A Unified Adaptive Co-identification Framework for High-D Expression Data*

Shuzhong Zhang[1], Kun Wang[3], Cody Ashby[3],
Bilian Chen[2], and Xiuzhen Huang[3,**]

[1] University of Minnesota, Minneapolis, MN 55455, USA
zhangs@umn.edu
[2] Xiamen University, Xiamen 361000, China
chenbilian_158@hotmail.com
[3] Arkansas State University, Jonesboro, AR 72467, USA
{kun.wang,cody.ashby}@smail.astate.edu, xhuang@astate.edu

Abstract. High-throughput techniques are producing large-scale high-dimensional (e.g., 4D with genes vs timepoints vs conditions vs tissues) genome-wide gene expression data. This induces increasing demands for effective methods for partitioning the data into biologically relevant groups. Current clustering and co-clustering approaches have limitations, which may be very time consuming and work for only low-dimensional expression datasets. In this work, we introduce a new notion of "co-identification", which allows systematical identification of genes participating different functional groups under different conditions or different development stages. The key contribution of our work is to build a unified computational framework of co-identification that enables clustering to be high-dimensional and adaptive. Our framework is based upon a generic optimization model and a general optimization method termed Maximum Block Improvement. Testing results on yeast and *Arabidopsis* expression data are presented to demonstrate high efficiency of our approach and its effectiveness.

1 Introduction

While genome data is relatively static, gene expression, which reflects gene activity, is highly dynamic. Gene expression of the cell could be used to infer the cell type, state, stage, and cell environment and may indicate a homeostasis response or a pathological condition and thus relate to development of new medicines, drug metabolism, and diagnosis of diseases [33,27,8]. High-throughput gene expression techniques, such as microarray and next-generation sequencing, are generating huge amounts of high-dimensional genome-wide expression data (e.g., data in 2D matrices: genes vs conditions, or in 3D, 4D, or 5D: genes vs time points vs conditions vs tissues vs development stages vs stimulations). While the availability of these data presents unprecedented opportunities, it also

* This research is supported by grants from NIH NCRR (5P20RR016460-11) and NIGMS (8P20GM103429-11).
** Corresponding author.

T. Shibuya et al. (Eds.): PRIB 2012, LNBI 7632, pp. 59–70, 2012.

presents major challenges for extractions of biologically meaningful information from the large data sets. In particular, it calls for effective computational models, equipped with efficient solution methods, to categorize gene expression data into biologically relevant groups in order to facilitate further functional assessment of important biological and biomedical processes. Classical clustering and co-clustering analysis is a worthy approach in this endeavor.

Clustering is usually applied to partition expression data into groups. A lot of research has been conducted in clustering. Cf. [12] for classical clustering, where the author discussed two classes of clustering: hierarchical clustering and partitioning, and three popular clustering methods: hierarchical clustering [15], k-means clustering [35] and the self organizing map (SOM) method [34]. The classical clustering methods cluster genes into groups based on their similar expression on all the considered conditions. The concept of *co-clustering* was introduced to 2D expression data analysis by Cheng and Church [7]. The co-clustering method can cluster genes and conditions simultaneously. Subsequently, many co-clustering algorithms were developed, such as the plaid model approach [23], xMotif[28], BiMax [29], OPSM [3], Bicluster [9], BCC[2], and ROCC [11]. Different techniques improving co-clustering approaches were also developed [1,38,39]. Readers may refer to [26,17,9,11] for the ideas of different co-clustering algorithms and techniques and [29] for a comprehensive comparison of the popular co-clustering approaches. Recently there are approaches developed for 3D expression data clustering analysis [31,25,41,21]. However, for current clustering and co-clustering approaches, there are important issues to address:

How to develop a systematic method to be able to associate one item to multiple co-groups under different conditions or different development stages of high-dimensional gene expression data? Most classical clustering and co-clustering methods assign one element to one specific cluster or co-clusters. For gene expression analysis, it is important to associate genes/conditions with multiple clusters or co-clusters, inducing the concept of "soft" clustering which allows elements to be members of multiple groups. In [30], soft clustering is represented by a probabilistic distribution. There are methods considering co-cluster overlapping such as the ROCC approach in [11], which, however, only tries to merge some related co-clusters as a post-processing step, and the approach in [7], which allows overlaps but has introduced the masking problem (where the elements in a previously-discovered co-cluster are replaced by random numbers). Refer to [26] for different additive and multiple overlapping models for co-clustering.

In this work, our co-identification approach is different from the previous overlapping bicluster approaches: Our approach does not aim to overlap the biclusters as a post-processing step. Our approach will systematically identify the genes involved in different functional groups at different time points or conditions while the biclusters are being built all at the same time. Note that Lazzeroni and Owen [23] attempted to discover one co-cluster at a time in an iterative process where a plaid model is obtained [26].

How to naturally determine the number of clusters and co-clusters? Classical clustering and co-clustering methods usually rely on the predetermined numbers

as the numbers of clusters and co-clusters. There are methods for estimating the number of clusters or co-clusters in a data set, such as the SVD method [9], the gap statistic or similarity matrix [4,37,14,24], which are, however, not related to the clustering process. In this work we develop an adaptive method to determine the number of co-clusters while the co-clusters are being formed.

2 Methods

We build the computational approach for co-identification based on block optimization, and develop new algorithms from a general scheme which we termed as Maximum Block Improvement (MBI), for naturally growing the size of co-groups and encouraging the degree of co-identification.

The Co-clustering Problem. To illustrate the ideas, consider the conventional co-clustering formulation [40]. Suppose that $A \in \Re^{n_1 \times n_2 \times \cdots \times n_d}$ is an d-dimensional tensor. Let $I_j = \{1, 2, \cdots, n_j\}$ be the set of indices on the j-th dimension, $j = 1, 2, ..., d$. We wish to find a p_j-partition of the index set I_j, say $I_j = I_1^j \cup I_2^j \cup \cdots \cup I_{p_j}^j$, where $j = 1, 2, ..., d$, in such a way that each of the *sub-tensor* $A_{I_{i_1}^1 \times I_{i_2}^2 \times \cdots \times I_{i_d}^d}$ is as tightly packed up as possible, where $1 \leq i_j \leq n_j$ and $j = 1, 2, ..., d$. The notion that plays an important role in our model is the so-called *mode product* between a tensor X and a matrix P. Suppose that $X \in \Re^{p_1 \times p_2 \times \cdots \times p_d}$ and $P \in \Re^{p_i \times m}$. Then, $X \times_i P$ is a tensor in $\Re^{p_1 \times p_2 \times \cdots \times p_{i-1} \times m \times p_{i+1} \times \cdots \times p_d}$, whose $(j_1, j_2, \cdots, j_{i-1}, j_i, j_{i+1}, \cdots, j_d)$-th component is defined by

$$(X \times_i P)_{j_1, j_2, \cdots, j_{i-1}, j_i, j_{i+1}, \cdots, j_d} = \sum_{\ell=1}^{p_i} X_{j_1, j_2, \cdots, j_{i-1}, \ell, j_{i+1}, \cdots, j_d} P_{\ell, j_i}.$$

Let $X_{j_1, \cdots, j_{i-1}, j_i, j_{i+1}, \cdots, j_d}$ be the value of the co-cluster $(j_1, \cdots, j_{i-1}, j_i, j_{i+1}, \cdots, j_d)$ with $1 \leq j_i \leq p_i$, $i = 1, 2, ..., d$. Let an assignment matrix $Y^j \in \Re^{n_j \times p_j}$ for the indices for j-th array of tensor A be:

$$Y_{ik}^j = \begin{cases} 1, \text{ if } i \text{ is assigned to the } k\text{-th partition } I_k^j; \\ 0, \text{ otherwise.} \end{cases}$$

Then, we introduce a *proximity* measure $f(s) : \Re \to \Re_+$, with the property that $f(s) \geq 0$ for all $s \in \Re$ and $f(s) = 0$ if and only if $s = 0$. The co-clustering problem can be formulated as

$$\begin{aligned} (CC) \min \; & \sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} \cdots \sum_{j_d=1}^{n_d} \\ & f\left(A_{j_1, \cdots, j_d} - (X \times_1 Y^1 \times_2 \cdots \times_d Y^d)_{j_1, \cdots, j_d}\right) \\ \text{s.t. } & X \in \Re^{p_1 \times p_2 \times \cdots \times p_d}, \; Y^j \in \Re^{n_j \times p_j} \\ & \text{is a row assignment matrix, } j = 1, 2, ..., d \end{aligned}$$

We may consider a variety of proximity measures. For instance, if $f(s) = |s|^2$ then (CC) can be written as

$$\begin{aligned} (CC_1) \min \; & \left\| A - X \times_1 Y^1 \times_2 Y^2 \times_3 \cdots \times_d Y^d \right\|_F \\ \text{s.t. } & X \in \Re^{p_1 \times p_2 \times \cdots \times p_d}, \; Y^j \in \Re^{n_j \times p_j} \\ & \text{is a row assignment matrix, } j = 1, 2, ..., d, \end{aligned}$$

A well-known approach to the above problem is the *block descent method* [5], which, though simple to implement, fails to converge to a stationary point (local optimum). Recently this issue of convergence was resolved in [6] and the authors proposed an enhanced search algorithm termed the *maximum block improvement* (MBI) method. This method is highly effective and easy to implement, according to our experience in the co-clustering analysis for gene expression data, alongside its excellent theoretical convergence properties [40,6].

An Adaptive Co-identification Model. The power of the MBI method is now extended to solve a much more complex model - the co-identification model, where even the size of a block becomes a variable. This degree of flexibility is exactly needed in the analysis of the gene expression data, since any presumed knowledge, such as the total number of co-clusters and the number of times a gene is allowed to assign to co-clusters, would risk the blockage of key information from being revealed. By introducing the needed flexibility one has to deal with the newly introduced complications: optimization will naturally select only *one* assignment in a group, and will like to have *as many as possible* groups, notwithstanding the flexibility. To circumvent the difficulty, an enhanced model can be as follows:

$$\begin{aligned}(CI)\ \min\ & \textstyle\sum_{j_1=1}^{n_1}\sum_{j_2=1}^{n_2}\cdots\sum_{j_d=1}^{n_d} \\ & f\left(A_{j_1,\cdots,j_d} - (X\times_1 Y^1\times_2\cdots\times_d Y^d)_{j_1,\cdots,j_d}\right) \\ & +\lambda(p_1,p_2,...,p_d) - \mu(\textstyle\sum_{i,k} Y^1_{ik},...,\sum_{i,k} Y^d_{ik}) \\ \text{s.t.}\ & X\in\Re^{p_1\times p_2\times\cdots\times p_d}, \\ & Y^j\in\{0,1\}^{n_j\times p_j}\ \text{with}\ \textstyle\sum_{i,k} Y^j_{ik}\geq 1,\ j=1,...,d,\end{aligned}$$

where $\lambda(p_1,p_2,...,p_d)$ is a penalty function, intended to punish the possible abuse of more groups for identification, and $\mu(\sum_{i,k} Y^1_{ik},...,\sum_{i,k} Y^d_{ik}))$ is an *incentive* function, intended to encourage the identification of similar data in a group, without restricting to only one data per row. Some immediate choices of a penalty function include $\lambda(p_1,p_2,...,p_d) = c_1 p_1\cdots p_d$ or $\lambda(p_1,p_2,...,p_d) = c_1\sum_{i=1}^d p_i$, where c_1 is a positive constant. Similarly, choices of incentive function include: $\mu(y_1,...,y_d) = c_2\sum_{i=1}^d y_i$, where $c_2 > 0$ is another parameter. The purpose of introducing such penalty and incentive functions is to: (a) encourage the adaptiveness in the choices of the groups; and (b) avoid the introduction of too many unnecessary groups. Notice that in the new model, even the dimensions $(p_1,...,p_d)$ become a part of the decision. Such optimization models have rarely been studied in the optimization literature; however, they perfectly fit in the realm where the power of the MBI method would extend, and they are very relevant for the gene expression data analysis. The features of the new model are summarized in the following.

– By replacing co-clusters, we work with the new notion of *co-groups*, which will lead to the new *co-identification* model, allowing assigning one element to multiple co-groups under different conditions.

  The new model will characterize and model the information of different groups of genes being regulated by different transcription factors at different conditions, or the same group of genes at different conditions being regulated

by a different group of transcription factors, or the same group of genes involving in different networks and pathways.

– We develop an adaptive scheme to naturally grow the size of *co-groups*.

  Our model will naturally systematically search for the number of co-clusters for every specific gene expression datasets. One idea is to apply the MBI approach [6,40] to conduct local search for the values of the parameters $p_1, ..., p_d$ to control the number of co-groups. Methods like higher-order principal components analysis [36] and higher-order singular value decomposition (HOSVD) [22] can be applied to set the initial values of the parameters $p_1, ..., p_d$ as the start point of the local search.

– We develop a general optimization scheme for the co-identification model.

  Please refer to Figure 1 for our generic algorithm for the co-identification model based on the MBI method. The general optimization scheme of MBI is suitable for not only 2D but also multiple- and high- dimensional (3D, 4D, 5D) gene expression data.

Our co-identification model could accommodate different evaluation and objective functions. Therefore, different co-clustering approaches previously developed in the literature could be considered as special cases of our approach. Besides L_1, L_2, L_∞ [40], our model could use the Bergman divergence functions [2], where the authors chose the appropriate Bregman divergence based on the underlying data generation process or noise model. For classical clustering, Euclidean distance and Pearson correlation are both reasonable distance measures, with Euclidean distance being more appropriate for log ratio data, and Pearson correlation working better for absolute-valued data [16,10,12].

3 Results and Discussion

To simplify the testing, in the following experiments, we separate the determination of the number of co-groups from the co-identification analysis. Our approach is implemented using C++. The figures are generated using MATLAB. The testing is mainly performed on a regular PC (3GB Mem, 64bit Windows7). The running-time testing is conducted on a server (PowerEdge 2950III, 32GB Mem). We use both synthetic and real datasets to validate and evaluate the co-identification model and the generic MBI algorithm. We give a brief description of the real datasets we use to test our algorithm in this section. The 2D dataset is the yeast gene expression dataset with 2884 genes and 17 conditions. The detailed information about this dataset can be found in [7,35]. The 3D dataset is the Arabidopsis thaliana abiotic stress gene expression from [18,32]. We extract a file which has 2395 genes, 5 conditions (cold, salt, drought, wound, and heat), with each condition containing 6 time points. Due to space limit, some detailed testing results on synthetic datasets and some identified co-groups from other real datasets are not shown here.

Testing Results for Determining the Number of Co-groups. We test the MBI approach for determining the number of co-groups: We first randomly

Generic co-identification algorithm

Input: $A \in \Re^{n_1 \times n_2 \times \cdots \times n_d}$ is an d-dimensional tensor, which holds the d-dimensional gene expression data set. Parameters p_1, p_2, ..., p_d, are all positive integers, $0 < p_i \leq n_i$, $1 \leq i \leq d$.
Output: $p_1 \times p_2 \times \cdots \times p_d$ co-groups of A.

Main Variables: A non-negative integer k as the loop counter;
A $p_1 \times p_2 \times \cdots \times p_d$-tensor X with each entry a real number as the artificial central point of each of the co-groups;
A $n_i \times p_i$-matrix Y_i as the assignment matrix with $\{0, 1\}$ as the value of each entry, $1 \leq i \leq d$.

Begin
[**A.**] Start with some initial values for p_1, p_2, ..., p_d in order to control the number of co-groups, conduct local search on each p_i, $1 \leq i \leq d$.

[**A.0**] *(Initialization).* $Y^0 = X$; Choose a feasible solution $(Y_0^0, Y_0^1, Y_0^2, \cdots, Y_0^d)$ and compute the initial objective value $v_0 := f(Y_0^0, Y_0^1, Y_0^2, \cdots, Y_0^d) + c(p_1, p_2, ..., p_d; Y_0^1, Y_0^2, \cdots, Y_0^d)$. Set the loop counter $k := 0$.
[**A.1**] *(Block Improvement).* For each $i = 0, 1, 2, \ldots, d$, solve

$$\begin{aligned}(G_i)\ \max\ & f(Y_k^0, Y_k^1, \cdots, Y_k^{i-1}, Y^i, Y_k^{i+1}, \cdots, Y_k^d) + \\ & c(p_1, ..., p_d; Y_k^1, \cdots, Y_k^{i-1}, Y^i, Y_k^{i+1}, ..., Y_k^d) \\ \text{s.t.}\ \ & Y^i \in \Re^{n_j \times p_j} \text{ is an assignment matrix,}\end{aligned}$$

and let

$$\begin{aligned}y_{k+1}^i := & \arg\max f(Y_k^0, Y_k^1, ..., Y_k^{i-1}, Y^i, Y_k^{i+1}, ..., Y_k^d) \\ & + c(p_1, ..., p_d; Y_k^1, ..., Y_k^{i-1}, Y^i, Y_k^{i+1}, ..., Y_k^d) \\ w_{k+1}^i := & f(Y_k^0, Y_k^1, ..., Y_k^{i-1}, y_{k+1}^i, Y_k^{i+1}, ..., Y_k^d) \\ & + c(p_1, ..., p_d; Y_k^1, ..., Y_k^{i-1}, Y^i, Y_k^{i+1}, ..., Y_k^d).\end{aligned}$$

[**A.2**] *(Maximum Improvement).* Let $w_{k+1} := \max_{1 \leq i \leq d} w_{k+1}^i$ and $i^* = \arg\max_{1 \leq i \leq d} w_{k+1}^i$. Let

$$\begin{aligned}Y_{k+1}^i &:= Y_k^i,\ \forall\ i \in \{0, 1, 2, \cdots, d\} \backslash \{i^*\} \\ Y_{k+1}^{i^*} &:= y_{k+1}^{i^*} \\ v_{k+1} &: := w_{k+1}.\end{aligned}$$

[**A.3**] *(Stopping Criterion).* If $|v_{k+1} - v_k| > \epsilon$, set $k := k + 1$, and go to Step 1; Otherwise, set $V_{p_i} = v_{k+1}$.

[**B.**] According to the assignment matrices $Y_{k+1}^1, Y_{k+1}^2, \cdots, Y_{k+1}^d$ corresponding to the maximum V_{p_i}, $1 \leq i \leq d$, print the $p_1 \times p_2 \times \cdots \times p_d$ co-groups of A.
End

Fig. 1. Co-identification Algorithm Based on Maximum Block Improvement

generate some starting points, say the values of $(p_1, ..., p_d)$, and then we conduct a local improvement strategy, meaning that we try to increase or decrease each p_i value until no more improvement is possible locally. We refer the reader to Table 1 for our testing on the effectiveness of the proposed local search strategy.

Table 1. Testing of the Maximum Block Improvement strategy on the 2D yeast dataset from [35]. The initial objective function value is -25900; the first column: the initial p values; the second column: the new p values after the local search; the third column: the objective function values with the initial p values and with the new p values respectively; and the last column: the running time for the local search.

Initial $p_{1,0}$, $p_{2,0}$	New p_1, p_2	Obj-value (Initial value -25900)	Run Time (seconds)
20,10	25,16	-7192.42, -6810.89	646.91
5,8	13,9	-9291.28, -7498.96	341.67
97,10	96,11	-6706.33, -6220.54	364.64
68,8	69,11	-6763.02, -6337.49	441.18
32,9	35,15	-6967.74, -6591.19	624.38
20,11	25,16	-7202.46, -6810.89	609.70
19,4	28,6	-7480.57, -7041.12	487.72
51,3	50,5	-7157.43, -6808.80	224.95
65,9	64,11	-6736.08, -6413.91	349.50
43,1	42,5	-7888.83, -6832.58	271.23
6,3	11,5	-8918.90, -7677.44	202.40
2,4	11,5	-15973.50, -7677.44	280.21

Coherent Groups from 3D Arabidopsis Gene Expression Data. To demonstrate the effectiveness of our algorithm in search for coherent patterns from gene expression datasets, Figure 2 provides several exemplary 3D co-groups identified from the 3D Arabidopsis dataset. We present here the co-groups with a small number of genes, which shows clear coherent expression pattern over a series of time points and under different conditions. These co-groups clearly facilitate further functional analysis of the genes. The analysis of 3D *Arabidopsis* dataset in [32] has generated three biologically relevant co-cluster module types: 1) modules with genes that are co-regulated under several conditions are the most prevalent ones, 2) Coherent modules with similar responses under all conditions occurred frequently, too, 3) A third module type, which covers a response specific to a single condition was also detected, but rarely. Especially for the third module type, refer to the top-left pattern of Figure 2, which shows the two Arabidopsis genes are co-regulated at all 3 conditions: cold, salt, drought, but are differently expressed at the condition heat. The two genes are: 250296_at and 245955_at. The following information of the two genes is from http://www.arabidopsis.org/: gene 250296_at: 17.6 kDa class II heat shock protein (HSP17.6-CII), identical to 17.6 kDa class II heat shock protein SP:P29830 from (Arabidopsis thaliana); gene 245955_at: glycosyl hydrolase family 1 protein, contains Pfam PF00232 : Glycosyl hydrolase family 1 domain, TIGRFAM TIGR01233: 6-phospho-beta-galactosidase, similar to beta-glucosidase 1

(GI:12043529) (Arabidopsis thaliana). *Gene 250296_at (green-colored on the figure) with significantly high expression at the heat condition, which is identified by our approach, is confirmed coding for a heat-shock protein.*

Results from Yeast Gene Expression Data. We apply our co-identification approach to the analysis of the Saccharomyces cerevisiae gene expression data collected in [35]. This data contains the expression of 2884 genes under 17 conditions. From our co-identification analysis, we identify genes that are co-listed in two or more co-groups, such as YER068W, YDR103W, YGL130W, YJR129C, YLR425W, YOR383C and YLL004W. This information could be used to predict the functions of unknown genes from the known functions of the genes in the same co-groups. This information could also lead to identification of previously undetected novel functions of genes. Specifically we have checked the function information (http://www.yeastgenome.org) of the following two genes which are involved in more than one co-group.

Gene ORC3/YLL004W: Subunit of the origin recognition complex, which directs DNA replication by binding to replication origins and is also involved in transcriptional silencing. We find out three pathways from KEGG Pathway Database (http://www.genome.jp/kegg/) in which this gene are involved.

Gene STE5/YDR103W: Pheromone-response scaffold protein that controls the mating decision; binds Ste11p, Ste7p, and Fus3p kinases, forming a MAPK cascade complex that interacts with the plasma membrane and Ste4p-Ste18p; allosteric activator of Fus3p. Our approach identifies two co-groups in which Gene STE5/YDR103W is involved. The two co-groups are biologically significant with low p-values: Co-group#41 (6 genes: STE5 ADA2 AFG3 MOT2 PHO23 DSS4), zinc ion binding, with p-value 0.001735, Co-group#62 (4 genes: STE5 MOT2 ORC3 RAD52), pheromone response, mating-type determination, sex-specific proteins, with p-value 0.007773 (The p-value information is obtained from the website of Funcspec: http://funspec.med.utoronto.ca/).

Running Time Analysis. Our approach is highly efficient and could be applied to 2D, 3D and higher-dimensional gene expression data. When testing on 3D datasets (genes vs time points vs conditions), the running time of our approach increases linearly with the number of genes, the number of time points, or the number of conditions. We conduct our running-time testing on the Arabidopsis thaliana abiotic stress 3D gene expression datasets from [32]. We use a file which has 2395 genes, 5 conditions (cold, salt, drought, wound, and heat), with each condition containing 6 time points. Especially when we keep the number of genes (2395) and increase the second dimension for the number of time points, or the third dimension for the number of the conditions, we increase the size of the dataset significantly, however, the running time of our algorithm still has only a linear increase (Figure 3). The performance of our algorithm is very robust. The testing results demonstrate the high efficiency of our algorithm. In contrast, other existing methods for 3D co-clustering such as *TriCluster* [41], the running time is exponential with the number of time points, or the number of conditions. Other existing methods usually do not work for gene expression data of four- or higher- dimensions.

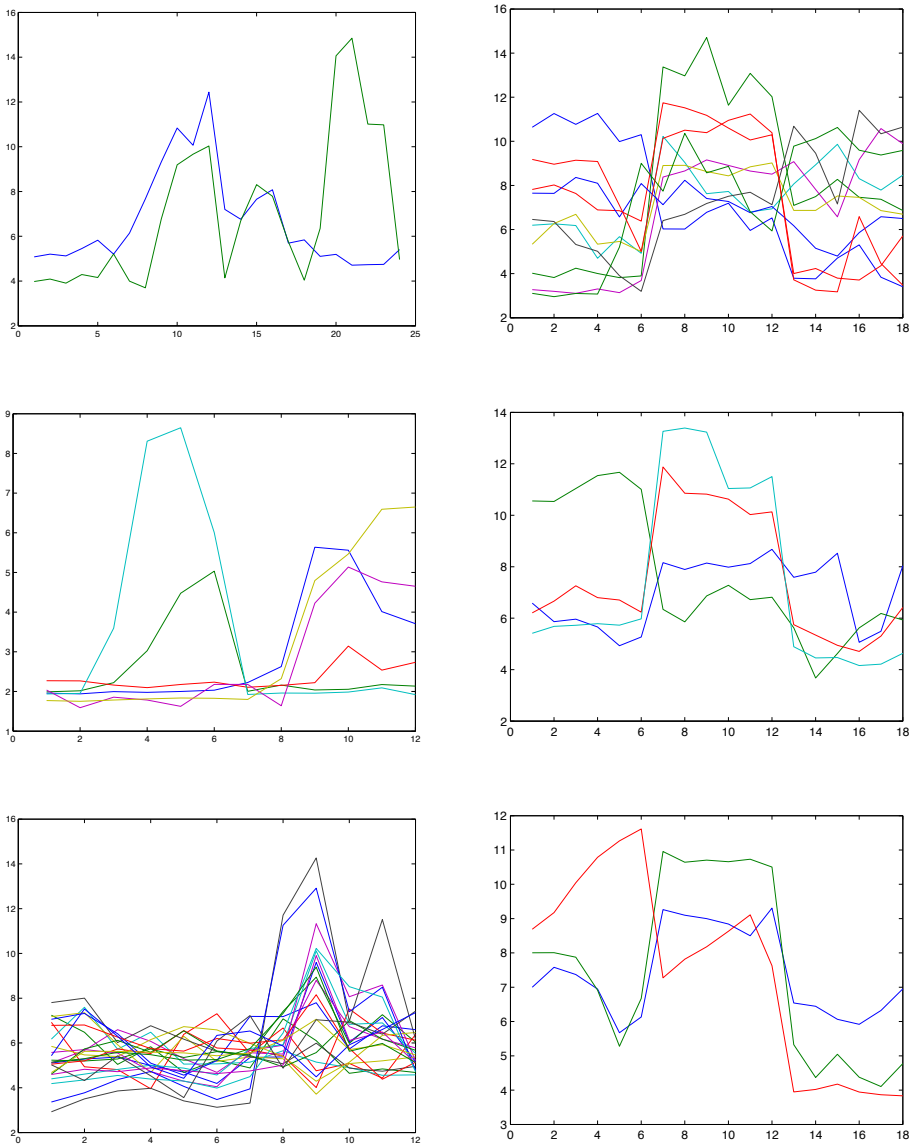

Fig. 2. Exemplary 3D co-groups (with no. of genes x 6 time points x no. of conditions) generated from the 3D Arabidopsis dataset. Genes have different expression patterns at different conditions. The x-axis represents the different number of time points (with every 6 time-points in one condition), while the y-axis represents the values of the gene expression level. Each curve corresponds to the expression of one gene. For example, the co-group at the top-left shows the clear expression patterns of 2 genes at 4 conditions (cold, salt, drought and heat).

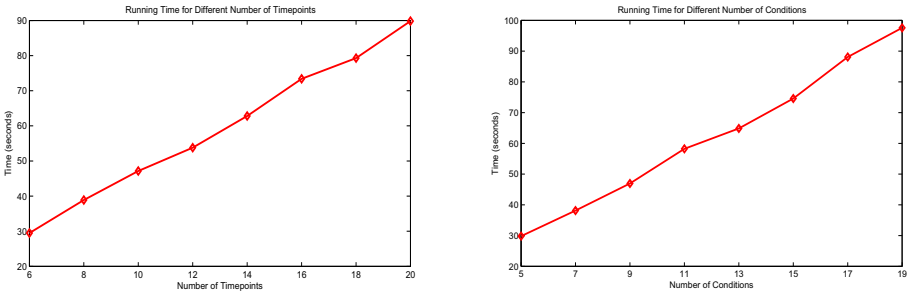

Fig. 3. Evaluation of our approach on the 3D Arabidopsis dataset (2396 genes x 6 timepoints x 5 conditions). The parameters to control the number of co-groups for all these evaluations are set the same: $p_1 = 100, p_2 = 2, p_3 = 3$. For testing on different number of genes (this figure not shown due to space limit), the sizes of the 8 groups are 300x6x5, 600x6x5, 900x6x5, 1200x6x5, 1500x6x5, 1800x6x5, 2100x6x5, 2400x6x5 (these datasets are truncated from the original dataset). For testing on different number of time points (Figure on the left), the sizes of the 8 groups are 2396x6x5, 2396x8x5, 2396x10x5, 2396x12x5, 2396x14x5, 2396x16x5, 2396x18x5, 2396x20x5 (except for the first group which is the original dataset, the other 7 groups contain added repetitive time points). For testing on different number of conditions (Figure on the right), the sizes of the 8 groups are 2396x6x5, 2396x6x7, 2396x6x9, 2396x6x11, 2396x6x13, 2396x6x15, 2396x6x17, 2396x6x19 (except for the first group which is the original dataset, the other 7 groups contain added repetitive conditions).

4 Summary

In this work, for complex high-dimensional gene expression data clustering analysis, we introduce the new notion of co-identification, so that we may assign one element to different *groups* or *co-groups* to enable systematical identification of multiple functions of one gene or the involvement in multiple functional groups of one element under different conditions or different development stages. We build a scalable model not only for 2D but also for high-dimensional gene expression data, and develop a general adaptive scheme based on Maximum Block Improvement to solve the model, which could naturally grow the size of the co-groups and encourage the degree of co-identification. We apply the unified adaptive co-identification analysis to real gene expression datasets, which shows the high efficiency and effectiveness of the approach. The running time of our approach increases linearly with the number of genes, the number of time points, or the number of conditions. When applied to real gene expression data, our approach could lead to identification of previously undetected novel functions of genes. Our approach has identified a differentially expressed heat-shock gene that are co-regulated with other genes under three other conditions (cold, salt, and drought) of 18 time points from the *Arabidopsis* dataset (this type of patterns are considered important and rare by the EDSIA method [32]), and also identified genes that participate different biologically significant functional groups from the yeast dataset. The co-identification analysis could possibly enrich further functional study of important biological processes, which may lead to new insights into genome-wide gene expression of the cell.

Our approach is a unified systematic approach, which could be used for high-dimensional gene expression data analysis (as well as for high-dimensional data analysis for applications of other fields). There are few current approaches which efficiently work for datasets with dimensions greater than 3. Our approach is general enough to embrace many other clustering and biclustering methods proposed in the literature as special cases. Especially our framework could apply as evaluation or objective functions the 6 different schemes listed in [2]. Our co-identification model provides the framework for incorporating additional ideas from approximation [20,19], parameterization [13], randomization and probabilistic analysis, or approaches combined with statistic and greedy strategies.

References

1. Aguilar-Ruiz, J.S.: Shifting and scaling patterns from gene expression data. Bioinformatics 21, 3840–3845 (2005)
2. Banerjee, A., et al.: A generalized maximum entropy approach to bregman coclustering and matrix approximation. JMLR 8, 1919–1986 (2007)
3. Ben-Dor, A., et al.: Discovering local structure in gene expression data: the order-preserving submatrix problem. In: RECOMB 2002, pp. 49–57 (2002)
4. Ben-Hur, A., et al.: A stability based method for discovering structure in clustered data. In: Proc. of PSB (2002)
5. Bertsekas, D.P.: Nonlinear Programming. Athena Scientific, Belmont (1999)
6. Chen, B., et al.: Maximum block improvement and polynomial optimization. SIAM Journal on Optimization 22, 87–107 (2012)
7. Cheng, Y., Church, G.M.: Biclustering of expression data. In: Proc. Int. Conf. Intell. Syst. Mol. Biol., vol. 8, pp. 93–103 (2000)
8. Cheung, A.N.: Molecular targets in gynaecological cancers. Pathology 39, 26–45 (2007)
9. Cho, H., et al.: Minimum sum-squared residue co-clustering of gene expression data. In: Proc. SIAM on Data Mining, pp. 114–125 (2004)
10. Costa, I.G., et al.: Comparative analysis of clustering methods for gene expression time course data. Genet. Mol. Biol. 27, 623–631 (2004)
11. Deodhar, M., et al.: Hunting for Coherent Co-clusters in High Dimensional and Noisy Datasets. In: IEEE Intl. Conf. on Data Mining Workshops (2008)
12. D'haeseleer, P.: How does gene expression clustering work? Nature Biotechnology 23, 1499–1501 (2005)
13. Downey, R.G., Fellows, M.R.: Parameterized Complexity. Springer (1999)
14. Dudoit, S., Fridlyand, J.: A prediction based resampling method for estimating the number of clusters in a data set. Genome Biology 3, 1–21 (2002)
15. Eisen, M.B., et al.: Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. 95, 14863–14868 (1998)
16. Gibbons, F.D., Roth, F.P.: Judging the quality of gene expression-based clustering methods using gene annotation. Genome Res. 12, 1574–1581 (2002)
17. Hochreiter, S., et al.: FABIA: factor analysis for bicluster acquisition. Bioinformatics 26, 1520–1527 (2010)
18. Kilian, J., et al.: The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. The Plant Journal 2, 347–363 (2007)

19. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Review 51, 455–500 (2009)
20. Jegelka, S., Sra, S., Banerjee, A.: Approximation Algorithms for Tensor Clustering. In: Gavaldà, R., Lugosi, G., Zeugmann, T., Zilles, S. (eds.) ALT 2009. LNCS, vol. 5809, pp. 368–383. Springer, Heidelberg (2009)
21. Jiang, D., et al.: Mining coherent gene clusters from gene-sample-time microarray data. In: Proc. ACM SIGKDD, pp. 430–439 (2004)
22. Lathauwer, D., et al.: A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. 21, 1253–1278 (2000)
23. Lazzeroni, L., Owen, A.B.: Plaid models for gene expression data. Statistica Sinica 12, 61–86 (2002)
24. Lee, M., et al.: Biclustering via Sparse Singular Value Decomposition. Biometrics 66, 1087–1095 (2010)
25. Li, A., Tuck, D.: An Effective Tri-Clustering Algorithm Combining Expression Data with Gene Regulation. Gene Regulation and Systems Biology 3, 49–64 (2009)
26. Madeira, S.C., Oliveira, A.L.: Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans. Comput. Biology Bioinform. 1, 24–45 (2004)
27. Magic, Z., et al.: cDNA microarrays: identification of gene signatures and their application in clinical practice. J. BUON 12(suppl.1), S39–S44 (2007)
28. Murali, T., Kasif, S.: Extracting conserved gene expression motifs from gene expression data. In: Pacific Symposium on Biocomputing, vol. 8, pp. 77–88 (2003)
29. Prelic, A., et al.: A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics 22, 1122–1129 (2006)
30. Snider, N., Diab, M.: Unsupervised Induction of Modern Standard Arabic Verb Classes. In: HLT-NAACL, New York (2006)
31. Strauch, M., et al.: A Two-Step Clustering for 3-D Gene Expression Data Reveals the Main Features of the Arabidopsis Stress Response. J. Integrative Bioinformatics 4, 54–66 (2007)
32. Supper, J., et al.: EDISA: extracting biclusters from multiple time-series of gene expression profiles. BMC Bioinformatics 8, 334–347 (2007)
33. Suter, L., et al.: Toxicogenomics in predictive toxicology in drug development. Chem. Biol. 11, 161–171 (2004)
34. Tamayo, P., et al.: Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA 96, 2907–2912 (1999)
35. Tavazoie, S., et al.: Systematic determination of genetic network architecture. Nat. Genet. 22, 281–285 (1999)
36. Tucker, L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika 31, 279–311 (1966)
37. Tibshirani, R., et al.: Estimating the Number of Clusters in a Dataset via the Gap Statistic. J. Royal Stat. Soc. B 63, 411–423 (2001)
38. Wang, H., et al.: Clustering by pattern similarity in large data sets. In: Proc. KDD 2002, pp. 394–405 (2002)
39. Xu, X., et al.: Mining shifting-and-scaling co-regulation patterns on gene expression profiles. In: Proc. ICDE 2006, pp. 89–98 (2006)
40. Zhang, S., Wang, K., Chen, B., Huang, X.: A New Framework for Co-clustering of Gene Expression Data. In: Loog, M., Wessels, L., Reinders, M.J.T., de Ridder, D. (eds.) PRIB 2011. LNCS, vol. 7036, pp. 1–12. Springer, Heidelberg (2011)
41. Zhao, L., Zaki, M.J.: Tricluster: an effective algorithm for mining coherent clusters in 3D microarray data. In: Proc. ACM SIGMOD, pp. 694–705 (2005)