δ-TRIMAX: Extracting Triclusters and Analysing Coregulation in Time Series Gene Expression Data

Anirban Bhar[1], Martin Haubrock[1], Anirban Mukhopadhyay[2], Ujjwal Maulik[3],
Sanghamitra Bandyopadhyay[4], and Edgar Wingender[1,*]

[1] Department of Bioinformatics, Medical School, Georg August University of Goettingen, Goldschmidtstrasse 1, D-37077 Goettingen, Germany
{anirban.bhar,martin.haubrock,edgar.wingender}
@bioinf.med.uni-goettingen.de

[2] Department of Computer Science and Engineering, University of Kalyani, Kalyani-741235, India
anirban@klyuniv.ac.in

[3] Department of Computer Science and Engineering, Jadavpur University, Kolkata-700032, India
umaulik@cse.jdvu.ac.in

[4] Machine Intelligence Unit, Indian Statistical Institute, Kolkata-700108, India
sanghami@isical.ac.in

Abstract. In an attempt to analyse coexpression in a time series microarray gene expression dataset, we introduce here a novel, fast triclustering algorithm δ-TRIMAX that aims to find a group of genes that are coexpressed over a subset of samples across a subset of time-points. Here we defined a novel mean-squared residue score for such 3D dataset. At first it uses a greedy approach to find triclusters that have a mean-squared residue score below a threshold δ by deleting nodes from the dataset and then in the next step adds some nodes, keeping the mean squared residue score of the resultant tricluster below δ. So, the goal of our algorithm is to find large and coherent triclusters from the 3D gene expression dataset. Additionally, we have defined an affirmation score to measure the performance of our triclustering algorithm for an artificial dataset. To show biological significance of the triclusters we have conducted GO enrichment analysis. We have also performed enrichment analysis of transcription factor binding sites to establish coregulation of a group of coexpressed genes.

Keywords: Time series gene expression data, Tricluster, Mean-squared residue, Affirmation score, Gene ontology, KEGG Pathway, TRANSFAC.

1 Introduction

In the context of genomics research, the functional approach is based on the ability to analyze genome-wide patterns of gene expression and the mechanisms

[*] Corresponding author.

B. Raphael and J. Tang (Eds.): WABI 2012, LNBI 7534, pp. 165–177, 2012.

by which gene expression is coordinated. Microarray technology and other high-throughput methods are used to measure expression values of thousands of genes over different samples/experimental conditions. In recent years the microarray technology has been used to measure in a single experiment expression values of thousands of genes under a huge variety of experimental conditions across different timepoints. This kind of dataset can be referred to as time series microarray dataset. Because of the large data volume, computational methods are used to analyze such datasets. Clustering is one of the most common methods for identifying coexpressed genes [6]. This kind of analysis is facilitative for constructing gene regulatory networks in which single or groups of genes interact with other genes. Besides this, coexpression analysis also reveals information about some unknown genes that form a cluster with some known genes.

A clustering algorithm is used to group genes that are coexpressed over all conditions/samples or to group experimental conditions over all genes based on some similarity/ dissimilarity metric. However clustering may fail to find the group of genes that are similarly expressed over a subset of samples/experimental conditions i.e. clustering algorithms are unable to find such local patterns in the gene expression dataset. To deal with that problem, biclustering algorithms are used. A bicluster can be defined as a subset of genes that are coexpressed over a subset of samples/experimental conditions. The first biclustering algorithm that was used to analyse gene expression dataset was proposed by Cheng and Church and they used a greedy search heuristic approach to retrieve largest possible bicluster having mean squared residue (MSR) under a predefined threshold value δ (δ-bicluster) [4]. But nowadays, biologists are eager to analyze 3D microarray dataset to answer the question: "*Which genes are coexpressed under which subset of experimental conditions/samples across which subset of time-points*?" Biclustering is not a proper method to answer this question. So, in this case we need some other clustering technique that can deal with the problem. Hence the term *Triclustering* has been defined and a tricluster can be delineated as a subset of genes that are similarly expressed across a subset of experimental conditions/ samples over a subset of time-points. Zhao and Zaki proposed a triclustering algorithm *TRICLUSTER* that is based on graph-based approach. They defined coherence of a tricluster as $\frac{max(e_{ib}/e_{ia}, e_{jb}/e_{ja})}{min(e_{ib}/e_{ia}, e_{jb}/e_{ja})} - 1$, where e_{ia}, e_{ib} denote the expression values of two columns a and b respectively for a row i. A tricluster is valid if it has a ratio below a maximum ratio threshold ϵ [23].

Here we propose an efficient triclustering algorithm *δ-TRIMAX* that aims to cope with noisy 3D gene expression dataset and is less sensitive to input parameters. The normalization method does not influence the performance of our algorithm, as it produces the same results for both normalized and raw datasets. Here we propose a novel extension of MSR [4] for 3D gene expression data and use a greedy search heuristic approach to retrieve triclusters, having MSR values below a threshold δ. Hence the triclusters can be defined as δ-tricluster. The performance of the proposed algorithm is demonstrated on a synthetic dataset as well as a real-life dataset.

2 Methods

2.1 Definitions

Definition 1 (Time Series Microarray Gene Expression Dataset). *We can model a* time series microarray gene expression dataset *(D) as a $G \times C \times T$ matrix and each element of D (d_{ijk}) corresponds to the expression value of gene i over jth sample/experimental condition across time-point k, where $i \in (g_1,g_2,...,g_G)$, $j \in (c_1,c_2,...,c_C)$ and $k \in (t_1,t_2,...,t_T)$.*

Definition 2 (Tricluster). *A* tricluster *is defined as a submatrix M(I,J,K) = [m_{ijk}], where $i \in I$, $j \in J$ and $k \in K$. The submatrix M represents a subset of genes (I) that are coexpressed over a subset of conditions (J) across a subset of time-points (K).*

Definition 3 (Perfect Shifting Tricluster). *A Tricluster M(I,J,K) = m_{ijk}, where $i \in I$, $j \in J$ and $k \in K$, is called a* perfect shifting tricluster *if each element of the submatrix M is represented as: $m_{ijk} = \mathit{\Gamma} + \alpha_i + \beta_j + \eta_k$, where $\mathit{\Gamma}$ is a constant value for the tricluster, α_i, β_j and η_k are shifting factors of ith gene, jth samples/experimental condition and kth time-point, respectively. As noise is present in microarray datasets, the deviation from actual value and expected value of each element in the dataset also exists. For this deviation, every tricluster is not a perfect one.*

Cheng and Church proposed an algorithm for retrieving large and maximal biclusters that have mean squared residue score (MSR) below a threshold δ in 2D microarray gene expression dataset. They also showed that MSR of a perfect δ-bicluster and perfect shifting bicluster is zero ($\mathbf{S} = \delta = 0$) [4,6]. Now extending this idea, here we present a novel definition of Mean Squared Residue score for 3D microarray gene expression datasets. The MSR ($\mathbf{S}$) of a perfect shifting tricluster becomes also zero, where each element $m_{ijk} = \mathit{\Gamma} + \alpha_i + \beta_j + \eta_k$. For delineating new MSR score ($\mathbf{S}$), at first we need to define the residue score:

Let the mean of ith gene (m_{iJK}): $m_{iJK} = \frac{1}{|J||K|}\sum_{j\in J,k\in K} m_{ijk}$, the mean of jth sample/experimental condition (m_{IjK}): $m_{IjK} = \frac{1}{|I||K|}\sum_{i\in I,k\in K} m_{ijk}$, the mean of kth time-point (m_{IJk}): $m_{IJk} = \frac{1}{|I||J|}\sum_{i\in I,j\in J} m_{ijk}$, and the mean of tricluster (m_{IJK}): $m_{IJK} = \frac{1}{|I||J||K|}\sum_{i\in I,j\in J,k\in K} m_{ijk}$. Now the mean of the tricluster can be considered as the value of constant i.e. $\mathit{\Gamma} = m_{IJK}$. We can define the shifting factor for the ith gene (α_i) as the difference between m_{iJK} and m_{IJK} i.e. $\alpha_i = m_{iJK} - m_{IJK}$. Similarly, we can define shifting factor for the jth condition (β_j) as $\beta_j = m_{IjK} - m_{IJK}$ and shifting factor for the kth time-point (η_k) can be defined as $\eta_k = m_{IJk} - m_{IJK}$. Hence we can define each element of a perfect shifting tricluster as $m_{ijk} = \mathit{\Gamma} + \alpha_i + \beta_j + \eta_k = m_{IJK} + (m_{iJK} - m_{IJK}) + (m_{IjK} - m_{IJK}) + (m_{IJk} - m_{IJK}) = (m_{iJK} + m_{IjK} + m_{IJk} - 2m_{IJK})$. But usually noise is evident in microarray gene expression dataset. Therefore to evaluate the difference between the actual value of an element (m_{ijk}) and its expected value, obtained from above equation, the term *"residue"* can be used [6]. Thus the residue of a tricluster (r_{ijk}) can be defined as follows: $r_{ijk} = m_{ijk} - (m_{iJK} + m_{IjK} + m_{IJk} - 2m_{IJK}) = (m_{ijk} - m_{iJK} - m_{IjK} - m_{IJk} + 2m_{IJK})$.

Definition 4 (Mean Squared Residue). *We define the term Mean Squared Residue MSR(I,J,K) or* ***S*** *of a tricluster M(I,J,K) to estimate the quality of a tricluster i.e. the level of coherence among the elements of a tricluster as follows:*

$$\begin{aligned} \boldsymbol{S} &= \frac{1}{|I||J||K|} \sum_{i \in I, j \in J, k \in K} r_{ijk}^2 \\ &= \frac{1}{|I||J||K|} \sum_{i \in I, j \in J, k \in K} (m_{ijk} - m_{iJK} - m_{IjK} - m_{IJk} + 2m_{IJK})^2. \end{aligned} \tag{1}$$

Lower residue score represents larger coherence and better quality of a tricluster.

2.2 Proposed Method

The method proposed here (*δ-TRIMAX*) aims to find largest and maximal triclusters in a 3D microarray gene expression dataset. It is an extension of Cheng and Church biclustering algorithm [4] that deals with 2-D microarray datasets. In contrast, our algorithm is capable to mine 3D gene expression dataset. There is always a submatrix in an expression dataset that has a perfect MSR(I,J,K) or **S** score i.e. **S** = 0 and this submatrix is each element of the dataset. But as mentioned above, our algorithm finds maximal triclusters having **S** score under a threshold δ, hence we have used a greedy heuristic approach to find triclusters. Our algorithm therefore starts with the entire dataset containing all genes, all samples/experimental conditions and all time-points.

```
Algorithm I (δ-TRIMAX):
Input. D, a matrix that represents 3D microarray gene expression dataset, $\lambda > 1$, an input parameter for multiple node deletion algorithm, $\delta \geq 0$, maximum allowable MSR score.
Output. All possible δ-triclusters.
Initialization. Missing elements in D ← random numbers, D' ← D
repeat
    a. D'$_1$ ← Results of Algorithm II on D' using $\delta$ and $\lambda$. If the no. of genes (conditions/samples and/or no. of time-points) is 50 (This value can be chosen experimentally. Large value increases the execution time of the algorithm as it then executes more number of iterations), then do not apply Algorithm II on genes (conditions/samples and/or time-points).
    b. D'$_2$ ← Results of Algorithm III on D'$_1$ using $\delta$.
    c. D'$_3$ ← Results of Algorithm IV on D'$_2$.
    d. Return D'$_3$ and replace the elements that exist in D' and D'$_3$ with random numbers.
until (there is no gene in δ-tricluster)
```

Initially, our algorithm removes genes or conditions or time-points from the dataset to accomplish largest diminishing of score **S**; this step is described in the following section in which a node corresponds to a gene or experimental condition or time-point in the 3D microarray gene expression dataset.

Algorithm II (Multiple Node Deletion):
Input. D, a matrix of real numbers that represents 3D microarray gene expression dataset; $\delta \geq 0$, maximum allowable MSR threshold, $\lambda > 1$, threshold for multiple node deletion. The value of λ is set experimentally to optimize the speed and performance (to avoid falling into local optimum) of the algorithm.
Output. M_{IJK}, a δ-tricluster, consisting of a subset(I) of genes, a subset(J) of samples/ experimental conditions and a subset of time-points, having MSR score (**S**) less than or equal to δ.
Initialization. I $\leftarrow$ {set of all genes}, J $\leftarrow$ {set of all experimental conditions/ samples} and K $\leftarrow$ {set of all time-points} and to M(I,J,K) $\leftarrow$ D

```
repeat
  Calculate $\mathrm{m}_{iJK}$, $\forall$ i $\in$ I; $\mathrm{m}_{IjK}$, $\forall$ j $\in$ J; $\mathrm{m}_{IJk}$, $\forall$ k $\in$ K; $\mathrm{m}_{IJK}$ and S.
  if S $\leq \delta$ then
    return M(I,J,K)
  else
    Delete genes i $\in$ I that satisfy the following inequality

      $\frac{1}{|J||K|}\Sigma_{j\in J,k\in K}(m_{ijk} - m_{iJK} - m_{IjK} - m_{IJk} + 2m_{IJK})^2 > \lambda\mathbf{S}$

    Recalculate $\mathrm{m}_{iJK}$, $\forall$ i $\in$ I; $\mathrm{m}_{IjK}$, $\forall$ j $\in$ J; $\mathrm{m}_{IJk}$, $\forall$ k $\in$ K; $\mathrm{m}_{IJK}$ and S
    Delete samples/experimental conditions j $\in$ J that satisfy the following inequality

      $\frac{1}{|I||K|}\Sigma_{i\in I,k\in K}(m_{ijk} - m_{iJK} - m_{IjK} - m_{IJk} + 2m_{IJK})^2 > \lambda\mathbf{S}$

    Recalculate $\mathrm{m}_{iJK}$, $\forall$ i $\in$ I; $\mathrm{m}_{IjK}$, $\forall$ j $\in$ J; $\mathrm{m}_{IJk}$, $\forall$ k $\in$ K; $\mathrm{m}_{IJK}$ and S
    Delete time-points k $\in$ K that satisfy the following inequality

      $\frac{1}{|I||J|}\Sigma_{i\in I,j\in J}(m_{ijk} - m_{iJK} - m_{IjK} - m_{IJk} + 2m_{IJK})^2 > \lambda\mathbf{S}$

  end if
until (There is no change in I, J and/or K)
```

The complexity of this algorithm is O(max(m,n,p)) where m, n and p are the number of genes, samples and time-points in the 3D microarray dataset.

In the second step, we delete one node at each iteration from the resultant submatrix, produced by Algorithm II, until the score **S** of the resultant submatrix is less than or equal to δ. This step results in a δ-tricluster.

```
Algorithm III (Single Node Deletion):
Input. D, a matrix of real numbers that represents 3D microarray gene expression dataset; $\delta \geq 0$, maximum allowable MSR threshold.
Output. M$_{IJK}$, a $\delta$-tricluster, consisting of a subset(I) of genes, a subset(J) of samples/experimental conditions and a subset of time-points, having MSR score (S) less than or equal to $\delta$.
Initialization. I ← {set of all genes in D}, J ← {set of experimental conditions/samples in D} and K ← {set of time-points in D} and to M(I,J,K) ← D
Calculate m$_{iJK}$, ∀ i ∈ I; m$_{IjK}$, ∀ j ∈ J; m$_{IJk}$, ∀ k ∈ K; m$_{IJK}$ and S.
while S > $\delta$ do
  Detect gene i ∈ I that has the highest score
      $\mu(i) = \frac{1}{|J||K|} \Sigma_{j \in J, k \in K} (m_{ijk} - m_{iJK} - m_{IjK} - m_{IJk} + 2m_{IJK})^2$
  Detect sample/experimental condition j ∈ J that has the highest score
      $\mu(j) = \frac{1}{|I||K|} \Sigma_{i \in I, k \in K} (m_{ijk} - m_{iJK} - m_{IjK} - m_{IJk} + 2m_{IJK})^2$
  Detect time-point k ∈ K that has the highest score
      $\mu(k) = \frac{1}{|I||J|} \Sigma_{i \in I, j \in J} (m_{ijk} - m_{iJK} - m_{IjK} - m_{IJk} + 2m_{IJK})^2$
  Delete gene or sample/experimental condition or time-point that has highest $\mu$ score and modify I or J or K.
  Recalculate m$_{iJK}$, ∀ i ∈ I; m$_{IjK}$, ∀ j ∈ J; m$_{IJk}$, ∀ k ∈ K; m$_{IJK}$ and S.
end while
Return M(I,J,K)
```

The complexity of first and second steps is O(mnp) as those will iterate (m+n+p) times. The complexity of selection of best genes, samples and time-points is O(log m + log n + log p). So it is suggested to use algorithm II before algorithm III.

As the goal of our algorithm is to find maximal triclusters, having MSR score (**S**) below the threshold δ, the resultant tricluster M(I,J,K) may not be the largest one. That means some genes and/or experimental conditions/samples and/or time-points may be added to the resultant tricluster T produced by node deletion algorithm, so that the MSR score of new tricluster T' produced after node addition does not exceed the MSR score of T. Now the third step of our algorithm is described below.

```
Algorithm IV (Node Addition):
Input. D, a matrix of real numbers that represents $\delta$-tricluster, having a subset of genes (I), a subset of experimental conditions/samples (J) and a subset of time-points (K).
```

```
Output. $M_{I'J'K'}$, a $\delta$-tricluster, consisting of a subset of genes (I') , a subset of samples/experimental conditions (J') and a subset of time-points (K'), such that $I \subset I'$, $J \subset J'$, $K \subset K'$ and MSR(I',J',K') $\leq$ MSR of D.
Initialization. M(I,J,K) $\leftarrow$ D
repeat
    Calculate $m_{iJK}$, $\forall$ i; $m_{IjK}$, $\forall$ j; $m_{IJk}$, $\forall$ k; $m_{IJK}$ and S.
    Add genes i $\notin$ I that satisfy the following inequality
        $\frac{1}{|J||K|}\Sigma_{j\in J,k\in K}(m_{ijk} - m_{iJK} - m_{IjK} - m_{IJk} + 2m_{IJK})^2 \leq \mathbf{S}$
    Recalculate $m_{IjK}$, $\forall$ j; $m_{IJk}$, $\forall$; $m_{IJK}$ and S
    Add samples/experimental conditions j $\notin$ J that satisfy the following inequality
        $\frac{1}{|I||K|}\Sigma_{i\in I,k\in K}(m_{ijk} - m_{iJK} - m_{IjK} - m_{IJk} + 2m_{IJK})^2 \leq \mathbf{S}$
    Recalculate $m_{iJK}$, $\forall$ i; $m_{IJk}$, $\forall$ k; $m_{IJK}$ and S
    Add time-points k $\notin$ K that satisfy the following inequality
        $\frac{1}{|I||J|}\Sigma_{i\in I,j\in J}(m_{ijk} - m_{iJK} - m_{IjK} - m_{IJk} + 2m_{IJK})^2 \leq \mathbf{S}$
until (There is no change in I, J and/or K)
I' $\leftarrow$ I, J'$\leftarrow$ J, K' $\leftarrow$ K
Return I', J', K'
```

The complexity of this algorithm is O(mnp) as each step iterates (m+n+p) times.

3 Results and Discussion

3.1 Results on Simulated Dataset

We have produced one simulated dataset SMD of size 2000 × 30 × 30. At first we have implanted three perfect shifting triclusters of size 100 × 6 × 6, 80 × 6 × 6 and 60 × 5 × 5 into the dataset SMD and then implanted three noisy shifting triclusters of the same size mentioned before into it. To estimate the degree of similarity between the implanted and obtained triclusters, we define *affirmation score* in the same way as Prelic et. al. defined for two sets of biclusters [6,7]. So, overall average affirmation score of T_1 with respect to T_2 is as follows, where ($SM^*{}_G(T_1, T_2)$) is the average gene affirmation score, ($SM^*{}_C(T_1, T_2)$) is the average sample affirmation score and ($SM^*{}_K(T_1, T_2)$) is the average time-point affirmation score of T_1 with respect to T_2:

$$SM^*(T_1, T_2) = \sqrt{(SM^*_G(T_1, T_2) \times SM^*_C(T_1, T_2) \times SM^*_T(T_1, T_2))} \qquad (2)$$

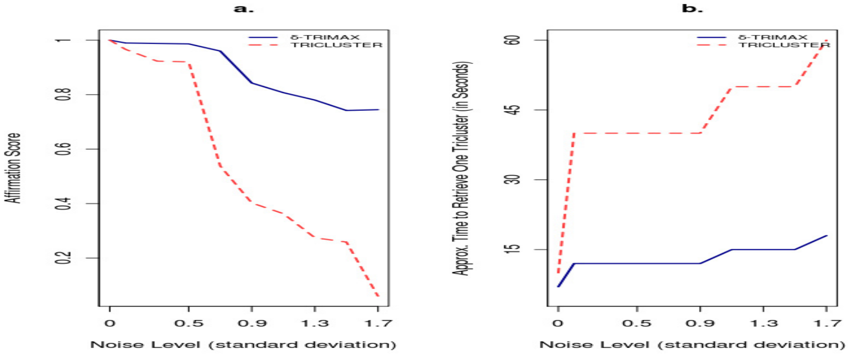

Fig. 1. a. Comparison of Affirmation scores produced by δ-TRIMAX and *TRICLUSTER* algorithm. b. Comparison of running time of δ-TRIMAX and *TRICLUSTER* algorithm on the synthetic dataset.

Suppose, we have two sets of triclusters T_{im} and T_{res} where T_{im} represents the set of implanted triclusters and T_{res} corresponds to the set of triclusters retrieved by any triclustering algorithm. Hence $\mathrm{SM}^*(T_{im}, T_{res})$ denotes how well the triclustering algorithm finds the true triclusters that have been implanted into the dataset. This score varies from 0 to 1 (if $\mathrm{T}_{im} = \mathrm{T}_{res}$).

In this case we have assigned 0.35 and 1.0005 to the parameters δ and λ, respectively. The value of δ varies from one dataset to another dataset. Here we have first implanted three perfect shifting triclusters into the dataset SMD and then added noise to those triclusters with different standard deviations (σ = 0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5, 1.7). To have an idea about the δ value, we have first clustered the genes over all time-points and then the time-points over the subset of genes for each gene cluster in each sample plane using the K-means algorithm. Then we have computed the MSR value(S) of the submatrix, considering a randomly selected sample plane, gene and time-pont cluster for 100 times. Then we have taken the lowest value as the value of δ. For these noisy datasets, we have assigned 3.75 and 1.004 to the parameters δ and λ, respectively. In Figure 1 we have compared the performance of our algorithm with that of the *TRICLUSTER* algorithm [23] in terms of affirmation score using the artificial dataset. Our δ-TRIMAX algorithm performs better than *TRICLUSTER* algorithm for the noisy dataset. For perfect additive triclusters, performances of both these algorithms are comparable with each other.

3.2 Results on Real-Life Dataset

Datasets for Genome-Wide Analysis of Estrogen Receptor Binding Sites: This dataset contains 54675 affymetrix probe-set ids, 3 biological replicates and 4 time-points. In this experiment MCF7 cells are stimulated with 100 nm estrogen for 0, 3, 6 and 12 hours and the experiments are performed in triplicate. This dataset can be downloaded from the following webpage publicly:

`http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE11324`. This has been used for discovering of cis-regulatory sites in previously uninvestigated regions and cooperating transcription factors underlying estrogen signaling in breast cancer [11]. We assign 0.012382 and 1.2 to δ and λ respectively. From Figure 2, we observe that the genes in tricluster 4 have similar expression profiles over all three samples across 0, 6 and 12 hours but not at 3 hour. Our algorithm results in 115 triclusters that cover 96.37% of all genes, 100% of all samples and 100% of all time-points.

Biological Significance. We have used the *GOstats* package [22] in R to perform GO and pathway enrichment analysis for genes belonging to each tricluster. We have adjusted the p-values using Benjamini-Hochberg FDR method [1] and considered those terms as significant ones that have a p-value below a threshold of 0.05. The smaller p-value represents higher significance level. We have found statistically enriched GO terms for genes belonging to each tricluster. Additionally to analyse the potential coregulation of coexpressed genes, we have done transcription factor binding site (TFBS) enrichment analysis using the TRANSFAC library (version 2009.4) [8] that contains eukaryotic transcription factors, their experimentally-proven binding sites, and regulated genes. Here we used 42,544,964 TFBS predictions that have high affinity scores and are conserved between human, mouse, dog and cow (Haubrock et al., in preparation). Out of

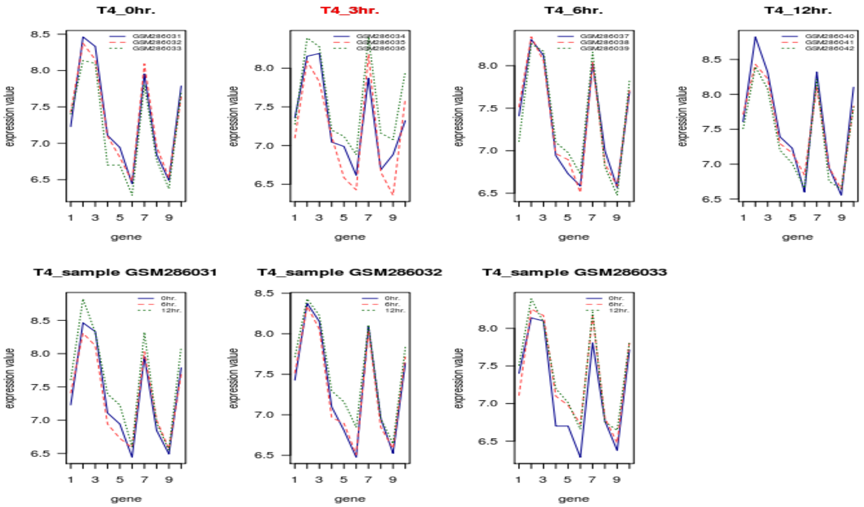

Fig. 2. a.The figures in first row show the expression profiles of genes *ESR1, HOXA11, FAM71A, SPEF2, IFIH1, FPR2, SPAG9, NCF4, ADAM3A, CCNYL1*, respectively of tricluster 4 over all samples; The red-colored time-point (3 hr.) is not a member of this tricluster. b. The figures in second row show the expression profiles of the same genes across 0, 6 and 12 hour.

Table 1. TRANSFAC Matrices for Triclusters, having statistically enriched TFBS for real-life dataset. Transcription factors of red, green, blue, orange, brown and violet colored matrices are Helix-turn-helix, β-Sheet binding to DNA, Zinc-coordinating DNA-binding, basic, all-α-helical DNA-binding and Immunoglobulin fold domain factors, respectively.

Tricluster (no. of genes)	5 most significant TRANSFAC matrices (in ascending order of p-values)	FDR-BY corrected p-value of top-most matrix
Tricluster 3 (875)	VNCX_02, VMSX1_02, V$PAX4_02, V$POU3F2_01, V$TBP_01	4.29e-08
Tricluster 1 (4477)	VNCX_02, VHDX_01, V$BCL6_01, V$ZNF333_01, V$DLX2_01	1.27e-05
Tricluster 26 (3177)	V$E2F_Q2, V$ZF5_01, V$USF2_Q6, V$SP1_Q6_01, V$KID3_01	2.99e-05
Tricluster 4 (3482)	V$BCL6_01, V$HOXA10_01, VSRY_01, VNKX23_01, V$WT1_Q6	9.51e-05
Tricluster 2 (2186)	V$CHCH_01, V$MOVOB_01, VMAZ_Q6, VPAX4_03, V$CACD_01	0.0001
Tricluster 12 (476)	VSRY_02, VNCX_02, V$BCL6_01, V$HB24_01, V$HOXA10_01	0.002
Tricluster 17 (999)	V$CREB_01, V$CREBATF_Q6, V$SP1_Q6_01, V$ATF3_Q6, V$CREBP1CJUN_01	0.004
Tricluster 50 (182)	V$ETF_Q6	0.006
Tricluster 18 (260)	V$STAT1STAT1_Q3	0.042
Tricluster 31 (2465)	V$SP1_Q6_01	0.046

these 42 million conserved TFBS we have selected the best 1% for each TRANSFAC matrix individually to identify the most specific regulators (transcription factors). We have used hypergeometric test [9] and Benjamini Yekutieli-FDR method [2] for p-value correction to find over-represented binding sites (p-value ≤ 0.05) in the upstream regions of genes belonging to each tricluster. Here we rank the triclusters in ascending order of corrected p-values of transcription factor matrices. Triclusters 3 and 1 are found to be the top ranked triclusters as per our ranking. Genes belonging to triclusters 3 and 1 are enriched with GOBP terms *cyclic nucleotide catabolic process* (GO: 0009214) and *multicellular organismal process* (GO: 0032501), respectively. We also observe enrichment of KEGG pathway terms *Retinol metabolism (KEGG: 00830)* and *Neuroactive ligand receptor interaction* (KEGG: 04080) for triclusters 3 and 1 respectively.

3.3 Discussion

Table 1 shows the list of triclusters where we have found statistically meliorated TFBS. Figure 3 demonstrates the interactions between regulators for all triclusters that have statistically enriched TFBS. Homeodomain factors MSX2, TLX1, DLX2, DLX5, BARX2 were found to play an important role in a breast cancer cell line [10, 16, 17, 20]. PBX1 was reported as a pioneer factor in ERα-positive breast cancer cell line [12]. Rel homology region factor NFAT serves as a pro-invasive, pro-migratory and an inhibitor of cell motility in a breast cancer cell line [13, 14, 21]. TEA domain factor ETF, E2F related factor E2F1 (proliferative marker), zinc finger factors SP1 are also found to be instrumental in a breast cancer cell line [15, 18, 19]. *ANAPC1, MCM7, WEE1, E2F1, E2F4, TFDP1, TFDP2* genes are found to be coexpressed in tricluster 26 and participate in *Cell cycle* pathway. *ANAPC1, MCM7, WEE1* genes also share conserved TFBS for V$E2F_Q2 matrix and *E2F1, E2F4, TFDP1, TFDP2* are coding genes of the same TRANSFAC matrix. The CREB-related factors CREB1, ATF and Jun-related factors JunB and JunD play an important role in estrogen receptor-mediated regulation in a breast cancer cell line [3, 5]. Fig. 2 in

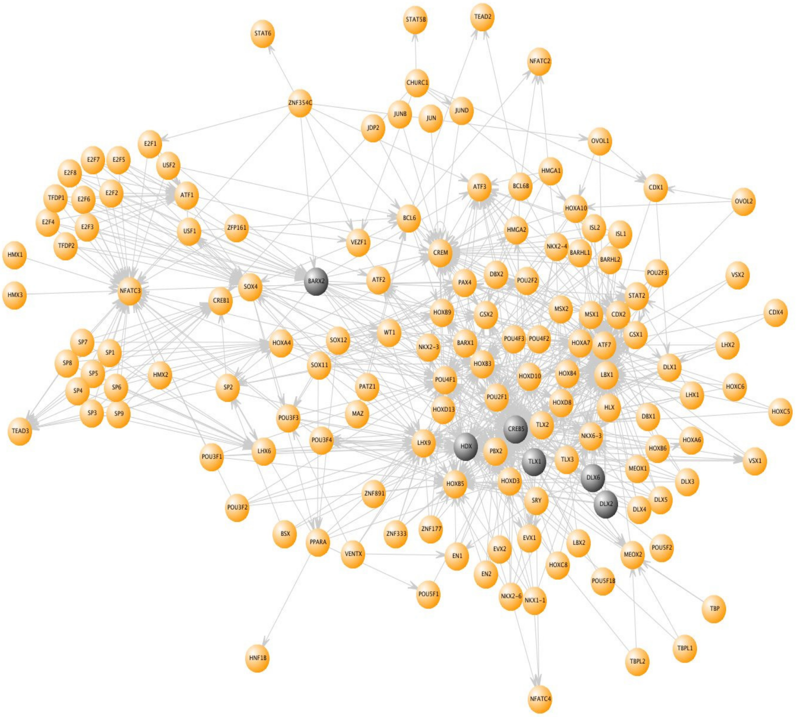

Fig. 3. Gene Regulatory Network for TFBS enriched triclusters

supplementary file (Supp1.pdf) shows the regulatory behavior of transcription factors BARX2, TLX1, DLX2, DLX6, HDX and CREB5 over all time-points. BARX2, member of tricluster 26 have been found to be a regulator of tricluster 3 and 1 and activates HDX and CREB5 at 6 hrs., whereas BARX2, TLX1 and DLX2 jointly act as activators of CREB5 at 6 hrs. We have also observed that a group of coexpressed genes can be regulated by a transcription factor that itself is a member of another tricluster. We can also observe that the expressions of DLX2 and DLX6 are mutually exclusive and it could be the fact that both DLX6 and DLX2 play the role of activator of CREB5, as they belong to the same sub family. The supplementary file is available at `http://www.bioinf.med.uni-goettingen.de/services/talks/wabi_2012/`.

4 Conclusion

From the above work, we can conclude that the proposed δ-TRIMAX triclustering algorithm is able to retrieve large and coherent groups of genes, having an MSR score below a threshold δ. Genes belonging to each tricluster are coexpressed over a subset of samples/ experimental conditions and across subset of time-points. The results show that the proposed triclustering algorithm is able to find functionally enriched sets of coexpressed genes. In case of the artificial dataset our algorithm outperformed the *TRICLUSTER* algorithm. To extend this work, we plan to use a genetic algorithm (GA) that will yield triclusters that are optimized with respect to homogeneity and volume.

References

1. Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society 57(1), 289–300 (1995)
2. Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics 29(4), 1165–1188 (2001)
3. Chen, D., et al.: JunD and JunB integrate prostaglandin E2 activation of breast cancer-associated proximal aromatase promoters. Mol. Endocrinol. 25(5), 767–775 (2011)
4. Cheng, Y., Church, G.M.: Biclustering of expression data. In: Proc. Int. Conf. Int. Syst. Mol. Biol., pp. 93–103 (2000)
5. Chhabra, A., et al.: Expression of transcription factor CREB1 in human breast cancer and its correlation with prognosis. Oncology Reports 18(4), 953–958 (2007)
6. Mukhopadhyay, A., et al.: A novel coherence measure for discovering scaling biclusters from gene expression data. Journal of Bioinformatics and Computational Biology 7(5), 853–868 (2009)
7. Prelic, A., et al.: A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics 22, 1122–1129 (2006)
8. Wingender, E., et al.: The TRANSFAC system on gene expression regulation. Nucleic Acids Res. 29(29), 281–283 (2001)
9. Boyle, E.I., et al.: GO::TermFinder-open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. Bioinformatics 20(18), 3710–3715 (2004)
10. Lanigan, F., et al.: Homeobox transcription factor muscle segment homeobox 2(Msx2) correlates with good prognosis in breast cancer patients and induces apoptosis in vitro. Breast Cancer Research 12(R59) (2010)
11. Carroll, J.S., et al.: Genome-wide analysis of estrogen receptor binding sites. Nature Genetics 38(11) (November 2006)
12. Magnani, L., et al.: PBX1 genomic pioneer function drives ERα signaling underlying progression in breast cancer. PLOS Genetics 7(11) (November 2011)
13. Fougere, M., et al.: NFAT3 transcription factor inhibits breast cancer cell motility by targeting the Lipocalin 2 gene. Oncogene 29(15), 2292–2301 (2010)
14. Yoeli-Lerner, M., et al.: Akt blocks breast cancer cell motility and invasion through the transcription factor NFAT. Molecular Cell 20(4), 539–550 (2005)
15. Khan, S., et al.: Role of specificity protein transcription factors in estrogeninduced gene expression in mcf-7 breast cancer cells. Journal of Molecular Endocrinology 39, 289–304 (2007)
16. Tommasi, S., et al.: Methylation of homeobox genes is a frequent and early epigenetic event in breast cancer. Breast Cancer Research 11(R14) (2009)
17. Lee, S.Y., et al.: Homeobox gene Dlx-2 is implicated in metabolic stress-induced necrosis. Molecular Cancer 10(113) (2011)
18. Zhang, S.Y., et al.: E2F-1: a proliferative marker of breast neoplasia. Cancer Epidemiology, Biomarkers & Prevention 9, 395–401 (2000)

19. Maeda, T., et al.: TEF-1 transcription factors regulate activity of the mouse mammary tumor virus LTR. Biochemical and Biophysical Research Communications 296(5), 1279–1285 (2002)
20. Stevens, T.A., et al.: BARX2 and estrogen receptor-alpha (ESR1) coordinately regulate the production of alternatively spliced esr1 isoforms and control breast cancer cell growth and invasion. Oncogene 25, 5426–5435 (2006)
21. Jauliac, S., et al.: The role of NFAT transcription factors in integrin-mediated carcinoma invasion. Nature Cell Biology 4(7), 540–544 (2002)
22. Falcon, S., Gentleman, R.: Using GOstats to test gene lists for GO term association. Bioinformatics 23(2), 257–258 (2007)
23. Zhao, L., Zaki, M.J.: TRICLUSTER: An effective algorithm for mining coherent clusters in 3D microarry data. In: SIGMOD (June 2005)