RESEARCH ARTICLE

Cancer Detection in Microarray Data Using a Modified Cat Swarm Optimization Clustering Approach

Pandi M*, Balamurugan R, Sadhasivam N

Abstract

Objective: A better understanding of functional genomics can be obtained by extracting patterns hidden in gene expression data. This could have paramount implications for cancer diagnosis, gene treatments and other domains. Clustering may reveal natural structures and identify interesting patterns in underlying data. The main objective of this research was to derive a heuristic approach to detection of highly co-expressed genes related to cancer from gene expression data with minimum Mean Squared Error (MSE). **Methods:** A modified CSO algorithm using Harmony Search (MCSO-HS) for clustering cancer gene expression data was applied. Experiment results are analyzed using two cancer gene expression benchmark datasets, namely for leukaemia and for breast cancer. **Result:** The results indicated MCSO-HS to be better than HS and CSO, 13% and 9% with the leukaemia dataset. For breast cancer dataset improvement was by 22% and 17%, respectively, in terms of MSE. **Conclusion:** The results showed MCSO-HS to outperform HS and CSO with both benchmark datasets. To validate the clustering results, this work was tested with internal and external cluster validation indices. Also this work points to biological validation of clusters with gene ontology in terms of function, process and component.

Keywords: Cancer diagnosis- gene treatments- genomics- gene expression data- Cat Swarm Optimization

Asian Pac J Cancer Prev, **18 (12)**, 3451-3455

Introduction

DNA microarray enables the researchers to analyze the expression of many genes in a single reaction quickly and in an efficient manner (Shalon et al., 1996). A typical DNA microarray analysis involves a multi-step procedure: Specific genes are represented by fabrication of microarrays which has properly designed oligonucleotides; hybridization of cDNA populations onto the microarray; scanning hybridization signals and image analysis; transformation and normalization of data; and analyzing data to detect differentially expressed genes as well as the sets of genes that are co regulated. The gene expression matrix is a processed data obtained after the normalization. Each row in the matrix corresponds to a particular gene and each column could either correspond to an experimental condition or a specific time point at which expression of the genes has been measured (Bilban et al., 2002; Smyth et al., 2003). The expression level for a gene across different experimental conditions is cumulatively called the gene expression profile, and the expression level of each gene under an experimental condition is cumulatively called the sample expression profile (Cho and Won 2003; Androulakis et al., 2009). An expression profile of a gene or an experimental condition is thought of as a vector and can be represented in vector space (Liu et al., 2017).

Gene expression profiling provides many ways to study about the gene expression patterns (Alon et al., 1999; Pandi and Premalatha 2015). 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 (Yu et al., 2017; Balamurugan et al., 2016) is helpful to understand gene regulation, gene function and cellular processes. While considering the case of gene expression data, the elements are genes. There is no previously defined class label for clustering. Clustering of gene expression data helps to understand gene functions and regulations network and assists in the diagnostics of disease conditions and effects of medical treatment.

In the case of partitional and hierarchical, the solutions may be local optimum or may not be necessarily the global solution. This makes worse when the solution space is very large.

The number of ways of sorting N objects into K groups is given by Liu (1968).

$$Q(N,K)=\frac{1}{K!}\sum_{i=0}^{K}(-1)^{i}\binom{K}{i}(K-i)^{N} \qquad (1)$$

For example, for Q(25,5) there are 2,436,684,974,110,751 ways of sorting 25 objects into 5 groups. If the number of clusters is unknown the objects

*Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Erode, India. *For Correspondence: mpandi123@gmail.com*

can be sorted $\sum_{k=1}^{K} o(N,K)$ ways. For 25 objects this is over 4×1018. Clearly, it is impractical for an algorithm to exhaustively search the solution space to find the optimal solution. Furthermore traditional clustering algorithms search relatively a less subset of the solution space. As a result, the probability of success of these methods is small and it requires for an algorithm with the potential to search large solution spaces effectively. Contrary to the localized searching of the traditional algorithm, the global optimization algorithm (Kang and Geem, 2004) performs a globalized search in the entire solution space.

Materials and Methods

Problem statement

The clustering problem is expressed as follows: The set of M genes G = {G_1, G_2, ...,G_N} is to be clustered. The genes are to be grouped into non-overlapping clusters C= {C_1, C_2, ..., 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, ..., 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 follows:

Given,

1. A set of genes G = {G_1, G_2, ...,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 →{1, 2, ...K} and maxi*m*izes the objective function.

The global maximization problem can be defined as follows (Paradalos et al., 2001): Given $f : S \rightarrow \Re$ where $S \subseteq \Re^N$ and *N* is the dimension of the search space S. Find $y \in S$ such that $f(y) \geq f(z), \forall z \in S$

The variable y is called the global maximizer of f and f(y) is called the global maximum. The process of finding the global optimal solution is known as global optimization (Gray et al., 1997). A true global optimization algorithm will find y regardless of the selected starting point $z_o \in S$

(Van den Bergh and Engelbrecht, 2002). The variable *yL* is called the local maximizer of L because *f (yL)* is the largest value within a local neighborhood, L. Mathematically speaking, the variable *yL* is a local maximizer of the region L if $f(y_L) \geq f(z), \forall z \in L$ where $L \subset S$. For clustering, two measures of cluster quality are used. One type of measure allows comparing different sets of clusters without reference to external knowledge and is called an internal quality measure. The other type of measures evaluates how well the clustering is working by comparing the groups produced by the clustering techniques to known classes. This type of measure is called as external quality measure. Internal criterion function focuses on producing a clustering solution that optimizes a particular criterion function that is defined over the genes. These genes are part of each cluster and do not take into account the genes assigned to different clusters. The proposed work applies the 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.

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

where

N is the number of Genes;

G = {G_1 ,G_2 , ,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

A Combined Cat Swarm Optimization with Harmony Search for Cancer Gene Expression Data Clustering

The performances of the metaheuristic algorithms are mainly dependent on two properties of the algorithm: diversification and intensification, also mentioned as exploration and exploitation. Like other swarm optimization techniques, the philosophy of CSO is "to follow the leader." In CSO, the seeking mode provides local search whereas the tracing mode searches globally. If the fitness of the current best cat is improved by some means, the convergence of CSO would be improved. However, it will be better if the current best cat is allowed to search locally. It is therefore suggested that the current best cat is mandatorily selected for the seeking mode. If it happens, the current best cat may upgrade its fitness, and later on this will positively influence the movement of all the cats going through the tracing mode. Moreover, it can also avoid possible local trappings. Although the basic CSO algorithm demonstrates good local optimal search ability in optimization problems, but it has the problem of premature convergence.

Therefore, the CSO is improved by balanced intensification and diversification. In the proposed work the Cat optimization algorithm is combined with conventional harmony search to cluster the gene expression data. Recently, nature-inspired algorithms are well capable of solving numerical optimization problems more efficiently. HS algorithm has been successfully applied to a wide range of applications such as structural optimization, design optimization of water distribution networks, and vehicle routing. Like evolutionary algorithms, it generates a population of candidate solutions and then iteratively improves on the candidate population by adding and removing individual candidates. Unlike most evolutionary algorithms, it does not update the entire solution population at every iteration but only changes one individual at a time. Here, If the solution stagnates for designated number of iterations then run harmony search for few number of iterations for seeking mode cats. Consider cats as harmonies to precede harmony search. This maintains critical diversity in the population for more

iterations making the early convergence to local optima much less likely.

Procedure of MCSO-HS

1. Create N cats in the process.
2. Initialize the velocities of each cat randomly and decide the mixture ratio to choose number of cats in seeking mode and tracing mode
3. Calculate the fitness value of each cat and keep the best cat in memory.
4. Move the cats according to their flags, if catk is in seeking mode, apply the cat to the seeking mode process, otherwise apply it to the tracing mode process.
5. Re-pick number of cats and set them into tracing mode according to MR, then set the other cats into seeking mode.
6. If the solution stagnates for designated number of iterations then run harmony search for few number of iterations for seeking mode cats. Consider cats as harmonies to proceed harmony search.
7. If the number of iterations completed, terminate the process, and otherwise repeat Step3 to Step5.

Results

Datasets

The experiments are conducted on two well–known preprocessed gene expression datasets namely Leukaemia Cancer, Breast Cancer. The data set is collected from the broad institute database (http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi/).

Figures 1 and 2 correspondingly show the fitness values obtained from HS, CSO and MCSO-HS for Leukaemia Cancer and Breast Cancer datasets. The results show that the proposed MCSO-HS algorithms outperform existing CSO and HS methods in both Cancer gene expression data sets. Figures 3 and 4 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 (Vito Di Gesú et al., 2005). Obtained clustering results are verified after conducting several statistical and biological significance tests. The results reveal that for both datasets the proposed methods attain

Table 1. Parameter and its Value for Benchmark Datasets

Parameter	Value
No. of Cats (N)	100
SMP	20
SRD	10
CDC	20
SPC	0 or 1
Harmony memory considering rate (HMCR)	0.9
Pitch Adjustment Rate (PAR)	0.3
Harmony memory size(HMS)	100
Number of iteration(NI)	200
Cluster size	1 to 15

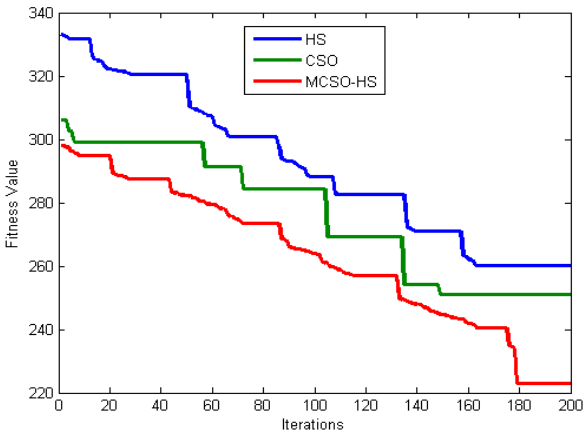

Figure 1. Convergence of MCSO-HS and CSO on LeuKaemia Cancer Dataset for 5 Clusters

the maximum Figure of Merit (FOM) and minimum Adjusted Rand (AR) index values.

Comparative Analysis of Adjusted Rand and FOM Validation Index

Table 2 depicts the internal and external validation index results for Leukaemia and Breast Cancer gene expression datasets of the proposed method that is compared with the well known existing methods. The high value of adjust Rand index shows that the cluster has co-expressed genes while a low value of FOM indicates that it is highly correlated. In Leukaemia Cancer dataset, GenClust-Random seems to be related to K-means-Random. Indeed, the relation is quite strong for FOM. As for the adjusted Rand index, the minimum values of the two algorithms are in many circumstances quite close. Such a relation is less pronounced for the maximum values, where sometimes one of the two algorithms dominates the other. Next, Cast and CSO outperform the Avlink in FOM index. Compared with all the other methods CSO and proposed MCSO-HS algorithm returns significant cluster with minimum FOM and maximum adjusted Rand index. MCSO-HS is much better than CSO in external measure on the both datasets.

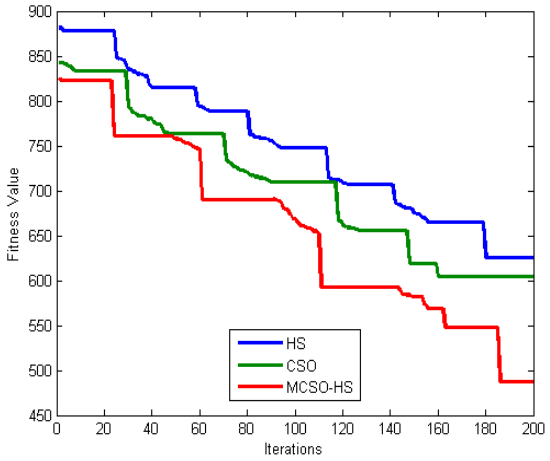

Figure 2. Convergence of MCSO-HS and CSO on Breast Cancer Dataset for 5 Clusters

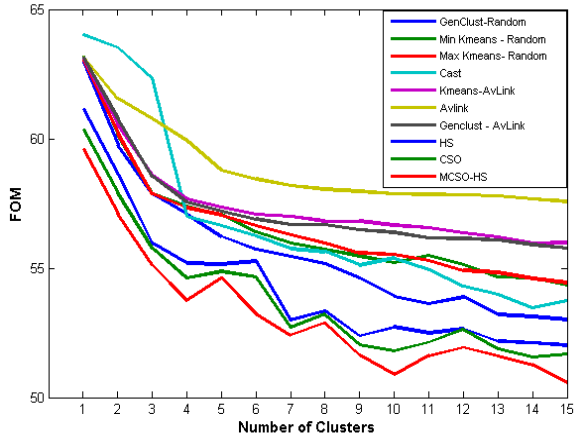

Figure 3. Plot of Number of Clusters Versus FOM Index on Leukaemia Cancer Dataset

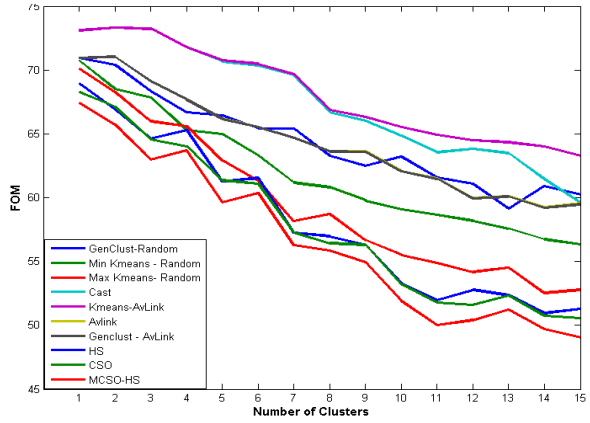

Figure 4. Plot of Number of Clusters Versus FOM Index on Breast Cancer Dataset

Biological annotation for Breast Cancer data using GOTermFinder toolbox

In order to identify the biological annotations for the clusters, we use GOTermFinder which is a tool available in the Saccharomyces Genome Database (SGD) (http://www.yeastgenome.org/cgi-bin/GO/goTerm Finder.pl). GOTermFinder is designed to search for the significant shared GO terms of the groups of genes and provides users with the means to identify the characteristics that the genes may have in common. Table 3 lists the significant shared GO terms used to describe the set of genes in each cluster for the process, function and component ontologies. Only the most significant terms are shown. For example to the cluster 3, the genes are mainly involved in binding activity. The tuple (n = 41, p = 1.8×10^{-7}) represents that out of 83 genes in cluster 3, 41genes belong to binding activity function, and the statistical significance is given by the p-value of p = 1.8×10^{-7}. Figure 5 shows the biological

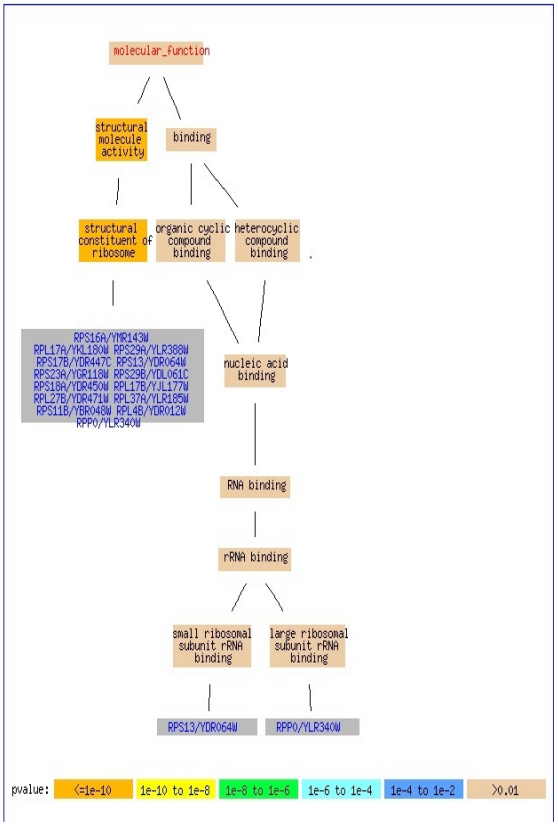

Figure 5. Gene Ontology Biological Process of Breast Cancer Data (10 Genes)

Table 2. Comparative Analysis on Leukaemia and Breast Cancer Data

Method	Leukaemia Cancer (Fifth Clusters)		Breast Cancer (Fifteenth Clusters)	
	Adjusted Rand	FOM	Adjusted Rand	FOM
Genclust random	0.47	57.05	0.51	57.49
Min kmeans –random	0.44	57.05	0.38	55.73
Max kmeans-random	0.49	57.05	0.51	55.73
Cast	0.529	56.66	0.68	50.21
Kmeans-Avlink	0.508	57.36	0.62	59.49
Avlink	0.559	58.78	0.52	62.27
GenClust-Avlink	0.518	57.21	0.8	59.33
HS	0.671	56.35	0.83	49.21
CSO	0.78	55.94	0.85	47.94
MCSO-HS	0.891	54.03	0.92	45.86

Table 3. Significant GO Terms for Three Clusters on Breast Cancer Data

Cluster No.	No. of Genes	Process	Function	Component
3	83	cell cycle process (n=38, p=1.9×10^{-8})	binding activity (n= 41, p=1.8×10^{-7})	intracellular organelle (n=62, p=3.3×10^{-6})
4	78	mitotic cell cycle process (n=42, p=6.5×10^{-7})	hydrolase activity (n=48, p=3.2×10^{-6})	cell part (n=58, p=1.9×10^{-4})
5	131	single-organism process (n=93, p=1.1×10^{-3})	transferase activity (n=81, p=1.8×10^{-2})	intracellular part (n=1,344, p=2.9×10^{-1})

network of the cluster for ten genes, the false discovery rate (FDR) is very low (0.0006) and it is zero in many occasions. Further the corresponding p-value is very small ($p = 0.00465$) which shows that there is a very less probability to obtain the gene cluster in random. These results mean that the proposed MCSO-HS clustering approach can find biologically meaningful clusters.

Discussion

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 cancer gene expression data analysis and is an important task in bioinformatics research. The better understanding of functional genomics is obtained by extracting the patterns hidden in gene expression data. It is handled by clustering which reveals natural structures and identify interesting patterns in the underlying data. In the proposed work clustering gene expression data is done through Modified CSO algorithm to identify the Cancer detected gene expression data. The Modified CSO method is achieved by the hybridization of HS with CSO and gives better results compared with existing methods. The performance of CSO and MCSO-HS is analyzed with two cancer gene expression benchmark data sets.

Statement conflict of Interest

The authors whose names are listed immediately below certify that they have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Author names:

1. Pandi M
2. Balamurugan R
3. Sadhasivam N

References

Alon U, Barkai N, Notterman DA, et al (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide array. *Proc Natl Acad Sci U S A*, **96**, 6745-50.

Balamurugan R, Natarajan AM, Premalatha K (2016). A modified harmony search method for biclustering microarray gene expression data. *Int J Data Min Bioinf*, **16**, 269-9.

BilBan M, Buehler LK, Head S, Desoye G, Quaranta V (2002). Normalizing DNA microarray data. *Mol Biol*, **4**, 57-4.

Cho SB, Won HH (2003). Machine learning in DNA microarray analysis for cancer classification. Proc Asia-Pac Bioinf Conf Bioinf, Adelaide, Australia, pp 189-8.

Gray P, Hart WE, Painton L, et al (1997). A survey of global optimization methods, Technical report, Sandia national laboratory, pp 323-45.

Kang SL, Geem ZW (2004). A new structural optimization method based on the Harmony search Algorithm. *Comput Struct*, **82**, 781-8.

Liu D, Wang YL, Gao CH, et al (2017). Regularized non-negative matrix factorization for identifying differential genes and clustering samples: A survey. *IEEE/ACM Trans Comput Biol Bioinf*, **99**, 1-10.

Liu GL (1968). Introduction to combinatorial mathematics. McGraw Hill, New York, pp 220-43.

Pandi M, Premalatha K (2015). Clustering microarray gene expression data using enhanced harmony search. *Int J Bio-Inspired Comput*, **7**, 296-6.

Paradalos JHS, Rajasekaran PM, Reif JDP, Rolim D (2001). Handbook of randomized computing: vols. I and II. Comb Optim, 9, Kluwer Academic Publishers, pp 923-41.

Shalon D, Smith SJ, Brown PO (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res*, **6**, 639–5.

Smyth GK, Yang YH, Speed T (2003). Statistical issues in cDNA microarray data analysis. *Methods Mol Biol*, **224**, 111-6.

Tung T, Nguyen, Richard S, Nowakowski, Androulakis P (2009). OMICS. *A J Integr Biol*, **13**, 219-7.

Van den Bergh F, Engelbrecht AP (2004). A cooperative approach to particle swarm optimization. *IEEE Trans Evol Comput*, **10**, 225-9.

Vito Di G, Raffaele G, Giosu LB, Alessandra R, Davide S (2005). GenClust: A genetic algorithm for clustering gene expression data'. *BMC Bioinf*, **280**, 1-11.

Yu X, Yu G, Wang J (2017). Clustering cancer gene expression data by projective clustering ensemble. *PLoS One*, **12**, 101-5.