged if the IR is larger than the rt. The conventional chi-square distribution table is used to calculate the chi values. The degree of freedom for the chi value is the product of the total number of classes minus one and the total number of intervals minus one The technique is the same as the OCDD algorithm (Wong et al., 2004).

Figure Legend

- Figure 1(a): The flow chart of the ECA.
- Figure 1(b:)The chart of the probabilistic mutation.
- Figure 2: A normal line graph of the expression pattern of three representative genes for cluster 1 of a single simulation in the ECA algorithm for the gastric cancer dataset.
- Figure 3: A normal line graph of the expression pattern of three sample genes in cluster 2 of a single simulation CMI algorithm for the gastric cancer dataset

- Figure 4: A normal line graph of the expression pattern of three example genes in cluster 3 of a single k-means simulation for the gastric cancer dataset

Figures

Figure 1 (a)

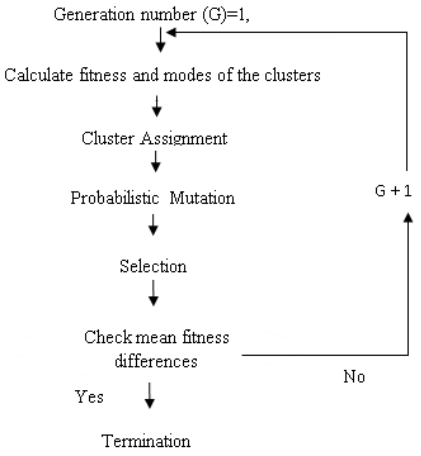


Figure 1 (b)

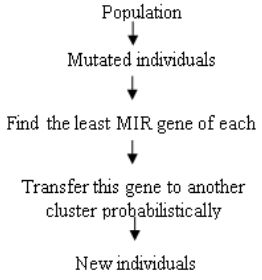
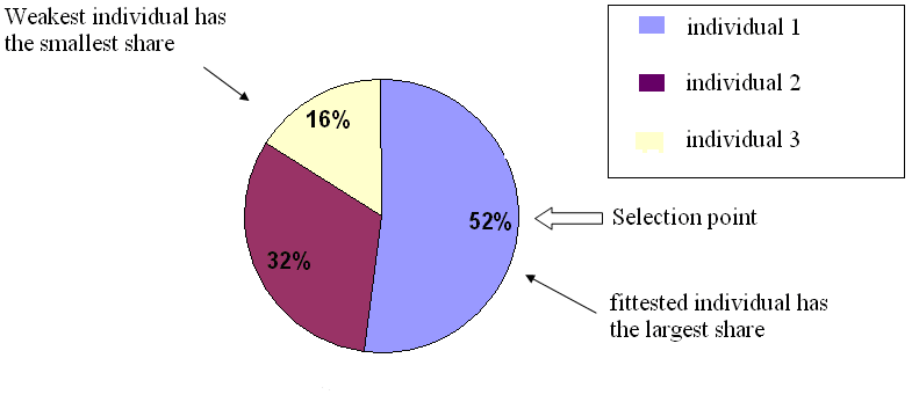


Fig 1(c)



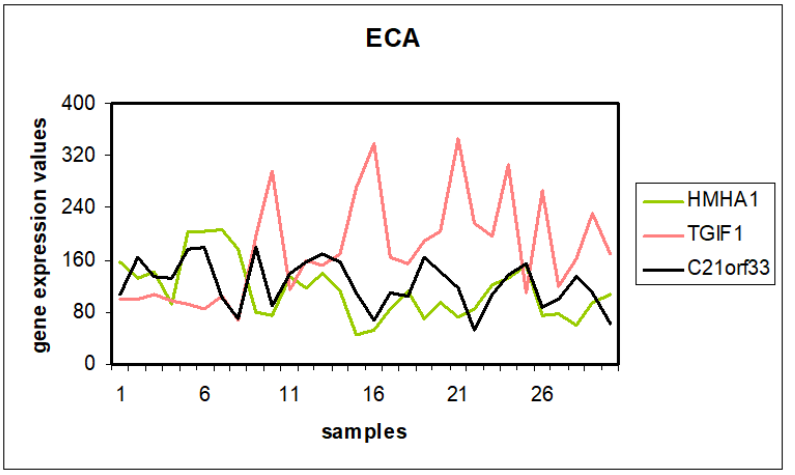


Figure 2

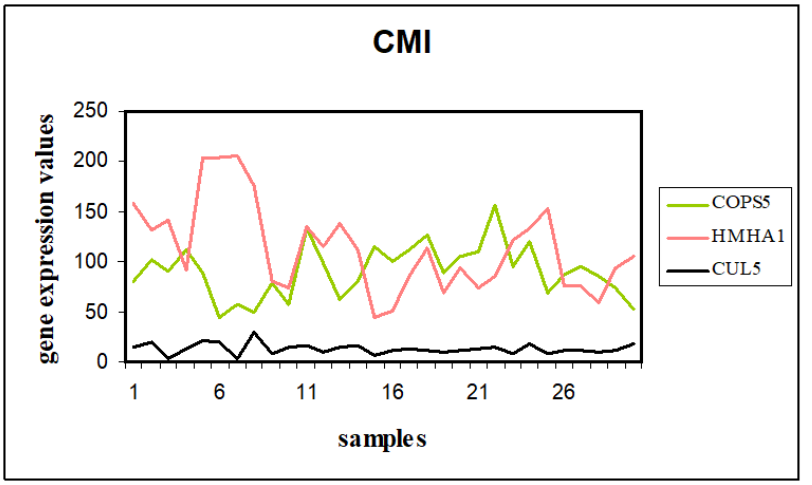


Figure 3

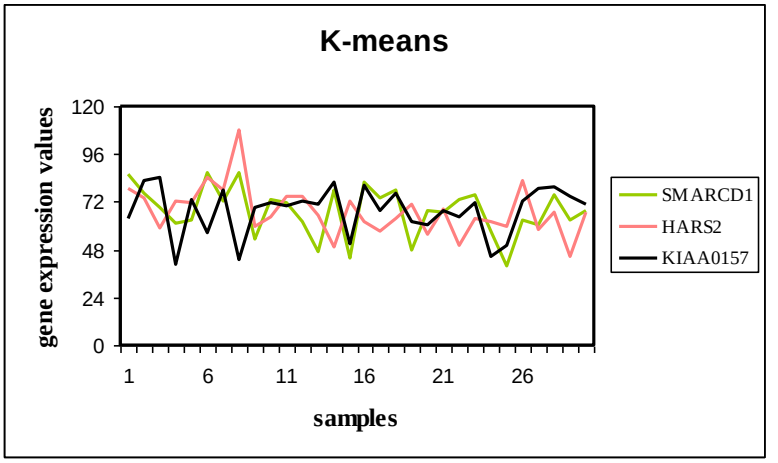


Figure 4

Table Legend

Table 1: The number of genes in each dataset, the total number of samples, the number of sick samples, the number of healthy samples, the number of clusters, and the lowest and maximum gene expression levels.

Gastric cancer dataset (GDS1210)

- Table 2: The gastric cancer dataset. It illustrates the classification accuracy of the two top-ranking genes for the ECA, the CMI, and the k-means. Individual classification accuracy for each of the ten simulations as well as the average classification accuracy for each method, are shown.
- Table 3: The ECA outperforms the CMI and k-means in the gastric cancer dataset. The difference in percentages with regard to the ECA is computed using the average classification accuracies.
- Table 4: The ECA, CMI, and k-means performances in terms of the classification accuracy for the gastric cancer dataset. In the appropriate columns, the number of simulations with the associated classification accuracy for one to four top-ranking genes is presented.
- Table 5: The ECA, CMI, and k-means have the classification accuracy of common genes for 1,2, and 3 top-ranking genes, respectively.
- Table 6: The list of genes chosen by the ECA (a) based on their presence as common genes across simulations and (b) based on the C4.5 classification tree (for the gastric cancer dataset)

Dataset on colon cancer

- Table 7: The ECA, CMI, and k- means classification accuracy for the top four genes in the colon cancer dataset Individual classification accuracy for each of the ten simulations as well as the average classification accuracy for each method is found.

- Table 8: The ECA outperforms the CMI and k-means in the colon cancer dataset. The % difference with regard to the ECA is found using the average classification accuracies of the related algorithms.

Tables

Table 1

	gastric cancer	colon cancer
Genes	7129	2000
Samples	30	62
Diseased samples	22	40
Healthy samples	8	22
Clusters	4	10
Gene expression value (min)	0.1	5.82
Gene expression value (max)	14,237	20,903

Table 2

Simulation no	Top 4 genes per cluster		
	ECA	CMI	K-means
1	75.35	81.33	61.83
2	93.88	72.56	71.08
3	98.65	85.34	72.78
4	97.27	78.15	63.88
5	94.69	82.76	72.96
6	99.02	74.82	74.38
7	85.08	91.91	68.05
8	79.22	75.23	71.83
9	75.39	71.59	74.38
10	95.11	61.88	60.77
Average	89.36	77.55	69.19

Table 3

Top genes per cluster	ECA percentage better than	
	CMI	k-means
1	8.7	3.13
2	7.37	15.15
3	15.7	32.66
4	11.80	26.25
5	8.48	28.58

Table 4

Genes per cluster	classification accuracy in percentage								
	>60 & <=80			>80 & <=90			>90 & <100		
	ECA	CMI	k-means	ECA	CMI	k-means	ECA	CMI	k-means
Top 1	6	8	3	3	3	6	3	-	1
Top 2	4	5	9	-	5	-	7	-	-
Top 3	3	2	10	1	1	-	10	7	2
Top 4	3	4	8	-	2	5	10	8	1

Table 5

genes per cluster	common genes			classification accuracy in percentage		
	ECA	CMI	k-means	ECA	CMI	k-means
1	1	2	2	79.61	75.67	70.38
2	5	5	4	85.44	72.39	73.77
3	7	6	3	99.19	93.37	71.17

Table 6

Simulation no	TOP 4 genes per cluster		
	ECA	CMI	k-means
1	88.10	59.06	60.22
2	74.97	69.76	65.52
3	74.81	62.52	54.87
4	68.43	70.54	73.52
5	87.48	62.29	69.74
6	89.74	65.9	60.52
7	76.59	72.85	66.85
8	86.86	81.68	65.74
9	86.79	82.64	66.78
10	83.43	66.75	53.67
Average	81.72	69.39	63.74

Table 7

Top genes per cluster	ECA better percentage than	
	CMI	k-means
1	17.88	11.94
2	11.80	8.36
3	15.13	10.59

4	18.78	14.78

Table 8

Simulation no	Top 2 gene per cluster		
	ECA	CMI	k-means
1	66.22	77.26	70.91
2	67.22	68.57	49.13
3	78.57	46.13	38.78
4	73.57	43.13	39.13
5	73.57	39.48	52.17
6	62.17	27.39	34.78
7	72.57	63.87	43.48
8	69.57	44.78	39.13
9	69.57	73.57	60.91
10	72.57	34.78	34.43
Average	70.56	51.89	46.28

Table 9

Top genes per cluster	ECA better percentage than	
	CMI	k-means
1	29.28	33.11
2	33.78	17.53
3	26.55	14.35
4	8.15	5.00
5	18.69	3.59