Hybrid Genetic Algorithm and Simulated Annealing for Clustering Microarray Gene Expression data

M Pandi¹, T Sivakumar², N Senthil Madasamy³ and N Sadhasivam⁴
¹, ², ³ CSE Department, Dr. Mahalingam college of Engineering and Technology, Pollachi.
⁴CSE Department, Sri Indu college of Engineering and Technology, Ibrahimpatan, Telangana.

Email: mpandi123@gmail.com

Abstract. Gene expression is the process by which information in gene is used to create proteins. The gene expression studies generate large amount of data. These data, referred to as the gene expression matrix, represent the expression levels for thousands of genes recorded at a few time instances. A typical microarray experiment involves the hybridization of an mRNA molecule to the DNA template from which it is originated. Many DNA samples are used to construct an array. The amount of mRNA bound to each site on the array indicates the expression level of the various genes. This number may run in thousands. All the data is collected and a profile is generated for gene expression in the cell. Clustering is a process of partitioning a set of meaningful subclasses called clusters. Clustering is a key step in the analysis of gene expression data. Genetic Algorithms are a family of computational models inspired by evolution. The searching capability of genetic algorithms is exploited in order to search for appropriate cluster center in feature space such that a similarity metric of resulting clusters is optimized. The chromosome which are represented as strings of real numbers, encode the centers of fixed number of clusters. The experiment results are demonstrated on real data sets and the performance of GA is evaluated in comparison with the state-of-the art algorithm K-Means with use of internal validation criteria.

1. Introduction

1.1 Gene Expression

A Gene is a unit of heredity in a living organism. It normally resides on a stretch of DNA that codes for a type of protein or for an RNA chain that has a function in the organism. All living things depend on genes, as they specify all proteins and functional RNA chains. Genes hold the information to build and maintain an organism’s cell and pass genetic traits to offspring. The vast majority of living organism encodes their genes in long strands of DNA. The most common form of DNA is in a cell is in a double helix structure, in which two individual DNA strands twist around each other in a right-handed spiral. DNA consists of a chain made from four types of nucleotide subunits such as adenine, cytosine, guanine, and thymine. In this structure, the base pairing rules specify that guanine pairs with cytosine and adenine pairs with thymine [1][23].
Gene expression occurs in two steps: a. transcription, b. translation. The process of genetic transcription produces a single stranded RNA molecule known as messenger RNA from DNA. The process of translation produces a defined sequence of amino acids from mRNA. To measure the gene expression levels, number of techniques has developed. In earlier days, using low-level techniques such as Northern blot, Western blot, and Reverse transcription using PCR is used. These techniques are able to monitor only expression level of single gene. Nowadays, Microarray technology is used. Two main types of microarrays are: (a) cDNA arrays: spotted onto surface (b) oligonucleotide arrays: created on surface. Microarray technology is used to monitor the thousands of genes simultaneously. Some of the high-level techniques are SAGE, DNA microarray, RNA-Seq were developed.

1.2 Microarray Technology

Microarray technology now allows looking at many genes at once and determining which are expressed in a particular cell type. DNA molecules representing many genes are placed in discrete spots on a microscope slide. This is called a microarray. Thousands of individual genes can be spotted on a single square inch slide. The RNA molecules are then “labeled” by attaching a fluorescent dye that allows us to see them under a microscope, and added to the DNA dots on the microarray. Due to a phenomenon termed base-pairing RNA will stick to the gene it came from. After washing away all of the unstuck RNA will look at the microarray under a microscope and see which RNA remains stuck to the DNA spots. From this know which gene each spot represents, and the RNA only sticks to the gene that encoded it, we can determine which genes are turned on in the cells [2]. The genes that are expressed differently in the two tissues may be in causing the disease. Microarray technology measures the level of mRNA expression. Some of the application of the microarrays is gene discovery, disease diagnosis, drug discovery, taxological research. Microarray technology will help researchers to learn more about many different diseases, including heart disease, mental illness and infectious disease to name only a few. With the help of microarray technology, however, they will be able to further classify the type of cancers based on the patterns of gene activity in the tumor cells.

![Gene expression data matrix](image1)

**Figure 1.** Gene expression data matrix

Gene expression dataset is represented in the form of matrix. A dataset (e.g., from microarray experiments) is normally given as a rectangular m x n matrix A, where each column represents a condition and each row represents a gene:

\[ A = (a_{ij})_{mn} \quad (1) \]

The rest of the paper is organized as follows: Section 2 describes the literature review on Gene expression data clustering. The overview Genetic Algorithm (GA) is given in Section 3. Section 4 presents the Modified...
Genetic Algorithm with Simulated Annealing (GA-SA) algorithm for gene expression data clustering. The experiment results are analyzed and demonstrated in Section 5.

2. Literature Review
Genetic Algorithm is the optimization search technique. GAs has been used in an attempt to optimize a specified objective function related to a clustering problem [3]. Genetic algorithm have been applied to a wide range of industrial applications but research shows that standard GAs have some defects such as unsatisfactory local searching ability and premature convergence. The article proposes a genetic algorithm to overcome these shortcomings. The simulated result shows that the hybrid algorithm helps the practical system achieve a better performance [4][24]. A systematic framework for assessing the results of clustering algorithms has been proposed. This methodology is to apply a clustering algorithm to the data from all but one experimental condition [5]. Microarrays have become a central tool in biological research [6]. A key step in the analysis of gene expression data is the identification of groups of genes that manifest similar expression patterns. This translates to the algorithmic problem of clustering genes based on their expression patterns. [7] Used Clustering as a key step in the analysis of gene expression data. Despite the widespread use of artificial intelligence techniques in bioinformatics and, more generally, data analysis, there are very few clustering algorithms based on the genetic paradigm, yet that paradigm has great potential in finding good heuristic solutions to a difficult optimization problem such as clustering. On the basis of analyzing the major advantages and disadvantages of the chaos, the chaos genetic simulated annealing hybrid algorithm is presented [2] according to the complementary advantages strategy. Proximity measure plays an important role in making a clustering technique effective [8]. Finally, since evaluation of the effectiveness of the clustering techniques over gene data requires validity measures and data sources for numeric data. Jorge Parraga-Alava et al.,[9] proposed a new method to find out the subset of matrices which are tightly co-regulated under the similar expression patterns. Xiaofeng Yang et al., [10] proposed semisupervised clustering technique called “Semi-FeaClustMOO” is demonstrated on five publicly available benchmark gene-expression datasets. Comparison results with the existing techniques for gene-expression data clustering again reveal the superiority of the proposed technique. Statistical and biological significance tests have also been carried out. Sudipta Acharya et al.,[11][25] introduced Simultaneous feature selection and unsupervised clustering for gene-expression data in multiobjective optimization framework.

3. Clustering Gene Expression Data using Genetic Algorithm
3.1 Chromosome Representation
A chromosome is an organized structure of DNA and protein found in cells. It is a single piece of coiled DNA containing many genes, regulatory elements and other nucleotide sequences. Chromosomes also contain DNA-bound proteins, which serve to package the DNA and control its functions. Genes joined together to form a string of values called chromosome. It contains the heredity information for the cell. Set of genes usually in the form of a string.

| 1 | 2 | 3 | 1 | 2 | 1 | ... | ... | ... | ... |
|---|---|---|---|---|---|-----|-----|-----|-----|
| G1 | G2 | G3 | G4 | G5 | G6 | ... | ... | ... | Gn |

Figure 2. Representation of chromosome
Cluster 1 \rightarrow \{G_1, G_4, G_6 \ldots\}
Cluster 2 \rightarrow \{G_2, G_5 \ldots\}
Cluster 3 \rightarrow \{G_3\}

3.2 Fitness Function Evaluation
The fitness of an organism is how much it can reproduce before it dies. Let \( X = \{x_1, x_2, \ldots, x_n\} \) be a set of elements, where each element is a d-dimensional vector. In our case, each gene is an element \( x \in X \), and \( x_i \) is the value of its expression level under experimental condition \( i \). Given a subset \( Y = \{y_1, y_2, \ldots, y_m\} \) of \( X \), let \( c(Y) \) denote the centroid of \( Y \) and let its variance be

\[
\text{VAR}(Y) = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{d} (y_{i,j} - c(y_{i,j}))^2
\]

Given an integer \( k \), we are interested in finding a partition \( \mathcal{P} \) of \( X \) into \( k \) classes \( C_0, C_1, \ldots, C_{k-1} \) so that the total internal variance is minimized.

\[
\text{VAR}(\mathcal{P}) = \sum_{i=0}^{k-1} \text{VAR}(C_i)
\]

3.3 Selection
During each successive generation, a proportion of the existing population is selected to breed a new generation. Individual solutions are selected through a fitness-based process, where fitter solutions (as measured by a fitness function) are typically more likely to be selected. Certain selection methods rate the fitness of each solution and preferentially select the best solutions \cite{12}\cite{26}. Other methods rate only a random sample of the population, as the former process may be very time-consuming.

Roulette wheel selection is used for this process. It is also known as fitness proportionate selection. The individual is selected on the basis of fitness. The probability of individuals to be selected increases with the fitness of the individual greater or lesser than its competitor’s fitness.

3.4 Crossover
In genetic algorithms, crossover is a genetic operator used to vary the programming of a chromosome or chromosomes from one generation to the next. It is analogous to reproduction and biological crossover, upon which genetic algorithms are based. Cross over is a process of taking more than one parent solutions and producing a child solution from them. Single point crossover is used here. Given two parents, single-point crossover will generate a cut-point and recombines the first part of first parent with the second part of the second parent to create one offspring. Single-point crossover then recombines the second part of the first parent with the first part of the second parent to create a second offspring.

3.5 Mutation
In genetic algorithms of computing, mutation is a genetic operator used to maintain genetic diversity from one generation of a population of algorithm chromosomes to the next. It is analogous to biological mutation. Mutation alters one or more gene values in a chromosome from its initial state. In mutation, the solution may change entirely from the previous solution. Hence GA can come to better solution by using mutation. Mutation occurs during evolution according to a user-definable mutation probability. This probability should be set low. If it is set too high, the search will turn into a primitive random search \cite{13}\cite{19}. Random search mutation takes a single candidate and randomly changes some aspects of it. Mutation operator performs occasional random changes of the value of genes.
3.6 Evolution
GA’s are a subclass of Evolutionary Computing. Evolution is based on “survival of the fittest”. Elitist selection is used for evaluation. The fit members of each generation are guaranteed to be selected.

4. Modified Genetic Algorithm for Clustering Gene Expression Data
Genetic algorithm suffer from the premature suboptimal convergence or stagnation which occurs when some poor individuals attract the population due to a local optimum or bad initialization, it prevents further exploration of the search space. One of the causes of this problem is that a very fit chromosome is generally sure to be selected for mating and since offspring resemble their parents, chromosomes become too similar (i.e. population loses diversity). Hence, the population will often converge before reaching the global optimal solution, resulting in premature convergence. Premature convergence can be prevented by

- Using sub populations: The population of chromosome is divided into separate subpopulations. Each subpopulation is evolved independent of the other subpopulations for a user-specified number of generations. Then, a number of chromosomes are exchanged between the subpopulations. This process helps in increasing diversity and thus preventing premature convergence.
- Re-initializing some chromosomes: A few chromosomes are re-initialized from time to time in order to add diversity to the population.
- Increase the mutation probability: Mutation aids in exploring new areas in the search space and increase diversity. Therefore increasing $P_m$ will help in preventing premature convergence.

In general, any mechanism that can increase diversity will help in preventing premature convergence. The proposed work uses applies modified GA to obtain better optimal solution [14].

To avoid premature convergence or stagnation, genetic algorithm is hybrid with simulated annealing technique. The 10% of chromosome is selected from the population, and simulated annealing technique is applied on the selected chromosomes.

4.1 Simulated Annealing (SA)
Annealing is the physical process if it’s done in the system it is called as the simulated annealing. It mainly deals with heating of a metal at the maximum temperature at these temperature particles/atoms moves freely then gradually its cooled until it reaches its freezing point at this state we can get the shape of the metal however its need gradual cooling is done to avoid the imperfection shape of the metal [15][27].

A solid is heated in a hot bath, increasing the temperature up to a maximum value. At this temperature, all material is in liquid state and the particles arrange themselves randomly. As the temperature of the hot bath is cooled gradually, all the particles of this structure will be arranged in the state of lower energy [16][22].
4.2 Simulated Annealing Algorithm

- **Initial Solution**
  - Generated using an heuristic
  - Chosen at random
- **Neighborhood**
  - Generated randomly
  - Mutating the current solution
- **Acceptance**
  - Neighbor has lower cost value
  - Neighbor has higher cost value is accepted with the probability \( p \)
- **Stopping Criteria**
  - Maximum CPU time
  - Solution with a lower value than threshold
  - Maximum number of iterations without improvement
  - Maximum total number of iterations

4.3 GA Hybrid with SA

From the result of GA individual solution will be taken. The main property of simulated annealing is that, it accepts the worse solution too. The individual solution will be taken as the Initial solution. The fitness value for initial solution should be found. From the initial solution, next solution will be generated the temperature value will be set to high initially and the fitness value will be found for the next solution. Then the value of the temperature is decreased gradually. It is decreased by using \( \text{Temperature} = \text{Temperature} \times 0.9 \).

Initial solution and the next solution value will be compared. If next solution is better than the initial solution. Then the next solution will be taken as the current solution i.e. \( \text{current} = \text{next} \). Comparison is based on the minimum value only. If the next solution obtained is the worse solution. Then the random number will be generated. The random number is applied to the value is found by subtracting the random value and the next value \cite{17,21} (which is found as the worst).

5. Experimental Result

The modified GA is tested on three datasets of gene expression data. Table 1 shows three Microarray Gene Expression Dataset for clustering.

| Name of the dataset       | Number of genes | Number of conditions |
|---------------------------|-----------------|----------------------|
| Standard Yeast Cell       | 384             | 17                   |
| CNS RAT                   | 112             | 17                   |
| Blood Monocytes           | 2329            | 139                  |
Figure 3. Comparisons of GA and SA in Standard Yeast Cell Dataset

Figure 4. Comparisons of GA and SA in CNS_RAT Cell Dataset
5.1 Comparison of Fitness Value

Comparing GA and modified GA for all the data sets the modified GA outperforms GA. In modified GA the diversity in the population is created when the stagnation occurs.

Table 2. Fitness Values of Standard Yeast Cell Dataset

| No. of Cluster | GA     | GA with SA |
|---------------|--------|------------|
| 3             | 1904.1 | 1830.33    |
| 4             | 1415.63| 1360.65    |
| 5             | 1132.34| 1086.96    |
| 6             | 939.12 | 840.54     |
| 7             | 806.93 | 789.53     |
| 8             | 704.52 | 672.87     |
Table 3. Fitness Values of CNS_Rat Dataset

| No. of Cluster | GA     | GA with SA |
|----------------|--------|------------|
| 3              | 829.13 | 786.75     |
| 4              | 627.97 | 610.06     |
| 5              | 393.03 | 384.49     |
| 6              | 274.37 | 268.42     |
| 7              | 182.92 | 174.65     |
| 8              | 94.27 | 89.14      |

Table 4. Fitness Values of Blood Monocytes Dataset

| No. of Cluster | GA                      | GA with SA          |
|----------------|-------------------------|---------------------|
| 3              | 3521078375.07           | 3412965732.81       |
| 4              | 2861386966.29           | 2761382354.29       |
| 5              | 2132268122.34           | 2091386966.29       |
| 6              | 1811542196.12           | 1711542196.12       |
| 7              | 1294295430.93           | 1094295430.93       |
| 8              | 186192109.52            | 166192109.52        |

6. Conclusion
DNA micro-arrays experiments are an important tool for monitoring and analyzing gene expression profiles of thousands of genes simultaneously. But, the large number of genes has greatly increased the challenges of comprehending and interpreting the resulting mass of data, which often contains millions of measurements. So, clustering techniques are used to tackle this challenge. Cluster analysis arranges the samples and genes. In this work the standard yeast, CNS_RAT, and blood monocytes data sets are used for clustering. Clustering in Genetic algorithm suffers from the premature suboptimal convergence or stagnation. To avoid stagnation in GA the SA is incorporated. It causes diversity in the population and avoids premature convergence. The modified GA outperforms simple GA technique for the datasets.

References
[1] S. Bandyopadhyay and U. Maulik, 2002, “Genetic clustering for automatic evolution of clusters and application to image classification”, Pattern Recognition, Vol.35, No.6, pp.1197-1208.
[2] Bai, G. Xiuguang and Y. Huaying, 2009, “A chaos genetic simulated annealing hybrid algorithm,” Grey System Intelligent Server, Vol.1, No.1, pp.1467-1479.
[3] C. Murthy and N. Chowdhury, 1996, “In search of optimal cluster using genetic algorithms”, Pattern Recognition, Vol.17, No.8, pp.285-832.
[4] Taishong and Huang, 2000, “Simulated annealing genetic hybrid algorithm and its application”, IEEE Intelligent Control Automation, Vol.1, No.1, pp.641-645.

[5] K.Y. Yeung, D.R.Haynor and W.L.Ruzzo, 2001, “Validating clustering for gene expression data”, BMC Bioinformatics, Vol.17, No.10, pp.309-318, 2001.

[6] Gat-Viks, R. Sharan and R. Shamir, 2003, “Scoring clustering solutions by their biological relevance”, BMC Bioinf., Vol.19, No.4, pp.2381-2389.

[7] Pandi M & Premalatha K, 2014, ‘An Advanced nelder mead simplex method for clustering gene expression data’, World Academy of Science, Engineering and Technology, International Journal of Computer, Information, Systems and Control Engineering, Vol.8, No.4, pp.588-596.

[8] D. G. Vito, G. Raffaele, L. B. Giosu, R. Alessandra, and S. Davide, 2005, “GenClust: A genetic algorithm for clustering gene expression data”, BMC Bioinformatics, Vol. 280, No.6, pp. 1-11.

[9] Pandi M & Premalatha K, 2016, ‘A Cuckoo Search with Differential Evolution for Clustering Microarray Gene Expression Data’, World Academy of Science, Engineering and Technology, International Journal of Computer, Information, Systems and Control Engineering, vol.10, no.2, pp.334-339.

[10] Nagi and Sajid, 2011, “Gene expression data clustering analysis: A survey,” IEEE Transaction on Knowledge and Data Engineering, Vol. 16, No.11, pp.1-12.

[11] J. P. Alava, M. I. Ponta, 2016, "A bi-objective model for gene clustering combining expression data and external biological knowledge", Computational Conference (CLEI) 2016 XLII Latin American, pp. 1-12.

[12] X. Yang, Feng Zhao, Tiejun Chen, 2018, "NLOS Identification for UWB Localization Based on Import Vector Machine", AEU – Int. J. Electron. Commun., Vol.11, No.4, pp.123-131.

[13] S. Acharya, S. Saha and N. Nikhil, 2017, "Unsupervised gene selection using biological knowledge: application in sample clustering", BMC Bioinformatics, Vol. 18, No.4, pp.34-39.

[14] Pandi M, Balamurugan R, Sadhasivam N, 2017 ‘Cancer Detection in Microarray Data Using a Modified Cat Swarm Optimization Clustering Approach’ Asian Pacific journal of Cancer Prevention, Vol. 18, No.12, pp. 3451 – 3455

[15] S. Monti, P. Tamayo, J. Mesiro and T. Golub 2003, “Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data”, Machine Learning, Vol.52, No. 4, pp.91-118.

[16] P. Hansen and B. Jaumard, 1997, “Cluster Analysis and Mathematical Programming” Proceedings of the 16th International Symposium on Mathematical Programming, Vol.79, No.3, pp.191-215.

[17] R. Priscilla and S. Swamynathan, 2011, “An Efficient Hybrid Clustering Approach of Microarray Gene Expression Data”, Proc. IEEE International Conference on Information and Communication Technology., pp.552-557.

[18] A.K. Jain, M.N. Murty, P.J Flynn, 1999, “Data Clustering: a Review” ACM Computational Survey, Vol.12, No.6, pp.264-323.

[19] Yi and Huawei, 2011, “A new optimization algorithm based on improved simulated annealing and genetic algorithm”, IEEE E -Bus. E –Government, Vol.1, No.1, pp.1-4.

[20] Guo, Ping. “Gene Expression Data Cluster Analysis”, IEEE Trans. Knowl. Data Eng., Vol.1, No.1, pp.99-102, 2009.

[21] Pandi M & Athul Jose, 2014, ‘An Efficient Shuffled Frog Leaping Algorithm for Clustering of Gene Expression Data’, International Journal of Computer Applications, vol.5, no.1, pp. 439-445.
[22] Pandi M & Premalatha K, 2015, ‘Clustering microarray gene expression data using enhanced harmony search’, International Journal of Bio-Inspired Computation (IJBIC), vol.7, no.5, pp. 296-306

[23] S. Bose, C. Das, A. Banerjee, M. Chattopadhyay and S. Chattopadhyay, "An Ensemble Filtering and Supervised Clustering based Informative Gene Selection Algorithm in Microarray Gene Expression Data," 2020 4th International Conference on Computational Intelligence and Networks (CINE), Kolkata, India, 2020, pp. 1-7, doi: 10.1109/CINE48825.2020.234391.

[24] Y. Yang, P. Yin, Z. Luo, W. Gu, R. Chen and Q. Wu, 2019, "Informative Feature Clustering and Selection for Gene Expression Data," in IEEE Access, vol. 7, pp. 169174-169184, doi: 10.1109/ACCESS.2019.2952548.

[25] H. Wang, P. Sham, T. Tong and H. Pang, 2020, "Pathway-Based Single-Cell RNA-Seq Classification, Clustering, and Construction of Gene-Gene Interactions Networks Using Random Forests," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 6, pp. 1814-1822, doi: 10.1109/JBHI.2019.2944865.

[26] X. Chen, J. Z. Huang, Q. Wu and M. Yang, 2019, "Subspace Weighting Co-Clustering of Gene Expression Data," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 2, pp. 352-364, doi: 10.1109/TCBB.2017.2705686.

[27] B. Belean, R. Gutt, C. Costea and O. Balacescu, 2020,"Microarray Image Analysis: From Image Processing Methods to Gene Expression Levels Estimation," in IEEE Access, vol. 8, pp. 159196-159205, doi: 10.1109/ACCESS.2020.3019844.

[28] Sadhasivam N, Balamurugan R, Pandi M, 2017 ‘Cancer Diagnosis Epigenomics Scientific Workflow Application Scheduling in Cloud Computing Environment Using Improved PSO Algorithm’ Asian Pacific Journal of Cancer Research. Vol. 19, No.1, pp. 243 – 246.