World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:10, No:2, 2016

A Cuckoo Search with Differential Evolution for Clustering Microarray Gene Expression Data

M. Pandi, K. Premalatha

***Abstract*—**A DNA microarray technology is a collection of microscopic DNA spots attached to a solid surface. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously or to genotype multiple regions of a genome. Elucidating the patterns hidden in gene expression data offers a tremendous opportunity for an enhanced understanding of functional genomics. However, the large number of genes and the complexity of biological networks greatly increase the challenges of comprehending and interpreting the resulting mass of data, which often consists of millions of measurements. It is handled by clustering which reveals the natural structures and identifying the interesting patterns in the underlying data. In this paper, gene based clustering in gene expression data is proposed using Cuckoo Search with Differential Evolution (CS-DE). The experiment results are analyzed with gene expression benchmark datasets. The results show that CS-DE outperforms CS in benchmark datasets. To find the validation of the clustering results, this work is tested with one internal and one external cluster validation indexes.

***Keywords*—**DNA, Microarray, genomics, Cuckoo Search, Differential Evolution, Gene expression data, Clustering.

I. Introduction

A microarray experiment evaluates a large number of DNA sequences consisting of genes, cDNA clones or expressed sequence tags under different conditions. These conditions may be a time based or tissue samples based. A gene expression data set from a micro-array experiment can be represented by a real-valued expression matrix [1]. In this matrix, rows represent expression pattern of genes, columns represent expression profile of samples or experimental conditions. Datasets are represented as set of genes G = {g_1, g_2, g_3… g_n}, where g_i represents i^{th} gene in the data set and w_{ij} represents expression profile of i^{th} gene at j^{th} samples/ conditions [2]. Fig. 1 represents the gene expression data matrix with n genes and m samples/conditions vector.

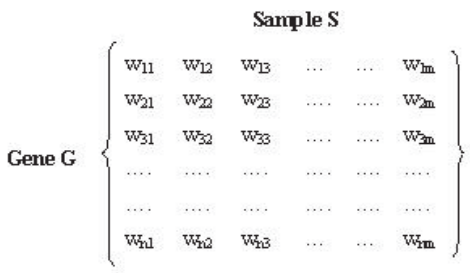

Fig. 1 Gene expression data matrix

M.Pandi is with the Bannari Amman Institute of Technology, Sathyamangalam, Erode, Tamilnadu, India (e-mail: mpandi123@gmail.com).

Clustering is a process of partitioning a dataset into separate groups based on any similarity measure and each cluster contains similar data items [3]. The objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). Gene expression profiling provides many ways to study about the gene expression patterns [1]. Co-expressed genes can be identified by the cluster analysis of gene expression data. The main step in analyzing gene expression data is to identify the group of genes that are having the similar expression pattern. Clustering of gene expression data is helpful to understand gene regulation, gene function and cellular processes [4].

Cuckoo Search is a meta-heuristic search method proposed by Xin-She et al. [5]. It mimics the parasitic breeding behaviour of cuckoos. Cuckoos do not breed their eggs. A cuckoo relies on other birds to host its egg. To do so, a cuckoo first selects a random nest and lays its eggs there. The host bird after finding an alien egg may destroy it or abandon the nest. In order to avoid the detection, cuckoos emulate the size, colour and shape of the eggs of hosting bird. The search for a nest by cuckoos follows the Levy flight distribution. The cuckoo search algorithm characterizes all these behaviours.

The rest of the paper is organized as follows: Section II describes the literature review on Gene expression data clustering. The overview CS is given in Section III. Section IV presents the CS-DE algorithm for gene expression data clustering. The experiment results are analyzed and demonstrated in Section V.

II. Literature Review

Fazel Famili et al. [6] proposed evaluation and optimization of clustering in gene expression data analysis. This work introduced new cluster quality method called stability. Tseng et al. [7] proposed a comparative review of gene clustering in expression profile. This paper compared simulated data with a real data. Vito Di Gesú et al. [8] proposed genetic algorithm for clustering of gene expression data called GenClust. The performance was evaluated based on real dataset and have used internal and external validation techniques. Ma et al. [9] proposed a novel evolutionary algorithm called evolutionary clustering (EvoCluster). It encodes an entire cluster grouping in a chromosome so that each gene in the chromosome encodes one cluster. Kustra R. et al. [10] introduced clustering expression data that permits integration of various biological data sources through combination of corresponding dissimilarity measures. This work reviews about genomic data fusion and validating results from clustering expression data. Kerr G et al. [11] conducted a review on techniques of

clustering gene expression data. This work mentions about the limitations and addresses them and provides a framework for the evaluation of clustering in gene expression analyses. Zhihua Du [12] proposed a new clustering algorithm for clustering gene expression data called PK means. This method incorporates Particle Pair Optimizer (PPO), K means and Fuzzy Kmeans for clustering which provide a more accurate result. Wei Liu et al. [13] proposed a novel methodology for finding the regulation on gene expression data. This work helps to find feature subset to build the classifier for gene expression data analysis. Principal component analysis was employed to construct the classifier. Rui Xu et al. [14] conducted a review on clustering algorithm in biomedical research. The work provides an overview of the status quo of clustering algorithms, to illustrate examples of biomedical applications based on cluster analysis, and to help biomedical researchers to select the most suitable clustering algorithms for their own applications.

Nagi et al. [15] had done a survey on gene expression data clustering analysis. This work mentions about various approaches to gene expression data analysis using clustering techniques. This work also discusses about the performance of various existing clustering algorithms under each of these approaches and proximity measures. Salome et al. [16] proposed an efficient clustering of gene expression data. This work introduced methods to improve the searching and the clustering performance in genomic data from commonly used clustering techniques. Jaskowiak et al. [17] investigated about the choice of proximity measures for the clustering of microarray data by evaluating the performance of 16 proximity measures in 52 data sets from time course and cancer experiments. This work mentions about commonly employed measures, such as Pearson, Spearman, and Euclidean distance.

III. Cuckoo Search

Cuckoo Search is an optimization technique developed by Yang and Deb based on the obligate brood parasitism of cuckoo species by laying their eggs in the nests of other host birds [18]. If a host bird discovers the eggs which are not its own, it will either throw these foreign eggs away or simply abandon its nest and build a new nest elsewhere. Each egg in a nest represents a solution, and a cuckoo egg represents a new solution. The better new solution (cuckoo) is replaced with a solution which is not so good in the nest. In the simplest form, each nest has one egg [19], [20].

A new solution was generated by Levy flight. The breeding behaviour of cuckoos can be summed up in three rules [5]: (i) each cuckoo lays one egg at a time and places it in a randomly selected nest; (ii) nests with high quality eggs would be carried to nest level production; (iii) the number of nests is fixed and the probability of discovery of cuckoo egg by the host bird is pa[0, 1]. The host bird either destroys the egg or abandons the nest and builds a new nest. The algorithm for CS is given below:

Pseudo Code for CS

```
Generate an initial population of n host nests;
while (t<MaxGeneration) or (stop criterion)
Get a cuckoo randomly (say, i) and replace its solution
By performing Levy flights;
Evaluate its fitness Fi
Choose a nest among n (say, j) randomly;
 if (Fi <  Fj)
        Replace j by the new solution;
 end if
 A fraction (pa) of the worse nests is abandoned and
 new ones are built;
 Keep the best solutions/nests;
 Rank the solutions/nests and find the current best;
 Pass the current best to the next generation;
end while
```

While generating the new solution x(t+1) for a cuckoo i, a Levy flight [5] is performed using (1):

$$x_i(t+1) = x_i(t) + \alpha \oplus \text{Levy}(\delta) \quad (1)$$

The symbol $\oplus$ is an entry-wise multiplication. Basically Levy flights provide a random walk while their random steps are drawn from a Levy distribution [5] for large steps as given in (2):

$$\text{Levy} \sim u = t^{-\delta} \quad (2)$$

This has an infinite variance with an infinite mean. Here, the consecutive jumps of a cuckoo essentially form a random walk process which obeys a power-law step-length distribution with a heavy tail.

IV. A Combined Cuckoo Search with Differential Evolution for Gene Expression Data Clustering

A. Problem Statement

The clustering problem is expressed as follows:

The set of M genes $G = \{G_1, G_2, \ldots, G_N\}$ is to be clustered. The genes are to be grouped into non-overlapping clusters $C = \{C_1, C_2, \ldots, C_K\}$ (C is known as a clustering), where K is the number of clusters, $C_1 \cup C_2 \cup \ldots \cup C_K = G$, $C_i \neq \varphi$, and $C_1 \cap C_2 = \varphi$ for i≠j.

Assuming $f : G \times G \rightarrow \Re^+$ is a measure of distance between genes. Clustering is the task of finding a partition $\{C_1, C_2, \ldots, C_K\}$ of G such that $\forall i, j \in \{1, \ldots, K\}, j \neq i, \forall x \in C_i : f(x, O_i) \geq f(x, O_j)$ where Oi is one cluster representative of cluster Ci.

The goal of clustering is stated as:

Given,

1. A set of genes $G = \{G_1, G_2, \ldots, G_N\}$,
2. A desired number of clusters K, and
3. An objective function or fitness function that evaluates the quality of a clustering, the system has to compute an assignment $g : G \rightarrow \{1, 2, \ldots, K\}$ and minimizes the objective function.

The proposed work applies global searching strategies for identifying optimal clusters in the exhaustive search space.

Typical objective function in clustering formalizes the goal of achieving high intra-cluster similarity, where genes within a cluster are similar, and low inter-cluster similarity, where genes from different clusters are dissimilar.

This is an internal criterion for the quality of a clustering. It is formulated by minimizing a formal objective function Mean Squared Error (MSE) distortion.

$$MSE(P) = \sum_{i=1}^{N} \left\| G_i - C_{p(i)} \right\|^2 \quad (3)$$

where N is the number of Genes; $G = \{G_1, G_2, \ldots, G_N\}$ is a set of N gene samples; P = { p(i) | i = 1,…N } is class label of G; C = { cj | j = 1,…, K} are K cluster centroids.

B. Egg Representation

Each egg is represented as candidate solution for the problem. The proposed work represents the whole partition of the genes in an egg of length N, where N is the size of the gene expression data. Each position in the cuckoo's egg is a label where the gene belongs to. In particular, if the number of cluster is K, each note value of egg is an integer value between 1 to K clusters. An example of egg representation is given in Fig. 2. The egg represents G1 is present in cluster #1, G2 is present in cluster #2, G3 is present in cluster #1 and so on.

$$[1 \quad 2 \quad 1 \quad 3 \quad \ldots \quad 2]$$

Fig. 2 Cluster representation

At the initial stage, the random number is generated between 0 and 1 and K is the number of clusters. Let v be the generated random number then the cluster value v' is

$$v' = \text{int}(vK) + 1 \quad (4)$$

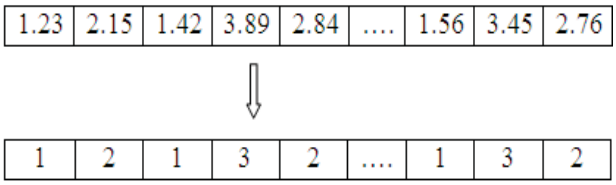

Fig. 3 Representation of egg for clustering

C. Combined Cuckoo Search with Differential Evolution

In the proposed work Differential Evolution (DE) algorithm is combined with conventional CS to cluster the gene expression data. Here, the fractions of worst nests are destroyed and new eggs for the nests are generated by using agent position generation of Differential Evolution. The global search area is enhanced through DE. The three benefits of DE are given below:

1) Finding the true global minimum regardless of the initial parameter values
2) Fast convergence
3) Using few control parameters.

Pseudo Code for CS-DE

```
Generate an initial population of n host nests;
  While (t<MaxGeneration) or (stop criterion)
    Get a cuckoo randomly (say, i) and replace its
     solution by performing using performing Levy
     flights;
        Evaluate its fitness Fi
          Choose a nest among n (say, j) randomly;
              if (Fi <  Fj)
                 Replace j by the new solution;
              end if
          A fraction (pa) of the worse nests is abandoned and
          new ones are built by Differential evolution;
        Keep the best solutions/nests;
          Rank the solutions/nests and find the current best;
          Pass the current best to the next generation;
          end while
```

Pseudo Code for Differential Evolution

1. Initialize the random solution x_i
2. Calculate the objective function value $f(x_i)$ for all x_i.
3. Select three points x_{r1}, x_{r2} and x_{r3} from population and generate perturbed individual using
   $v_i = x_{r1} + F(x_{r2} - x_{r3})$
4. Recombine each target vector x_i with perturbed individual generated
   $$u_i = \begin{cases} v_i & \text{if rand}(0,1) < P_c \\ x_i & \text{otherwise} \end{cases}$$
5. Calculate the objective function value for u_i.
6. Choose better of the two function value at target and trial point u_i and x_i for next generation.
7. Check whether convergence criterion is met if yes then stop; otherwise got step 3

V. Experimental Analysis

A. Datasets

The experiments are conducted on two well–known preprocessed gene expression datasets namely Yeast Cell Cycle (YCC) and Pheripheral Blood Monocytes (PBM). The YCC data set is part of that studied by [21]. The complete data set contains the expression levels of roughly 6000 yeast ORFs over 79 conditions. This preprocessed data set is consisting of 698 genes and 72 conditions. Next, Pheripheral Blood Monocytes data set was used by [22] to test their clustering algorithm. It contains 2329 cDNAs with a fingerprint of 139 oligos. This of data matrix gives 2329 genes and 139 conditions.

B. Validation Index Measures

One of the most important issues in cluster analysis is the evaluation of clustering results to find the partitioning that best fits the underlying data. This is the main subject of cluster validity. Here two measures are taken to validate the clustering results namely Figure of Merit (FOM) and Adjusted Rand (AR).

C. Figure of Merit

It is an internal measure used for this research work. For a given data set, let R denote the raw data matrix. Assume that R has dimension n×m, which means each row corresponds to a gene and each column corresponds to an experimental condition. Assume that a clustering algorithm is given the raw matrix R with column e excluded. Assume also that, with that

reduced data set, the algorithm produces k clusters C_0... C_{k-1}. Let R(g,e) be the expression level of gene g and mi(e) be the average expression level of condition e for genes in cluster C_i. The 2-norm FOM with respect to k clusters and condition e is defined as:

$$FOM(e,k) = \sqrt{\frac{1}{n}\sum_{i=0}^{k-1}\sum_{x\in Ci}\left(R(x,e) - mi(e)\right)^2} \quad (5)$$

Notice that FOM (e, k) is essentially a root mean square deviation. The aggregate 2-norm FOM for k clusters is then:

$$FOM(k) = \sum_{e=1}^{m} FOM(e,k) \quad (6)$$

D.Adjusted Rand Index

The expected value of the Rand Index of two random partitions does not take a constant value (e.g. zero). Thus Hubert and Arabie proposed an adjustment [23] which assumes a generalized hypergeometric distribution as null hypothesis: the two clusterings are drawn randomly with a fixed number of clusters and a fixed number of elements in each cluster (the number of clusters in the two clusterings need not be the same). Then the adjusted Rand Index is the (normalized) difference of the Rand Index and its expected value under the null hypothesis. It is defined as follows [24]:

$$R_{adj}(C,C') = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}\binom{m_{i,j}}{2} - t3}{\frac{1}{2}(t1+t2) - t3} \quad (7)$$

This index has expected value zero for independent clusterings and maximum value 1 (for identical clusterings). The significance of this measure has to be put into question because of the strong assumptions it makes on the distribution. Meila [7] notes, that some pairs of clustering may result in negative index values.

TABLE I
PARAMETERS AND THEIR VALUES FOR BENCHMARK DATASETS

Parameter	Value
Number of nests	50
Number of iterations	200
p_a	0.3
α	1
δ	1.5
Number of clusters	3 to 18
P_c	0.5

Figs. 4 and 5 correspondingly show the results obtained from CS and CS-DE for YCC and PBM datasets. In order to evaluate the performance of proposed CS-DE method, it has been applied for two publicly available real life gene expression data sets namely Yeast Cell Cycle (YCC) and Pheripheral Blood Monocytes (PBM). The results show that the proposed CS-DE algorithms outperform existing CS method in both gene expression data sets. Figs. 6-9 show that results obtained by the proposed technique are also compared with GenClust random, Min kmeans-random, Max kmeans-random, Cast, Kmeans-Avlink, Avlink and GenClust-Avlink [8]. Obtained clustering results are verified after conducting several statistical and biological significance tests. The results reveal that for both datasets the proposed methods attain the maximum Figure of Merit (FOM) and minimum Adjusted Rand (AR) index values.

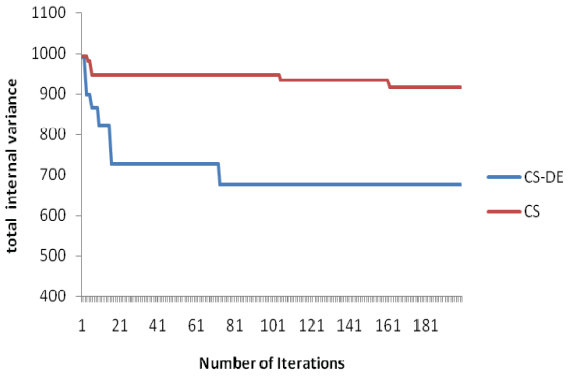

Fig. 4 Convergence of CS-DE and CS on YCC dataset for 5 clusters

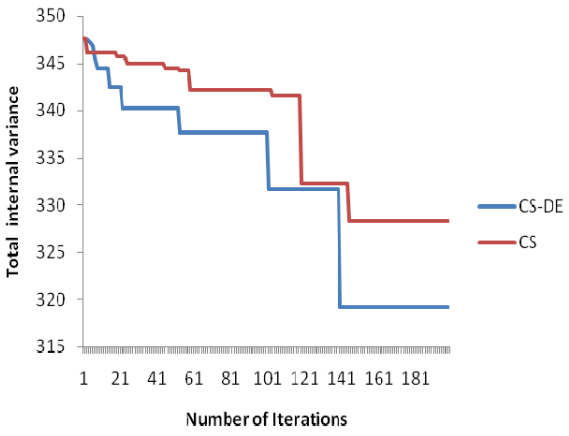

Fig. 5 Convergence of CS-DE and CS on PBM dataset for 5 clusters

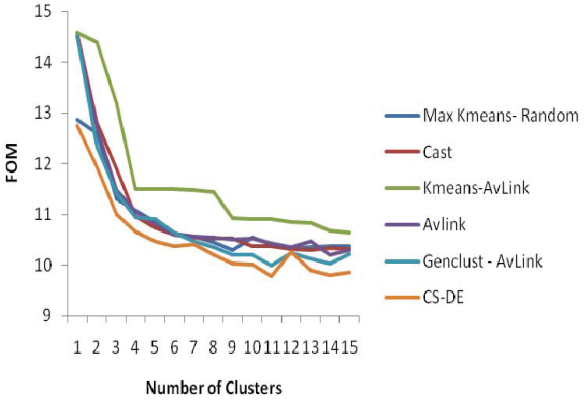

Fig. 6 Plot of number of clusters versus FOM index on YCC

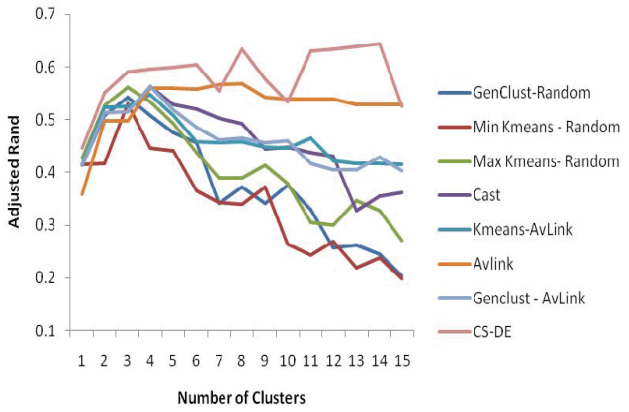

Fig. 7 Plot of number of clusters versus AR index on YCC

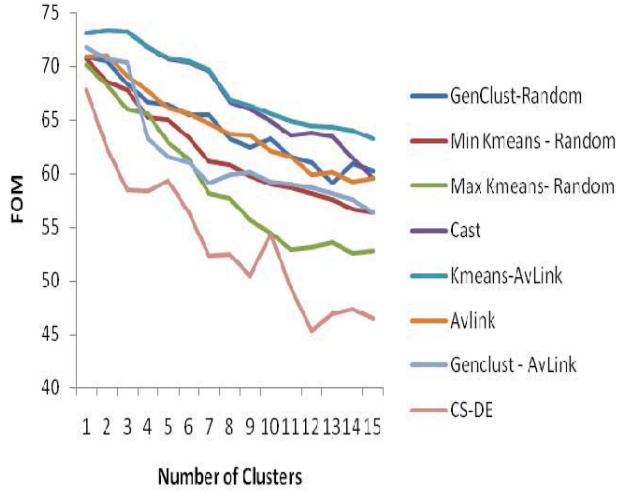

Fig. 8 Plot of number of clusters versus FOM index on PBM

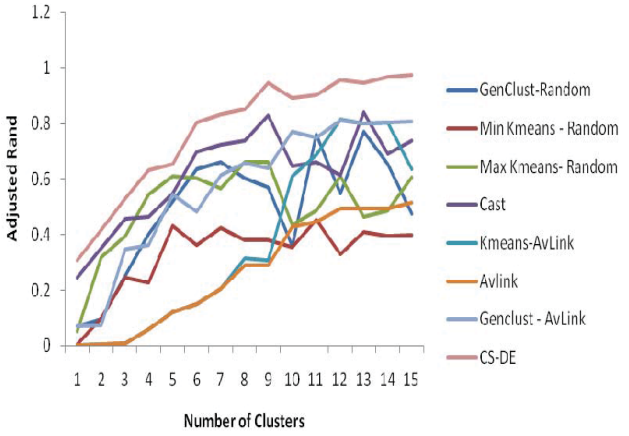

Fig. 9 Plot of number of clusters versus AR index on PBM

VI. Conclusion

Microarrays are useful to simultaneously monitor the expression profiles of thousands of genes under various experimental conditions. Identification of gene cluster is the main goal in gene expression data analysis and is an important task in bioinformatics research. In this work the gene expression data are clustered using CS and CS-DE. To avoid stagnation in CS, it is combined with DE. Best solutions in each nest are calculated and ranked. The nests with worst solutions are destroyed and are replace by DE. The performance of CS and CS-DE is analyzed with two gene expression benchmark data sets. The results show that CS-DE outperforms CS in both benchmarks datasets.

References

[1] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Array", *In Proc. of the Natl. Acad. Sci. U.S.A.,* Vol. 96, No. 12, pp. 6745-6750, 1999.

[2] C. Ding, "Analysis of Gene Expression Profiles: Class Discovery and Leaf Ordering", In *Proc. of the Int. Conf. Comput. Mol. Biol. (RECOMB)*, Berlin, Germany, pp. 27-136, 2002.

[3] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data Clustering: A Review", *ACM Comput. Surv.*, Vol. 31, pp. 264-323, 1999.

[4] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering Gene Expression Patterns", J. Comput. Biol., Vol. 6, No 3, pp. 281-297, 1999.

[5] X.S. Yang, & S. Deb, "Cuckoo search via Levy flights" *Proc. of World Congress on Nat. & Biologically Inspired Comput.*, pp. 210 – 214, 2009.

[6] F. Fazel, L. Ganming, L. Ziying, "Evaluation and optimization of clustering in gene expression data analysis", *BMC Bioinf.,* Vol. 20, No. 10, pp. 1535-1545, 2004.

[7] L. Nazareth, P. Tseng, "Gilding the Lily: A Variant of the Nelder-Mead Algorithm Based on Golden-Section Search", *Comput. Optim. Appl*, Vol. 22, no. 1, pp. 133–144, 2002.

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

[9] P.C.H. Ma, K.C.C. Chan, X. Yao, and D.K.Y. Chiu, "An evolutionary clustering algorithm for gene expression microarray data analysis", *IEEE Trans. Evol. Comput.,* Vol. 10, No. 3, pp. 296-314, 2006.

[10] R. Kustra, "A factor analysis model for functional genomics", *BMC Bioinf.,* Vol. 216, No. 7, pp. 1-13, 2006.

[11] G. Kerr, H. J. Ruskin, M. Crane, and P. Doolan, "Techniques for clustering gene expression data", *Comput. Biol. Med.,* Vol. 38, pp. 283–293, 2007.

[12] D. Zhihua, W. Yiwei, J. Zhen, "PK-means: A new algorithm for gene clustering", *Comput. Biol. Chem.,* Vol. 32, pp.243-247, 2008.

[13] L. Wei, B. Wang, G. Jarka, M. Elaine, and Z. Jian, "A novel methodology for finding the regulation on gene expression data", *Proc. Nat. Sci.*, Vol. 19, pp. 267-272, 2009.

[14] Rui Xu and D.C. Wunsch," Clustering Algorithms in Biomedical Research: A *Review", IEEE Rev. Biomed. Eng.,* Vol. 3, pp. 120 – 154, 2010.

[15] Sajid Nagi, D.K. Bhattacharyya, and J.K. Kalita, "Subspace Clustering in Gene Expression Data Analysis: A Survey, in Machine Intelligence: Recent Advances", *Narosa Publ., Delhi*, pp. 211-219, 2011.

[16] J. Jacinth Salome and R.M. Suresh, "Efficient Clustering for Gene Expression Data", *Int. J. Comput. Appl.,* Vol. 47, pp. 30-35, 2012.

[17] P. A. Jaskowiak and R.J.G.B Campello, "Comparing correlation coefficients as dissimilarity measures for cancer classification in gene expression data", *Proc. Braz. Symp. Bioinf. Brasilia. Braz*, pp. 1-8, 2011.

[18] N. Arulanand, S. Subramanian and K. Premalatha "An Enhanced Cuckoo Search for Optimization of Bloom Filter in Spam Filtering," *Global J. comp. Scie. Tech.,* vol. 12, no. 1, Jan. 2012.

[19] N. Arulanand, S. Subramanian and K. Premalatha "A Comparison study of cuckoo-bat search for Optimization of Bloom Filter in Spam Filtering," *Int. J. Bio-Inspired Comput.,* vol. 4, no. 2, pp.89-99, June 2012.

[20] X.S. Yang, and S. Deb, "Engineering optimisation by Cuckoo search" *Int. J. Math. Modeil. Numer. optim.,* vol. 1, no. 4, pp. 330-343, Dec. 2010.

[21] P. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and Futcher, (1998) "Comprehensive identification of cell cycle regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization", Mol. Biol. Cell, Vol. 9, pp - 3273-3297, 1998.

[22] P. Gray, W.E. Hart, L. Painton, C. Phillips, M. Trahan, and J. Wagner, "A Survey of Global Optimization Methods", *Tech. Rep., Sandia Nat. Lab*, 2000.

[23] L. Hubert, P. Arabie, P. "Comparing partitions" *Journal of Classification*, 2:193–218, 1985.
[24] Kuncheva, I. Ludmila, Hadjitodorov, T. Stefan, "Using Diversity in Cluster Ensembles" *IEEE SMC Int. Conf. on Sys.*, pp. 345-353, 2004.