Tchagang *et al. BMC Bioinformatics* 2012, **13**:54
http://www.biomedcentral.com/1471-2105/13/54

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Mining biological information from 3D short time-series gene expression data: the OPTricluster algorithm

Alain B Tchagang[1*], Sieu Phan[1], Fazel Famili[1], Heather Shearer[2], Pierre Fobert[2], Yi Huang[2,3], Jitao Zou[2], Daiqing Huang[2], Adrian Cutler[2], Ziying Liu[1] and Youlian Pan[1]

Abstract

Background: Nowadays, it is possible to collect expression levels of a set of genes from a set of biological samples during a series of time points. Such data have three dimensions: gene-sample-time (GST). Thus they are called 3D microarray gene expression data. To take advantage of the 3D data collected, and to fully understand the biological knowledge hidden in the GST data, novel subspace clustering algorithms have to be developed to effectively address the biological problem in the corresponding space.

Results: We developed a subspace clustering algorithm called Order Preserving Triclustering (OPTricluster), for 3D short time-series data mining. OPTricluster is able to identify 3D clusters with coherent evolution from a given 3D dataset using a combinatorial approach on the sample dimension, and the order preserving (OP) concept on the time dimension. The fusion of the two methodologies allows one to study similarities and differences between samples in terms of their temporal expression profile. OPTricluster has been successfully applied to four case studies: immune response in mice infected by malaria (*Plasmodium chabaudi*), systemic acquired resistance in *Arabidopsis thaliana*, similarities and differences between inner and outer cotyledon in *Brassica napus* during seed development, and to *Brassica napus* whole seed development. These studies showed that OPTricluster is robust to noise and is able to detect the similarities and differences between biological samples.

Conclusions: Our analysis showed that OPTricluster generally outperforms other well known clustering algorithms such as the TRICLUSTER, gTRICLUSTER and K-means; it is robust to noise and can effectively mine the biological knowledge hidden in the 3D short time-series gene expression data.

Background

Clustering of co-expressed genes has been an active data mining topic and advanced in parallel with the development of microarray technology [1]. There is a vast amount of literature on clustering algorithms developed for microarray data analysis [1]. Microarray gene expression data can be classified into two categories: steady state and time-series gene expression data [2]. Time-series gene expression data are widely used to study the dynamic behaviour of various biological processes in the cell [3-5]. They can be classified into two categories (relative to the clustering algorithms design for their analysis): short time-series corresponding to 3-8 time points [6], and long time-series corresponding to more than 8 time points. Short time-series are the most abundant type of time-series data in the literature [6]. Short-time series data are usually very noisy. Algorithms that are designed to analyze either steady state data or long time-series data do not perform well on short time-series data due to their relatively small number of time points [5-7]. Hence, it is necessary to develop algorithms that can be used specifically for their analysis.

Pioneering clustering algorithms such as K-means [8], Hierarchical clustering [9], and Self Organizing Map [10] identify full space clusters. Unfortunately, in many applications, subspace clusters are more meaningful than full space clusters [11]. Biclustering algorithms were recently proposed to find subgroups of genes that

* Correspondence: alain.tchagang@nrc-cnrc.gc.ca
[1]Knowledge Discovery Group, Institute for Information Technology, National Research Council Canada, 1200 Montréal Road, Ottawa, ON K1A 0R6, Canada
Full list of author information is available at the end of the article

exhibit similar behaviour across subsets of samples, experimental conditions, or time points [11-15]. Nowadays, it is possible to collect expression levels of a set of genes for a given set of biological samples during a series of time points. Such data have three dimensions, gene-sample-time (GST), and thus are called 3D gene expression data. To take advantage of the 3D data collected, and to fully understand the biological knowledge hidden in the GST data, we have to move beyond the full space clustering concepts and develop algorithms that can effectively address the problem in the corresponding 3 dimensions [16-18]. A 3D cluster consists of a subset of genes of similar expression profiles along a segment of time-series, in a subset of samples. This kind of coherent clusters may contain information that could be used to identify useful phenotypes, potential genes related to these phenotypes and their interaction/regulation.

Although subspace clustering algorithms are biologically more meaningful than full space clustering algorithms, the identification of full space clusters are less costly compared to subspace clusters. In fact, most subspace clustering algorithms have been shown to be NP-complete [11], thus making their identification computationally expensive. This is due to the fact that in full space clustering, one usually looks for clusters across the entire dataset at once, whereas in subspace clustering one looks for clusters across all possible subsets of the data space.

Most clustering models, including those used in subspace clustering described above, define similarity among different objects by *distances* over either all or only a subset of the dimensions [1]. However, distance functions, such as Euclidean distance, Manhattan distance, and cosine distance are not always adequate in capturing correlations among the objects. In fact, strong correlations may still exist among a set of objects even if they are far apart from each other as measured by the distance functions [1]. Pioneering works on triclustering algorithm relied on graph-based approaches and similarity measures to mine triclusters [16-18]. For example, the triclustering algorithm of [16] mines the largest triclusters satisfying a constant multiplicative or additive relationship between the expression levels in a cluster. Such a strict constraint considerably limits the capability of an algorithm to identify useful patterns and may not be able to fully cope with noise when dealing with time-series gene expression data in general. Several of the algorithms designed for the analysis of long time-series - do not work well on short time-series due to over fitting [6]. Also, most pioneering triclustering algorithms described in the literature only focus on the similarities between the biological samples and do not consider differences between them.

In this paper, we developed an order preserving 3D clustering algorithm named OPTricluster, for 3D short time-series data mining. A 3D short time series corresponds to GST with 2-5 samples and 3-8 time points. In fact most of the 3D time series gene expression data in the GEO database [19] have less than 5 samples and less than 9 time points. OPTricluster is able to identify triclusters (3D clusters) with coherent evolution from a given 3D dataset using a combinatorial approach on the sample dimension, and the order preserving (OP) concept on the time dimension. We say that a matrix is order preserving if there exists a permutation of its columns such that its rows are monotonic functions [13]. The OP concept is a well established mathematical theory, and has been used by several other authors in the past to tackle the problem of gene expression data analysis [13,14,20,21].

The integration of the two methodologies (combinatorial and OP) allows one to study similarities and differences between samples in terms of temporal expression profile and/or differential expression. Basically, the novel triclustering algorithm is able to mine triclusters of genes in a subset of samples, with expression level having same directions (i.e. increase, decrease and/or stay constant similarly) across the time-series experiments. OPTricluster takes into account the sequential nature of the time-series and the noisy nature of the data through the OP concept. The OP model focuses on the similarity in the relative order of the time points, rather than on the distance between the actual expression levels as in several other clustering models [7-10]. This approach is potentially more robust to the stochastic nature of the gene expression levels, and to the variation caused by the measurements procedures.

Results

OPTricluster is applied to analyze four different 3D gene expression datasets: immune response in mice infected by malaria (*Plasmodium chabaudi*), systemic acquired resistance (SAR) in *Arabidopsis thaliana*, similarities and differences between inner and outer cotyledon in *Brassica napus* during seed development, and *Brassica napus* whole seed development. We used these four datasets to show how OPTricluster can be used to tackle a variety of problems in Computational Biology. For example, application of OPTricluster to the mouse dataset shows how it can be used to study similarities and differences between biological samples in terms of temporal expression profiles. Application to the *Arabidopsis thaliana* SAR dataset shows how it can be used not only to study similarities and differences between biological samples, but also to infer the transcription network, that is, the relationship between the transcription factors (TFs) and their target genes. Application to the *Brassica napus* cotyledon dataset shows how OPTricluster can be used to study the spatial similarities and differences between biological samples in term of temporal expression profile, whereas its application to the *Brassica napus* whole seed dataset shows how OPTricluster can be used to handle classical short time-series dataset (2 dimensional dataset).

Implementation

OPTricluster is implemented entirely in Java and will work with any operating system supporting Java 1.6 or later. The executable OPTricluster Java package is available as Additional File 1, and its user manual as Additional File 2. Portions of the interface of OPTricluster are implemented using a third party library, the JFreeChart [22]. In a post processing step, OPTricluster also makes use of external Gene Ontology files. OPTricluster can download the Gene Ontology and gene annotation files directly from the Gene Ontology websites. In fact the GO analysis plug-in of the Gene Ontology Analysis (GOAL) [23] package that we recently developed is integrated into OPTricluster for biological evaluation of the clusters. A user of OPTricluster first specifies a tab delimited gene expression data file as input to OPTricluster. The file is uploaded and displayed as a table on the screen. The user can check their data and do other manipulations on the data once it is displayed. This includes visual check, plot and view the distribution of the input gene expression data, which may help in deciding the best ranking threshold input parameter (see Methodology Section). Next, the user selects the analysis parameters through the OPTricluster parameter input interface. Following this input phase, the OPTricluster executes and a new table will appear displaying the clustering results, where new columns are added to the initial table and they correspond to the ranking profile of each gene across experimental time points in each sample respectively. From this step, the user will either select between constant, conserved, and divergent patterns for further analysis as explained below.

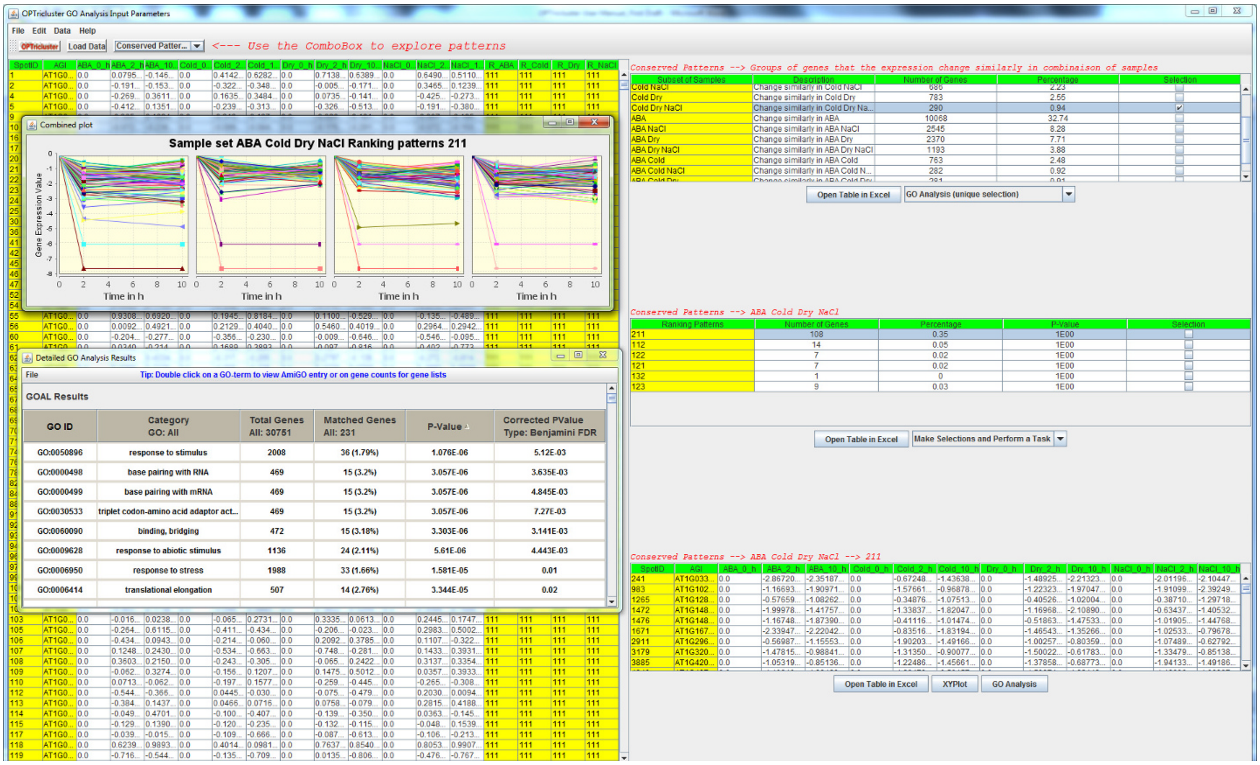
There are three types of patterns that the user can choose from: conserved patterns, divergent patterns, and constant patterns. Conserved patterns correspond to group of genes having same behaviour across experimental time points in subsets of samples. Divergent patterns correspond to group of genes that behave differently in at least one sample along the time point experiments. Constant patterns correspond to groups of genes that the expression levels do not change across experimental time points. Once the type of patterns is selected, a new table appears on the screen, describing the subset of samples the number of genes with the selected patterns in each subset of samples. By clicking on a subset of sample, another table appears showing how the genes are grouped in the selected patterns and in the corresponding subset of samples. Finally the user can view the genes in each group as a table by clicking on the corresponding ranking profile. Figure 1 shows an example of the analysis of a conserved pattern. We refer the reader to the OPTricluster user manual (Additional File 2) for more details relative to OPTricluster Software.

Note that, at each step of the analysis described above, OPTricluster offers the possibility to the user to perform several other tasks and gain more knowledge on the results obtained at the corresponding step. These manipulations include: open the table in Excel and use the Excel capabilities for other manipulations. Plot and view the pie chart or the bar chart of the results. Merge cluster with similar profile for further analysis. Obtain the difference between patterns in different subsets of samples. Perform and visualize the gene ontology enrichment analysis of the group of genes obtained at that particular step.

Simulation and robustness to noise

To test the robustness of OPTricluster to noise, we used the adjusted rand index (ARI). ARI has been used previously for clustering techniques comparison and robustness to noise [18]. ARI's values lie between 0 and 1, where larger value means higher similarity between the clustering results. If the experimental result is perfectly consistent with the domain knowledge (known triclusters embedded in the simulated dataset), the index value will be 1. If a clustering is no more than a random choice, the index will be 0.

We generated a synthetic 3D microarray dataset, consisting of N = 1000 genes, M = 4 samples, and L = 3 time points, with four order preserving triclusters across the entire four samples imbedded in it. Each tricluster had 20, 25, 30, and 100 genes respectively and there was no overlap between the four triclusters. Thus the domain knowledge corresponds to the four embedded triclusters. We added 0%, 2%, 5%, 10%, and 15% noise to the original dataset and computed the ARI values between the triclustering results on noisy datasets and the domain knowledge. Such a process was performed using OPTricluster a modified version of TRICLUSTER [16] gTRISCLUSTER [18], and the K-means algorithm. For the K-means algorithm the 3D dataset is converted into a 2D dataset as follows. The N × M × L dataset becomes N × ML dataset. For each level of noise, each algorithm was run five times and the mean of the ARI obtained were represented in Figure 2. It can be seen that the ARI values of OPTricluster are larger than those of modified gTRICLUSTER, modified TRICLUSTER, and the K-means algorithm. This means that the OPTricluster algorithm developed in this study performs better on order preserving patterns, and it is more robust to noise than modified TRICLUSTER, modified gTRICLUSTER, and the K-means algorithm. Comparative analyses were only performed by studying similarity (same OP patterns) between samples because TRICLUSTER, gTRICLUSTER, and the K-means algorithm cannot investigate differences between samples, for which, only the proposed OPTricluster does.



<---- Use the ComboBox to explore patterns

SpotID	AGI	ABA_0_h	ABA_2_h	ABA_10...	Cold_0...	Cold_2...	Cold_1...	Dry_0_h	Dry_2_h	Dry_10...	NaCl_0...	NaCl_2...	NaCl_1...	R_ABA	R_Cold	R_Dry	R_NaCl
1	AT1G0...	0.0	0.0795...	-0.146...	0.0	0.4142...	0.6282...	0.0	0.7138...	0.6389...	0.0	0.6490...	0.5110...	111	111	111	111
2	AT1G0...	0.0	-0.191...	-0.153...	0.0	-0.322...	-0.348...	0.0	-0.005...	-0.171...	0.0	0.3465...	0.1239...	111	111	111	111
4	AT1G0...	0.0	-0.269...	0.3611...	0.0	0.1635...	0.3484...	0.0	0.0735...	-0.141...	0.0	-0.425...	-0.273...	111	111	111	111
5	AT1G0...	0.0	-0.412...	0.1351...	0.0	-0.239...	-0.313...	0.0	-0.326...	-0.513...	0.0	-0.191...	-0.380...	111	111	111	111
55	AT1G0...	0.0	0.9308...	0.6920...	0.0	0.1945...	0.8184...	0.0	0.1100...	-0.529...	0.0	-0.135...	-0.489...	111	111	111	111
56	AT1G0...	0.0	0.0092...	0.4921...	0.0	0.2129...	0.4040...	0.0	0.5460...	0.4019...	0.0	0.2964...	0.2942...	111	111	111	111
60	AT1G0...	0.0	-0.204...	-0.277...	0.0	-0.356...	-0.230...	0.0	-0.009...	-0.646...	0.0	-0.546...	-0.095...	111	111	111	111
103	AT1G0...	0.0	-0.016...	0.0238...	0.0	-0.065...	0.2731...	0.0	0.3335...	0.0613...	0.0	0.2445...	0.1747...	111	111	111	111
105	AT1G0...	0.0	-0.264...	0.6115...	0.0	-0.411...	-0.434...	0.0	-0.206...	-0.023...	0.0	0.2983...	0.5002...	111	111	111	111
106	AT1G0...	0.0	-0.434...	0.0943...	0.0	-0.214...	-0.060...	0.0	0.2092...	0.3785...	0.0	0.1107...	-0.322...	111	111	111	111
107	AT1G0...	0.0	0.1248...	0.2430...	0.0	-0.534...	-0.663...	0.0	-0.748...	-0.281...	0.0	0.1433...	0.3931...	111	111	111	111
108	AT1G0...	0.0	0.3603...	0.2150...	0.0	-0.243...	-0.305...	0.0	-0.065...	0.2422...	0.0	0.3137...	0.3354...	111	111	111	111
109	AT1G0...	0.0	-0.062...	0.3274...	0.0	-0.156...	0.1207...	0.0	0.1475...	0.5012...	0.0	0.0357...	0.3933...	111	111	111	111
110	AT1G0...	0.0	0.0713...	-0.062...	0.0	-0.197...	0.1577...	0.0	-0.259...	-0.445...	0.0	-0.265...	-0.308...	111	111	111	111
112	AT1G0...	0.0	-0.544...	-0.366...	0.0	0.0445...	-0.030...	0.0	-0.075...	-0.479...	0.0	0.2030...	0.0094...	111	111	111	111
113	AT1G0...	0.0	-0.384...	0.1437...	0.0	0.0466...	0.0716...	0.0	0.0758...	-0.079...	0.0	0.2815...	0.4188...	111	111	111	111
114	AT1G0...	0.0	-0.049...	0.4701...	0.0	-0.100...	-0.407...	0.0	-0.139...	-0.350...	0.0	0.0363...	-0.145...	111	111	111	111
115	AT1G0...	0.0	-0.129...	0.1390...	0.0	-0.120...	-0.235...	0.0	-0.132...	-0.115...	0.0	-0.048...	0.1539...	111	111	111	111
117	AT1G0...	0.0	-0.039...	-0.015...	0.0	-0.109...	-0.666...	0.0	-0.087...	-0.613...	0.0	-0.106...	-0.213...	111	111	111	111
118	AT1G0...	0.0	0.6239...	0.9893...	0.0	0.4014...	0.0981...	0.0	0.7637...	0.8540...	0.0	0.8053...	0.9907...	111	111	111	111
119	AT1G0...	0.0	-0.716...	-0.544...	0.0	-0.135...	-0.709...	0.0	0.0135...	-0.806...	0.0	-0.476...	-0.767...	111	111	111	111

Sample set ABA Cold Dry NaCl Ranking patterns 211

Tip: Double click on a GO-term to view AmiGO entry or on gene counts for gene lists

GOAL Results

GO ID	Category GO: All	Total Genes All: 30751	Matched Genes All: 231	P-Value	Corrected PValue Type: Benjamini FDR
GO:0050896	response to stimulus	2008	36 (1.79%)	1.076E-06	5.12E-03
GO:0000498	base pairing with RNA	469	15 (3.2%)	3.057E-06	3.635E-03
GO:0000499	base pairing with mRNA	469	15 (3.2%)	3.057E-06	4.845E-03
GO:0030533	triplet codon-amino acid adaptor act...	469	15 (3.2%)	3.057E-06	7.27E-03
GO:0060090	binding, bridging	472	15 (3.18%)	3.303E-06	3.141E-03
GO:0009628	response to abiotic stimulus	1136	24 (2.11%)	5.61E-06	4.443E-03
GO:0006950	response to stress	1988	33 (1.66%)	1.581E-05	0.01
GO:0006414	translational elongation	507	14 (2.76%)	3.344E-05	0.02

Conserved Patterns --> Groups of genes that the expression change similarly in combinaison of samples

Subset of Samples	Description	Number of Genes	Percentage	Selection
Cold NaCl	Change similarly in Cold NaCl	686	2.23	□
Cold Dry	Change similarly in Cold Dry	783	2.55	□
Cold Dry NaCl	Change similarly in Cold Dry Na...	290	0.94	☑
ABA	Change similarly in ABA	10068	32.74	□
ABA NaCl	Change similarly in ABA NaCl	2545	8.28	□
ABA Dry	Change similarly in ABA Dry	2370	7.71	□
ABA Dry NaCl	Change similarly in ABA Dry NaCl	1193	3.88	□
ABA Cold	Change similarly in ABA Cold	763	2.48	□
ABA Cold NaCl	Change similarly in ABA Cold N...	282	0.92	□

Conserved Patterns --> ABA Cold Dry NaCl

Ranking Patterns	Number of Genes	Percentage	P-Value	Selection
211	108	0.35	1E00	□
112	14	0.05	1E00	□
122	7	0.02	1E00	□
121	7	0.02	1E00	□
132	1	0	1E00	□
123	9	0.03	1E00	□

Conserved Patterns --> ABA Cold Dry NaCl --> 211

SpotID	AGI	ABA_0_h	ABA_2_h	ABA_10_h	Cold_0_h	Cold_2_h	Cold_10_h	Dry_0_h	Dry_2_h	Dry_10_h	NaCl_0_h	NaCl_2_h	NaCl_10_h
241	AT1G033...	0.0	-2.86720...	-2.35187...	0.0	-0.67248...	-1.43638...	0.0	-1.48925...	-2.21323...	0.0	-2.01196...	-2.10447...
983	AT1G102...	0.0	-1.16693...	-1.90971...	0.0	-1.57661...	-0.96878...	0.0	-1.22323...	-1.97047...	0.0	-1.91099...	-2.39249...
1265	AT1G128...	0.0	-0.57659...	-1.08262...	0.0	-0.34876...	-1.07513...	0.0	-0.40526...	-1.02004...	0.0	-0.38710...	-1.29718...
1472	AT1G148...	0.0	-1.99978...	-1.41757...	0.0	-1.33837...	-1.82047...	0.0	-1.16968...	-2.10890...	0.0	-0.63437...	-1.40532...
1476	AT1G148...	0.0	-1.16748...	-1.87390...	0.0	-0.41116...	-1.01474...	0.0	-0.51863...	-1.47533...	0.0	-1.01905...	-1.44768...
1671	AT1G167...	0.0	-2.33947...	-2.22042...	0.0	-0.83516...	-1.83194...	0.0	-1.46543...	-1.35266...	0.0	-1.02533...	-0.79678...
2911	AT1G296...	0.0	-0.56987...	-1.15553...	0.0	-1.90203...	-1.49166...	0.0	-1.00257...	-0.80359...	0.0	-1.07489...	-0.62792...
3179	AT1G320...	0.0	-1.47815...	-0.98841...	0.0	-1.31350...	-0.90077...	0.0	-1.50022...	-0.61783...	0.0	-1.33479...	-0.85138...
3885	AT1G420...	0.0	-1.05319...	-0.85136...	0.0	-1.22486...	-1.45661...	0.0	-1.37858...	-0.68773...	0.0	-1.94133...	-1.49186...

Figure 1 OPTricluster software. Example of OPTricluster interface showing a conserved cluster and its Gene Ontology analysis.

Immune responses in mice infected by malaria

The goal of this study is to examine whether immune responses to malaria (*Plasmodium chabaudi*) infection differ between the sexes and are altered by the presence of gonadal steroids. To tackle this problem, we used a 3D short time-series gene expression dataset downloaded from the Gene Expression Omnibus website [19], (accession number: GSE4324). This dataset has N = 33935 probes and it corresponds to the gene expression profiles of mice response to *Plasmodium chabaudi* [24],

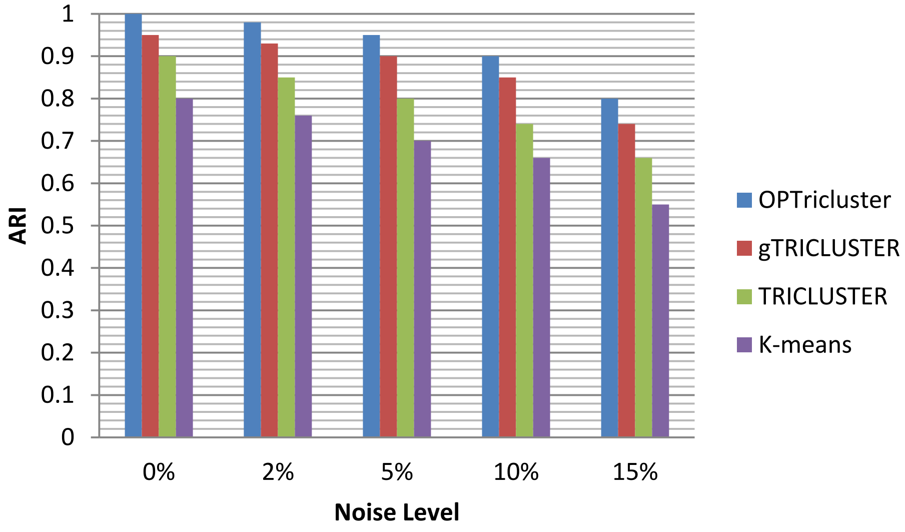


Figure 2 Comparative analysis of clustering algorithms. Simulation results showing the comparative evaluation of OPTricluster, the modified gTRICLUSTER, modified TRICLUSTER, and the K-means algorithm.

representing two disease states (*P. chabaudi* infected and non-infected), two genders (male and female), two protocols (intact and gonadectomized), and four time point experiments: 0, 3, 7, and 14 days after inoculation (DAI). We refer to the four biological samples used in this study as follows: intact male (I_M), intact female (I_F), gonadectomized male (G_M), and gonadectomized female (G_F). For ethical issue pertaining to this dataset, see reference [24].

After data pre-processing and normalization, we ended up with 5783 significant probes, corresponding to 5063 unique genes. The three dimensions of the data are: G (N = 5783 probes), S (M = 4 samples: I_M, I_F, G_M, and G_F), and T (L = 4 time points: 0, 3, 7, and 14 DAI). We set the input parameters to: minimum number of genes in a cluster $I_m = 1$, minimum number of samples in a cluster $J_m = 1$, and differential expression threshold (or ranking threshold) $\delta = 0.31$. With the minimum number of samples in a tricluster set to 1, the algorithm generated 2^4-1 = 15 combinations of samples ({I_M, I_F, G_M, G_F}, {I_M, I_F, G_M}, {I_M, I_F, G_F}, {I_M, G_M, G_F}, {I_F, G_M, G_F}, {I_M, I_F}, {I_M, G_M}, {I_M, G_F}, {I_F, G_M}, {I_F, G_F}, {G_M, G_F}, {I_M}, {I_F}, {G_M}, {G_F}).

Figure 3 shows the set of genes in which the expression level changed similarly by the infection across the time series, in one and a combination of two or more samples. Among the 3943 probes conserved in the four biological samples {I_M, I_F, G_M, G_F}, 3516 genes are unchanged, whereas 427 (Figure 3) changed similarly in all four samples. These 427 genes are further clustered into six groups (Figure 4) with 12 or more genes. Clearly, the genes in Figure 4 have similar behaviour in the four samples and across the entire time series. Most of these 427 genes may play the role of housekeeping. In other words, they represent the set of genes that are co-expressed regardless of the experimental condition to maintain basic cellular function. Indeed, Gene Ontology (GO) analysis of these six clusters (Figure 5), showed that they are involved in similar molecular function and biological processes, such as protein and DNA binding, transcription regulation, cell cycle and basic metabolism.

The results shown in Figure 3 suggest that intact males (I_M) have the highest number of genes (1778) that the expression level changed following pathogen attack. This indicates that I_M is probably more vulnerable to *Plasmodium* infection compared to the other three phenotypes. This is consistent with the phenotypical observation made in [24]. Indeed the Gene Ontology analysis shows that I_M have more genes involved in the GO biological processes, such as, cell death (GO:0008219), programmed cell death (GO:0012501), apoptosis (GO:0006915), than I_F (Figure 6). On the other hand, these same results showed that gonadectomy of males altered the sex-associated differences, suggesting that sex steroid hormones may modulate immune responses to infection [24].

In terms of differences between the four samples tested, our analysis identified genes that were unique to only one and a combination of two or more samples. For example we identified 251, 266, 216, and 249 genes unique to I_M, I_F, G_M, and G_F respectively. These genes may be the origin of the differences between the four samples after *Plasmodium* infection. Thus, they may represent potential targets or biomarkers to be used not only to understand the differences between the samples, but also to develop novel therapeutic means.

Systemic acquired resistance in *Arabidopsis thaliana*

We applied OPTricluster to study similarities and differences in defence mechanism of *Arabidopsis thaliana* against pathogenesis. The goal of the study is to understand the roles of NPR1 and some of TGA family TFs during systemic acquired resistance in *Arabidopsis*. The 3D microarray data used here was obtained using Affymetrix *Arabidopsis* Genechip consisting of 22810 probes. The Columbia wild-type (W), mutant *npr1* (P), double mutant *tga1 tga4* (Z_1), and triple mutant *tga2 tga5 tga6* (Z_2) were treated with salicylic acid (SA) for 0, 1, and 8 hours. After data pre-processing and normalization, we ended up with 3945 significant genes. We set the Columbia wild-type as our baseline and took the $\log_2$ ratio of the mutant gene expression levels over the wild-type at respective time points.

Given the $N \times M \times L$ gene expression matrix, our goal is to identify the set of genes that are controlled by the TFs at a given time point, to study similarities and differences between them, and to infer a temporal transcriptional regulatory network controlling SAR in *A. thaliana*. OPTricluster generated 2^4-1 = 15 combinations of samples. Below we present some of the results obtained by OPTricluster. Figure 7 for example shows an example of divergent patterns. The expression levels of these genes are relatively unchanged (constant) in three samples and behave differently in one. For example, in the first row, significant changes are visible within WT, whereas the other three genotypes stay relatively constant (within threshold of ± 0.5) across the three time point. Clearly, since the expression level of these genes stay constant in three of the four experimental conditions and change considerably in only one of them, these genes may represent potential targets of the TFs tested in this study.

Figure 8 shows a wiring diagram of the genetic network of SAR in *Arabidopsis thaliana* at 0 h, 1 h, and 8 h inferred using the OPTricluster algorithm (Equation 4 in the Method Section). For example, our analysis showed that only 23, 66, and 73 genes are either down- or up-regulated by the combined action of the three sets of TFs at 0, 1, and 8 h respectively. The number of NPR1 targeted genes is less than that of TGA1 TGA4 and TGA2 TGA5 TGA6 at 0 h. But at 8 h, it is the reverse situation where

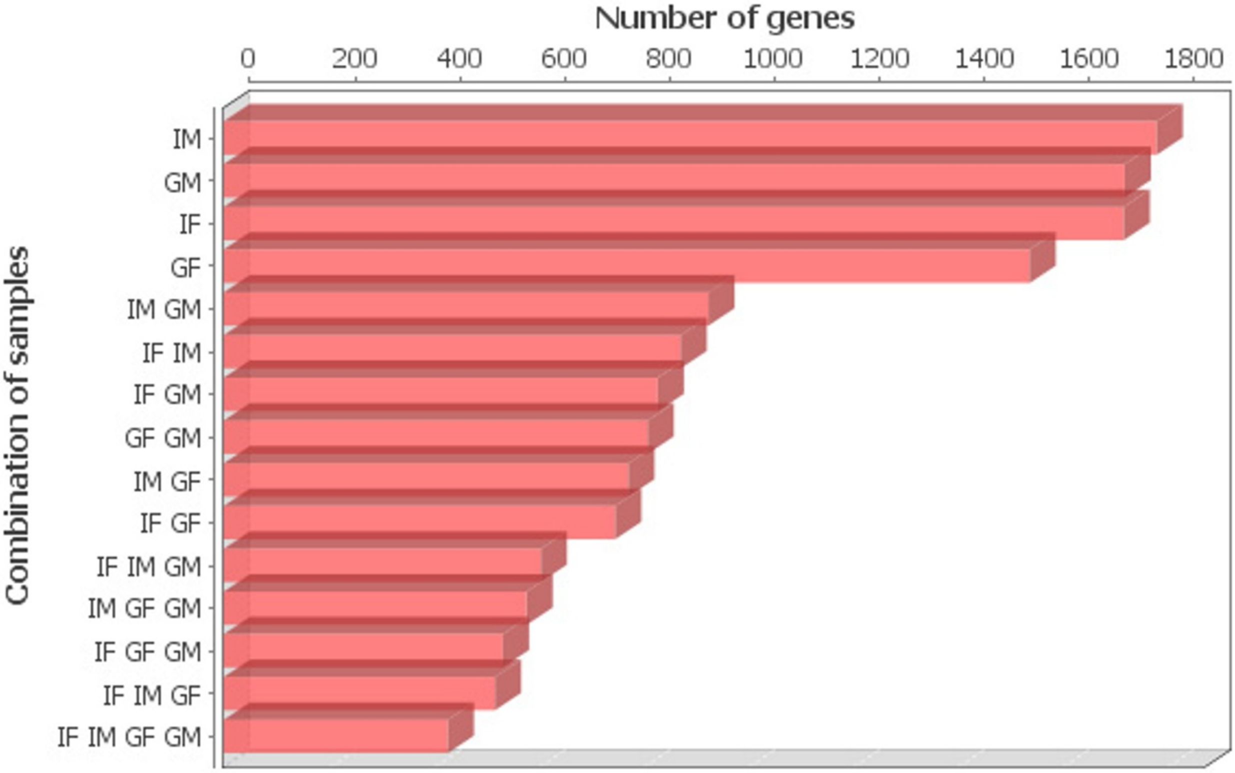


Figure 3 Similarities among samples in mouse data. Similarities in the gene expression data of mice response to *Plasmodium chabaudi* infection. The x-axis corresponds to the subset of samples and the y-axis the number of genes that behave similarly in this subset of samples.

the number of NPR1 targeted genes is higher than those regulated by TGA1 TGA4 and TGA2 TGA5 TGA6, respectively. This is consistent with the fact that NPR1 gene expression in the Columbia wild type was initially moderate but drastically increased at 1 hour and increased further until 8 hours after SA treatment (Additional file 3).

Gene Ontology analysis reveals that several of the genes that are regulated by the three sets of TFs (NPR1, TGA1 TGA4, and TGA2 TGA5 TGA6) at 8 h (Figure 8) are associated with response to stimulus (GO:0050896; p-value = 1.1e-08), stress (GO:0006950; p-value = 1.7e-05), abiotic stimulus (GO:0009628; p-value = 3.1e-05), biotic stimulus (GO:0009607; p-value = 1.3e-03), and defence mechanism (GO:0006952; p-value = 5.0e-02). These correspond to the fact that the plants were treated by SA, which mimic pathogen infection. They also confirm the fact that the TFs tested in this study are known to play major roles in plant defence mechanism [25]. In this application of OPTricluster, we used similarities in gene expression profiles of *Arabidopsis thaliana* with single, double or triple mutations of key transcription factors in the defence signalling network. We studied the network dynamics over a time series after treatment with salicylic acid (SA), which mimics a pathogen infection. We found that most SA-responsive genes were affected by at least one mutation and that most affected genes fit one of a few patterns of regulation. We then provided a first glimpse into the temporal pattern of the gene regulatory network during systemic acquired resistance in *Arabidopsis*.

Similarities and differences between inner and outer cotyledon in *brassica napus* during seed development

Canola (*Brassica napus*) has two cotyledons; one embraces the other. The outer cotyledon is bigger and has higher oil content than the inner one. There exists similarity and differences in the development of and the metabolic processes between the two cotyledons. The objective of this analysis is to unravel the spatial similarity and differences in gene expression profiles between the two cotyledons in order to identify genes that play important roles in seed development and oil production. After obtaining the respective cotyledons, DNA microarray experiments were performed on each cotyledon using the DNA Combimatrix 90 K chip, developed by Plant Biotechnology Institute, NRC [26]. The output yielded two time-series data matrices X_I and X_O representing the gene expression level in the inner and outer cotyledons, respectively. The time-series has six time points, 20, 22, 24, 26, 28, and 30 days after pollination (DAP). After data pre-processing and normalization, significantly expressed genes were identified

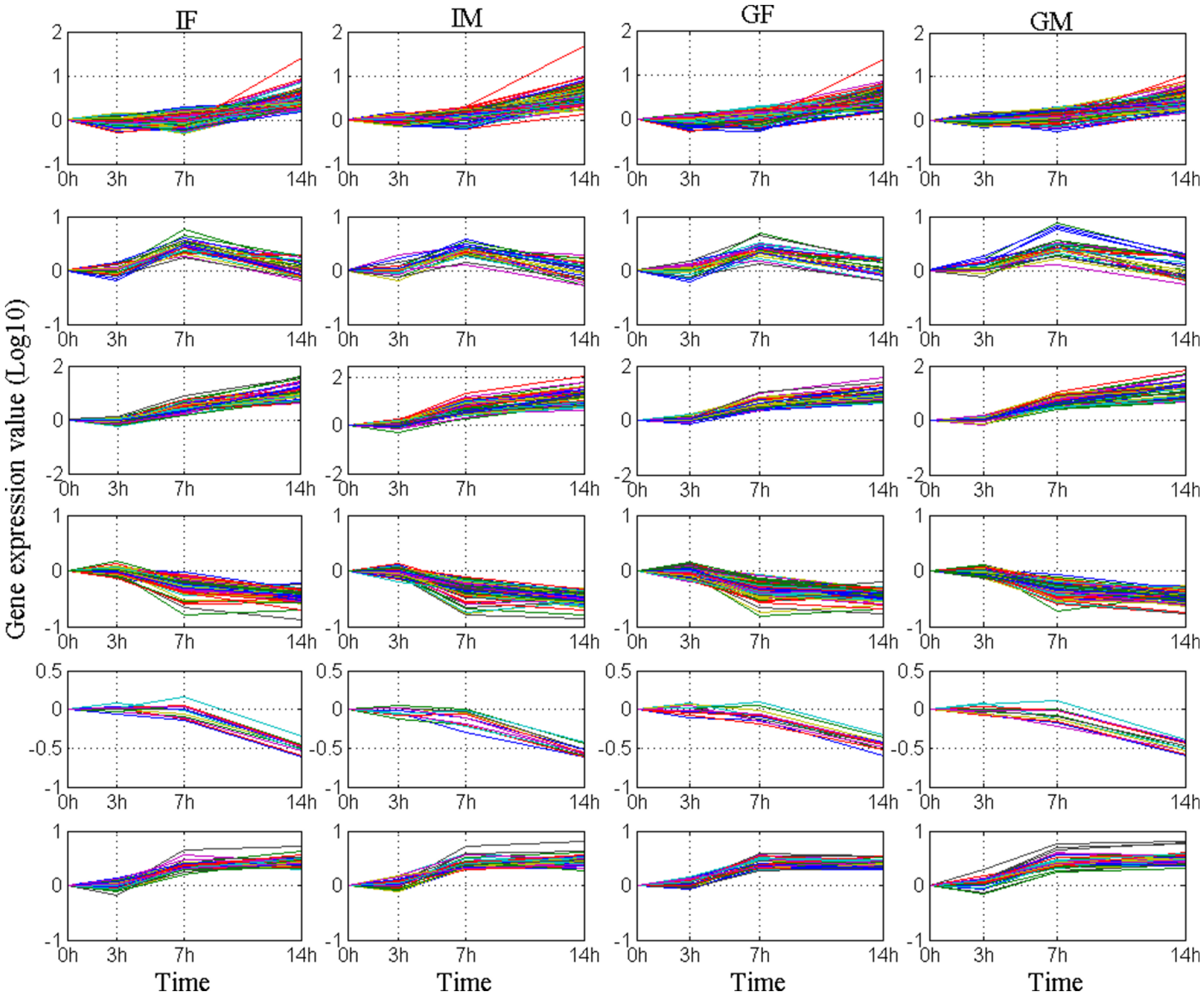


Figure 4 Example of similar patterns in mouse data. Set of genes with conserved expression profiles across the four biological samples: {I_F, I_M, G_F, G_M}. The x-axis corresponds to the experimental time points and the y-axis to the expression level of the genes across the time series and in the four samples. Note that each sample corresponds to the column chart.

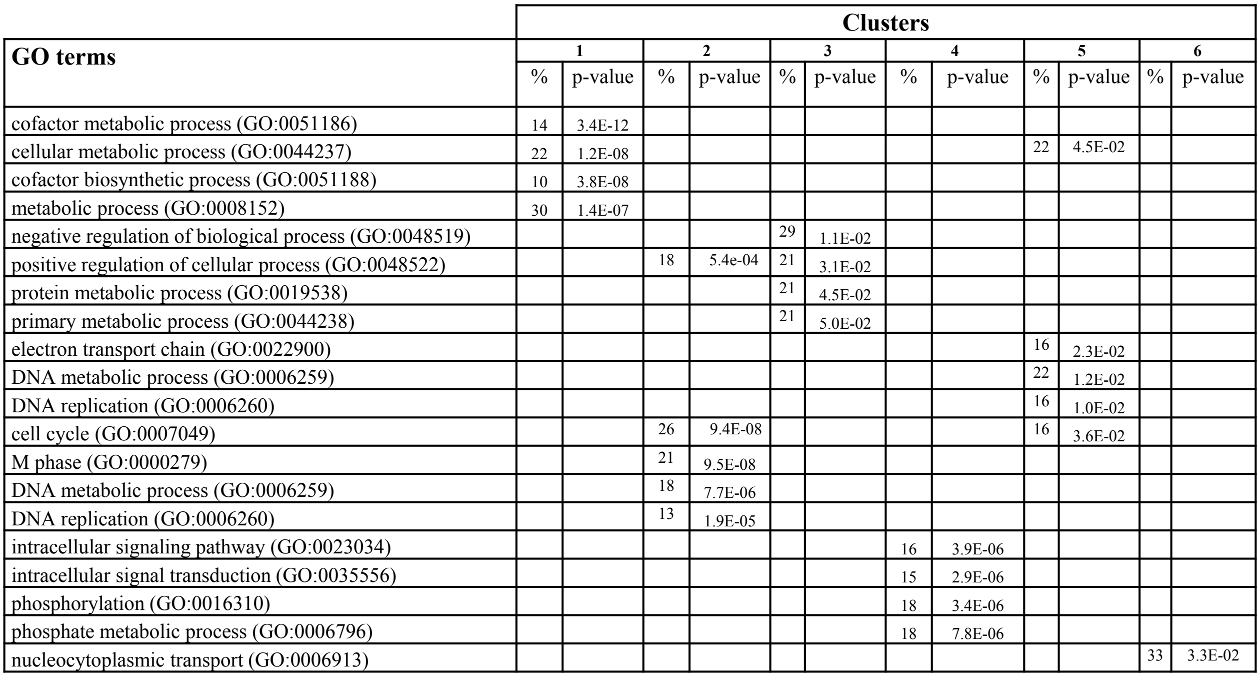
for this study. The three dimensions considered in this analysis were G (N = 3945 genes), S (M = 2 biological samples: *I* and *O*), and T (L = 6 time points: 20, 22, 24, 26, 28, and 30 DAP).

With the minimum number of samples in a tricluster set to 1 and a threshold of ~ 1.5 fold change, the algorithm generated 2^2-1 = 3 combinations of samples ({*I, O*}, {*I*}, {*O*}*)*. The subset of sample {*I, O*} yielded similar patterns between the inner and outer cotyledons, whereas the equations {*O*} - {*I, O*}, or {*I*} - {*I, O*} yielded patterns specific to outer or inner cotyledon, respectively. Analysis reveals 33 genes depicting the main difference between the two samples across the six time point experiments (Figure 9) and several others across subsets of the six time points (Additional file 4). Among the 33 genes, 17 and 16 were highly expressed in inner compared to outer, and outer compared to inner, respectively.

Figure 9 shows that the expression level of the genes that are highly expressed in the inner compared to the outer cotyledon decreases with time whereas the ones that are highly expressed in the outer compare to the inner cotyledon increase with time. In general, in terms of the direction of the gene expression profile (Figure 10 for example), OPTricluster did not depict significant differences between the inner and the outer cotyledons. The main difference between the inner and the outer cotyledons resides in the amplitude of the gene expression level.

Brassica napus whole seed development

We also analyzed another in house Combimatrix 94 K *Brassica napus* microarray dataset, which corresponds to the gene expression profiles during seed development. It has eight times points: 10, 15, 20, 25, 30, 35, 40, and



GO terms	Clusters											
	1		2		3		4		5		6	
	%	p-value	%	p-value	%	p-value	%	p-value	%	p-value	%	p-value
cofactor metabolic process (GO:0051186)	14	3.4E-12										
cellular metabolic process (GO:0044237)	22	1.2E-08							22	4.5E-02		
cofactor biosynthetic process (GO:0051188)	10	3.8E-08										
metabolic process (GO:0008152)	30	1.4E-07										
negative regulation of biological process (GO:0048519)					29	1.1E-02						
positive regulation of cellular process (GO:0048522)			18	5.4e-04	21	3.1E-02						
protein metabolic process (GO:0019538)					21	4.5E-02						
primary metabolic process (GO:0044238)					21	5.0E-02						
electron transport chain (GO:0022900)									16	2.3E-02		
DNA metabolic process (GO:0006259)									22	1.2E-02		
DNA replication (GO:0006260)									16	1.0E-02		
cell cycle (GO:0007049)			26	9.4E-08					16	3.6E-02		
M phase (GO:0000279)			21	9.5E-08								
DNA metabolic process (GO:0006259)			18	7.7E-06								
DNA replication (GO:0006260)			13	1.9E-05								
intracellular signaling pathway (GO:0023034)							16	3.9E-06				
intracellular signal transduction (GO:0035556)							15	2.9E-06				
phosphorylation (GO:0016310)							18	3.4E-06				
phosphate metabolic process (GO:0006796)							18	7.8E-06				
nucleocytoplasmic transport (GO:0006913)											33	3.3E-02

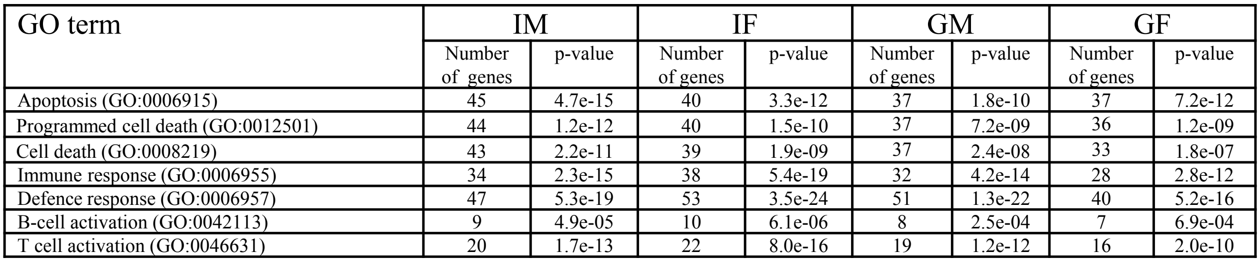
Figure 5 Gene Ontology analysis of mouse clusters. GO analysis of the 6 clusters that are affected similarly in the four biological samples {I_M, I_F, G_M, G_F}.

45 DAP. Each time point has six biological replicates. After data pre-processing and normalization, we ended up with 10865 probes, from which we took the mean of the replicates. Figure 11 show examples of patterns identified using OPTricluster.

Biological evaluation of these clusters showed that they are highly enriched under seed development; fatty acid and lipid metabolism; fatty acid and lipid biosynthesis; lipid storage, transport, and localization; oil content and biosynthesis, with *p*-values < 1.0e-03 (Additional file 5). These results demonstrate that seed development and fatty acid synthesis are highly correlated. Indeed, this is consistent with the fact that in higher plants, the biosynthesis of most fatty acids and lipids is physiologically coupled with seed development [27]. Further analyses showed that most of these patterns are highly positively or negatively correlated with the expression profile of transcription factors such as: LEC1, WRI1, FUS3, ABI3, and ABI5. This observation is consistent with the fact that the LEC1 function is partially dependent on ABI3, FUS3, and WRI1 in the regulation of fatty acid and fatty acid derived complex lipid [27-29]. It is also consistent with the fact that WRI1 plays a significant role during oil accumulation in maturing seed, that WRI1 is a prerequisite for fatty acid synthesis, and is important for normal embryo development [30]. Most of these results suggest that, the genes involved in these patterns may represent potential targets that could be used for the genetic improvement of oil production in *B. napus*.

Discussion

In this paper, we developed a subspace clustering algorithm OPTricluster, for 3D short time-series data



GO term	IM		IF		GM		GF	
	Number of genes	p-value	Number of genes	p-value	Number of genes	p-value	Number of genes	p-value
Apoptosis (GO:0006915)	45	4.7e-15	40	3.3e-12	37	1.8e-10	37	7.2e-12
Programmed cell death (GO:0012501)	44	1.2e-12	40	1.5e-10	37	7.2e-09	36	1.2e-09
Cell death (GO:0008219)	43	2.2e-11	39	1.9e-09	37	2.4e-08	33	1.8e-07
Immune response (GO:0006955)	34	2.3e-15	38	5.4e-19	32	4.2e-14	28	2.8e-12
Defence response (GO:0006957)	47	5.3e-19	53	3.5e-24	51	1.3e-22	40	5.2e-16
B-cell activation (GO:0042113)	9	4.9e-05	10	6.1e-06	8	2.5e-04	7	6.9e-04
T cell activation (GO:0046631)	20	1.7e-13	22	8.0e-16	19	1.2e-12	16	2.0e-10

Figure 6 Gene Ontology analysis of genes affected in Mouse. GO analysis of the genes affected in each sample after *Plasmodium* infection.

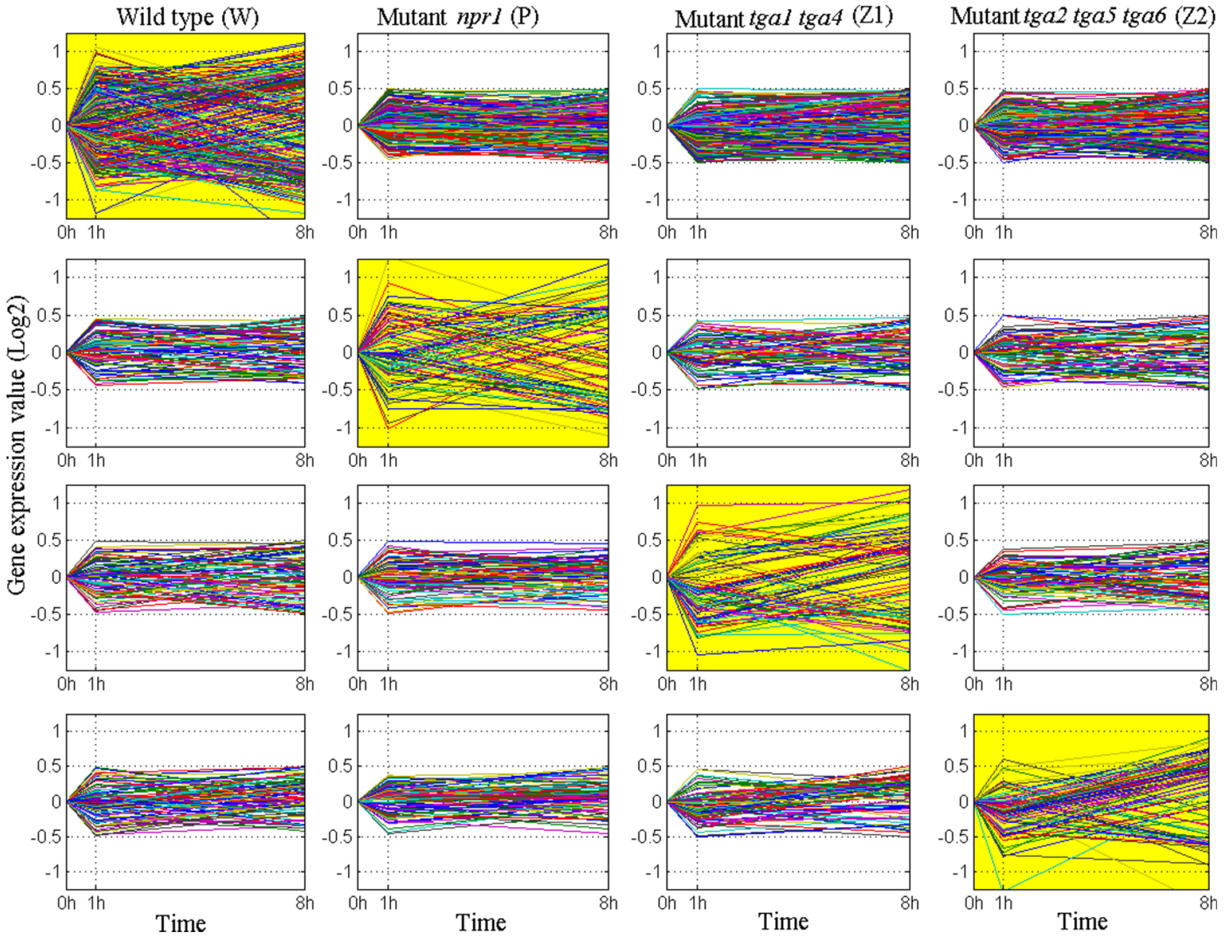


Figure 7 Example of Divergent Patterns. Example of genes that may be specific to the W, P, Z_1, and Z_2 respectively. The x-axis corresponds to the experimental time points and the y-axis to the expression level of the genes across the time series and in the four samples. Note that each sample corresponds to the column chart: W, P, Z_1, and Z_2 respectively.

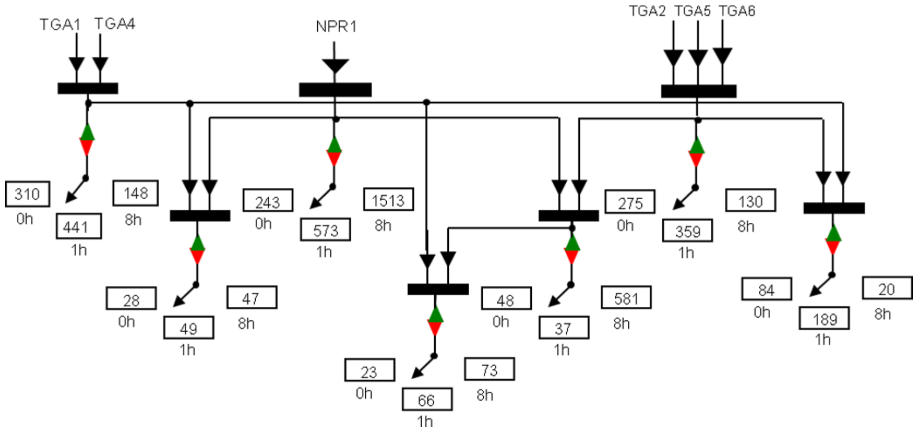


Figure 8 SAR transcription networks. SAR transcriptional network at 0, 1, and 8 h. In the diagrams, ▬ = action could be inclusive or, ▼ = information flow, red inverted triangle symbol = up regulation, and green triangle symbol = down regulation. Numbers in boxes represent the number of genes regulated by the corresponding groups of TFs at respective time points.

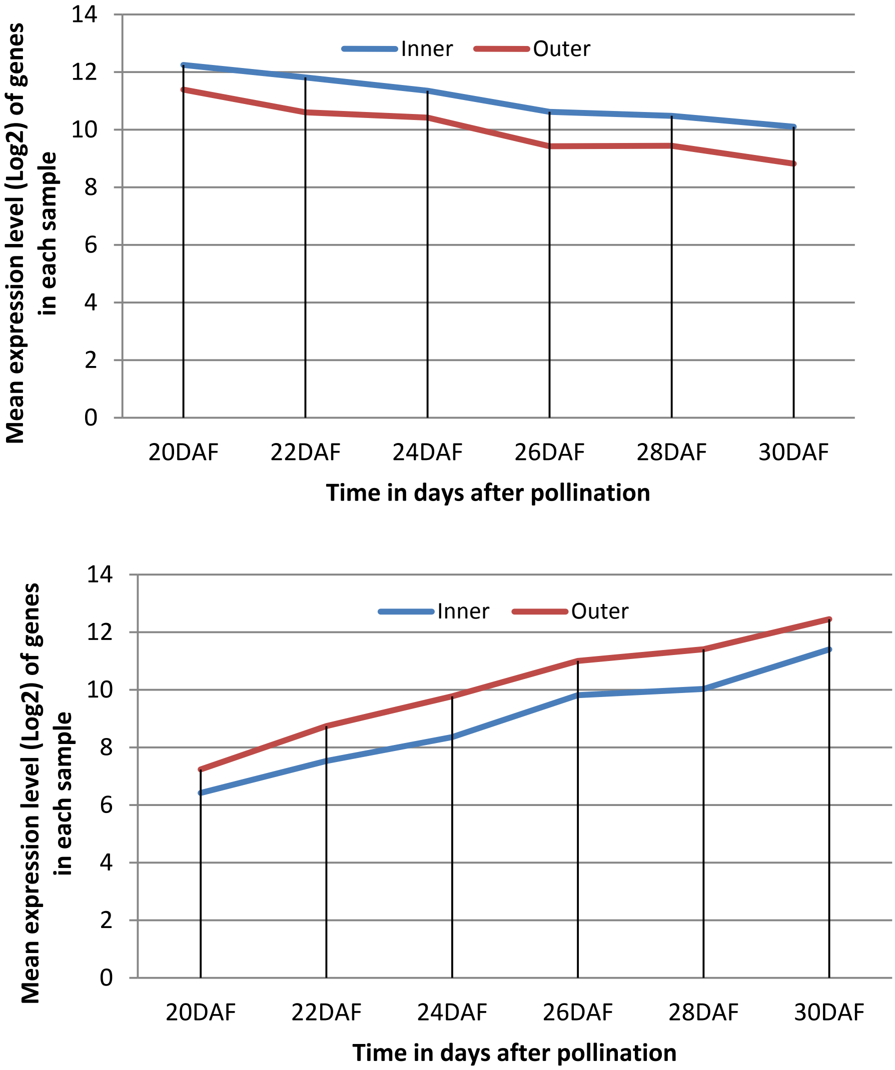


Figure 9 Differences between inner and outer cotyledons. Gene expression differences between inner and outer cotyledons. The x-axis corresponds to the experimental time points and the y-axis to the mean of the expression level of genes in each sample. Each line represents a sample.

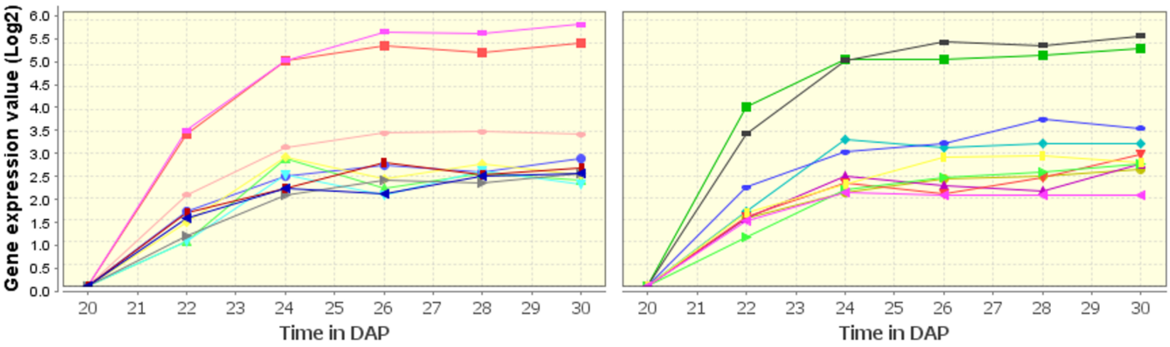


Figure 10 Example of conserved patterns in the cotyledon datasets. Example of group of genes with similar expression profile in the inner and the outer cotyledons dataset.

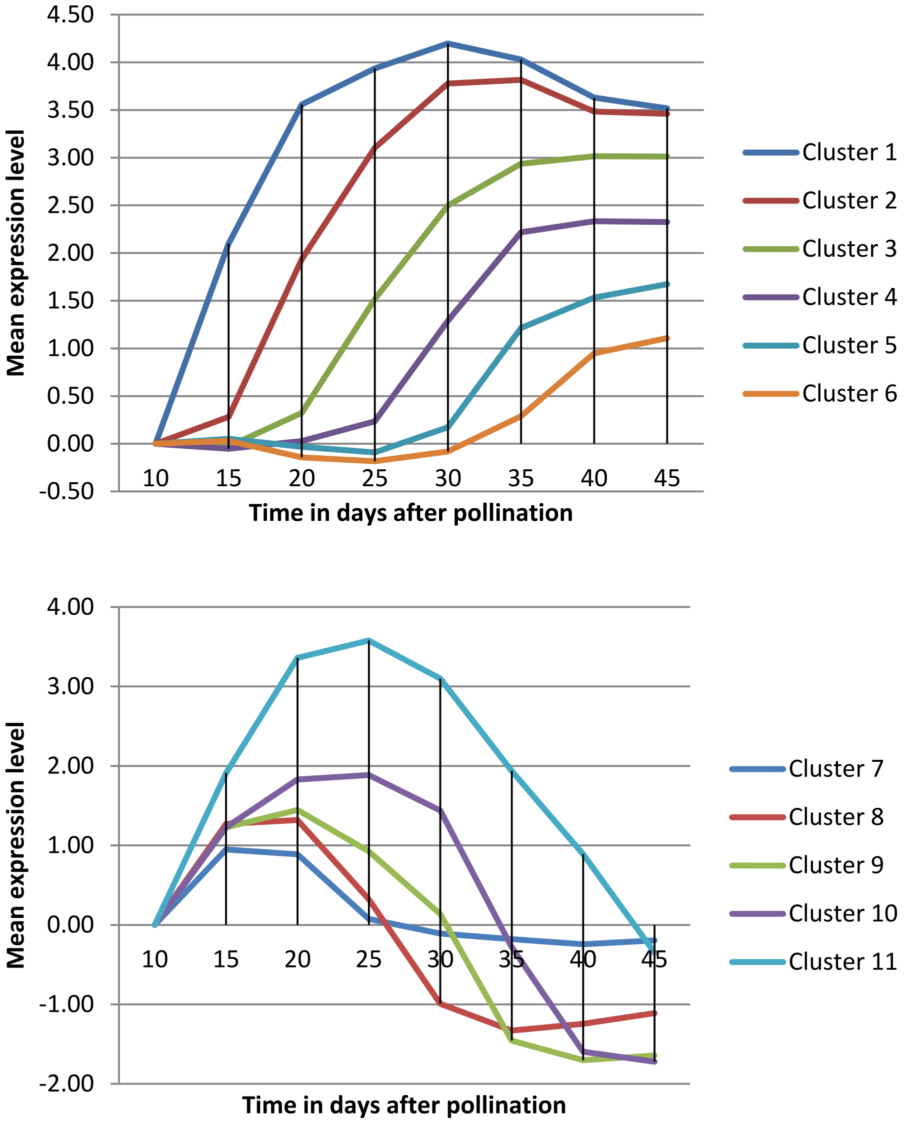


Figure 11 Whole seed Brassica napus clusters. Example of patterns identified in *Brassica napus* whole seed dataset using OPTricluster. The x-axis corresponds to the experimental time points whereas the y-axis to the mean of the expression level of the genes in each cluster. Each curve corresponds to a cluster.

mining. OPTricluster is able to identify triclusters (3D clusters) with coherent evolution from a given 3D dataset using a combinatorial approach on the sample dimension, and the OP concept on the time dimension. The amalgamation of the two methodologies, combinatorial and order preserving allows one to study similarities and differences between samples in terms of temporal expression profile and/or differential expression. The combinatorial approach on the sample dimension is necessary because the goal of the algorithm is not only to study similarities between biological samples, but also to study differences between them at the single and multiple samples level. Because of this enumerative approach, it makes OPTricluster computational heavy when one is dealing with an increasing number of samples. As a result, we have restrictions for analyzing 3D short time series genes expression data. That is 3D gene expression with ~3-8 time points and ~2-5 samples. For longer 3D time series, OPTricluster will still work, but the computational complexity of the algorithm will increase exponentially with the increase in the number of time points and/or the number of samples.

Because of this combinatorial approach in the sample dimension and the OP concept in the time dimensions, the algorithm is able to mine clusters of genes in a subset of samples, with expression level having same directions across a time-series experiment. OPTricluster takes into account the sequential nature of the time-series and the noisy nature of the data through the OP concept. The results of Gene Ontology analysis and statistical analysis show that OPTricluster is robust to noise and is able to detect the similarities and differences between samples in terms of temporal expression profile of relevant functional categories. This is due to the fact that the OP model focuses on the similarity of the relative order over the time-series, rather than on the similarity of the actual expression levels based on a distance measure as in several other clustering models. This approach is potentially more robust to the stochastic nature of the gene expression values, and to the variation caused by the measurement procedure. This is clearly demonstrated in our comparison among the four clustering methods (Figure 2). The OP model has been accepted as a biologically meaningful clustering approach, capturing the general tendency of gene expressions across a subset of conditions [5,11-14,20,21].

Furthermore, OPTricluster also offers the possibility to analyze the genes that the expression profiles change similarly across different samples for a given time point. This is easily done within the current version of the OPTricluster algorithm by simply swapping the time dimension with the sample dimension.

One relevant application of OPTricluster is the identification of genes that may play the role of housekeeping. This is revealed by the study of similarities among the samples. Indeed, several of the genes with same behaviour across all the samples and over the entire time series are usually not affected by phenotypes (Figure 4). Thus these genes may represent genes that are needed by all cells of the organism to fulfill basic cellular functions. As has been shown in the literature, housekeeping genes have many applications. They can be used as transcription and expression controls in laboratory experiments, they can be used to infer the set of basic cellular functions, and they are essential to characterizing normal and diseased states [31]. One has to keep in mind that in medicine, genetic diseases that are linked to housekeeping genes are more likely to affect multiple organs; in microbiology, housekeeping genes of pathogens play a role in enhancing virulence; and in evolutionary biology studies, housekeeping genes diverge more slowly than other genes [31]. Therefore, their detection may enhance the creation of new drug targets and can also be used for subspecies discrimination.

The study of differences among samples in terms of co-expression is among the most important applications of the OPTricluster algorithm. This feature does not exist in the other three algorithms tested in this study. In the mice response to *Plasmodium* infection for example, OPTricluster was able to identify a set of genes specific to each sample. This group of genes can be used as biomarkers for the differentiation between these samples, thus for the detection of diseases, and subsequently use as targets to design drugs that are specific to diseases at their subtype level.

In the *Arabidopsis* data, OPTricluster identified genes that may only be regulated by one TF, or a combination of two or more TFs. This kind of information can be used to identify transcription networks as we did in this analysis, and subsequently to get more insights between TFs and gene regulation. Figure 7 for example show that only a small number of genes in each or multiple mutant genotypes overlaps across different time points. This observation suggests that most of the genes that participate in SAR have impulse behaviour; different sets of genes are targeted by the corresponding transcription factors at different time points during a response to pathogen infection. On the other hand, these results show that the true behaviour of the underlying biological process is captured at different time points, with each time point containing a unique piece of information that should be integrated in order to get the whole picture underlying the signalling pathway during SAR. Therefore, studies, such as [32], that focus on a single time point to infer genetic information and generalize the results to describe the pathogen-host interaction, could miss out important information.

In the cotyledons dataset, OPTricluster depicted differences in the inner and the outer cotyledons in terms of direction and magnitude of the gene expression level, and thus yielded more biological information relative to the spatial distribution of gene expression at the molecular, cell, tissue, organ, or system level. The analysis of the *Brassica* whole seed dataset showed that OPTricluster can not only be used to study similarities and differences among samples in a 3D short time series, but also, can be used to tackle classical short time-series problems. In several of these studies OPTricluster offered more biological insights to the problem compared to prior computational approaches used to tackle similar problems.

Conclusions

We developed a subspace clustering algorithm for 3D short time-series gene expression data analysis. The developed algorithm is used to identify statistically and biologically significant clusters from 3D gene expression data, to study similarities and differences between samples in terms of co-expression and/or differential expression. Biological and statistical results obtained showed that OPTricluster is robust to noise and is able to detect the temporal expression profile of relevant functional categories in terms of similarities and differences in samples. The OPTricluster is implemented in Java and it is available as Additional File 1 to this manuscript.

Methodology

Our objective is to develop a subspace clustering algorithm that is able to cope with the sequential nature of a time-series, along with the noisy nature of a dataset. We also aim at developing an algorithm that can be used to identify biologically significant subspace clusters from a given 3D short time-series gene expression dataset. This would allow one to study similarities and differences between two or more biological samples in terms of temporal expression profile.

Definitions

Given a set of N genes $G = \{g_1, . . ., g_n, ..., g_N\}$, a set of M biological samples $S = \{s_1, . . .,s_m, ..., s_M\}$, and a series of L time points $T = \{t_1, . . ., t_l, ..., t_L\}$, a 3D microarray gene expression dataset or gene-sample-time (GST) dataset is a real-values $N \times M \times L$ matrix, $A = \{a_{nml}\}$. Each entry a_{nml} represents the expression level of gene g_n in sample s_m at time t_l. We shall also refer to the 3D data as a set: $A = \{G, S, T\}$. In the sequel, we denote the expression level of gene g_n in sample s_m across the time points as $f_{nm}(T)$, which is a row vector. The expression profile of gene g_n in all samples and across the time-series $f_n(s,T)$, which is a 2D matrix. Thus the 3D gene expression matrix can be viewed as a set of 2D matrices in the horizontal plane: Equation 1.

$$A = \begin{bmatrix} f_1(s,T) \\ f_2(s,T) \\ \vdots \\ f_n(s,T) \\ \vdots \\ f_N(s,T) \end{bmatrix} \tag{1}$$

A 3D cluster or tricluster is a 3D submatrix $C = \{c_{ijk}\}$ of A, or a subset $C = \{I, J, K\}$, with $I \subseteq G$, $J \subseteq S$, and $K \subseteq T$, such that the content of $C = \{c_{ijk}\}$ ($i \in I$, $j \in J$, and $k \in K$), verifies a desired pattern: constant, coherent values, or order preserving.

Problem statement

Given the 3D matrix A as defined above and an ordering ranking threshold δ, find all triclusters $C = \{I, J, K\}$, with minimum number of genes I_{min}, minimum number of samples J_{min}, such that the content of each $C = \{c_{ijk}\}$ is order preserving over a given segment of a time-series. Since we are dealing with short time-series in this study, we set $K = T$. Thus we are interested in patterns that increase, decrease, or stay constant coherently across the entire time-series experiments. Note that constant patterns and coherent values are subsets of order preserving clusters.

OPTricluster

Our triclustering algorithm identifies triclusters of genes with expression level having same direction across the time series experiments in subsets of samples. OPTricluster takes into consideration the sequential nature of the time-series and is able to cope with the effect of noise through the order preserving approach. Basically, for a given subset of samples, we say that a tricluster is order preserving if there exists a permutation of the time points such that the expression levels of the genes are monotonic functions. Such a matrix is defined by Equation 2 below. In this example, it is obvious that whatever permutations of its columns that we do, the variation between the values in its rows across the columns will always follow the same patterns.

$$\begin{bmatrix} 2 & 8 & 5 & 3 \\ 0 & 7 & 4 & 1 \\ 1 & 5 & 3 & 2 \end{bmatrix} \tag{2}$$

Furthermore, there exists a permutation of its columns (1, 4, 3, 2) or (2, 3, 4, 1) such that the values of its rows are monotonic increasing or decreasing function, respectively.

In all, after the data pre-processing and normalization, OPTricluster has five main steps (Figure 12). First, OPTricluster performs the gene expression data quantization. Second, it ranks the expression level of the genes across the time-dimension in all the samples for a given δ. Third, it identifies the set of distinct coherent 3D patterns in the 3D dataset. Fourth, triclusters of coherent patterns are formed by assigning genes with similar ranking along the time-dimension and across subsets of samples to the same group, then divergent patterns are identified. Finally, statistical significance and biological evaluation of the triclusters identified are performed.

Gene expression matrix quantization

The first step of OPTricluster which is in fact optional consists of performing the gene expression data quantization. This is due to the fact that we are not only working with noisy data, but also DNA experimental data contains missing values. Many techniques are available in the literature to deal with noise through data quantization and to recover missing values by imputation [13]. OPTricluster uses the following approach for data quantization. Given the ordering and ranking threshold δ, for each row (gene) and a given sample, OPTricluster computes $E = (b_E - b_0)/\delta$, where $b_0 = min(f_{nm}(T))$, $b_E = max(f_{nm}(T))$, and $\delta \neq 0$. Note that E is always rounded to the next integer of $(b_E - b_0)/\delta$. Then the interval $[b_0, b_E]$ is divided into E equal intervals: $[b_0, b_E] = [b_0, b_1[\text{U} ...\text{U} [b_{e-1}, b_e [\text{U} ...\text{U} [b_{e-1}, b_E]$, where $b_e = b_0 + e\delta$, and $e = 1$ to E. Finally, a new expression level

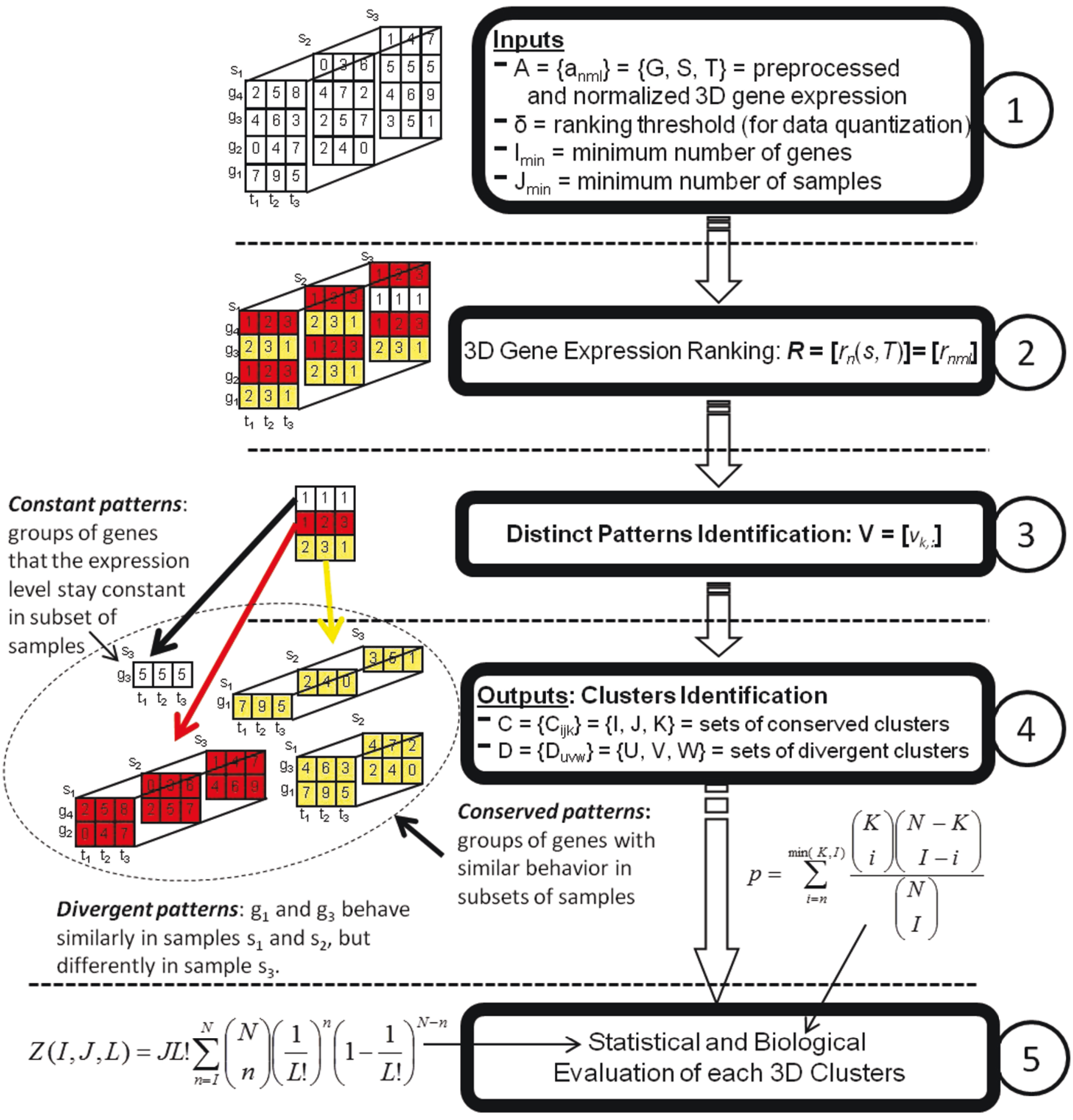

Inputs
- A = {a_{nml}} = {G, S, T} = preprocessed and normalized 3D gene expression
- δ = ranking threshold (for data quantization)
- I_{min} = minimum number of genes
- J_{min} = minimum number of samples
1
3D Gene Expression Ranking: $R = [r_n(s,T)] = [r_{nml}]$
2
Distinct Patterns Identification: V = [$v_{k,:}$]
3
Outputs: Clusters Identification
- C = {C_{ijk}} = {I, J, K} = sets of conserved clusters
- D = {D_{uvw}} = {U, V, W} = sets of divergent clusters
4
Statistical and Biological Evaluation of each 3D Clusters
5
Constant patterns: groups of genes that the expression level stay constant in subset of samples
Conserved patterns: groups of genes with similar behavior in subsets of samples
Divergent patterns: g_1 and g_3 behave similarly in samples s_1 and s_2, but differently in sample s_3.
$p = \sum_{i=n}^{\min(K,I)} \frac{\binom{K}{i}\binom{N-K}{I-i}}{\binom{N}{I}}$
$Z(I,J,L) = JL!\sum_{n=I}^{N}\binom{N}{n}\left(\frac{1}{L!}\right)^{n}\left(1-\frac{1}{L!}\right)^{N-n}$
s_1 s_2 s_3 g_1 g_2 g_3 g_4 t_1 t_2 t_3

Figure 12 OPTricluster illustration Example illustrating the different steps of the OPTricluster algorithm.

of the corresponding gene in the considered sample at the given time point is obtained using Equation 3.

$$\alpha_e = (b_e + b_{e-1})/2 \qquad (3)$$

Specifically, if the expression level f_{nml} of the corresponding gene g_n in the given sample s_m and at a given time points t_l falls in the interval $[b_{e\text{-}1}, b_e[$, then it is quantized to the centroid α_e of that interval. Subsequently, each centroid α_e is assigned a ranking order (see below).

Note that, for a given microarray data, if the distribution of the expression levels of most genes in each sample and across the time points are densely located in a certain region, depending on the δ value, the quantization scheme defined above will likely put the densely related genes in the same cluster, usually the constant ones. Thus the value of δ should be chosen based on the following criteria: distribution of the gene expression data, the desired fold change between the expression levels of the genes across the time series, and based on the expected level of noise in the microarray data. Since the later is difficult to measure in real applications, one can also run the algorithm several times for a given value of δ, with perturbation values below and above it, and consider only the clusters that the content does not change a lot.

3D rank expression matrix

OPTricluster then uses the quantized 3D gene expression matrix to generate a 3D rank expression matrix in the second step. A 3D rank expression matrix is an $N \times M \times L$ matrix, $R = [r_n(s,T)] = [r_{nml}]$, in which every row along the time-dimension for a given sample is a vector of the ranks of the corresponding expression values in A, in an increasing or decreasing order. For example, if the expression levels of gene g_n in sample s_m along the time-dimension is $f_{nm}(T) = [1.5, 3, 0.5]$ at $\delta \leq 1.0$, then, the corresponding row in the rank matrix would be $r_{nm}(T)$ = [1-3] in the increasing order. The ranking matrix of the example of an order preserving matrix defined by Equation 2 above will be:

$$\begin{bmatrix} 1\,4\,3\,2 \\ 1\,4\,3\,2 \\ 1\,4\,3\,2 \end{bmatrix} \tag{4}$$

Without lost of generalities, one can easily see from this example (Equation 4) that the rows of the ranking matrix of an order preserving matrix will always be identical. In fact, this property is exploited below for the identification of order preserving patterns from a 3D short time series gene expression matrix. Note that, if more than two entries have the same value, they are given the same ranking. For example $[0.5, 3, 0.5]$ would be $[1, 2, 1]$. This approach allows the identification of constant patterns.

There are several advantages associated with this transformation. First, it avoids the use of greedy algorithms, probabilistic approaches, and exhaustive permutations along the time-dimension, and thus speeds up the computation time. Second, it is obvious that for any $k > 1$ rows $f_{nm}(T)$ of similar ranking $r_{nm}(T)$, under any permutation of the time points, their order is always preserved (example in Equation 4). Thus they will always belong to the same coherent tricluster. Since the rank is conserved under any permutation along the time-dimension of the 3D gene expression matrix and given that we are dealing with multiple samples at the same time, the probability that a random pattern might be picked up in a cluster is very low.

Identification of distinct patterns

In its third step, OPTricluster identifies the set of distinct 3D coherent patterns that can be found in the 3D gene expression matrix A. Given the 3D gene expression matrix $A = [f_n(s,T)]$ as defined above, with a set of genes $G = \{g_1, \ldots, g_n, \ldots, g_N\}$, a set of biological samples $S = \{s_1, \ldots, s_m, \ldots, s_M\}$ and a series of time points $T = \{t_1, \ldots, t_l, \ldots, t_L\}$, we define the biological sample space Ω as the set of all possible combination of the samples, and Γ their number. That is, if $S = \{s_1\ s_2\ s_3\}$ as in Figure 12, then the biological sample space would be $\Omega = \{\{s_1, s_2, s_3\}, \{s_1, s_2\}, \{s_1, s_3\}, \{s_2, s_3\}, \{s_1\}, \{s_2\}, \{s_3\}\}$ and $\Gamma = 7$. Recall that $r_n(s,T)$ corresponds to a row vector of the ranks of the nth gene, in sample s and across the entire time series T. Hence, for each combination $\Omega_i \in \Omega$, the exact number h_i of distinct order preserving triclusters that can be found in the 3D dataset is the number of distinct 2D $r_n(\Omega_i,T)$ matrices of its corresponding 3D ranked matrix R. Thus, the set of 3D distinct order preserving patterns, V, can be identified by considering R as a set of 2D matrices $r_n(\Omega_i,T)$, that is, $R = \{r_1(\Omega_i,T), r_2(\Omega_i,T), \ldots, r_n(\Omega_i,T), \ldots, r_N(\Omega_i,T)\}$, and identify all distinct $r_n(\Omega_i,T)$ in it. From the above definitions, one can easily show that the exact number Λ of order preserving triclusters in the 3D gene expression matrix is:

$$\Lambda = \sum_{i=1}^{\Gamma} h_i \tag{5}$$

Recall that h_i is the number of distinct 2D $r_n(\Omega_i,T)$ (rank matrices)corresponding to each $\Omega_i \in \Omega$ as defined above.

Conserved clusters identification

Once the exact number of distinct 3D order preserving patterns has been identified, for each $\Omega_i \in \Omega$, OPTricluster assigns each gene to one of the h_i groups by comparing each distinct pattern v_k of V (V: set of 3D distinct order preserving patterns identified from the previous step) to $r_n(\Omega_i,T)$, and assign gene g_n to the order preserving tricluster $C\{k\}$ each time $r_n(\Omega_i,T) = v_k$. This approach is guaranteed to identify all order preserving triclusters of size $I \times J \times K$, with $I_{min} \leq I \leq N$, $J_{min} \leq J \leq M$, and $K = L$, where I_{min} and J_{min} are the minimum number of genes and samples in a tricluster, respectively.

Since the goal of the OPTricluster algorithm is to study the similarities and differences between samples in terms of the expression profile of all genes, $\mathrm{I_{min}}$ and $\mathrm{J_{min}}$ should be set to 1. In this case, the algorithm will identify all the conserved clusters and perform comparison between them at the single sample and single gene level in the divergent patterns identification step, as explained below.

Divergent patterns identification

The 3D procedure as presented above identifies sets of genes that behave similarly (same OP patterns) across the subsets of samples considered. The sets of divergent patterns D can be easily derived from the sets of conserved ones using Equation 5.

$$D_{pq} = C\{p\} - C\{q\} = \begin{cases} I_p - I_q \\ J_p - J_q \end{cases}, \; p \neq q \tag{6}$$

$C\{p\} = \{I_p, J_p, K_p\}$ and $C\{q\} = \{I_q, J_q, K_q\}$ are two conserved triclusters (similar OP patterns). Basically, Equation 5 identifies sets of genes that are co-expressed in the subset of sample in $C\{p\}$, but split and co-expressed differently in one or more samples in $C\{q\}$. For example, if clusters $C\{p\}$ and $C\{q\}$ have the same OP patterns, and if $C\{p\} = \{\{g_1, g_2, g_3, g_4\}, \{s_1, s_2\}, \{t_1, t_2, t_3\}\}$ and $C\{q\} = \{\{g_1, g_2\}, \{s_1, s_2, s_3\}, \{t_1, t_2, t_3\}\}$, then $D_{pq} = \{\{g_3, g_4\}, \{s_3\}, \{t_1, t_2, t_3\}\}$, meaning that genes $\{g_3, g_4\}$ have different behaviour in $\{s_3\}$ compared to $\{s_1, s_2\}$. The computational burden of this step is reduced because only triclusters with same OP or ranking patterns are compared. This is due to the fact that ranking patterns are associated with the expression profile and are unique to each cluster for a given subset of samples.

Statistical significance and complexity analysis

The statistical significance of each identified tricluster with I genes and J samples is assessed by computing the tail probability that a random dataset of size $N \times M \times L$ will contain an order preserving tricluster with I or more genes and J or more samples in it. In principle, the probabilistic description of the reference *3D* random matrix would be that of the observed noise in the microarray experiment [5,14,33]. Since this distribution is difficult to calculate in closed form, the upper bound of this tail probability is estimated using the same approach as in [5,13,14,33]. Assuming that we have a dataset with L time points that are independent and identically distributed according to the uniform distribution, the probability that a random gene-sample supports a given cluster is equal to the number of possible time points permutations or *1/L!*. Since the genes and samples are assumed to be independent, the probability of having at least I genes and J samples in the cluster is the I -tail of the *(N,(1/L!))* binomial distribution, i.e.:

$$P(X \geq I) = \sum_{n=I}^{N} \binom{N}{n} \left(\frac{1}{L!}\right)^{n} \left(1 - \frac{1}{L!}\right)^{N-n} \tag{7}$$

As there are *Ls = L!* ways to choose an OP tricluster of size L, the following expression *Z(I, J, L)* is an upper bound on the probability of having a tricluster of size L with I or more genes and J samples:

$$Z(I, J, L) = JL! \sum_{n=I}^{N} \binom{N}{n} \left(\frac{1}{L!}\right)^{n} \left(1 - \frac{1}{L!}\right)^{N-n} \tag{8}$$

We use this bound to estimate the significance of any given tricluster of size L with I genes and J samples. The best tricluster is the one with the largest statistical significance, i.e., the one with the *smallest Z(I, J, L)* Therefore, as long as that upper bound probability is smaller than any desired significance level, the identified tricluster in the real 3D gene expression matrix will be statistically significant.

The overall complexity of the triclustering algorithm is $O(N\Gamma\Lambda)$. Recall that the 3D short time-series gene expression data A is an $N \times M \times L$ matrix. The 3D rank matrix can be identified within $O(NML)$. The set of distinct 3D patterns can be identified with $O(N\Gamma)$. Finally, the set of coherent conserved triclusters can be identified within $O(N\Gamma\Lambda)$. In all, the complexity of the triclustering algorithm is $O(NML) + O(N\Gamma) + O(N\Gamma\Lambda)$, which is $O(N(ML + \Gamma + \Gamma\Lambda))$. Note that the complexity for identifying the sets of divergent patterns from convergent ones is negligible. Since $\Gamma\Lambda > \Gamma$ and $\Gamma\Lambda > ML$, the overall time complexity is $O(N\Gamma\Lambda)$.

Additional material

Additional file 1: OPTricluster Java package.

Additional file 2: OPTricluster user manual.

Additional file 3: Expression profile of NPR1 in different samples The axis corresponds to the time point experiments, the y-axis the expression level in Log2. Each curve corresponds to a sample.

Additional file 4: Statistic of differences between inner and outer cotyledons The x-axis corresponds to the combination of time points, the y-axis the number of genes.

Additional file 5: Gene Ontology analysis of whole seed *Brassica napus* clusters GO analysis of the 11 clusters in whole seed development *Brassica napus*.

Acknowledgements
This work is jointly supported by Genomics and Health Initiative, Institute for Information Technology, and Plant Biotechnology Institute of the National Research Council Canada. We thank William Leung for implementation of an earlier java version of the OPTricluster core algorithm, and Alex Gawronski for implementation of graphical user interface and tools for cluster merging.

Author details
[1]Knowledge Discovery Group, Institute for Information Technology, National Research Council Canada, 1200 Montréal Road, Ottawa, ON K1A 0R6, Canada. [2]Seed Systems Group, Plant Biotechnology Institute, 110 Gymnasium Place, Saskatoon, SK S7N 0W9, Canada. [3]Oil Crops Research, Institute, Chinese Academy of Agricultural Sciences, No. 2 Xudong 2nd Road, Wuhan, Hubei 430062, China.

Authors' contributions
Designed the algorithm and performed clustering: ABT. Implemented the algorithm in Matlab and Java: ABT. Conceived and designed experiments: ABT, SP, and YP. Performed the *Arabidopsis* and *Brassica* biological experiments and contributed the data: HS, PF, YH, JZ, DH, and AC. Analyzed the data: ABT, SP, ZL, and YP. Wrote the first draft manuscript: ABT. Commented on the manuscript with important intellectual contributions: SP, FF, PF, JZ, ZL, AC and YP. Revised the manuscript: ABT and YP. (All authors read and approved the final manuscript.)

Competing interests
The authors declare that they have no competing interests.

Received: 21 December 2011 Accepted: 4 April 2012
Published: 4 April 2012

References

1. Androulakis IP, Yang E, Almon RR: **Analysis of time-series gene expression data: methods, challenges, and opportunities.** *Annu Rev Biomed Eng* 2007, **9**:205-228.
2. Gardner TS, Faith JJ: **Reverse-engineering transcription control networks.** *Phys Life Rev* 2005, **2**:65-88.
3. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.** *Mol Biol Cell* 1998, **9(12)**:3273-3297.
4. Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP: **Gene expression during the life cycle of Drosophila melanogaster.** *Science* 2002, **297(5590)**:2270-2275.
5. Tchagang AB, Bui KV, McGinnis T, Benos PV: **Extracting biologically significant patterns from short time series gene expression data.** *BMC Bioinformatics* 2009, **10**:255.
6. Ernst J, Nau GJ, Bar-Joseph Z: **Clustering short time series gene expression data.** *Bioinformatics* 2005, **21(Suppl 1)**:i159-i168.
7. Shah M, Jacques Corbeil J: **A general framework for analyzing data from two short time-series microarray experiments.** *IEEE/ACM Trans Comput Biol Bioinform* 2011, **8(1)**:14-26.
8. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22(3)**:281-285.
9. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95(25)**:14863-14868.
10. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci USA* 1999, **96(6)**:2907-2912.
11. Madeira SC, Oliveira AL: **Biclustering algorithms for biological data analysis: a survey.** *IEEE/ACM Trans Comput Biol Bioinform* 2004, **1(1)**:24-45.
12. Tanay A, Sharan R, Shamir R: **Biclustering Algorithms: A survey.** In *In Handbook of Computational Molecular Biology. Volume 1.*. 1 edition. Edited by: Aluru S. New York: Chapman and Hall; 2005:26-27.
13. Tchagang AB, Pan Y, Famili F, Tewfik AH, Benos P: **Biclustering of DNA Microarray Data: Theory, Evaluation, and Applications.** In *In Handbook of Research on Computational and Systems Biology, Interdisciplinary Applications. Volume 1.*. 1 edition. Edited by: Liu L, Wei D, Li Y, Lei H. Hershey, Pennsylvania: IGI Publishing; 2011:148-186.
14. Ben-Dor A, Chor B, Karp R, Yakhini Z: **Discovering local structure in gene expression data: the order-preserving submatrix problem.** *J Comput Biol* 2003, **10(3-4)**:373-384.
15. Tchagang AB, Tewfik AH: **DNA microarray data analysis: a novel biclustering algorithm approach.** *EURASIP J Appl Signal Process* 2006, Article ID 59809, 12 pages.
16. Zhao L, Zaki MJ: **TRICLUSTER: an effective algorithm for mining coherent clusters in 3D microarray data.** *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data* 2005.
17. Araujo R, Trielli G, Orair G, Meira W Jr, Ferreira R, Guedes D: **ParTriCluster: A Scalable Parallel Algorithm for Gene Expression Analysis.** *SBAC-PAD'06: Proceedings of the 18th International Symposium on Computer Architecture and High Performance Computing 2006* .
18. Jiang H, Zhou S, Guan J, Zheng Y: **gTRICLUSTER: a more general and effective 3 d clustering algorithm for gene sample-time microarray data.** *BioDM'06: Proceedings of the 2006 international conference on Data Mining for Biomedical Applications* 2006.
19. **The Gene Expression Omnibus Database.** [http://www.ncbi.nlm.nih.gov/geo/].
20. Bleuler S, Zitzler E: **Order preserving clustering over multiple time course experiments.** *EvoWorkshops'05: Proceedings of the 3rd European conference on Applications of Evolutionary Computing* 2005.
21. Syeda-Mahmood T: **Order-Preserving Clustering and Its Application to Gene Expression Data.** *ICPR'04: Proceedings of the 17th International Pattern Recognition Conference* 2004.
22. **The Java chart library.** [http://www.jfree.org/jfreechart/].
23. Tchagang AB, Gawronski A, Bérubé H, Phan S, Famili F, Pan Y: **GOAL: a software tool for assessing biological significance of genes group.** *BMC Bioinformatics* 2010, **11**:229.
24. Cernetich A, Garver LS, Jedlicka AE, Klein PW, Kumar N, Scott AL, Klein SL: **Involvement of gonadal steroids and gamma interferon in sex differences in response to blood-stage malaria infection.** *Infect Immun* 2006, **74(6)**:3190-203.
25. Jones JD, Dangl JL: **The plant immune system.** *Nature* 2006, **444**:323-9.
26. **The DNA Combimatrix 90 K chip.** [http://www.brassicagenomics.ca/90koligoarray.html].
27. Ohlrogge JB: **Regulation of fatty acid synthesis.** *Annu Rev Plant Physiol Plant Mol Biol* 1997, **48.**
28. Mu J, Tan H, Zheng Q, Fu F, Liang Y, Zhang J, Yang X, Wang T, Chong K, Wang XJ, Zuo J: **LEAFY COTYLEDON1 is a key regulator of fatty acid biosynthesis in Arabidopsis.** *Plant Physiol* 2008, **148**:1042-1054.
29. Braybrook SA, Stone SL, Park S, Bui AQ, Le BH, Fischer RL, Goldberg RB, Harada JJ: **Genes directly regulated by LEAFY COTYLEDON2 provide insight into the control of embryo maturation and somatic embryogenesis.** *PNAS* 2006, **9(103)**:3468-3473.
30. Baud S, Mendoza MS, To A, Harscoët E, Lepiniec L, Dubreucq B: **WRINKLED1 specifies the regulation action of LEAFY COTYLEDON2 towards fatty acid metabolism during seed maturation in Arabidopsis.** *Plant J* 2007, **50**:825-838.
31. Eisenberg E, Levanon EY: **Human housekeeping genes are compact.** *Trends Genet* 2003, **19(7)**:362-365.
32. Wang L, Mitra RM, Hasselmann KD, Sato M, Lenarz-Wyatt L, Cohen JD, Katagiri F, Glazebrook J: **The genetic network controlling the Arabidopsis transcriptional response to Pseudomonas syringae pv. maculicola: roles of major regulators and the phytotoxin coronatine.** *Mol Plant Microbe Interact* 2008, **21**:1408-20.
33. Tewfik AH, Tchagang AB, Vertatschitsch L: **Parallel identification of gene biclusters with coherent evolution.** *IEEE Trans Signal Process* 2006, **54(6)**:2408-2417.

doi:10.1186/1471-2105-13-54
Cite this article as: Tchagang *et al.*: Mining biological information from 3D short time-series gene expression data: the OPTricluster algorithm. *BMC Bioinformatics* 2012 **13**:54.