Extracting the candidate genes for cancer from the microarray gene expression data by stochastic computation

Narayan Behera (✉ narayan.behera@astu.edu.et)

Adama Science and Technology University

Shruti Sinha

Institute of Bioinformatics and Applied Biotechnology

Research Article

Keywords: microarray gene expression data, mutual information, genetic algorithm, clustering, classification accuracy, cancer-causing genes

Posted Date: January 31st, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1229120/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Extracting the cancer genes from the microarray gene expression data by stochastic computation

***Narayan Behera^{1,2} and Shruti Sinha²**

¹Department of Applied Physics, School of Natural Sciences,

Adama Science and Technology University, Adama, P O Box 1888, Ethiopia

and

²Institute of Bioinformatics and Applied Biotechnology, Electronics City, Phase I,

Bengaluru – 560100, India

ABSTRACT

Microarray gene expression data provides a simultaneous gene expression profile of thousands of genes. A handful of genes among the thousands of genes play dominant roles in a disease process. It is impractical to find these faulty genes from experimental data. A computational approach is necessary to find the key genes and hence it is an important area of research in bioinformatics. A new algorithm is developed in this work to identify the candidate genes of a cancer from microarray gene expression data. Gene clustering enables identification of co-expressed genes in specified biological conditions. Many algorithms have inherent limitations to understand the accurate information. This model is a hybrid of clustering algorithm and genetic. A genetic algorithm utilizes the three biological principles of evolution, such as selection, recombination, and mutation to solve an optimization problem. An interdependence measure between the genes is based on a metric which can be mutual information. In many conventional algorithms, a Euclidean genetic distance measure

is used. This is basically the difference between gene expression values. Genes are similar if their expressions levels are very close. Similarity of the gene expression levels, and also positive and negative correlations between the genes are taken into account while clustering them. This property is used in mutual information which introduces a parameter called interdependence measure. This says roughly how the two genes are correlated. The genes having higher interdependence measures are the top candidate genes responsible for a diseased state. These top genes are believed to be faulty genes that contain the most diagnostic information for cancer. An analysis is done on gastric cancer, colon cancer, and brain cancer microarray gene expression datasets to pinpoint the faulty genes. In comparison with many existing computational tools, the top candidate genes found by this evolutionary computational model, are able to classify the known samples into cancerous and normal ones with high accuracy. The new model called as evolutionary clustering algorithm also creates more even-like distribution of genes in the clusters. Furthermore, the present computational tool is more coherent in clustering the genes having large differences in their expression values. This information-theoretic computational method can be potentially applied to the big data analysis from other sources.

Keywords: microarray gene expression data; mutual information; genetic algorithm; clustering; classification accuracy, cancer-causing genes

***Corresponding author email: narayan.behera@astu.edu.et**

INTRODUCTION

Microarray technology monitor genome-wide expression levels of genes in an organism. Genes are made of deoxyribonucleic acid (DNA). DNA molecules are constructed by sequentially mapping nucleotides into a linear sequence. There are four possible nucleotides. This sequence allows DNA to encode information. DNA has wonderful storage capacity much like an encoding device. Within the DNA, genes are sequences of different lengths. The genes within a cell contain the information for synthesizing the proteins, which make cells function. Proteins are the ultimate product of the gene expression process. Proteins are formed from 20 different subunits called amino acids. A microarray is a glass slide onto which DNA molecules are fixed systematically at specific spots. A microarray may contain thousands of spots. Each spot may uniquely correspond to a gene. The DNA in a spot may either be genomic DNA or short stretch of oligo-nucleotide strands for a gene. The spots are printed onto the glass slide by a machine. The DNA microarray technology measures simultaneously the interactions of thousands of genes as evident from their expression levels.

Differential gene expression analysis involves genes showing different expression levels under different experimental conditions. The experimental levels can be various tissue types, or can be different developmental stages of the organism. Typical analysis is normal-versus-diseased state investigations. Gene co-regulation is somewhat similar to differential gene expression. Generally, one looks at a single gene profile against the experimental conditions mentioned above. Gene co-regulation study compares the gene profiles with each other. One needs to identify the genes whose expression levels vary in a correlated fashion across the studied experimental conditions or different samples. There is a positive co-

regulation between two genes if the expression level of one gene increases when the other increases. A negative co-regulation exists if the expression level of one gene increases while the other decreases. Gene co-regulation analysis compares gene profiles of several genes.

Microarray experiments reveal the function of genes. The process basically involves the comparison of the desired genes' expression profiles. If the expression profile of a known gene matches with the expression profile of an unknown gene, the function of the unknown gene can be successfully inferred. The functions of the genes with highly similar expression profiles serve as tools for inferring the function of the new genes. Gene expression data are also an important tool for clinical diagnostics. They can discover expression patterns that are characteristics of a particular disease.

Microarray data analysis involves techniques from computer science, and statistics. A gene profile is a gene expression values for a single gene across many samples or experimental conditions. An array profile is a gene expression profile that describes the expression values for many genes under a single experimental condition or for a sample. There are several steps in microarray data analysis. The raw data is first processed to reduce the data volume. A gene data matrix contains numbers representing a particular gene's expression levels in different samples. Due to the imperfect instrument, the measured value deviates from the true expression level due to the measurement error. A measurement error has two components. A bias describes the tendency of the instrument to detect either too low or too high values. The measurement error is generally normally distributed. Small deviations are more frequent than the large ones. Normalization involves numerical methods to deal with measurement errors.

After the successful completion of the human genome project, knowing the functions of genes in various biological systems are a big challenge. The advent of microarray technology has revolutionized functional genomics. Now scientists can study thousands of

genes together and their expression profiles under different experimental conditions. This makes us understand how the gene products are regulated in normal and diseased conditions on a genome-scale. This global analysis allows to determine the cellular functions of genes, the nature of biochemical pathways, and the regulatory mechanisms in certain disease processes. A cell's genome has an implicit mechanism for transformation of gene information into proteins. The gene to protein transformation constitutes the “central dogma of molecular biology”. It is a two-step process. First, a gene makes a ribonucleic acid (RNA). This process is known as transcription. Second, in the process called translation, the RNA makes a protein. The information present in the DNA sequence of genes is transferred into an intermediate RNA sequence. The information in the RNA is then used for constructing proteins. The RNA occurring as an intermediate structure is referred to as messenger RNA (mRNA). The overall process consisting of transcription and translation is known as gene expression. Due to the RNA's inherent chemical instability, one prefers more stable complementary DNA (cDNA) made by reverse transcription. However, before the microarray is made, the cDNA is broken up into its individual strands. The RNA samples are hybridized to known cDNA probes on the arrays.

DNA microarray experiments involve the comparison of gene expressions in several different cells and in a cell exposed to different biological conditions (normal versus diseased). The genetic diseases like cancer are characterized by certain genes being inappropriately transcribed. A cDNA microarray study can show the transcription differences between a normal and a diseased phase. It can also reveal different patterns of abnormal transcription corresponding to different disease stages. Microarray experiments give an estimation of the absolute values of gene expression. The gene expression analysis for gastric cancer, colon cancer, and brain cancer datasets in this study has been done using microarray experiments. The literature on the experimental and computational analysis of microarray

data can be found (Berrar et al 2003).

Data mining procedure is explained in detail by Razak (2013). Liu (2009) has discussed computational data mining in cancer bioinformatics. A web platform SEMA for building models of cancer mechanisms from large cancer genomic data set is developed by Solmaz et al (2019). Herbola et al (2022) have discussed the basics of data mining techniques and their applications in genomics, proteomics, and medical analysis. Microarray data analysis have grown tremendously in recent years (Zhang et al 2009; Selvaraj and Natarajan 2011; Koschmieder et al 2012; Amaral et al 2018).

Gene clustering algorithms:

It is essential to cluster genes with similar functions by developing suitable algorithms. Clustering of microarray gene expression data to identify the genes with similar expression profiles is an important area of research. They can reveal the co-regulated and co-expressed genes under certain specified conditions. Genes need to be selected that can accurately classify the test samples into diseased and normal classes. In the microarray data, generally, the number of genes far exceeds the number of samples. This makes gene selection and sample classification difficult. A large number of genes can potentially create noise in the data. A small subset of genes containing good diagnostic information is necessary. This reduces the dimensionality of the data. These genes can then be used for diagnostic study.

Microarray technology puts the gene expression values of an entire genome onto a chip. Each data point provided by an experimenter lies in the high-dimensional space defined by the size of the genome under investigation. However, the sample size is severely limited in these experiments. So feature selection methods are essential. The goal of feature selection is to select relevant features and eliminate irrelevant ones. An important problem in the

microarray experiments is the classification of biological samples using the gene expression data. This problem is important in cancer research. Reliable classification of tumors is essential for the successful diagnosis and treatment.

Many clustering algorithms have been developed in the recent past. The K-means is a simple algorithm for clustering objects (Lloyd 1982). In this algorithm, the number of clusters, k , has to be specified a priori. It is unsupervised learning. The purpose is simply to cluster the data points and not compare with any benchmark performance. It classifies the genes based on certain attributes/features (i.e. gene expression values) into k distinct partitions. Each gene belongs to one cluster only. The range of gene expression values is extracted from the microarray data. The steps of the algorithm go as follows. Initially, the k centers are chosen randomly. These centers are randomly chosen different gene expression values. The squared distance between a data point and the corresponding cluster centroid is calculated. Then the sum for all data points is minimized. In this way unique clusters are generated. Each data point is put to the nearest cluster center. Then the cluster center is found by computing the average of the data point values from the cluster centre for each cluster. Iteration is continued until there is no change in the cluster centers. Hence k clusters are generated for grouping the similar genes. The genes within each cluster are similar in their expression values as found from the microarray data. The genes between the clusters are more dissimilar. Generally, the similar genes have the same functions in an organism.

Some other clustering algorithms are given without going into details. They do not enter directly into the present subject but enhances our understanding of clustering. Kohonen's Self-Organizing Map (SOM) is an unsupervised learning technique used for clustering (Kohonen 1990). The SOM can be visualized as a sheet-like neural-network array. The cells are specifically tuned to various input signal patterns in an orderly fashion. The learning process is unsupervised. This makes the SOM interesting for low-dimensional views of high-

dimensional data. The hierarchical clustering algorithm has simple visualization tool (Johnson 1967). The traditional representation of this hierarchy is a tree with individual elements at one end. The hierarchical clustering is subdivided into agglomerative methods. Fusions of the genes into groups are made in a series. Then the genes are successively placed into finer groupings. Biclustering clustering is a data mining technique that allows simultaneous clustering of the rows and columns of a matrix of data (Cheng and Church 2000). Given a set of m rows and n columns, the algorithm generates biclusters - a subset of rows that exhibit similar behavior across a subset of columns, or vice versa. All these clustering techniques are commonly used for clustering gene data from microarray samples (Eisen et al 1998; Heyer et al 1999; Golub et al 1999).

The clustering algorithms use the gene expression values. The attribute of a gene is defined as its expression level under a particular experimental condition. It can be a value in the time series data of its expression levels. The widely used distance measures are Euclidian distance and Pearson's correlation coefficient. The Euclidian distance(e) between two genes, x , and y , having n attributes is given by :

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

Here x_i is the value of i^{th} attribute of gene x and y_i is the value of the i^{th} attribute of gene y . Pearson's correlation coefficient (r) between two genes, x , and y , having n attributes is given by:

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

where $\bar{x}, \bar{y}$ are the means of expression values of genes x and y respectively, x_i is the value of i^{th} attribute of gene x and y_i is the value of the i^{th} attribute of gene y . A negative value of r

denotes a negative linear relationship between x and y . A positive value denotes a positive linear relationship. If the value of r is equal to zero, the genes are independent.

The conventional clustering algorithms use an Euclidian or a Pearson's correlation as distance measures to calculate the similarity between two genes. The distance measures have some limitations. Normally genes having similar profile shapes are correlated in their functions. A Euclidian distance measures the similarity based on the expression values but not on its profile shape. Hence the genes having the same expression profile shapes but differing by large values of gene expression have the chances of not getting clustered together. Genes differing by a small magnitude and having different expression profile shape may get clustered together. The Pearson's correlation is better than the Euclidian distance for its ability to define the relationship. But it is sensitive to outliers. The clustering algorithms are also dependent on the initial distribution of the genes in the clusters. They tend to give different results for a different initial distribution of genes chosen by an algorithm. Furthermore, a drawback of these algorithms is that the number of clusters has to be specified in the beginning.

Recently, a new distance measure based on information-theoretic approach has been proposed. This is Attribute Clustering Algorithm (ACA) (Au et al 2005). It eliminates some limitations in the conventional distance measuring methods. The ACA essentially uses the k-means algorithm idea. However, the genetic distance is an information-theoretic measure called interdependence redundancy measure. It takes into account the interdependence between any two genes. The interdependence redundancy measure between two genes, x , and y , having n attributes is given by:

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

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

and

$$E(x:y) = - \sum_{k=1}^g \sum_{l=1}^h P(v_k \wedge v_l) \log P(v_k \wedge v_l), \quad (5)$$

where $M(x:y)$ is the mutual information, $E(x:y)$ is the joint entropy, g is the number of intervals of x , h is the number of intervals of y , v_k is the k^{th} interval of x , v_l is the l^{th} interval of y , and $P(v_k \wedge v_l)$ is the probability of a genic value occurring in the interval v_k and v_l , and $P(v_k)$ is the probability of a value occurring within the interval v_k .

Although mutual information provides knowledge about the dependence of two genes, its value increases with the number of possible gene expression values. Hence to incorporate the correct distance measure between genes, mutual information is normalized by dividing it by the entropy measure. This normalized information measure is the Interdependence Redundancy (IR(x:y)) measure. It is calculated as shown in equation 3. It reflects the degree of deviation from independence for two genes.

A model called Clustering by Mutual Information (CMI) is studied by us. It is fundamentally similar to the one developed by Au et al (2005) except the fact that they use a different smoothing process (Behera et al 2018). But the initial search space of the gene distributions in appropriate clusters is limited. This means the possible gene distributions have inherent limitations. This can put a constraint on finding the best the best configuration

of gene clusters. There is a high possibility that the best solution in this way may not be an optimized one. In order to overcome this problem, a hybrid of evolutionary computation and clustering algorithm model is implemented (Behera et al 2018). Evolutionary computations are efficient search algorithms. They can be used for studying a large number of initial distribution of genes in the clusters. Keeping these factors, the Evolutionary Clustering Algorithm (ECA) is proposed (Behera et al 2018). The ECA is an extension of the CMI but it uses Darwinian evolutionary concepts. It generates better gene classification and selection by minimizing the limitations of the biased initial distribution of genes in the clusters. This model has shown promising results with a synthetic dataset as discussed below. The ECA correctly picks the gene distributions in the clusters.

This algorithm is applied to three sets of real data – gastric cancer (Tsutsumi et al 2002), colon cancer (Alon et al 1999) and brain cancer (MacDonald et al 2001) data derived from the public domain available in the net. They provide insight in understanding a disease process. The results are compared with the outcomes from some other popular clustering algorithms. The C4.5 is a decision tree algorithm. It is used for creating classifiers on the selected subset of candidate genes (Quinlan 1993). The C4.5 software is known for its learning efficiency and robustness (Elomaa 1994). This builds a decision tree from a set of data. It makes use of information theory. It is a decision tree classifier. It has been in use in data mining. Here it finds the top-ranking genes from a microarray data. They are used to classify the test samples as normal or diseased (Cho et al 2007). The higher the classification accuracy, the better is the algorithm.

METHODOLOGY

Discretization

In the real world generally most data are continuous. They need to be discretized for

any quantitative analysis. They usually have noise arising from the measurement errors or incorrect entry of the values. This noise can create a large number of intervals in the discretization process. But more intervals lead to a greater loss in information. These problems need to be taken care of. The gene expression values in a microarray data are typically continuous. They require to be discretized into proper intervals to compute an information measure.

The Optimal Class Dependent Discretization (OCDD) algorithm is used to discretize a continuous data. It produces nearly a global optimal solution (Wong et al 2004). The OCDD takes into account the interdependence between the classes and the gene expression values. Then it minimizes the information loss (Appendix 1). Each sample belongs to a particular class. In this context, we have two classes – normal and diseased samples. To overcome information loss, smoothing of the data and a statistical test (chi-square test) as described in Appendix 2 are performed. Smoothing is done to remove noise before performing discretization. Chi-square test reduces the number of intervals. Smoothing, chi-square test and the parameter values used in ECA are essentially the same as described in the OCDD algorithm.

Genetic Algorithm

A genetic algorithm is an evolutionary search technique to find true or approximate solutions to optimization problems. This algorithm has power to solve important optimization problems in science and engineering (Holland 1975; Goldberg 2008). The algorithms use techniques inspired by evolutionary biology such as mutation, selection, and crossover (also called recombination). Hence it is a class of evolutionary algorithms. It imitates the Darwinian idea that Nature is the best optimizer. A population of solutions evolves in the

process of Darwinian selection over many generations. The raw material for Darwinian evolution is mutation where new genes appear spontaneously as a result of statistical fluctuations. This happens due to the change of some nucleotides in the gene sequence. In the process of reproduction, recombination of the genes in the chromosomes takes place. So the new chromosomes are created. For the purpose of illustration, a haploid organism is considered here. A crossover point is chosen at random for a chromosome. The genes up to the crossover point in parent 1 and the genes from the crossover point to the end of the chromosome of parent 2 are combined together. It creates a new chromosome as offspring. Then when parent 1 and parent 2 are interchanged in the above process, one gets another new offspring. They create new individuals in the population that were not present in the previous generation. As the population grows in an environment, some populations which adapt poorly to the environment do not survive to reproduce. This is due to the inherent carrying capacity of an environment. Their genes are eliminated from the population. The individuals that adapt better reproduce more individuals. Their genes go to the next generation. So the genes that give higher fitness to the individuals increase in frequency in subsequent generations. The fitness of an individual is defined as the net number of genes it leaves behind to the next generation. The central idea is to define the fitness function appropriately corresponding to the research problem that needs optimization. This genetic algorithm has been successfully applied to diverse problems in evolutionary biology (Behera and Nanjindiah 1995, 1996, 1997, 2004), optimization of multiple protein sequence alignment (Behera et al 2017) and finding cancer genes from microarray gene expression data (Behera et al 2018). Here genetic algorithm will be used to find key genes responsible for cancer from microarray gene expression data by analyzing thousands of genes.

An individual is coded as a string of 1s and 0s. It represents a possible approximate solution to the optimization problem one is interested in. The genetic algorithm starts from a

population of randomly generated individuals (possible solutions). In successive generations it finds better solutions as the solutions evolve. In each generation, every individual in the population is modified (recombined and possibly mutated) with some probability to form a new individual. The target solution becomes better in each successive generation. Each individual is evaluated for its fitness. Multiple individuals are stochastically selected from the current population by fitness criteria to form a new population. The new population is then used in the next iteration of the algorithm. The algorithm terminates after maximum number of generations, or after a satisfactory fitness level of the population. Finally, the individual with the highest fitness provides the best solution to the optimization problem.

The Evolutionary Clustering Algorithm

To cluster the genes an information-theoretic measure is introduced here. It is based on the CMI algorithm defined above. The ECA uses the interdependence redundancy measure as the distance measure. Modes of clusters are taken into account to calculate the distance. A mode (M) of a cluster is defined as the expression value of the gene in the cluster that has the highest multiple interdependence redundancy (MIR). The MIR of a gene (x) is defined as the sum of its IR measure with all the other genes belonging to the same cluster as shown below.

$$MIR(x) = \sum_{(x,y \in C_k)} IR(x:y) \quad (6)$$

where y represents the genes in the same cluster as x. The genes are clustered on the basis of a gene's higher IR measure with the mode of a cluster. A flow diagram of the ECA is shown in Figure 1(a).

Creation of Individuals:

Each individual is represented as a two-dimensional array of numbers. One index represents the gene number from the microarray data. The other denotes the cluster number. It simply shows the genes belong to certain clusters in unique ways. A set of gene expression values which is equal to the number of clusters is chosen randomly. They are assigned as the modes for the clusters. The remaining genes are then assigned to the clusters. It is done on the basis of the genes' highest IR measures with the corresponding modes of the respective clusters. The other individuals in the population are created in a similar fashion. The total number of genes in an individual is a constant. This is true for all individuals. A diverse set of individuals having different types of gene distributions is generated. This phenomenon is implemented in the algorithm by changing the seed values. Here the population size is optimized to 300. This was obtained after studying the effects of population size on the final outcome (Behera et al 2018).

Mutation Operations

a) Cluster Assignment Operator

The Genetic K-Means algorithm has been introduced by combining the K-means model of clustering and the genetic algorithm concepts (Krishna and Murty 1999). This model gives the optimum result in commensurate with a convergence criterion. Here we have implemented a fusion of the K-mode clustering algorithm with an evolutionary algorithm in two steps. It identifies the mode of each cluster. It is the gene number (with certain expression value) having the maximum MIR in that cluster. Other genes are assigned to respective clusters having higher IR with the particular cluster mode. Then the new modes of the clusters and the fitness of the individual are calculated.

b) Probabilistic Mutation Operator

The mutation rate is a user-defined input. The best 5% of the population is retained to prevent loss of better fit individuals. This process increases the efficiency of the genetic algorithm. The remaining 95% of the individuals are stochastically selected for mutation. To select an individual for mutation, a random number is generated. The random number varies between zero and one. An individual from the total population is selected randomly for the mutation if the random number is less than the mutation rate. A mutation rate is generally a small number. An individual is defined as a set of clusters and in each cluster there are a set of genes. For the selected individual in a cluster that has at least five genes is considered. A gene having the least MIR value is transferred to another cluster of the same individual. The values of the IRs of the genes with the modes of the clusters in an individual are used to construct a roulette wheel. This is described below for details. The cluster to which a gene is transferred is selected stochastically using the roulette wheel selection. Then the new modes of the clusters and the fitness of the individual are evaluated. All the clusters undergo the mutation process. Then a mutated individual is created. Then the new population consists of the top 5% of the populations in fitness plus the mutated individuals and the normal ones. The unmutated individuals are taken from the sorted population having lower fitness value in the population. This is implemented to remove the bias due to the selection of the top 5% of the population, avoid premature convergence, bring diversity and allow the worse individuals to evolve. The probabilistic mutation rate is optimized to be 0.1 after analyzing the effects of mutation rates (Behera et al 2018). A flow chart of the steps of the operator is shown in Fig. 1(b)

Fitness Function:

An individual with the highest fitness value is the fittest individual in the population. The fitness of an individual is the sum of the multiple interdependency measures of the total

number of clusters. It is defined as:

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

where F is the individual fitness, i denotes the cluster number and R_i is the multiple interdependence redundancy measure of the cluster i .

Selection:

The roulette wheel selection method is carried out to select all the individuals for the next generation. It is a popular selection method used in a genetic algorithm. A roulette wheel is constructed from the relative fitness of each individual. It is represented in the form of a pie chart where the area occupied by each individual on the roulette wheel is proportional to its relative fitness. The sum of relative fitness (S) is calculated. A random number in the interval of 0 and S is generated. The corresponding relative fitness values are summed. When the sum is greater than the random number generated, the corresponding individual is selected. An individual with better fitness value will occupy a bigger area in the pie chart. So the probability of selecting this individual will also be higher. A pictorial representation of the roulette wheel selection is shown in Figure 1(c).

Termination:

The algorithm terminates when the change of the mean fitness for ten consecutive generations is less than 2%. As an example, at the n^{th} generation, the percentage differences between the mean fitness of the $(n-i)^{\text{th}}$ generation and the $(n-10)^{\text{th}}$ generation are calculated, where i varies from 0 to 9. If all these ten differences are less than 2%, the program execution is terminated, else it proceeds to the next generation.

RESULTS

The tests was performed on a 3.0 GHz dual-core Intel Xeon processor with 2 GB RAM. The whole work was implemented in Java version 1.6.0. It was divided into three parts – discretization, calculation of redundancy, and the development of the ECA. Time taken for discretization for a set of 7129 genes and 30 samples was 0.6469 minutes. The total number of intervals produced was 55337. For the same dataset, the calculation of redundancy took 1363.3997 minutes. One simulation of ECA took 48.6153 minutes. The equilibrium generation was in the ranges between 12 and 14. For the same set of data, one simulation of CMI took 0.7068 minutes and had 2 to 4 iterations for convergence. One simulation of k-means for 10000 iterations took 70.56 minutes. For a higher number of iterations, k-means showed the recurrence of the solution more than once. This meant the solution was likely to be a statistically optimal solution.

Synthetic data:

To analyze the efficiency of the algorithm for clustering data, studies on two different synthetic data are performed initially. Each dataset contains 200 samples and 20 genes. The gene expression values vary from 0.0 to 1.0 such that the domain is divided into two intervals. The two intervals are $[0.0, 0.5]$ and $(0.5 - 1.0]$. The first dataset comprises of two clusters and three classes as defined below. The range of values of genes G_1 and G_2 are used to define the values of other genes. G_1 defines the values of genes G_3 to G_{11} such that G_3 to G_6 is in the same range as G_1 and G_7 to G_{11} are in the opposite range. Similarly, G_2 defines the values of G_{12} to G_{20} . The values of G_{12} to G_{15} are in the same range as of G_2 and that of G_{16} to G_{20} is in the opposite range. The samples for which G_1 and G_2 range from 0.0 to 0.5

are assigned class label 1. Class label 2 is assigned to the samples for which G_1 and G_2 ranges from 0.5 to 1.0 and the rest of the samples are assigned class label 3. The second dataset comprises four clusters and five classes and the expression values vary from 0 to 10. The method adopted for generating the second dataset is similar to the method described for the first dataset. The equal width discretization is used to discretize the synthetic datasets (Cios and Kurgan 2004). For both these datasets, it is seen that for all the simulations, the ECA correctly identifies gene distribution in the clusters. But the CMI is able to group on an average 65% genes correctly (Behera et al 2018).

Real data:

To evaluate the performance of the ECA, three gene expression data sets are used. They are gastric cancer dataset (Tsutsumi et al 2002), colon cancer dataset (Alon et al 1999) and brain cancer (MacDonald et al 2001) dataset. Descriptions of all the datasets are given in Table 1. For each dataset, 50 simulations of the CMI are taken. The number of clusters vary from 2 to 20 for each of these simulations. The cluster showing the highest individual fitness becomes the optimal cluster of the dataset for that simulation. The smallest value is chosen because the data can be scattered for large value. The smallest value of the optimal cluster number among the 50 simulations is used as the optimal cluster number for the dataset. For all the algorithms, the same cluster number is used for all the simulations.

The ECA is compared with the k-means and the CMI. For the purpose of comparison, ten simulations for each of the algorithms are considered. For each simulation, a set of clusters is obtained. A subset of genes, is selected from each cluster for the classification studies. These are called candidate genes. They are composed of the top-ranking genes having higher fitness. In the case of the ECA and the CMI, the top-ranking genes are defined as the genes having the highest multiple redundancy measure in clusters. For the k-means,

they are the genes having the least Euclidian distance from the mean value of the cluster. For the classification accuracy calculation, the leave-one-out cross- validation (LOOCV) method is used. For LOOCV, the first sample is selected as the test set and the remaining samples as the training set. The top-ranking genes from the training set are used to predict the test sample whether diseased or normal. Then the second sample is used as the test sample. The remaining set is taken as the training set. The process is repeated by considering each individual sample as the test sample. The classification accuracy is calculated as the percentage of the test samples correctly predicted as either diseased or normal. A higher value of classification accuracy implies better efficiency of the algorithm to pinpoint the genes that contain the most diagnostic information.

SECTION A: Studies on gastric cancer dataset (GDS1210)

Comparison of the algorithms based on the classification accuracy of samples

The average classification accuracy of the top-ranking genes is calculated for each of the algorithms. Table 2 shows one instance of gastric cancer dataset. The percentage differences of classification accuracy are found of the ECA. They are compared with the CMI and the k-means. It clearly shows the improvement of the ECA over the other algorithms. They are given for gastric cancer dataset are shown in Table 3. The studies show that the ECA outperforms the CMI and the k-means.

For any stochastic computation when the seed for generating the random number is changed for a new simulation, the initial gene distributions in the clusters also change. This can lead to different evolutionary dynamics for each simulation such as the equilibrium generation number and mean fitness. Hence the average effects of different simulations corresponding to different initial gene distributions in the clusters are computed. A

comprehensive analysis of the algorithms on gastric cancer dataset shows the following results. The number of simulations that provide higher classification accuracy of test samples is more in the ECA than the CMI and the k-means. This proves a higher probability of finding the correct candidate genes happens in the ECA than the other algorithms. The performance of classification accuracies for the ten simulations of the ECA, the CMI, and the k-means are shown in Table 4. The table shows the number of simulations corresponding to different numbers of top-ranking genes and the resultant classification accuracy.

Analysis of classificatory genes

The decision tree built by C4.5 on the genes subset is analyzed. The cases show 96.67 % classification accuracy for only one gene, either TGIF1 or in some cases D26129. A literature survey shows that TGIF1's lower expression is correlated significantly with the metastasis of lymph node. It may inhibit the invasion and metastasis of gastric cancer via the downregulation of MMP9 and VEGF proteins (Liu et al 2006). D26129 is also a dominant gene influencing the progression of gastric cancer (Wang et al 2006). A study on the ranking of the above two classificatory genes by the CMI and the k-means is depicted. Table 5 shows the comparative ranking of the genes in all three algorithms. The two genes TGIF1 and D26129 are ranked very low by the k-means and comparatively lower by the CMI, too. In Table 5, the numbers below the name of each algorithm indicate the positions of the corresponding genes in the respective clusters (shown in square brackets) for three different simulations. All the 10 simulations of ECA have shown 96.67% classification accuracy with the top 3 ranking genes. Since either gene TGIF1 or D26129 can only achieve 96.67% accuracy value, either of the genes must be ranked among the top 3 ranking genes. It can be concluded that the ECA is effective in choosing out genes that contain significant diagnostic information for a particular disease.

A comparative study of the three algorithms based on the common genes found across all the simulations beyond a certain level of classification accuracy is shown. A set of top-ranking genes (i.e. top 1, 2 and 3), as shown in Table 6, is selected from each simulation. The genes recurring in all these simulations are the common genes. A list of the top-ranking common genes that is obtained using the ECA for the top 6 ranking genes is given in Table 7(a). Table 7(b) contains the genes that are found to be significant in the classification process done by the C4.5 decision tree. Those genes are used as the nodes in constructing the decision tree.

Comparison of algorithms based on the representative genes

The collective trend of the expression levels of a group of co-expressed genes is given by the degree of coherence. Similarity among the genes takes into account only the genetic distances from the cluster centre. But the interdependence factor considers the genetic distance as well as the negative and the positive correlations among the genes. For a pair of genes, the degree of coherence can be measured in the sense of similarity or positive/negative correlation. To understand the coherence pattern, three representative genes are selected from the top five genes. Their patterns are examined across different clusters for each algorithm. For the ECA and the CMI, the most significant genes in different clusters are selected based on the magnitude of their multiple interdependency measures. For the k-means, those genes having lesser distance from the mean of the clusters are chosen. The pattern of coherence of these genes for the algorithms - ECA, CMI, and k-means are shown in Figure 2, 3 and 4 respectively for a single simulation. The simulation is a random scenario. The fact holds good for a many simulations.

In the ECA, the HMHA1 and TGIF1 genes are interdependent as they show a

negative correlation in their overall expression profile shape. However, they are not similar because of the large genetic distance between them. The HMHA1 and C21orf33 genes are similar in their expression profile and show negative correlations in some regions of the profile. The TGIF1 and C21orf33 genes show regions of positive and negative correlations across the expression profile shape. The CMI also shows the interdependence even though the genes are not similar due to large scaling factors as shown by the HMHA1 and CUL5 genes. The HMHA1 and COPS5 genes are similar and also show negatively and positively correlated regions across the profile. But the k-means algorithm shows coherence only in the similarity sense. It does not take into account the scaling factors or interdependence for all the genes.

The conclusion is that while the Euclidean distance takes into consideration only similarity factor for clustering the genes, the interdependence redundancy measure can cluster the genes using both similarity and positive/negative correlations. The ECA is able to group the genes which show correlations in their overall expression profile shape as opposed to the small regions across the profile shape (determined by short Euclidean distance). Also, the ECA takes into consideration larger scaling factors (higher value in gene expression) in comparison to the CMI. From these results, it can be said that the ECA is more coherent in clustering genes as compared to the CMI. This shows that although the ECA and the CMI use the same distance measure, i.e., interdependence redundancy measure for clustering, the ECA utilizes better the interdependence measure compared to the CMI. The algorithm in the ECA is more powerful than the CMI.

Study of gene distribution in clusters

The clusters formed by the ECA consist of nearly the same number of genes (Behera et al 2018). Thus the genes in different clusters are evenly distributed. The top-ranking genes, thus selected, provide significant and insightful information. This is evident by a higher

accuracy level in classifying the test samples. Thus the cluster configuration obtained by the ECA is more reliable to select the top candidate genes. The k-means algorithm produces the most lopsided distribution gene distribution (Behera et al 2018). Compared to the k-means algorithm, the CMI shows even-distribution pattern in many cases. But its performance is poor in comparison to the ECA.

A randomly selected single simulation for each of the ECA, the CMI, and the k-means algorithms shows the following trend of gene distributions in different clusters. As an example, the ECA has partitioned 7129 genes into four different clusters as 1911, 1786, 1628 and 1804 whereas the CMI has clustered the same number of genes into 2280, 2752, 1767 and 331 genes. The k-means algorithm generates clusters with 458, 25, 6545 and 101 genes. The gene distribution in clusters shows a better performance of the ECA over the CMI and the k-means algorithms. This result holds good for many simulation samples chosen at random.

SECTION B: A brief study on colon cancer dataset

Comparison of the algorithms based on classification accuracy

The average classification accuracy for all the ten simulations are calculated for each of the algorithms for the total test samples. Table 8 shows one instance of the colon cancer dataset for the top 5 genes per cluster. To study the improvement of the ECA over the CMI and the k-means algorithms, the percentage difference of the classification accuracy is computed. The improvement of the ECA over the CMI and the k-means algorithms with respect to a few top-ranking genes per cluster for the colon cancer dataset is shown in Table 9. The comparative studies show that the ECA outperforms the CMI and the k-means algorithms.

This colon cancer dataset has been extensively studied by Au et al (2005). Here our aim is restricted to the analysis of the classification accuracy of the test samples.

SECTION C : A brief study on brain cancer [medulloblastoma metastasis] dataset (GDS232)

Comparison of the algorithms based on the classification accuracy

For each of the algorithms, the average classification accuracy of all the ten simulations is found. Table 10 shows one instance of medulloblastoma metastasis dataset for the top one gene per cluster. Better performance of the ECA over the other two algorithms is obtained. The analysis show that the performance of the ECA surpasses the CMI and the k-means algorithms. The percentage improvement of the ECA over the CMI and the k-means algorithms predicted by the few top-ranking genes per cluster for the gastric cancer dataset is shown in Table 11. A comprehensive analysis of the above three algorithms on brain cancer dataset shows that the number of simulations containing the top-ranking genes is more in the ECA than in the CMI or the k-means algorithms. This proves that there is a higher probability of finding the correct candidate genes of cancer in the ECA than the CMI and the k-means algorithms. Here we have focused on proving only the superior performance of the present algorithm over the CMI and the k-means algorithms. Therefore, we have not attempted to discover the exact number, names, and functions of the candidate genes that contain the diagnostic information for brain cancer.

CONCLUSION

The recent advancement in technology can obtain high-throughput microarray data containing the expression profiles of thousands of genes. This effort requires the extraction of a handful of important top-ranking genes that most likely contain the diagnostic information

out of thousands of genes. This analysis shows light on a disease process. Hence the microarray data clustering and selection for the candidate genes for cancer have gained much importance in the recent years. Here a new algorithm towards solving this problem has been developed.

This algorithm is an amalgamation of the clustering algorithm and evolutionary computation. This ECA is found to be a powerful software tool in finding the top-ranking candidate genes to classify the test samples as diseased or normal with high accuracy. The Cluster Assignment Operator in the ECA is a powerful operator and has been seen to speed up the convergence of this ECA without compromising on the efficiency of the algorithm. This means that within a short computational time, the best possible solution in terms of the right gene distribution in different clusters is obtained. The Probabilistic Mutation Operator in the ECA brings diversity in the population and hence avoids early convergence with a poor target solution (bad clustering of genes). During the development of the algorithm, other genetic algorithm operators were also investigated but were discontinued due to their limitations. For example, the traditional crossover operator was tried but found to be disruptive in nature. This ECA is a very typical evolutionary computation program that provides a good solution without a crossover operator. Generally, crossover genetic algorithm operators help for obtaining better optimized solutions in other evolutionary models. In another operator called a Swap Mutation Operator in the ECA, the genes having the least MIR from two random clusters are swapped. The outcome shows it can explore only a small diversity of gene clustering space. This is not very useful for the present purpose.

The ECA is more stable with respect to lopsided gene distribution across clusters in comparison to the CMI and the k-means algorithms. Also, the same set of genes get clustered together with higher probability in the case of the ECA in different simulations unlike the other two algorithms. The investigation is done by selecting the top-ranking genes from

different clusters in different simulations. It shows very high efficacy in clustering genes based on similarity in the expression profiles and on positive and negative correlation between the genes. Here the interdependence redundancy measure is a far better metric. Although the CMI uses the interdependence measure like the ECA, the ECA is better in clustering the genes and finding the top-ranking genes because of its exposure to a very large search space in the gene expression values. Hence the ECA is very effective in grouping the genes.

Microarray datasets generally have many genes. Classification of test samples by machine learning algorithms becomes difficult due to the high possibility of the presence of noise in such systems. A dimension reduction of the datasets to obtain the informative genes is crucial. For this purpose, a subset of genes is selected for each algorithm. It is seen that the ECA, by and large, outperforms the other two algorithms in the classification accuracy of the test samples via a small number of selected candidate genes. For gastric cancer, colon cancer, and brain cancer datasets, the classification accuracy is calculated by the top-ranking genes selected by the ECA. Certain classification accuracies, when the top-ranking genes alone are used, are higher than when the whole dataset is used. As an example, for the gastric cancer dataset, the classification accuracy is 93.33% when the whole dataset is taken into consideration. However, when TGIF1/D26129 alone is used, the classification accuracy is 96.67%. This shows that the ECA is able to select candidate genes properly and effectively. They can be further investigated their functional and diagnostic properties in connection to drug development.

The Attribute Clustering Algorithm (ACA) developed by Au et al essentially uses interdependence redundancy measures and the OCDD discretization process. They have shown its superior performance than biclustering, K-means, Self Organization Map (SOM) and some other conventional algorithms. Their ACA is not available in the public domain nor

they reveal the details of their discretization process such as the smoothing process and chi-square test. The CMI algorithm used here is similar to the ACA except that a somewhat different smoothing process is used. Performance of the ACA and the CMI are comparable. It is clearly proven that the ECA is superior to the CMI and the K-means algorithms. Hence it should be superior to biclustering, SOM and some others (as shown by Au et al). The most fundamental reason is that most conventional algorithms use only the Euclidean genetic distance measures in different ways. They don't use interdependence redundancy measure that directly takes into account positive and negative correlations between the genes. Therefore, the ECA is the most superior algorithm at the present for the microarray data analysis to find the right cancer genes for a disease process.

The present formalism can be developed for a three-class microarray gene expression data. For example, cancer microarray data can have samples of normal, mildly developed and fully diseased tissues. The discretization algorithm must be modified. The smoothing of the data can be further improved to make the machine learning algorithm more effective. Other possible mutation operators can, in principle, be studied to make the ECA more effective.

Acknowledgement: The authors acknowledge a funding from the Philips Research Asia, Bengaluru.

References

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

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.

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

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 protein multiple sequence alignment by genetic algorithm, *Procedia Computer Science*, Vol 108, 1135-1144.

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)

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, Uhm, 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).

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.

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.

Herbola A et al (2022) Chapter 27- Bioinformatics and biological data mining. *Bioinformatics: methods and applications*, Academic press, 457-471.

Holland, J H (1975) *Adaptation in natural and artificial systems*, University of Michigan Press, Ann

Arbor, MI, USA.

Heyer, L.J., Kruglyak S and Yooseph S (1999) Exploring expression data: identification and analysis of

coexpressed genes. *Genome Research* 9, 1106-1115.

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, 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.

Liu et al (2009) Computational data mining in cancer bioinformatics and cancer Epidemiology. *BioMed*

Research international. DOI.org//1155//2009//582697.

Lloyd, S.P., (1982) *Least Squares Quantization in PCM*. *IEEE Transactions on Information Theory*, vol.

28, no. 2, pp. 129-137

MacDonald, T.J., et al. (2001) Expression profiling of medulloblastoma: PDGFRA and the RAS/MAPK

pathway as therapeutic targets for metastatic disease. *Nat Genet*. Oct;29(2):143-52.

- 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.
- 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 discretization of continuous data. *Intelligent Data Analysis* 8(2): 151-170.
- Zhang Y et al (2009) Bioinformatics analysis of microarray data. *Methods Mol. Biol.* 573, 259-284

APPENDICES

Appendix A1: A brief overview of the OCDD algorithm

OCDD is applied to discretize the continuous data.

1. For each gene, an arbitrary partition P is assumed. The total number of classes is C . The intervals are denoted as $i_0 \dots i_{r-1} \dots i_R$ and classes are denoted as $l_1 \dots l_c \dots l_C$. A quanta matrix of $P \times C$ is prepared. The matrix values are defined as q_{rc} where it represents the total number of observed data points belonging to the class l_c and falling in the interval i_{r-1} and i_r . From this quanta matrix q_{+c} , q_{r+} , T and p_{rc} are calculated, where q_{+c} represent the total number of data points in the class l_c , q_{r+} represents the total number of data points in the interval i_{r-1} and i_r and T is the total number of observed data points. The interdependence redundancy measure, z , is initialized as $M(L:G_p)/E(L:G_p)$ where $M(L:G_p)$ denotes the mutual information and $E(L:G_p)$ denotes the joint entropy between the gene G having P partitions and class label L . An algorithm A1 (Wong et al., 2004) is used to calculate the new partition P' .
2. An overview of the algorithm A1:
 - a) A dynamic programming table of size $N \times N$ is created, where N is the number of unique values of gene G . The gene values (x) are arranged in ascending order, such that i is the position of x . The element in the n^{th} row and s^{th} column represents the value of $m(n, s)$.
 - b) The matrix values are filled as per the recursive equation. $X_{l,n}$ corresponds to the interval having the gene values from x_l to x_n . $\left| x_{1,n} \Delta l_c \right|$ represents the number of genes belonging to the interval $X_{l,n}$ and the class c . f is a numeric value ranging from 1 to $n-s$ in order to traverse through all possible interval boundaries for partition s .
 - c) The column number corresponding to the maximum value in the last row denotes the optimal number of partitions. The interval boundaries are found by traceback.

$$m(n, s) = \min_{1 \leq f \leq n-s} \left(m(n-f, s-1) + \sum_{c=1}^C \left(\frac{\left| x_{n-f+1, n} \wedge l_c \right|}{T} \right) \frac{\left| x_{n-f+1, n} \wedge l_c \right|}{\left| x_{n-f+1, n} \right| \left| l_c \right|} \frac{\left| x_{n-f+1, n} \wedge l_c \right|}{T} \frac{\left| x_{n-f+1, n} \wedge l_c \right|}{T} \right) \quad (A.1)$$

$$m(n, 1) = \sum_{c=1}^C \frac{\left| x_{1, n} \wedge l_c \right|}{T} \frac{\left| x_{1, n} \wedge l_c \right|}{\left| x_{1, n} \right| \left| l_c \right|} \frac{\left| x_{1, n} \wedge l_c \right|}{T} \frac{\left| x_{1, n} \wedge l_c \right|}{T} \quad (\forall 1 \leq n \leq T) \quad (A.2)$$

$$m(1, s) = \begin{cases} \sum_{c=1}^C \frac{\left| x_{1, 1} \wedge l_c \right|}{T} \frac{\left| x_{1, 1} \wedge l_c \right|}{\left| x_{1, 1} \right| \left| l_c \right|} \frac{\left| x_{1, 1} \wedge l_c \right|}{T} \frac{\left| x_{1, 1} \wedge l_c \right|}{T} & i \neq 1 \\ \infty & o = t \end{cases} \quad (A.3)$$

3. The new redundancy value, z' is calculated.
4. z and z' are compared. If $z = z'$, then the optimal partition is P' . Otherwise, z is initialized to z' and steps 2, 3 and 4 are repeated.

Appendix 2: Smoothing and Chi-square test method

For smoothing of the data, segment s_i is defined for any gene value x_i such that the segment consists of values from x_{i-w} to x_{i+w} where w is the width of the segment. For this segment, a ratio (r) is calculated as the frequency of the most frequently occurring class label and the frequency of the class label for x_i . If r is greater than some threshold value (t), the class

label is changed to the most frequently occurring class label. For discretization, we have used the default values for w as 5.0 and for t as 1.3. The smoothing and the parameter values used are essentially the same as described in the OCDD algorithm (Wong et al., 2004).

The Chi-square test is done to test the statistical significance of the interdependence between a gene and the class labels. For two neighboring intervals, it finds whether the frequency distributions among them and the class label are significantly interdependent. Accordingly one decides to merge the intervals. For this purpose, the IR between each gene and class label is computed. Let the joint entropies between a gene and class label be E . A product (p) between E and the total number of data points is calculated. The ratio between the chi value and twice of p is computed. It is called as rt . If the IR is greater than the rt , then the intervals are merged. Chi values are taken from the standard chi-square distribution table. The degree of freedom for the chi value is defined as the product of the total number of classes minus one and the total number of intervals minus one. The methodology is the same as described in the OCDD algorithm (Wong et al., 2004).

FIGURE LEGEND

Figure 1(a) shows the flow chart of the ECA.

Figure 1(b) depicts the flow chart of the probabilistic mutation.

Figure 1(c) shows the pictorial representation of roulette wheel selection.

Figure 2 presents the normal line graph of the expression pattern of 3 representative genes for cluster 1 of

a single simulation for the ECA algorithm for the gastric cancer dataset.

Figure 3 shows the normal line graph of the expression pattern of 3 representative genes for cluster 2 of a single simulation for the CMI algorithm for the gastric cancer dataset.

Figure 4 gives the normal line graph of the expression pattern of 3 representative genes for cluster 3 of a single simulation for the k-means algorithm for the gastric cancer dataset.

FIGURES

Figure 1 (a)

Figure 1 (b)

Fig 1(c)

Figure 2

Figure 3

Figure 4

TABLE LEGEND

Dataset description:

Table 1 contains the description of each dataset with respect to the number of a gene, the number of total samples, number of diseased samples, number of healthy samples, the number of clusters, minimum and maximum gene expression values.

Gastric cancer dataset (GDS1210) [Table 2 – Table 7] :

Table 2 shows the classification accuracy of the two top-ranking genes for the ECA, the CMI and the k-means for the gastric cancer dataset. The individual classification accuracy for all the 10 simulations and the average classification accuracy for each algorithm are shown.

Table 3 shows the improvement of the ECA over the CMI and the k-means for the gastric cancer dataset. The average classification accuracies of the corresponding algorithms are used to compute the difference of the percentages with respect to the ECA.

Table 4 depicts the performance of the ECA, the CMI, and the k-means in terms of classification accuracy for the gastric cancer dataset. The number of simulations showing the corresponding classification accuracy for one to six top-ranking genes are shown in the

respective columns.

Table 5 presents the positioning of the two classificatory genes (TGIF1 and D26129) in different algorithms for 3 random simulations (gastric cancer dataset). The numbers represented below the name of each algorithm indicate the position of the corresponding gene in the respective clusters, which is shown in square brackets.

Table 6 gives the classification accuracy of common genes for 1,2,3 and 4 top-ranking genes for the ECA, the CMI and the k-means.

Table 7 shows the list of genes which are selected by the ECA (a) as found in common genes across simulations and (b) by using the C4.5 classification tree(for the gastric cancer dataset)

Colon cancer dataset [Table 8, Table 9]:

Table 8 shows the classification accuracy of the top-ranking five genes for the ECA, the CMI, and the k- means for the colon cancer dataset. The individual classification accuracy for all the 10 simulations and the average classification accuracy for each algorithm are shown.

Table 9 shows the improvement of the ECA over the CMI and the k-means for the colon cancer dataset. The average classification accuracies of the corresponding algorithms are used to compute the percent difference with respect to the ECA.

Brain cancer [medulloblastoma metastasis] dataset (GDS232) [Table 10, Table 11] :

Table 10 shows the classification accuracy of top-ranking one gene for the ECA, the CMI, and k- means for the medulloblastoma metastasis dataset. The individual classification accuracy for all the 10 simulations and the average classification accuracy for each algorithm are shown.

Table 11 shows the improvement of the ECA over the CMI and the k-means for the medulloblastoma metastasis dataset. The average classification accuracy of the corresponding algorithms is used to compute the difference of percentage with respect to the ECA.

TABLES

Table 1

	gastric cancer	colon cancer	brain cancer
Number of genes	7129	2000	2059
Number of total samples	30	62	23
Number of diseased samples	22	40	10
Number of healthy samples	8	22	13
Number of clusters	4	10	6
Minimum gene expression value	0.1	5.82	0.5
Maximum gene expression value	14,237	20,903	1,84,262

Table 2

Simulation s	Top 2 genes per cluster		
	ECA	CMI	K-means
1	73.33	83.33	63.33
2	96.67	73.33	70.00
3	96.67	86.67	73.33
4	96.67	76.67	63.33
5	96.67	83.33	73.33
6	96.67	76.67	73.33
7	80.00	93.33	70.00
8	76.67	73.33	73.33
9	73.33	73.33	73.33
10	96.67	63.33	63.33
Average	88.34	78.33	69.66

Table 3

Top genes per cluster	% improvement of ECA over	
	CMI	k-means
1	4.4	1.99
2	11.32	21.15
3	5.17	29.66
4	5.86	24.14
5	6.55	25.18

Table 4

Genes per cluster	% classification accuracy								
	>60 & <=80			>80 & <=95			>95 & <100		
	ECA	CMI	k-means	ECA	CMI	k-means	ECA	CMI	k-means
Top 1	5	7	3	3	3	7	2	-	-
Top 2	4	6	10	-	4	-	6	-	-
Top 3	-	2	10	-	1	-	10	7	-
Top 4	-	2	7	-	1	3	10	7	-
Top 5	-	2	10	-	2	-	10	6	-
Top 6	-	2	10	-	1	-	10	7	-

Table 5

D26129			TGIF1		
Position of genes [Cluster No.]			Position of genes [Cluster No.]		
ECA	CMI	k-means	ECA	CMI	k-means
1 [3]	812[1]	412 [3]	1 [3]	3 [1]	5854 [1]
1 [3]	706 [1]	414 [3]	2 [1]	38 [3]	5854 [4]
2[1]	7 [3]	412 [1]	1 [1]	25 [1]	5854 [4]

Table 6

genes per cluster	common genes			% classification accuracy		
	ECA	CMI	k-means	ECA	CMI	k-means
1	1	2	2	76.67	76.67	73.33
2	4	5	3	80.00	73.33	76.67
3	6	8	4	96.67	96.67	76.67

Table 7(a)

Genes	Functions
TGIF1	Low expression is significant in metastasis; may inhibit invasion by downregulation of two other genes; has functions in transcription corepressor activity and regulation of transcription
X90840	mRNA for axonal transporter of synaptic vesicles
IGFBP4	Role in cancer poorly defined; overexpression leads to reduced growth in some cancers; seems to be highly expressed in a study on gastric cancer; locally available IGFs appear to stimulate mitogenesis; involved in DNA metabolism, cell proliferation, regulation of cell growth and signal transduction
HMHA1	Histocompatibility minor (HA-1); expression found in many solid tumor cells; the function involves intracellular signaling cascade also
CUL5	Shows moderate expression in gastric cancer; has a role in cell growth and proliferation also
QSCN6 (QSOX1)	Potentially involved in cell growth and highly induced in quiescent cells; thus inhibition of QSCN6 could play a role in cancer
COPS5	Translation initiator factor activity involved in the regulation of progression through the cell cycle
UBE4A	Involved in ubiquitin-dependent protein catabolic processes; get downregulated in diffuse-type gastric cancer
LOC4403 45	A hypothetical protein
CCL4	Has roles in cell motility and signal transduction
ACAT1	Acetyl Co-A c-acetyltransferase activity; involved in fatty acid metabolism

Table 7(b)

Genes	Functions
TGIF1	Refer Table 8(a)
D26129	Associated with the progression of gastric cancer

Table 8

Simulation	TOP 5 genes per cluster		
	ECA	CMI	k-means
1	87.10	58.06	61.29
2	70.97	67.74	64.52
3	75.81	64.52	54.84
4	66.13	72.58	77.42
5	85.48	61.29	67.74
6	88.71	62.9	64.52
7	74.19	69.35	69.35
8	82.26	80.64	67.74
9	87.09	80.64	67.74
10	79.03	69.35	51.61
Average	79.68	68.71	64.68

Table 9

Top genes per cluster	% improvement of the ECA over	
	CMI	k-means
1	11.88	8.94
2	9.89	7.36
3	12.13	12.55
4	14.74	15.58
5	13.76	18.83

Table 10

Simulation	Top 1 gene per cluster		
	ECA	CMI	k-means
1	65.22	78.26	73.91
2	65.22	69.57	39.13
3	69.57	39.13	34.78
4	69.57	39.13	39.13
5	69.57	43.48	52.17
6	52.17	17.39	34.78
7	69.57	60.87	43.48
8	69.57	34.78	39.13
9	69.57	69.57	73.91
10	69.57	34.78	30.43
Average	66.96	48.7	46.09

Table 11

Top genes per cluster	% improvement of the ECA over	
	CMI	k-means
1	27.28	31.17
2	34.78	15.53
3	23.57	13.38
4	5.15	0.00
5	14.69	5.59