JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 34, 1617-1631 (2018)
DOI: 10.6688/JISE.201811_34(6).0014

TrioCuckoo: A Multi Objective Cuckoo Search Algorithm for Triclustering Microarray Gene Expression Data

P. SWATHYPRIYADHARSINI AND K. PREMALATHA
Department of Computer Science and Engineering
Bannari Amman Institute of Technology
Sathyamangalam, TamilNadu, 638402 India
E-mail: {swa.pspd@gmail.com; kpl_barath@yahoo.co.in}

Analyzing time series microarray dataset is a challenging task due to its three dimensional characteristic. Clustering techniques are applied to analyze gene expression data to extract group of genes under the tested samples based on a similarity measure. Biclustering appears as an evolution of clustering due to its ability to mine subgroups of genes and conditions from the data set, where the genes exhibit highly correlated patterns of behavior under certain experimental conditions. Triclustering contains a subset of genes that contains information related to the behavior of some genes from under some conditions over certain time periods. In this work,TrioCuckoo, a multi objective cuckoo search algorithm is proposed to extract co-expressed genes over samples and times with two different encoding representation of triclustering solution. TrioCuckoo is evaluated using two real life datasets such as the breast cancer and PGC-1 alpha time course datasets. The experimental analyses are conducted to identify the performance of the proposed work with existing triclustering approaches and Particle Swarm Optimization (PSO).The proposed work identifies the key genes which are involved in the breast cancer. The gene ontology, functional annotation and transcription factor binding site analysis are performed to establish the biological significance of genes belonging to the resultant cluster for Breast cancer.

Keywords: tricluster, cuckoo search, multi-objective optimization, gene ontology, breast cancer, microarray gene expression data, PSO, time course data analysis

1. INTRODUCTION

Microarray technology and other high-throughput methods are used to measure the expression values of thousands of genes over different samples or experimental conditions. The activity of all genes measured for a number of biological replicates at each time point is referred to as three-dimensional datasets. The time series datasets in microarray technology has been used to measure in a single experiment, the expression values of thousands of genes under a huge variety of experimental conditions across different time points. Due to its huge volume of data, several computational methods are needed to analyze such datasets.

Clustering is one of the unsupervised approaches to identify the coexpressed genes. Clustering algorithms aims to maximize similarity within the clusters as well as to minimize similarity between the clusters, based on a distance measure. The traditional clustering algorithms fail to find the group of genes that are similarly expressed over subset of experimental conditions. This problem is solved by biclustering algorithms [1]. A bicluster can be defined as a subset of genes that are coexpressed over a subset of experi-

Received April 10, 2017; revised June 19, 2017; accepted August 10, 2017.
Communicated by Tzung-Pei Hong.

mental conditions. The Triclustering deals with three dimensional (3D) datasets which is used to find group of genes coexpressed under subset experimental conditions across subset of time points.In this work, TrioCuckoo algorithm is used to extract the subset of genes that have similar expression values over a subset of samples in subset of time points with minimum MSR value. In the proposed work two different encoding schemes are used to represent the triclustering solution.

The rest of the paper is structured as follows. A review of the latest works on the triclustering can be found in Section 2. Section 3 presents the methodology of the TrioCuckoo optimization algorithm. Section 4 describes the triclustering, problem statement, proposed encoding representations and multi objective functions. In Section 5, the experimental results for the real life datasets are discussed and compared with the other algorithms and also its biological significances are analysed. Section 6 presents the conclusions.

2. RELATED WORK

Cheng and Church [1] proposed the first biclustering algorithm that was used to analyse gene expression datasets and it used a greedy search heuristic approach to retrieve largest possible bicluster having Mean Squared Residue (MSR) under a predefined threshold value δ (δ-bicluster). Feng *et al.* [2] proposed a time-frequency based full-space algorithm using a measure of functional correlation set between time course vectors of different genes. Zhao *et al.* [3] proposed the tricluster algorithm that aims to retrieve groups of genes that have similar expression profiles over a subset of samples during a subset of time points. It also provides set of metrics to assess the quality of the triclusters and tested the approach on real life microarray datasets [3]. Jiang *et al.* [4] proposed gTricluster algorithm to mine biologically meaningful coherent gene clusters using Spearman rank correlation similarity measurement and extended the clique search technique for the third dimension [4]. Yin *et al.* [5] has given a new definition of coherent cluster for time series gene expression data called ts-cluster. The ts-cluster algorithm is able to detect a significant amount of clusters of biological significance [5]. Hu and Bhatnagar (2010) presented a 3-Clustering algorithm that searches for meaningful combinations of biclusters in two related datasets. This algorithm extracts tricluster from two independent biclusters such that the standard deviations in each bicluster obey an upper bound and it has maximum overlap between the two biclusters.

Tchagang *et al.* [6] proposed a triclustering algorithm OPTricluster for mining short time series gene expression datasets. OPTricluster effectively mines time series gene expression data having approximately 3-8 time points and 2-5 samples. Thus, it mines only the short temporal datasets. According OPTricluster, genes belonging to a tricluster must have constant, coherent or order preserving expression patterns over a subset of samples during a subset of time points [6]. Kuo *et al.* [7] proposed a method with backward frequent itemsets to mine time delayed gene regulation patterns. The interactions between genes in any length of time delay is identified and the patterns obtained by the method are used for gene function prediction [7]. Bhar *et al.* [17] proposed δ-TRIMAX, a triclustering algorithm by introducing a novel Mean Squared Residue score (MSR) to mine a 3D gene expression dataset and that each tricluster must have an MSR score below a threshold δ [8]. This algorithm identifies the hub genes which are responsible for

the estrogen responsive cancers. Also the gene ontology and pathway analysis are done in order to process the genes properties. Aviles *et al.* [8] also proposed TriGen algorithm which implements genetic algorithm for mining triclusters in temporal gene expression data. This algorithm implements the genetic algorithm, an optimization technique in order to retrieve the triclusters [9]. Bhar *et al.* [9] proposed EMOA-δ-TRIMAX, multi-objective optimization algorithm by implementing genetic algorithm [10]. Liu *et al.* (2015) proposed fuzzy triclustering algorithm to mine triclusters based on the membership function for each dimension but it has computational efforts [11]. Guigoures *et al.* [11] also applied triclustering approach to track patterns in time-varying graphs [12].

3. MULTIOBJECTIVE TRIOCUCKOO ALGORITHM

Cuckoo Search (CS) is an optimization algorithm which is inspired from the breeding parasitism of cuckoo species. Some cuckoo species lay their eggs in the nest of other host birds in obligating its breeding parasitism [13]. If a host bird discovers the eggs which are not their own, it will either throw these foreign eggs away or simply abandon its nest and build a new nest elsewhere[14]. It initiates with number of nests in which each egg in the nest represents a solution. A new solution is generated by Lévy flight [15]. It aims to produce better solutions by replacing the worst solutions in the nest. Each cuckoo lays one egg at a time and throws down its egg in a randomly chosen nest. The best nests with good quality of eggs will be carried on to the next generation. The number of host nests is fixed, and a host can discover a foreign egg with a probability $p_a \in [0, 1]$. The host bird then throw away the egg or abandon the nest and also it finds the worst nest which is to be replaced.

CS is only applied on the two dimensional datasets so far for biclustering. In this work, TrioCuckoo subspace clustering is performed on three dimensional dataset using Cuckoo search. Trio refers to group of three dimensions such as genes, samples and time. Each tricluster has the genes that are coexpressed under certain samples and time points. TrioCuckoo as similar to CS uses a balanced combination of a local and global random walk. The local random walk is given in Eq. (1).

$$x_i(t+1) = x_i(t) + \alpha s \otimes H(p_a - \epsilon) \otimes (x_j(t) - x_k(t)) \tag{1}$$

Where $x_j(t)$ and $x_k(t)$ are two different solutions selected randomly by random permutation, $H(p_a - \epsilon)$ is a heavyside function, ϵ is a random number drawn from uniform distribution and s is the step size. The global random walk can be carried out using the Lévy flight. When generating new solutions $x_i(t+1)$ for a cuckoo subset matrix i, a Lévy flight is performed using the following Eqs. (2) and (3), which is the stochastic equation for random walk.

$$x_i(t+1) = x_i(t) + \alpha \oplus Lévy(s, \lambda) \tag{2}$$

$$Lévy(s, \lambda) = \frac{\lambda T(\lambda) \sin(\pi\lambda/2)}{\pi} \frac{1}{s^{1+\lambda}} (s \gg s_0 > 0) \tag{3}$$

The step size $\alpha > 0$ is related to the scale of the problem of interest but in most cases $\alpha = 1$ is maintained. $x_i(t+1)$ is the next location which depends on $x_i(t)$ is the current location and the second term in the equation is the transition probability. The symbol $\oplus$ is an entry-wise multiplication which is similar to those used in PSO, but the Lévy flight based random walk is more efficient to explore the search space through longer step length. 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.

Algorithm 1: Pseudo code for TrioCuckoo
Generate an initial population of *n* host nests representing genes, samples and time points;
while t < (*MaxGeneration*) or (*stop criterion*)
Get a cuckoo subset matrix of Triclusters randomly (i) and replace its solution by applying Lévy flights;
Evaluate its fitness F_i
Choose a nest among *n*(*j*) randomly;
if ($F_i < F_j$)
Replace *j* by the new solution;
end if
A fraction (p_a) of the worse nests is abandoned and new ones are built;
Keep the best solutions/Triclusters;
Rank the solutions and find the current best Tricluster;
Pass the current best Tricluster to the next generation;
end while

4. TRICLUSTERING MICROARRAY GENE EXPRESSION DATA

4.1 Triclustering

Traditional clustering algorithms work in the full dimensional space, which consider the value of each object in all the dimensions and try to group the similar objects together. Biclustering does not have such a strict constraint. Some points are similar in several dimensions then they will be clustered together in that subspace. Biclustering enables to identify the co-expression patterns of a subset of genes that might be relevant to a subset of the samples of interest. In addition to biclustering along the gene-sample dimensions, there has been a lot of interest in mining gene expression patterns across time. Hence, Triclustering finds the subset of genes that are similarly expressed across a subset of experimental conditions or samples over a subset of time points.

4.2 Problem Statement

For a given set of *n* genes $G = \{g_1, g_2, .., g_n\}$, a set of m biological conditions or samples $S = \{s_1, s_2, ..., s_m\}$ and a series of k time points $T = \{t_1, t_2, \ldots, t_k\}$, a *GST* microarray dataset is a real valued three dimensional $n \times m \times k$ cuboid, $D = G \times S \times T = \{d_{ijk}\}$. Each cell value d_{ijk} represents the expression level of gene g_i, in sample s_j at time point t_k. The existing work on the three dimensional data clustering investigates to find genes that

are coherent on a subset of the samples during the whole time series. Thus the extracted tricluster which has subset matrix of $G \times S \times T$ have the same genes which are restricted to the different samples and time points. Therefore, this work aims to find all the genes that are coherent on different samples and time points.

4.3 Encoding

Each egg in a nest is represented by a binary string with three parts. An egg encodes a possible tricluster. A time series gene expression dataset has G number of genes, S number of samples and T number of time points. Therefore a nest has the first m bits corresponding to the genes, the next n bits corresponding to the samples and the last k bits corresponding to the time points. Each string is represented by $m+n+k$ bits that have a value either 1 or 0. If the value is 1 then the corresponding gene or sample or time point is present in the tricluster. For example a gene expression dataset having 10 genes 6 samples and 4 time points, a string {10101110010101011011} represents that genes $\{g_1, g_3, g_5, g_6, g_7, g_{10}\}$, samples $\{s_2, s_4, s_6\}$ and time points $\{t_1, t_3, t_4\}$ are the members of thetricluster as shown in Fig. 1.

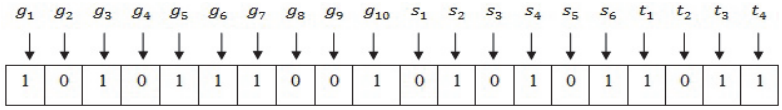


Fig. 1. Encoding representation-1 of an egg.

In the above method, the set of genes $\{g_1, g_3, g_5, g_6, g_7, g_{10}\}$ are presented for same set of samples $\{s_2, s_4, s_6\}$ at the same set of time points $\{t_1, t_3, t_4\}$. Alternatively the tricluster with subset of genes for different samples at different time points is also encoded. Fig. 2 shows tricluster encoding representation genes as rows, time periods as columns, and samples as depth representation and Fig. 3 shows tricluster encoding representation genes as rows, samples as columns, and time periods as depth time the coherent triclustering encoding representation of genes, samples and time points with respect to different samples and different time points respectively.

4.4 Fitness Function

Tricluster is given as $C[I, J, L] = [c_{ijl}]$ where $i \in I, j \in J$ and $l \in L$. The cuboid C represents subset of genes which have similar expression values over a subset of samples during subset of time points.

Mean Square Residue (MSR) of the tricluster can be modelled as

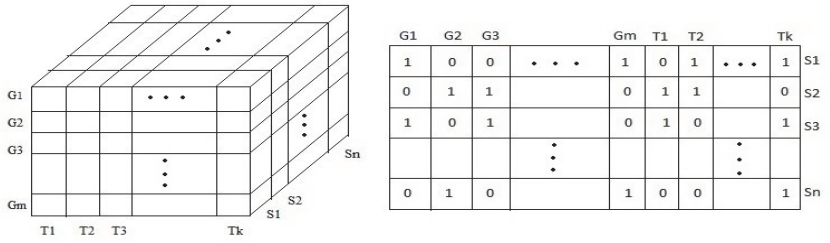


Fig. 2. Encoding representation-2 (Gene-Time-Sample) of an egg.

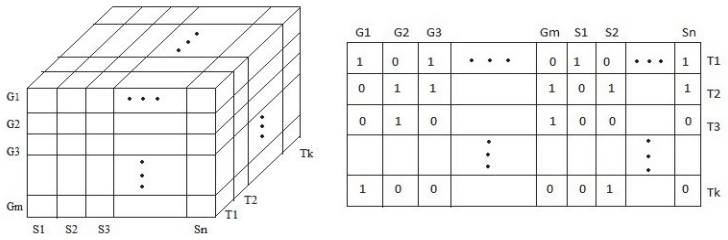


Fig. 3. Encoding Representation-3 (Gene-Sample-Time) of an egg.

$$MSR = \frac{\sum_{g \in G, s \in S, t \in T} r_{gst}^2}{|G| \times |S| \times |T|} \tag{4}$$

$$r_{gst} = TS_v(t,g,s) + M_{GS}(t) + M_{GT}(s) + M_{ST}(g) - M_{GT}(s,t) - M_S(g,t) - M_T(g,s) - M_{GST} \tag{5}$$

Where $M_{GS}(t)$ is the mean of genes under samples at a time point, $M_{GT}(s)$ is the mean of the genes over time under a sample, $M_{ST}(g)$ is the mean of a gene in time under the samples, $M_G(s, t)$ is the mean of the genes under a sample and a time point, $M_S(g, t)$ is the mean of the values of a gene at a time point under samples, $M_T(g, s)$ is the mean of a gene under a sample at all time points and M_{GST} is the mean value of all values in the tricluster.

The first objective is to calculate the MSR value of the tricluster which is given in Eqs. (4) and (5).

$$f_1 = MSR \tag{6}$$

The low MSR value denotes there is strong coherence in the tricluster. This includes only the trivial tricluster when there is no fluctuation. The row variance is calculated in order to include the non trivialtricluster. The second objective function is to calculate the row variance which is given in Eqs. (7) and (8) in which a_{ij} is the value of a gene in the tricluster, a_{ij} is the mean of ith row in tricluster for all j conditions and a_{ij} is the mean of ith row for all k time points.

$$f_2 = \frac{1}{|I|} \sum_{i \in I} \text{var}_i \tag{7}$$

$$\text{var}_i = \frac{1}{|J|} \sum_{j \in J} (a_{ij} - a_{iJ})^2 + \frac{1}{|K|} \sum_{j \in J} (a_{ik} - a_{iK}) \tag{8}$$

The third objective function is the volume of the tricluster which is given in the following Eq. (9).

$$f_3 = \frac{|I| \times |J| \times |K|}{|G| \times |S| \times |T|} \tag{9}$$

Where $(|I| \times |J| \times |K|)$ is the volume of the tricluster and $(|G| \times |S| \times |T|)$ is the volume of the dataset. The objective function is to be maximized in order to have increased size

of the tricluster.

The aim of this work is to find the triclusters which should have a lower MSR score and a higher variance and a larger volume of the tricluster. Thus the first objective function is to be minimized and the second and third objective function is to be maximized in order to accomplish the goals. Therefore the optimal solution of the objective function is different from each other. Pareto optimal solutions solve this problem by considering set of constraints to get the optimal solution [16]. It is based on the dominance criteria where a solution $x^{(1)}$ is said to dominate other solution $x^{(2)}$ if it holds the conditions such as the solution $x^{(1)}$ is no worse than $x^{(2)}$ in terms of all the objectives and the solution $x^{(1)}$ is strictly better than $x^{(2)}$ in at least one objective. The set of solutions which are not dominated by any others are called Pareto optimal front. Thus the solutions are selected based on the pareto optimal front. Fig. 4 shows the flowchart for the proposed work.

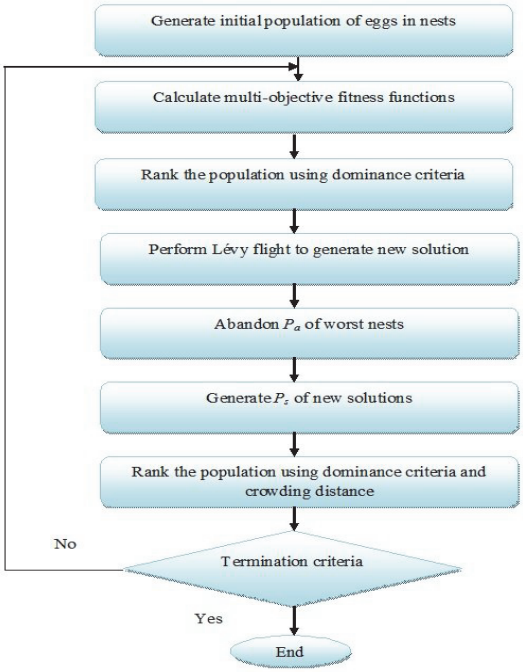


Fig. 4. Cuckoo search with multi objective function.

5. EXPERIMENTAL RESULTS

5.1 Datasets

The proposed method is implemented on two different real life datasets. First, the real life time series dataset GSE11324 is chosen from the Gene Expression Omnibus (GEO). It holds 54675 Affymetrix human genome U133 plus 2.0 probe ids, 3 samples and 4 time points (0, 3, 6, 12 hours). It aims at finding cis-regulatory sites responsible for conveying estrogen responses and to identify the cooperating transcription factors which

contribute to estrogen signaling in MCF7 breast cancer cells.

Second, the time course dataset PGC-1a oxidative phosphorylation (OXPHOS) gene expression data is chosen. It has the genes involved in OXPHOS exhibit reduced expression in skeletal muscle of diabetic and prediabetic humans. It holds 12490 genes, 4 samples and 4 time points (0, 1, 2, 3 days).

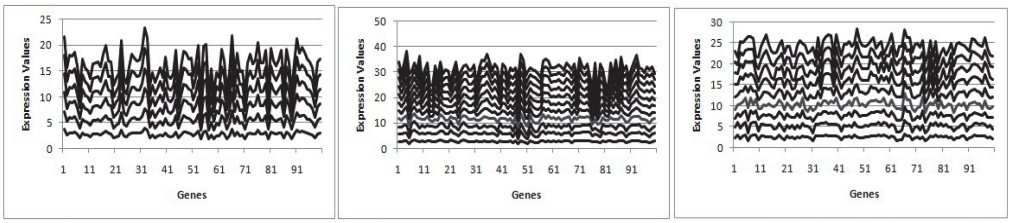
The TrioCuckoo and PSO algorithms are implemented in MATLAB version 7.14 (R2012a). The traditional CS algorithm [15] uses fixed value for Lévy distribution coefficient λ, probability of discovery rate of the eggs p_a and the step size α and the same values are assigned in TrioCuckoo. For implementing PSO, the same number of iterations and population size is used as in TrioCuckoo and $c_1 = 2.1$, $c_2 = 2.1$ and the inertia weight $\omega = 0.9$ is maintained for both the datasets [17].

Table 1.Parameters and values considered for TrioCuckoo.

Parameter	Value
Number of nest (n)	50
Discovery rate of alien eggs (p_a)	0.25
Step size (α)	1
Lévy distribution coefficient (λ)	1.5
Number of iterations	20

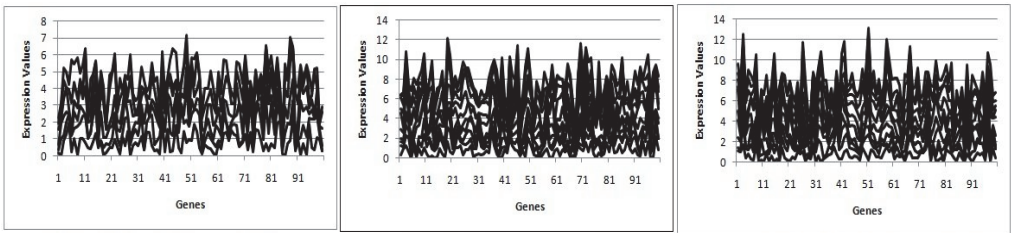
5.2 Performance Comparison

Figs. 5 (a)-(c) show the sample triclusters obtained from the proposed work using encoding representation-1, 2 and 3 respectively for the breast cancer dataset. Figs. 6 (a)-(c) show the sample triclusters obtained from the proposed work using encoding representation-1, 2 and 3 respectively for PGC-1 alpha time course dataset.



(a) Tricluster encoding representation-1. (b) Tricluster encoding representation-2. (c) Tricluster encoding representation-3.

Fig. 5. Sample Triclusters for breast cancer dataset.



(a) Tricluster encoding representation-1. (b) Tricluster encoding representation-2. (c) Tricluster encoding representation-3.

Fig. 6. Sample Triclusters for PGC-1 alpha time course dataset.

The performance of the proposed TrioCuckoo with different encoding representations is compared with PSO in terms of the fitness function. The MSR value of the TrioCuckoo is low when compared to PSO for both the datasets. Therefore, TrioCuckoo using encoding representations 2 and 3 performs better than others.

Table 2. Comparison with PSO.

Algorithm	MSR Value	
	Breast Cancer	PGC-1 alpha
TrioCuckoo using Encoding Representation – 1	0.000661	0.090521
TrioCuckoo using Encoding Representation – 2	0.000467	0.060725
TrioCuckoo using Encoding Representation – 3	0.000161	0.029356
PSO using Encoding Representation – 1	0.004924	0.109350
PSO using Encoding Representation – 2	0.000986	0.092593
PSO using Encoding Representation – 3	0.000798	0.063871

5.3 Comparison with Other Triclustering Algorithms

The performance of the proposed work with different representations is compared with other triclustering algorithms based on two validation indexes. The first measure is the Triclustering Quality Index (TQI) which is given in Eq. (10) [18].

$$TQI = \frac{MSR_i}{volume_i} \tag{10}$$

Where MSR_i is the mean squared residue of the tricluster i and $volume_i$ is the volume of the tricluster i. The volume of the ithtricluster is defined as ($|I_i| \times |J_j| \times |K_k|$) where $|I_i|$, $|J_j|$ and $|K_k|$ represent the number of genes, samples and time points of the ith tricluster. A lower TQI represents the better quality of the triclusters.

The second measure is the Statistical Difference from Background (SDB) score which signifies whether a set of n triclusters are statistically different from the background data matrix or not [18]. The SDB score is given in the Eq. (11).

$$SDB = \frac{1}{n}\sum_{i=1}^{n} \frac{MSR_I}{\frac{1}{r}\sum_{j=1}^{r} RMSR_j - MSR_i} \tag{11}$$

Where n is the total number of triclusters extracted by the algorithm. MSR_i represents the mean squared residue of the ith tricluster and $RMSR_j$ is the mean square residue of the jth random tricluster having the same number of genes, samples and time points as the ith resultant tricluster. The higher the value of the denominator denotes the better the quality of the resultant tricluster. Hence, lower SDB score signifies better performance of the algorithm. Table 3 shows the comparison of the performance of various algorithms in terms of SDB and TQI indexes.

5.4 Biological Significance

The biological significance of the genes which belongs to each of the triclusters is identified by performing Gene Ontology and KEGG pathway enrichment analysis and

Table 3. Performance comparison.

Algorithm	SDB	Average TQI
Tricluster using encoding representation – 1	0.25772	4.21E-09
Tricluster using encoding representation – 2	0.33203	3.27E-09
Tricluster using encoding representation – 3	0.20945	2.01E-09
δ-TRIMAX	0.46709	3.08E-05
TRICLUSTER	0.47753	3.35E-05

Transcription Factor Binding Site (TFBS) enrichment analysis. David Ontology tool which is freely available in the Internet is used for the analysis [19]. The *p*-values are adjusted using Benjamini Hochberg method [20]. The significant genes that have a *p*-value below the threshold of 0.05 are selected. The lower *p*-value represents the higher significance level. Thus, the statistically enriched GO terms belonging to each Tricluster is extracted.

5.4.1 Functional annotation

To find the estrogen influence in the breast cancer, the gene functions are analysed. The key genes from the triclusters are given as input to the ontology tool through which the functional process of the genes is identified. The functional annotation results in genes with genes that have smaller *p*-value that are ranked at top. The 100 genes from the tricluster are given as input for functional annotation. Table 4 provides the top three functional processes of the genes of Triclusters obtained with three different encoding representations. The experimental results show that the gene count percentages of tricluster uses encoding representation 2 and 3 gives better result than tricluster using encoding representation 1 and the *p*-value is also lower for encoding representation 2 and 3.

Table 4. Functional annotation.

Tricluster	Category	Term	Gene Count Percentage	*p*-value	Benjamini
Tricluster using encoding representation – 1	UP_KEYWORDS	Phosphoprotein	56.2	2.2E-04	3.7E-02
	UP_KEYWORDS	Alternative splicing	60.7	5.9E-04	5.0E-02
	UP_SEQ_FEATURE	splice variant	40.2	1.0E-03	2.8E-01
Tricluster using encoding representation – 2	UP_KEYWORDS	Chromosomal rearrangement	58.6	1.7E-04	2.6E-02
	UP_KEYWORDS		66.2	7.2E-04	6.9E-02
	UP_SEQ_FEATURE	Alternative splicing splice variant	45.7	2.1E-02	9.7E-01
Tricluster using encoding representation – 3	UP_SEQ_FEATURE	sequence variant	70.2	2.5E-04	5.2E-02
	UP_KEYWORDS	Polymorphism	70.2	7.3E-04	8.4E-02
	UP_SEQ_FEATURE	splice variant	45.6	1.6E-02	8.1E-01

5.4.2 TFBS enrichment analysis

The potential coregulations of coexpressed genes are analysed using Transcription Factor Binding Site (TFBS) analysis which is done through the same David ontology

tool. The Benjamini FDR method is used for *p*-value correction to find over-represented binding sites in the upstream regions of genes belonging to each tricluster. Table 5 shows the best five genes that have lower *p*-value from all the triclusters. Tricluster using encoding representation 1 covers 93.3% of genes that are entered. Tricluster using encoding representation 2 covers 94.3% of genes and Tricluster using encoding representation 3 covers 93.5% of genes.

Table 5. TFBS analysis.

Tricluster	Gene Term	*p*-value	Benjamini
Tricluster using encoding representation – 1	POU3F2	9.6E-05	1.7E-02
	FREAC3	2.5E-04	2.2E-02
	CDP	3.4E-04	2.0E-02
	CDC5	3.4E-04	1.5E-02
	CREBATF	3.9E-04	1.4E-02
	OCT1	4.1E-04	1.2E-02
Tricluster using encoding representation – 2	EN1	2.7E-03	3.8E-01
	PAX4	1.4E-02	7.2E-01
	USF	4.7E-02	5.8E-01
	NKX3A	4.8E-02	5.4E-01
	FOXC1	5.9E-02	5.6E-01
Tricluster using encoding representation – 3	SP1,4	3.5E-05	6.1E-03
	LHX3	1.2E-04	1.0E-02
	BRN2	4.2E-04	2.4E-02
	SOX9	6.2E-04	2.7E-02
	NKX25	9.5E-04	3.3E-02

The zinc finger transcription factors Sp1 and Sp4 play an important role in estrogen-inducedMCF-7 breast cancer cell line [21, 22]. The basic domain transcription factor CREB is important for malignancy in breast cancer cell. ATF1, ATF2, ATF3, ATF4, ATF5 (CREBATF) likewise play an important role in breast cancer cell [23]. POU2F1 and the TF associated with OCT1 03 is a helix-turn-helix domain transcription factor (Oct-1) has been reported to be estrogen-responsive of breast cancer [24]. The role of FOXC1 as a regulator of human breast cancer cells by activating NFκB signaling has been discovered in a recent years [25]. Also the biological functions, molecular functions and cellular component of the genes grouped in the best cluster are shown in the Table 6. Cell adhesion is the KEGG pathway terms that are observed to be associated with estrogen induced breast cancer [26]. Genes of the best tricluster in all the representation have almost similar properties for biological process, molecular functions and cellular component which are all related to the transcription binding sites of estrogen.

Fig. 7 shows the biological significance comparison of the proposed work with the existing algorithms. This is done by calculating the number of genes in the triclusters which hit the TFBS analysis with the lowest *p*-value. The proposed work has higher percentage of genes in the tricluster with lower *p*-value than the other algorithms. The inverse trend of genes hitting the TFBS analysis is observed with proposed work which has largest population at the lowest *p*-values whereas the other algorithms have increasing population with increasing *p*-values.

Table 6. Gene ontology for best Tricluster in each representation.

Clusters	Biological Process	Molecular Function	Cellular Component
Tricluster using encoding representation – 1	Cell proliferation, negative regulation of cell proliferation, viral process, cell adhesion, Positive regulation of transcription from RNA polymerase II promoter, heart deveopment	Protein binding, ATP binding, transcription factor activity, enzyme binding, protein complex binding, kinase activity, transcription factor binding	Cytosol, membrane, nucleoplasm, cytoplasm, extracellular exosome, protein complex, focal adhesion, nucleus, actin cytoskeleton, melanosome
Tricluster using encoding representation – 2	Transcription, DNA-templated, protein ubiquitination, protein transport, embryonic digit morphogenesis, multicellular organism development	Nucleic acid binding, metal ion binding, zinc ion binding, ubiquitin-protein transferase activity, liquase activity, transcription factor activity, sequence-specific protease activity,actin binding	Nucleoplsam, lysosomal membrane, ciliary basal body, cytoplasm, adherens junction, mitochondrial inner membrane, filopodium
Tricluster using encoding representation – 3	Regulation of transcription, DNA templated, Protein ubiquitination, miRNA mediated inhibition of translation, cilium morphogenesis	Zinc ion binding, Rho-quanyl-nucleotide exchange factor activity, Protein binding, ubiquitin-protein transferace activity, ubiquitin protein liquase activity	Cytoplasm, Nucleoplasm, adherens junction, axoneme, bicellular tight junction, lysosomal membrane, recycling endosome

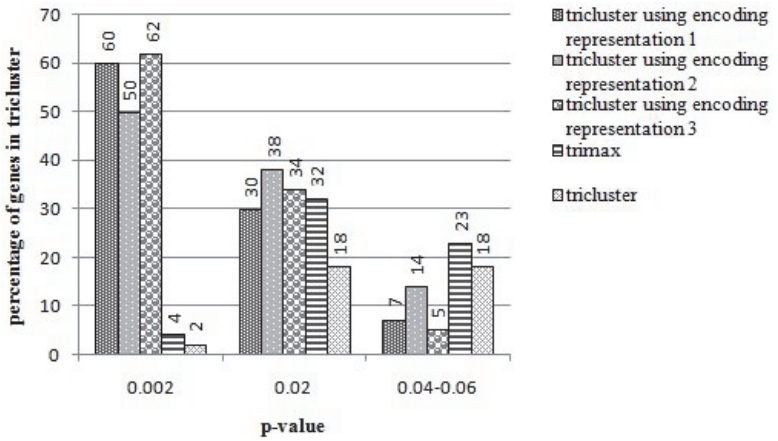


Fig. 7. Biological significance comparison.

6. CONCLUSION

In this work, TrioCuckoo, a multi objective cuckoo search is employed to identify the triclusters with three different encoding representations. It extracts group of genes which are similar on a three dimensional space, thus considering the gene, condition and

time aspect. The real life time series datasets, GSE11324 and PGC-1 aplha are used for experimental analysis. The proposed work is compared with PSO in terms of MSR value and it is also compared with other triclustering algorithms in terms of statistical difference from background and tricluster quality index and it outperforms the existing methods. The Biological significance and TFBS Enrichment Analysis of triclusters are analysed. The experiment results show that the proposed work gives better performance than existing algorithms. The tricluster encoding representation genes as rows, conditions as columns, and time as depth representation results outperforms other two representations. The results of gene ontology and TFBS enrichment analysis signifies that the resultant tricluster extracted the genes which are biologically significant. The function annotation ranks the genes with lower *p*-value indicating which genes are responsible for breast cancer. The proposed work identifies the top genesCDP, CDC5, CREBATF, OCT1, FREAC3, EN1, PAX4, USF, FOX5C1, SP1,SP4, LHX3, SOX9 and NKX25 present in the tricluster which are associated with the breast cancer.

REFERENCES

1. Y. Cheng and G. M. Church, "Biclustering of expression data," in *Proceedings of International Conference on Intelligent Systems Molecular Biology*, 2000, pp. 93-103.
2. J. Feng, P. E. Barbano, and B. Mishra, "Time-frequency feature detection for time course microarray data," in *Proceedings of ACM Symposium on Applied Computing*, 2004, pp. 128-132.
3. L. Zhao and M. J. Zaki, "TRICLUSTER: an effective algorithm for mining coherent clusters in 3D microarray data," in *Proceedings of ACM SIGMOD International Conference on Management of Data*, 2005, pp. 694-705.
4. H. Jiang, S. Zhou, J. Guan, and Y. Zheng, "gTRICLUSTER: A more general and effective 3D clustering algorithm for gene-sample-time microarray data," *Data Mining for Biomedical Applications*, Vol. 3916, 2006, pp. 48-59.
5. Y. Yin, Y. Zhao, B. Zhang, and G. Wang, *Mining Time-Shifting Co-regulation Patterns from Gene Expression Data*, Springer-Verlag Berlin Heidelberg, Vol. 4505, 2007, pp. 1-17.
6. A. B. Tchagang, S. Phan, F. Famili, H. Shearer, P. Fobert, and Y. Huang, "Mining biological information from 3D short time-series gene expression data: the OPTricluster algorithm," *BMC Bioinformatics*, Vol. 13, 2012, p. 54.
7. H. C. Kuo and P. C. Tsai, "Mining time-delayed gene regulation patterns from gene expression data," *GSTF Journal on Computing*, Vol. 2, 2012.
8. A. Bhar, M. Haubrock, A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and E. Wingender, "δ-TRIMAX: Extracting triclusters and analysingcoregulation in time series gene expression data," in *Proceedings of International Workshop on Algorithms in Bioinformatics*, Vol. 7534, 2012, pp. 165-177.
9. D. G. Aviles, C. R. Escudero, F. M. Alvarez, and J. C. Riquelme, "TriGen: A genetic algorithm to mine triclusters in temporal gene expression data," *Neurocomputing*, Vol. 132, 2014, pp. 42-53.

10. A. Bhar, M. Haubrock, A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and E. Wingender, "Multiobjectivetriclustering of time-series transcriptome data reveals key genes of biological processes," *BMC Bioinformatics*, Vol. 16, 2015, p. 200.
11. Y. Liu, T. Yang, and L. Fu, "A partitioning based algorithm to fuzzy tricluster," *Mathematical Problems in Engineering*, Vol. 2015, 2015.
12. R. Guigourès, M. Boullé, and F. Rossi, "Discovering patterns in time-varying graphs: a triclustering approach," *Advances in Data Analysis and Classification*, Springer Verlag, 2016, pp. 1-28.
13. I. Jr. Fister, X. S. Yang, D. Fister, and I. Fister, "Cuckoo search: A brief literature review," *Cuckoo Search and Firefly Algorithm*, *Studies in Computational Intelligence*, Vol. 516, 2014, pp. 49-62.
14. R. B. Payne, M. D. Sorenson, and K. Klitz, *The Cuckoos*, Oxford University Press, Oxford, 2005.
15. X. S. Yang and S. Deb, "Cuckoo Search via Lévy Flights," *World Congress on Nature & Biologically Inspired Computing*, 2009, pp. 210-214.
16. X. S. Yang and S. Deb, "Multiobjective cuckoo search for design optimization," *Computers & Operations Research*, Vol. 40, 2013, pp. 1616-1624.
17. J. Kennedy and R. C. Eberhart, "Particle swarm optimization," in *Proceedings of IEEE International Conference on Neural Networks*, 1995, pp. 1942-1948.
18. A. Bhar, M. Haubrock, A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and E. Wingender, "Coexpression and coregulation analysis of time-series gene expression data in estrogen-induced breast cancer cell," *Algorithms for Molecular Biology*, Vol. 8, 2013.
19. D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, Vol. 4, 2009, pp. 44-57.
20. Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society Series B.*, Vol. 57, 1995, pp. 289-300.
21. S. Khan, F. Wu, S. Liu, Q. Wu, and S. Safe, "Role of specificity protein transcription factors in estrogen-induced gene expression in MCF-7 breast cancer cells," *Journal of Molecular Endocrinology*, Vol. 39, 2007, pp. 289-304.
22. K. Kim, R. Barhoumi, R. Burghardt, and S. Safe, "Analysis of estrogen receptor α-Sp1 interactions in breast cancer cells by fluorescence resonance energy transfer," *Journal of Molecular Endocrinology*, Vol. 19, 2005, pp. 843-854.
23. J. K. Haakenson, M. Kester, and D. X. Liu, "The ATF/CREB family of transcription factors in breast cancer," *Targeting New Pathways and Cell Death in Breast Cancer*, Aft RL. ed., InTech, 2012.
24. C. Wang, J. Yu, and C. B. Kallen, "Two estrogen response element sequences near the PCNA gene are not responsible for its estrogen-enhanced expression in MCF7 cells," *PLOS ONE*, Vol. 3, 2008, pp. 3523.
25. J. Wang, P. S. Ray, M. S. Sim, X. Z. Zhou, K. P. Lu, A. V. Lee, X. Lin, S. P. Bagaria, A. E. Giuliano, and X. Cui, "FOXC1 regulates the functions of human basal-like breast cancer cells by activating NFKB signaling," *Oncogene*, Vol. 31, 2012, pp. 4798-4802.

26. M. Maynadier, P. Nird, J. M. Ramirez, A. M. Cathiard, N. Platet, M. Chambon, and M. Garcia, "Role of estrogens and their receptors in adhesion and invasiveness of breast cancer cells," *Advances in Experimental Medicine and Biology*, Vol. 617, 2008, pp. 485-491.

P. Swathypriyadharsini is currently a research scholar at Anna University, Chennai, Tamil Nadu, India. She completed her Master of Engineering in Computer Science Engineering (CSE) at Bannari Amman Institute of Technology, Erode, Tamil Nadu, India and Bachelor of Engineering in CSE at Avinashilingam University, Coimbatore, Tamil Nadu, India. Her research interests include data mining, soft computing and artificial intelligence.

K. Premalatha is currently working as a Professor in the Department of Computer Science Engineering at Bannari Amman Institute of Technology, Erode, Tamil Nadu, India. She completed her Ph.D. in CSE at Anna University, Chennai, India. She did her Master of Engineering in CSE and Bachelor of Engineering in CSE at Bharathiar University, Coimbatore, Tamil Nadu, India. Her research interests include data mining, networking, information retrieval and soft computing.