Visualizing Gene Expression Patterns: A Computational Approach with CAS Software

Kyriaki Tsilika[1] Menelaos Kavouras[2] & Athanasios Exadactylos[2]

[1] , *Laboratory of Operations Research, Department of Economics, University of Thessaly*

[2]*Department of Ichthyology and Aquatic Environment, University of Thessaly*

ktsilika@uth.gr, melios72@yahoo.gr, exadact@uth.gr

Abstract

Gene expression as a biological response to environmental stimulus and as a physiological process of development, is crucial for the study of life. Visualization is an important means to identify complex gene networks. The primary purpose of this study is to provide immediate connection to analytics of biological functions and visualization. We introduce a visual framework in the environment of a main computer algebra system (CAS), *Mathematica*, to picture the differences in gene expression. Our computer codes construct snapshots for gene expression patterns, with the advantage of being self-explanatory contrary to traditional approaches using charts, indices and numerals. They also provide a dynamic interface to facilitate comparisons among genes, indicate gene pairs at a glance and, possibly, help interpret the joint-or interaction-effects that arise. The programming codes along with their application in examples from selected case studies concerning genes involved in embryonic development of common sole (*Solea solea*) are our methodological contribution in the visualization of gene expression patterns. This work could assist researchers in biosciences with suggestions specific to gene expression patterns.

Keywords: Gene expression; common sole; gene expression pattern detection, tabular visualization, *Mathematica* computer software.

JEL Classification: C88; C63.

Οπτική Απεικόνιση των Προτύπων της Γονιδιακής Έκφρασης: Μια Υπολογιστική Προσέγγιση με CAS Λογισμικό

Κυριακή Τσιλίκα[1], Μενέλαος Κάβουρας[2] & Αθανάσιος Εξαδάκτυλος[2]

[1] *Εργαστήριο Επιχειρησιακών Ερευνών Τμήμα Οικονομικών Επιστημών, Πανεπιστήμιο Θεσσαλίας,*

[2]*Τμήμα Ιχθυολογίας & Υδάτινου Περιβάλλοντος, Πανεπιστήμιο Θεσσαλίας, Βόλος, Ελλάδα*

Περίληψη

Η γονιδιακή έκφραση ως βιολογική απόκριση σε περιβαλλοντικά ερεθίσματα και ως φυσιολογική αναπτυξιακή διαδικασία, είναι σημαντική για την κατανόηση και περιγραφή των ζωτικών λειτουργιών των οργανισμών. Η απεικόνιση ως μέσο περιγραφής των προρρηθέντων πολύπλοκων γονιδιακών δικτύων θεωρείται εκ των ων ουκ άνευ. Κύριος σκοπός της παρούσας μελέτης είναι να συνδέσει άμεσα τις βιολογικές λειτουργίες όπως η γονιδιακή έκφραση, με τον τρόπο απεικόνισής τους. Εδώ προτείνεται ένα υπολογιστικό πλαίσιο στο περιβάλλον ενός ισχυρού συστήματος υπολογιστικής άλγεβρας, του *Mathematica*, για τη δημιουργία οπτικών αναπαραστάσεων. Οι υπολογιστικοί κώδικες δημιουργούν στιγμιότυπα των προτύπων της γονιδιακής έκφρασης με αυτοεπεξηγηματική περιγραφή, σε αντίθεση με τις παραδοσιακές μεθόδους που χρησιμοποιούν διαγράμματα, δείκτες και αριθμούς. Η προτεινόμενη υπολογιστική προσέγγιση δύναται να αποτυπώσει οπτικά, σε στατικές και δυναμικές εικόνες, εν δυνάμει γονιδιακά ζεύγη ή δίκτυα, και πιθανώς να αναδείξει – ερμηνεύσει λανθάνουσες (υποκρύπτουσες) συνδέσεις ή αλληλεπιδράσεις. Οι υπολογιστικοί κώδικες και η εφαρμογή τους σε συγκεκριμένη πειραματική μελέτη η οποία αφορά σε γονίδια που εμπλέκονται στην εμβρυϊκή ανάπτυξη του είδους «γλώσσα η κοινή» (*Solea solea)*, αποτελούν μια μεθοδολογική συμβολή στην απεικόνιση των προτύπων της γονιδιακής έκφρασης. Η παρούσα εργασία αφορά στους ερευνητές των βιοεπιστημών που έχουν ως αντικείμενο τα πρότυπα της γονιδιακής έκφρασης.

Λέξεις κλειδιά: Γονιδιακή έκφραση; Γλώσσα η κοινή; Ανίχνευση Μοτίβων Γονιδιακής Έκφρασης; Μητρωική Απεικόνιση; Σύστημα υπολογιστικής άλγεβρας *Mathematica*.

JEL Κωδικοί: C88, C63.

1. Introduction

Over the years databases have progressed from theory, to small text files, to visual representations, to presently include research in genetic (genome-transcriptome) databases.

This work focuses on visual interfaces for gene expression data and especially for RT-qPCR data. We applied visualization to extract and/or verify the conclusions of gene expression analysis and facilitate comparisons when working with transcriptional data. We aimed at understanding different molecular/biomarker patterns which arise among different treatments. We created the computational framework to insert transcriptome data and handle them in a qualitative approach. For this purpose, a main computer algebra system, *Mathematica*, was used to generate visual outputs out of raw data. Specifically, in *Mathematica*'s computational environment, we created pattern constructs consisting of colored patches which described the differential gene expression per treatment. In this way, we provided a concise framework for synthesizing and displaying the RT-qPCR data. Our programming techniques develop three different aspects of related visualization: visual query by genes depicted in static images, comparative schemes for treatments accomplished by dynamic images and numerical data depiction using *Mathematica*'s dynamic visualization options.

Our visual interface was applied empirically by using data from six different embryos batches (eggs) of common sole (*Solea solea*) in order to generate versions of comparative snapshots, with controls added to allow the choice of two spawning seasons and the choice of six batches, three in each season. Our approach allowed for clear visual gene expression comparisons among treatments.

2. Literature Review

The great amount of information that new technologies provide and the complexity of life itself, primary due to nucleic acids proprieties and the high level of polymorphism that is observed, either genetic or phenotypic, require a more holistic overview in every particular type of problem.

A number of computational approaches have been applied to biological data, in order to facilitate their analysis and to attribute them more accessible and intuitive use. They involve a variety of programming languages, software packages, web applications i.e. Perl (Lang et al. 2015), Python (Boher et al. 2015), PHP & Java (Koch et al. 2015) and FuncTree (Uchiyama et al. 2015). Computational approaches of gene expression analysis with *Mathematica* are discussed in (Allen, 2013; Vilar and Saiz, 2010).

Graphical visualizations of gene expression patterns and networks can be found elsewhere (Uchiyama et al. 2015; Lang et al. 2015; Koch et al. 2015; Bohler et al. 2015). Related visualizations have been previously presented (Fails et al., 2006). Although not directly related to the biological field, a set of visualizations that build along the same schematic representations are already a fact (Halkos and Tsilika, 2014; Halkos and Tsilika, 2015a,b,c).

3. Methods and Data

3.1 Data Set Description

Solea solea is a promising species for aquaculture. Anomalies in embryonic stages are a serious setback during intensive rearing, entailing economical, biological and welfare issues.

Our example study utilized data from six different embryo batches (eggs) of common sole (*Solea solea*). Half of each collected from F1 cultured broodstock (already acclimatized for a long time period) and naturally fished breeders (again properly acclimatized), during winter and summer spawning period, respectively. In the following notations, winter period samples are associated with 5,7,9 treatments and summer period samples with E, E2, E3 treatments.

Six (6) target genes of hox family (A1a, A2a, A2b, A13a, B1a, B1b) regarding physiological embryo development were isolated and their expression were estimated 48 hours after spawning.

Table 1: Data matrix of Ct values[1] averages of three biological replications: rows E, E2, E3 correspond to summer spawning period while rows 5, 7, 9 correspond to winter spawning period

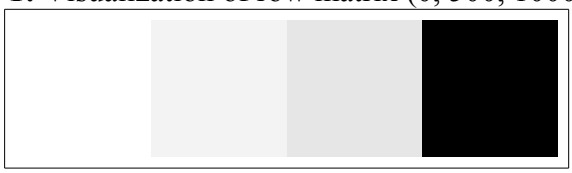
	A1a	**A2a**	**A2b**	**A13a**	**B1a**	**B1b**	**Reference: RPS4**
E	27.70	26.40	23.99	29.17	24.52	22.29	18.22
E2	25.98	24.35	23.70	26.54	24.62	20.86	16.46
E3	28.29	25.39	26.93	24.11	28.24	24.90	19.46
5	29.68	27.60	25.27	28.55	26.03	21.94	19.93
7	27.61	24.95	23.43	25.84	24.71	19.93	17.79
9	30.06	27.75	25.87	29.32	25.69	21.10	19.74

3.2 Methods

In our computational approach, the first step is the loading of experimental data; all experimental data are featured in a number of 1×n lists. Each cell identifies the average Ct value which stands for a specific gene expression level. Then, a matrix results from a list of the predefined lists. Appropriate built-in *Mathematica* functions generate plots, in which, the entries in the matrix are shown in a discrete array of squares.

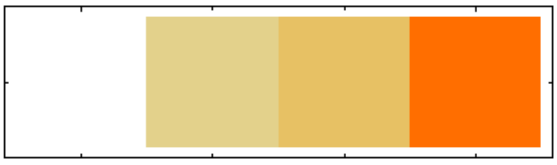
In black-and-white snapshots, white squares indicate the absence of gene expression while gray-shaded squares indicate gene expressions of variant levels. For example, the row matrix (0, 500, 1000, 10000) is depicted by **`ArrayPlot`** function in *Mathematica* as shown in Fig. 1.

Figure 1: Visualization of row matrix (0, 500, 1000, 10000) using gray-shaded squares



Alternatively, we can create colored snapshots using **`MatrixPlot`** function, where each colored cell illustrates a specific gene expression level. In the pattern construct for each treatment, a light colored cell indicates a low gene expression while cells having more intense colors indicate comparatively higher gene expression (Fig. 2)

Figure 2: Visualization of row matrix (0, 500, 1000, 10000) using colored squares

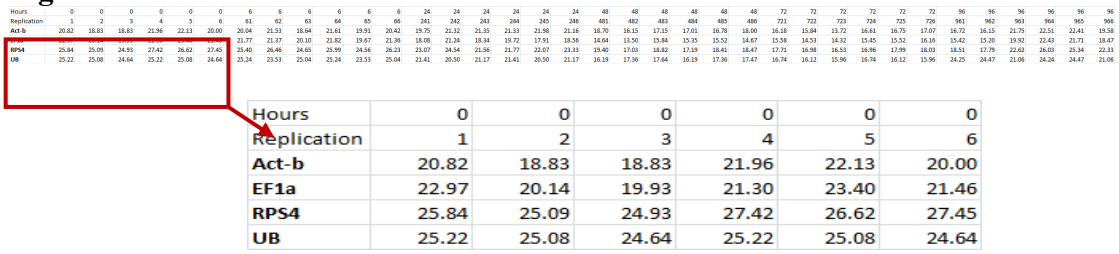


Additionally, the following expression pattern constructs of reference genes (Fig. 4) are created in *Mathematica.* Reference genes are important for RT-qPCR data normalization. Various visual outputs of RT-qPCR matrix (Fig. 3), corresponding to differential expression of four reference genes in sole eggs in a time period of 0, 6, 24, 48 72, 96 h after spawning can be constructed, but the tendencies or patterns observed in each reference are preserved. Each colored cell illustrates a different level of gene expression in each reference. The row

[1] Ct is the number of the cycles required for the fluorescent signal to cross the threshold. Ct levels are inversely proportional to the amount of target nuclear acid in the sample. High Ct values state for low expression

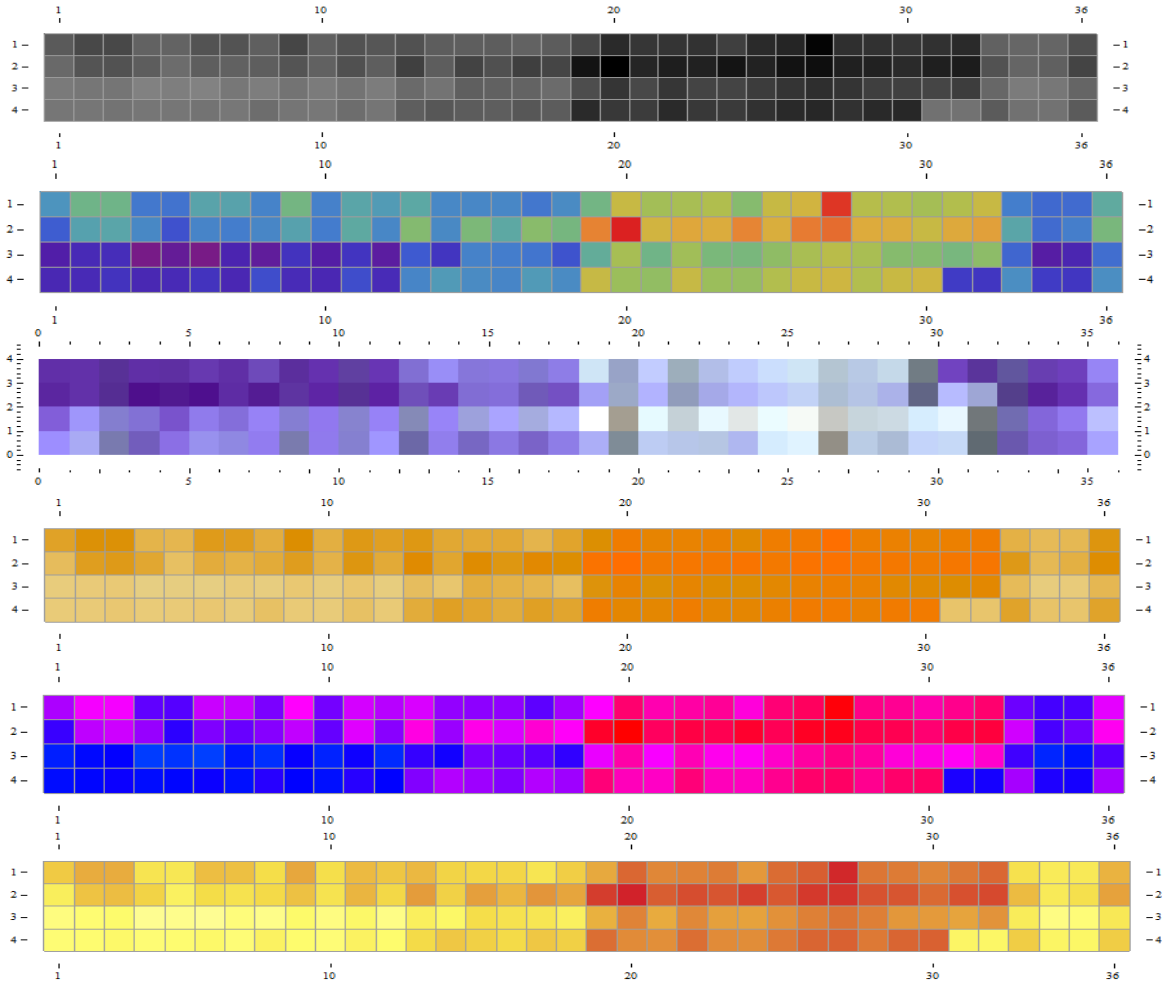
having the minimum color variation indicates stability in expression. Thus the gene with a uniform color pattern seems to be the best candidate for reference.

Figure 3: Row data in excel



Hours	0	0	0	0	0	0
Replication	1	2	3	4	5	6
Act-b	20.82	18.83	18.83	21.96	22.13	20.00
EF1a	22.97	20.14	19.93	21.30	23.40	21.46
RPS4	25.84	25.09	24.93	27.42	26.62	27.45
UB	25.22	25.08	24.64	25.22	25.08	24.64

Figure 4: A *Mathematica* output: several versions of the same plot give an overview of expression within these four genes of interest



4. Empirical Results

The *Mathematica* modules that follow were designed to record RT-qPCR data of par. 3.1 on lists, as well as to convert the indices associated to every gene expression level (Ct values in the technical terminology) to appropriate shades of colors/gray.

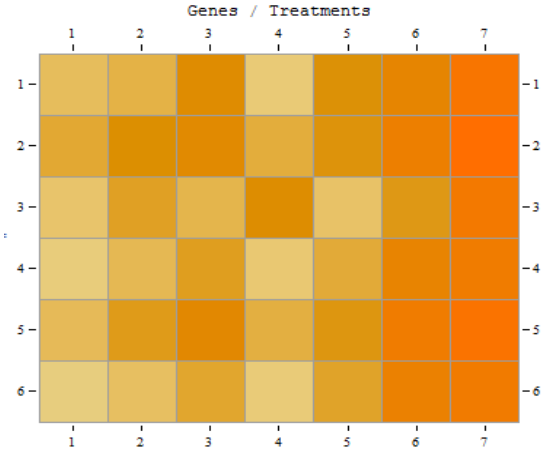
Data are inserted in lists of seven entries: the Ct values from E, E2, E3, 5, 7, 9 rows from Table 1. In *Mathematica*'s internal representation the example variables that must be set by the user are:

```
e:={27.7,26.4,23.99,29.17,24.52,22.29,18.22};
e2:={25.98,24.35,23.7,26.12,24.62,20.55,16.46};
e3:={28.29,25.39,26.93,24.11,28.24,24.9,19.46};
```

```
n5:={29.68,27.6,25.27,28.99,26.03,21.94,19.93};
n7:={27.61,24.95,23.43,26.19,24.71,19.93,17.79};
n9:={30.57,27.75,25.87,29.32,25.69,21.1,19.74};
```

The second step involves several visual schemes of the data (using inverse Ct values) were generated by the following *Mathematica* codes:

```
Labeled[MatrixPlot[{1/e, 1/e2, 1/e3, 1/n5, 1/n7, 1/n9}, FrameTicks -> All, Mesh ->
True, Frame -> True, FrameStyle -> Opacity[0], FrameTicksStyle -> Opacity[1]], {"Genes
/ Treatments"}, Top}]
```




```
Manipulate[MatrixPlot[{{1/n5, 1/n7, 1/n9}[[i]], {1/e, 1/e2, 1/e3}[[j]]}, FrameTicks ->
None,  Mesh -> True, Frame -> True, FrameStyle -> Opacity[0],  FrameTicksStyle ->
Opacity[1]], {i, 1, 3, 1}, {j, 1, 3, 1}]
```

or

```
Manipulate[ArrayPlot[{{1/n5, 1/n7, 1/n9}[[i]], {1/e, 1/e2, 1/e3}[[j]]},
FrameTicks -> None, Mesh -> True, Frame -> True, FrameStyle -> Opacity[0],
FrameTicksStyle -> Opacity[1]], {i, 1, 3, 1}, {j, 1, 3, 1}]
```

Fig. 5 illustrates the use of dynamic visualization options in Mathematica 8, comparing pairwise any winter with any summer treatment, in two different visual schemes.

Figure 5: A snapshot from *Mathematica*: comparing pairwise any winter with any summer treatment.

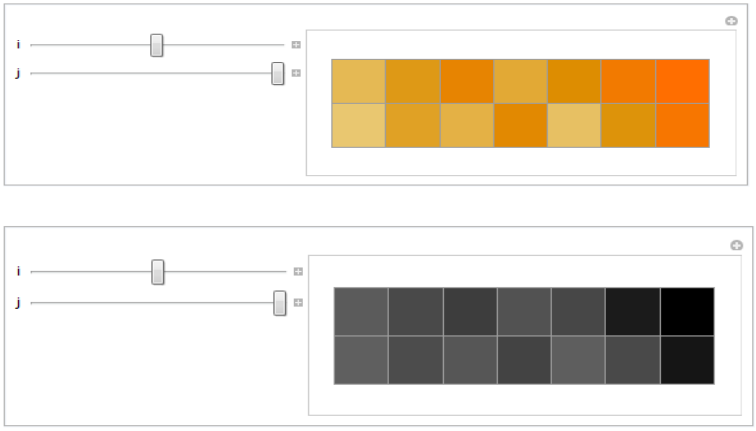


Fig. 6 illustrates the use of dynamic visualization options in Mathematica 8, comparing all treatments pairwise.

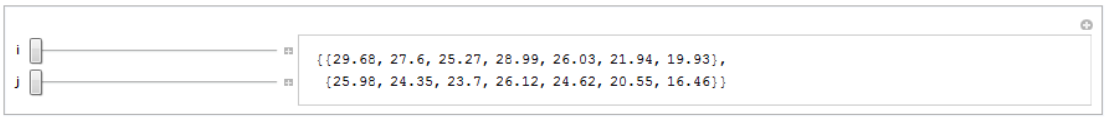
```
Manipulate[ MatrixPlot[{{1/n5, 1/n7, 1/n9, 1/e, 1/e2, 1/e3}[[i]], {1/n5, 1/n7, 1/n9, 1/e, 1/e2, 1/e3}[[j]]}, FrameTicks -> None, Mesh -> True, Frame -> True,  FrameStyle -> Opacity[0], FrameTicksStyle -> Opacity[1]], {i, 1, 6, 1}, {j, 1, 6, 1}]
```

Figure 6: A snapshot from *Mathematica*: comparing winter treatments 5 and 9

Moreover, we provided instant creation of a dynamic interface that allowed varying treatments per season, when comparing numerically different treatments pairwise and gaining useful insights from Ct values dataset (Fig. 7).

Figure 7: A snapshot from *Mathematica*: comparing winter treatment 5 with summer treatment E2

```
Manipulate[{{n5, n7, n9}[[i]], {e2, e3, e}[[j]]}, {i, 1, 3, 1}, {j, 1, 3, 1}]
```



5. Conclusions

This study focused on implementing computer codes of functional programming style in order to develop a tool that enables biologists to place their data in *Mathematica*'s computational environment and accurately depict their gene expression patterns in a variety of snapshots. Our programming codes make use of *Mathematica*'s built-in matrix functions and created an interface that automate the process of creating cognitively and aesthetically compelling representations of RT-qPCR data of different target genes and treatments. Our qualitative approach was applied at the empirical level, with actual estimated values related to the development of *Solea solea* in early life stages. The generated output makes clear that the present computational approach enables complex analyses and sorting strategies.

Acknowledgements

The sampling protocol was facilitated to IMARES, The Netherlands. The access to IMARES RECIRC was funded by the European Union's Seven Framework Programme (FP7/2007-2013) under grant agreement No 262336.

References

Allen T.D (2013). Detecting Differential Gene Expression Using Affymetrix Microarrays, The Mathematica Journal 15, dx.doi.org/doi:10.3888/tmj.15-11. Available from: http://www.mathematica-journal.com/2013/11/detecting-differential-gene-expression-using-affymetrix-microarrays/

Bohler A, Eijssen L.M.T, van Iersel M. P, Leemans C, Willighagen E.L, Kutmon M., Jaillard M. and Evelo C. T (2015). Automatically Visualise and Analyse Data on Pathways using PathVisioRPC from any Programming Environment, BMC Bioinformatics, 16:267. DOI 10.1186/s12859-015-0708-8.

Fails J. A, Karlson A, Shahamat L and Shneiderman B (2006). A Visual Interface for Multivariate Temporal Data: Finding Patterns of Events across Multiple Histories, VAST, 2006, IEEE Symposium on Visual Analytics Science and Technology 2006, IEEE Symposium on Visual Analytics Science and Technology 2006, pp. 167-174, doi:10.1109/VAST.2006.261421

Halkos G.E and Tsilika K. D (2014). Analyzing and Visualizing the Synergistic Impact Mechanisms of Climate Change Related Costs, Applied Mathematics and Computation 246, 586-596, DOI 10.1016/j.amc.2014.08.044.

Halkos G.E and Tsilika K. D (2015a). A Dynamic Interface for Trade Pattern Formation in Multi-regional Multi-sectoral Input-Output Modeling, Computational Economics, DOI 10.1007/s10614-014-9466-3.

Halkos G.E and Tsilika K. D (2015b). Trading Structures for Regional Economies in CAS Software, Computational Economics, DOI 10.1007/s10614-015-9515-6.

Halkos G.E and Tsilika K. D (2015c). Visualizing Trading Relationships in the Framework of Interregional Input-Output Analysis. 6th International Conference on Experiments / Process / System Modeling / Simulation / Optimization (6th IC-EpsMsO), pp. 550-556. (ISBN Vol. II: 978-618-80527-7-2)

Koch A, De Meyer T, Jeschke J and Van Criekinge Wim (2015). MEXPRESS: Visualizing Expression, DNA Methylation and Clinical TCGA Data, BMC Genomics, 16:636. DOI 10.1186/s12864-015-1847-z.

Lang S, Ugale A, Erlandsson E, Karlsson G, Bryder D and Soneji S (2015). SCExV: a Webtool for the Analysis and Visualisation of Single Cell qRT-PCR Data, BMC Bioinformatics, 16:320. DOI 10.1186/s12859-015-0757-z

Shapiro B.E, Levchenko A, Meyerowitz E.M, Wold B.J and Mjolsness E.D (2003). Cellerator: extending a computer algebra system to include biochemical arrows for signal transduction simulations, Bioinformatics, 19, 677-678.

Uchiyama T, Irie M, Mori H, Kurokawa K, Yamada T (2015). FuncTree: Functional Analysis and Visualization for Large-Scale Omics Data. PLoS ONE 10(5): e0126967. doi:10.1371/journal.pone.0126967.

Vilar J. M. G and Saiz L (2010). Gene Expression CplexA: a Mathematica package to study macromolecularassembly control of gene expression, Bioinformatics 26, 2060-2061.