

Boolean implication analysis

FERROLAB SEMINARS 2021

GIOVANNI MICALE

A solid orange horizontal bar spanning the width of the slide, located at the bottom.

Papers

StepMiner:

- D. Sahoo, D. L. Dill, R. Tibshirani, S. K. Plevritis: «Extracting binary signals from microarray time-course data». Nucleic Acids Research, 2007, Vol. 35, No. 11, pp. 3705-3712.

BooleanNet:

- D. Sahoo, D. L. Dill, A. J. Gentles, R. Tibshirani, S. K. Plevritis: «Boolean implication networks derived from large scale, whole genome microarray datasets». Genome Biology, 2008, 9:R 157.

StepMiner

Published online 21 May 2007

*Nucleic Acids Research, 2007, Vol. 35, No. 11 3705–3712
doi:10.1093/nar/gkm284*

Extracting binary signals from microarray time-course data

Debashis Sahoo¹, David L. Dill^{2,*}, Rob Tibshirani³ and Sylvia K. Plevritis⁴

¹Department of Electrical Engineering, ²Department of Computer Science, ³Department of Radiology and

⁴Department of Health Research and Policy and Department of Statistics, Stanford University

Received November 29, 2006; Revised March 2, 2007; Accepted April 11, 2007

ABSTRACT

This article presents a new method for analyzing microarray time courses by identifying genes that undergo abrupt transitions in expression level, and the time at which the transitions occur. The algorithm matches the sequence of expression levels for each gene against temporal patterns

the stimulus?’ and ‘When does the gene transition to up- or down-regulated?’

MATERIALS AND METHODS


StepMiner extracts three types of binary temporal patterns. The first type, shown in Figure 1(a and b),

StepMiner

Given time-series microarray expression data, find:

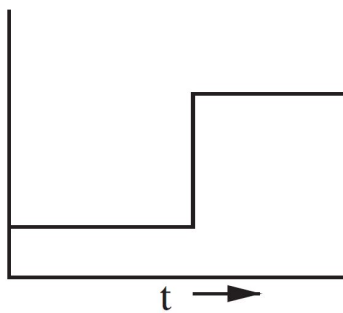
- a) Genes that undergo abrupt binary transitions (from low to high values or viceversa) in expression level;
- b) The time at which the transitions occur.

Answer to biological questions:

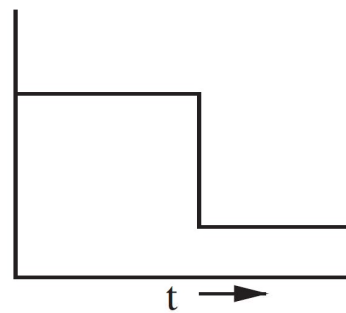
- a) Which genes are up-regulated or down-regulated as a result of a stimulus?
 - b) When does the gene transition to up- or down-regulated?
- 

StepMiner

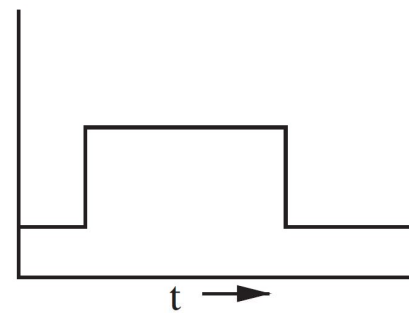
For each gene, the algorithm looks for temporal patterns having one or two transitions between two expression levels.



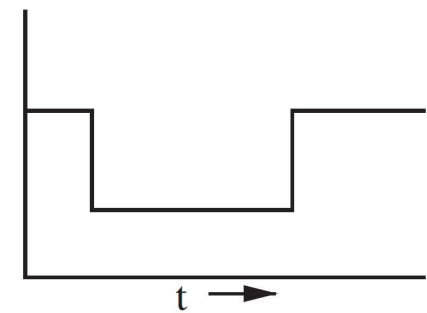
(a) One step (Up)



(b) One step (Down)



(c) Binary two step
(Up-Down)



(d) Binary two step
(Down-Up)

Finding the best one-step transition

Evaluates every possible placement of the transitions between time points.

Adaptive regression: choose the placement that gives the best fit.

Given a time series of n expression values X_1, X_2, \dots, X_n the best fit is the one that minimizes the Sum of Square Error (SSE) between the real values and the fitted values $\hat{X}_1, \hat{X}_2, \dots, \hat{X}_n$.

$$SSE = \sum_{i=1}^n (X_i - \hat{X}_i)^2$$

If k is the position of the transition, fitted values are computed as:

$$\hat{X}_i = \begin{cases} \frac{\sum_{j=1}^k X_j}{k} & \text{if } 1 \leq i \leq k \\ \frac{\sum_{j=k+1}^n X_j}{n-k} & \text{if } k+1 \leq i \leq n \end{cases}$$

Finding the best two-steps transition

The above procedure holds for two-steps transitions too.

Assuming that the first and the third segment of the two-steps function have the same value, the fitted values are computed as:

$$\hat{X}_i = \begin{cases} \frac{\sum_{j=1}^k X_j}{k} & \text{if } 1 \leq i \leq k_1 \text{ or } k_2 \leq i \leq n \\ \frac{\sum_{j=k+1}^n X_j}{n-k} & \text{if } k_1 + 1 \leq i \leq k_2 - 1 \end{cases}$$

where k_1 and k_2 are the transition points.

P-value of fitting

A p-value of fitting is calculated considering the following statistics:

$$F = \frac{\sum_{i=1}^n (\hat{X}_i - \bar{X})^2 / (m-1)}{\sum_{i=1}^n (X_i - \hat{X}_i)^2 / (n-m)}$$

where \bar{X} is the mean of the gene expression values.

Statistic F follows an F-distribution with $(m - 1, n - m)$ degrees of freedom.

If \mathcal{F}_{n-m}^{m-1} is a random variable with an F-distribution, the p-value corresponds to the tail of the F-distribution:

$$P = \Pr[\mathcal{F}_{n-m}^{m-1} > F]$$

Selecting the best model

Three possible outcome for each gene:

- 1) A significant one-step function exists that can fit its expression pattern;
- 2) A significant two-steps function exists that can fit its expression pattern;
- 3) No significant one- or two-steps functions exist that can fit its expression pattern;

To identify which is the best pattern describing a gene, we need to compute the F-statistics F_1 and F_2 for the one-step and the two-steps transitions, respectively, and their corresponding p-values of significance.

We also need to introduce another F-statistic F_{12} indicating the relative goodness of fit of a one-step vs a two-steps pattern.

Selecting the best model

Let $(m_1 - 1, n - m_1)$ and $(m_2 - 1, n - m_2)$ the degrees of freedom of F_1 and F_2 , respectively.

Let SSE_1 and SSE_2 the Sum of Square Error for the one-step and the two-steps transitions, respectively.

The F_{12} statistic is defined as:

$$F_{12} = \frac{(SSE_1 - SSE_2)/(m_2 - m_1)}{SSE_2/(n - m_2)}$$

Selecting the best model

The following algorithm is used to select the best model:

```
SelectBestModels(){  
    oneStep = F-Significant( $F_1$ ) && Not-F-Significant( $F_{12}$ )  
    twoStep = F-Significant( $F_2$ ) && NotIn(oneStep)  
    other = NotIn(oneStep, twoStep)  
}
```

FDR correction

In StepMiner a «false discovery» is found whenever the algorithm finds a one-step or a two-steps function for a gene, but the data contains no step.

The False Discovery Rate (FDR) in StepMiner is the ratio of false discoveries to true discoveries.

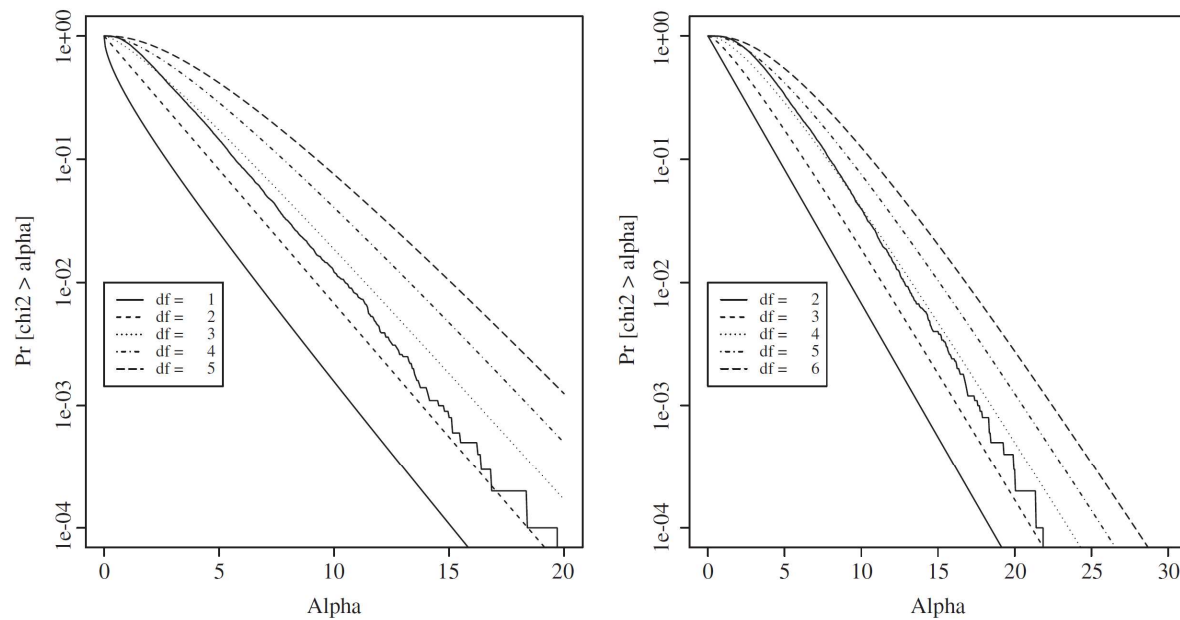
To estimate the FDR, many random permutations of the time points are computed and StepMiner is run on each permuted input.

FDR is estimated as the ratio between the average number of significant genes in the null model and the original number of significant genes.

The FDR can be adjusted by setting a P-value threshold in the fitting algorithm.

Degrees of freedom

Degrees of freedom m_1 and m_2 are derived from random simulations and approximated as 3 and 4, respectively.



Experiments on real data

<http://genomics-pubs.princeton.edu/DiauxicRemodeling/data.shtml>

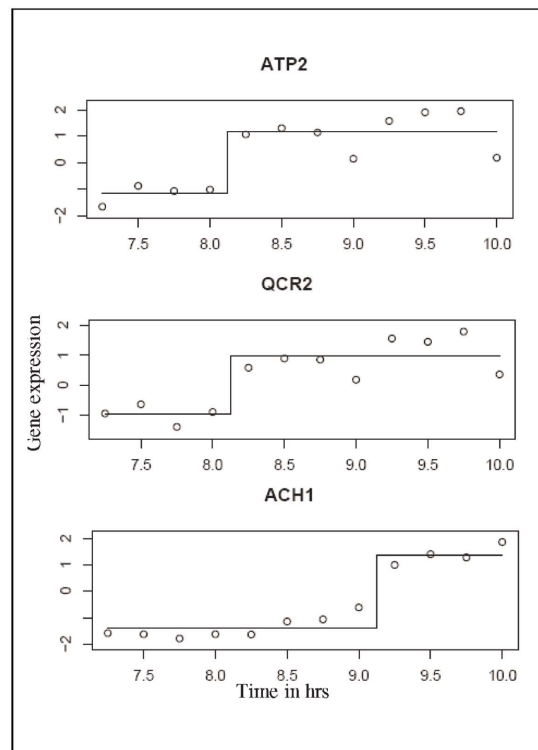
Gene expression levels in yeast during the diauxic shift in a glucose-limited culture.

The yeast utilizes fermentative metabolism when glucose is abundant.

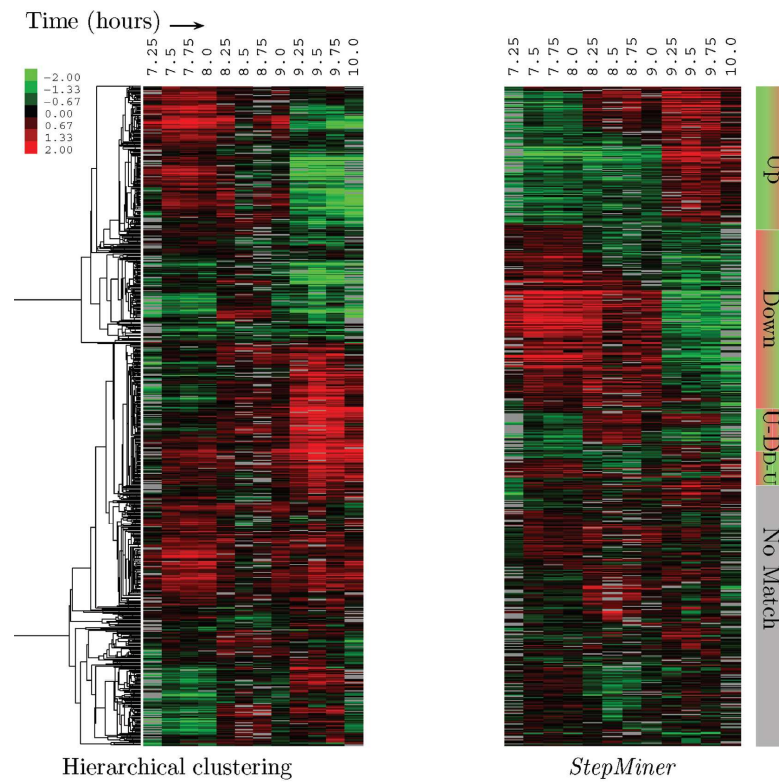
As the glucose is depleted, the metabolism shifts abruptly to oxidative metabolism.

RNA samples collected at 12 different timepoints every 15 minutes in 2284 genes.

Example of step functions



Comparison with hierarchical clustering



BooleanNet

Open Access

Method

Boolean implication networks derived from large scale, whole genome microarray datasets

Debashis Sahoo^{*}, David L Dill[†], Andrew J Gentles[‡], Robert Tibshirani[§] and Sylvia K Plevritis[‡]

Addresses: ^{*}Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA. [†]Department of Computer Science, Stanford University, Stanford, CA 94305, USA. [‡]Department of Radiology, Stanford University, Stanford, CA 94305, USA. [§]Department of Health Research and Policy and Department of Statistics, Stanford University, Stanford, CA 94305, USA.

Correspondence: David L Dill. Email: dill@cs.stanford.edu

Published: 30 October 2008

Genome Biology 2008, **9**:R157 (doi:10.1186/gb-2008-9-10-r157)

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2008/9/10/R157>

Received: 28 June 2008

Revised: 6 September 2008

Accepted: 30 October 2008

BooleanNet

Extract boolean implications between expression levels of genes in the form of «if-then» relationships from *very large* amounts of gene expression data (typically thousands of samples).

Boolean implications express relationships among genes that are invariant for the sample set.

To build implications, expression values of each gene are first discretized as «low», «intermediate» and «high» based on an automatically derived threshold.

Thresholds are derived individually for each gene.

Types of boolean implications

Asymmetric:

- a) Low-low: IF gene A is low THEN gene B is low;
- b) Low-high: IF gene A is low THEN gene B is high;
- c) High-low: IF gene A is high THEN gene B is low;
- d) High-high: IF gene A is high THEN gene B is high;

Each asymmetric implication has a contrapositive relationship (e.g. A low \Rightarrow B high is identical to B low \Rightarrow A high)

Symmetric:

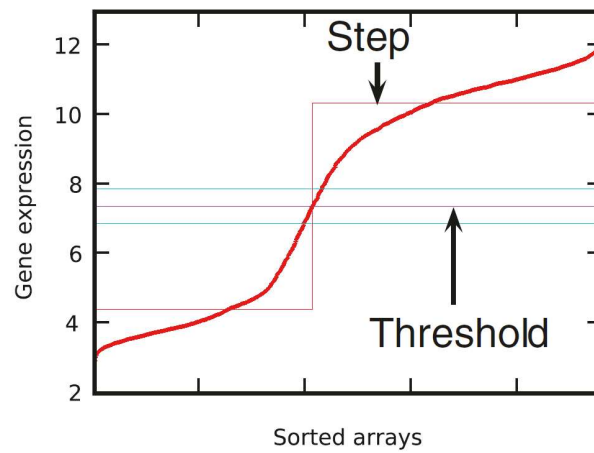
- a) Equivalent (correlation): IF gene A is low THEN gene B is low AND IF gene A is high THEN gene B is high;
- b) Opposite (anti-correlation): IF gene A is low THEN gene B is high AND IF gene A is high THEN gene B is low;

Discretization of expression values

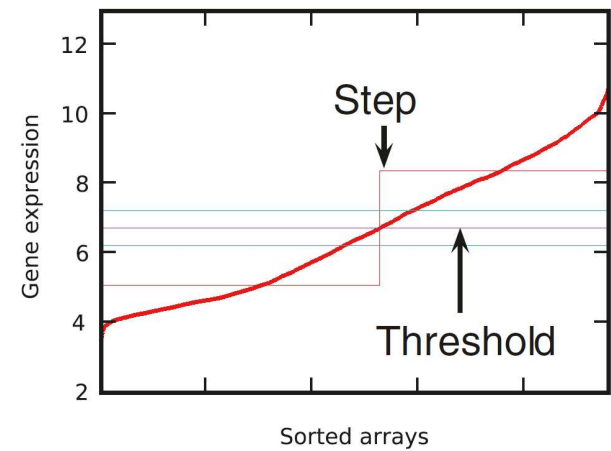
- 1) Order expression values of gene g from the lowest to the highest;
- 2) Fit a one-step rising function using StepMiner;
- 3) If k is the position of the transition, set the threshold T to the expression value of rank k .
- 4) Given a parameter σ (e.g. $\sigma = 0.5$ for log-transformed expression values) set:
 - a) All expression values $v < T - 0.5$ as «low»;
 - b) All expression values $v > T + 0.5$ as «high»;
 - c) All expression values v between $T - 0.5$ and $T + 0.5$ as «intermediate».

Examples

CDH1



CDC2



Discovery of boolean implications

Implications between two genes g_1 and g_2 are investigated by considering all possible combination of «low» and «high» values of the two genes across samples.

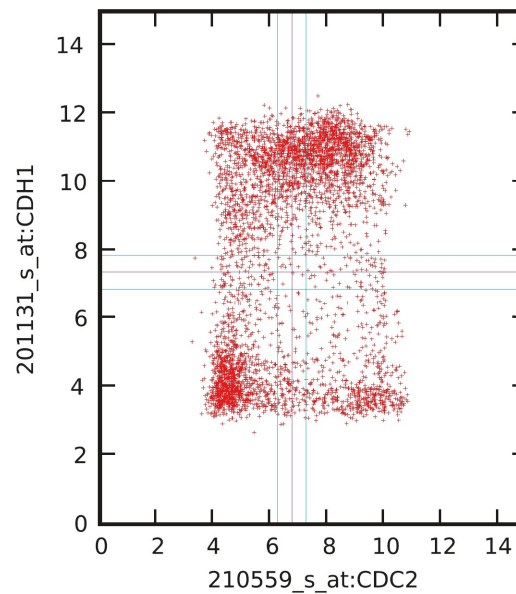
In this step of analysis, «intermediate» values are considered as potential noise and thus ignored.

All genes with at least 2/3 of expression values with an «intermediate» value are excluded.

Discovery of boolean implications

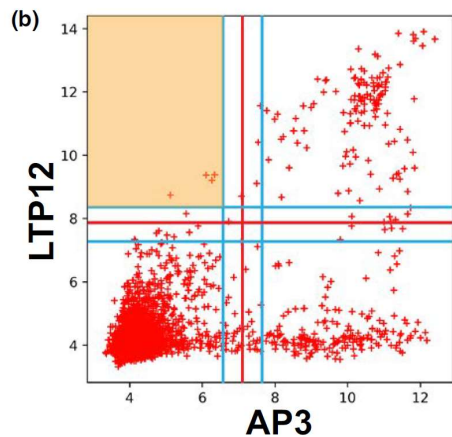
Suppose to plot all pairs of expression values of g_1 and g_2 in each sample in a scatter plot.

Four quadrants can be identified: «low-high», «high-high», «high-low» and «low-low».

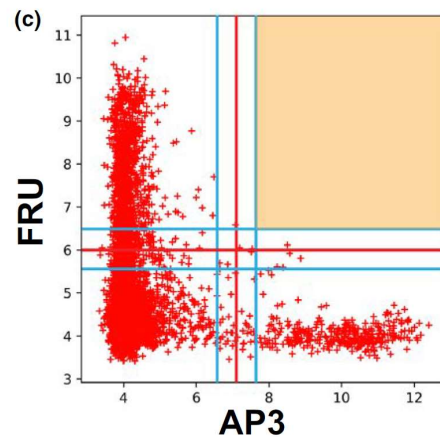


Asymmetric implications

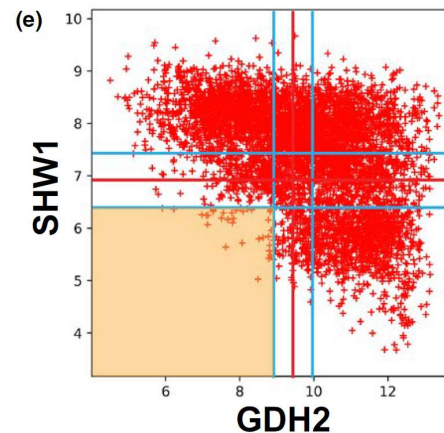
Check if one quadrant is significantly sparsely populated with points compared to the other quadrants.



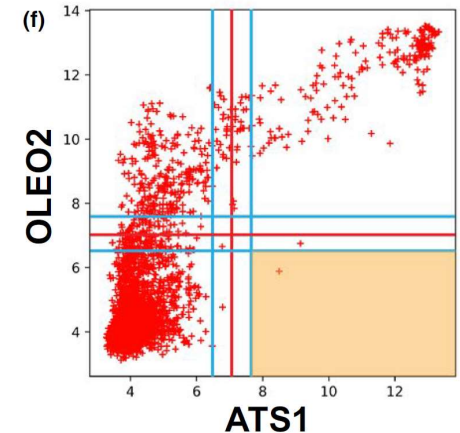
AP3 low => LTP12 low



AP3 high => FRU low



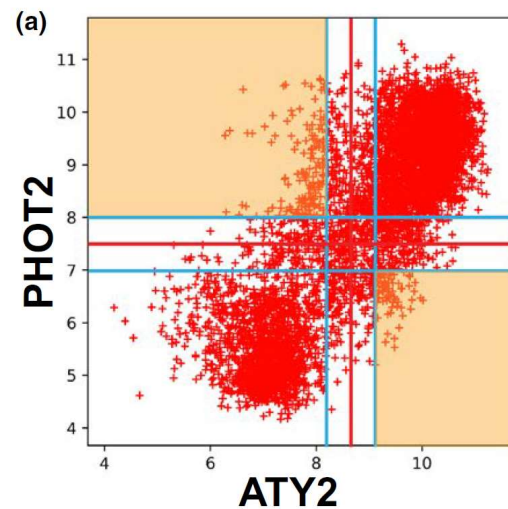
GDH2 low => SHW1 high



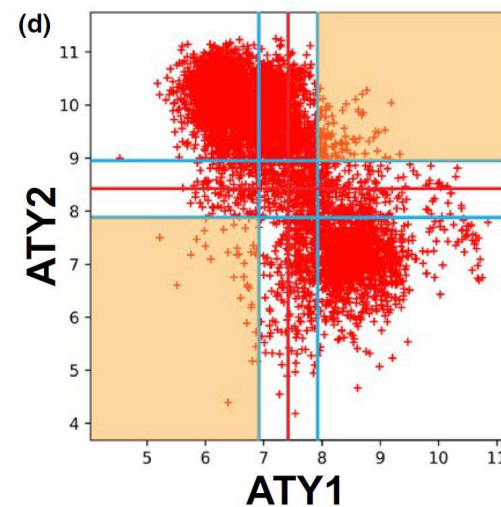
ATS1 high => LTP12 high

Symmetric implications

Check if two quadrants are significantly sparsely populated with points compared to the other quadrants.



ATY2 low \Rightarrow PHOT2 low AND
ATY2 high \Rightarrow PHOT2 high



ATY1 low \Rightarrow ATY2 high AND
ATY1 high \Rightarrow ATY2 low

Discovery of boolean implications

Suppose that the sparse quadrant is «low-low», potentially yielding an implication $A \text{ low} \Rightarrow B \text{ high}$.

First, the number of expression values in the sparse quadrant must be significantly less than the number that would be expected under an independence model, where we assume that genes A and B are independent.

Let:

- a_{00} the number of samples where both A and B are «low»;
- a_{01} the number of samples where A is «low» and B is «high»;
- a_{10} the number of samples where A is «high» and B is «low»;
- a_{11} the number of samples where both A and B are «high».

Discovery of boolean implications

Under the independence assumption, the expected number of points in the sparse quadrant is:

$$\widehat{a}_{00} = \left(\frac{nA_{low}}{total} \times \frac{nB_{low}}{total} \right) \times total = \frac{(nA_{low} \times nB_{low})}{total}$$

where:

- $total = a_{00} + a_{01} + a_{10} + a_{11};$
- $nA_{low} = a_{00} + a_{01};$
- $nB_{low} = a_{00} + a_{10};$

The statistic of the independence test is:

$$S_{00} = \frac{\widehat{a}_{00} - a_{00}}{\sqrt{\widehat{a}_{00}}}$$

Discovery of boolean implications

The observed values in the sparse quadrant are considered erroneous points.

A sparse quadrant must have a small number of erroneous points.

A maximum likelihood estimate of the error rate is computed as:

$$p_{00} = \frac{1}{2} \left(\frac{a_{00}}{a_{00}+a_{01}} + \frac{a_{00}}{a_{00}+a_{10}} \right)$$

The implication A low \Rightarrow B high is considered as significant iff:

- $S_{00} > sThr$;
- $p_{00} < pThr$.

Values used in the experiments reported in the BooleanNet paper are $sThr = 3$ and $pThr = 0.1$.

Discovery of boolean implications

A similar analysis is done to identify the other types of implications:

$$A \text{ low} \Rightarrow B \text{ low} \quad S_{01} > sThr, P_{01} < pThr$$

$$A \text{ high} \Rightarrow B \text{ high} \quad S_{10} > sThr, P_{10} < pThr$$

$$A \text{ high} \Rightarrow B \text{ low} \quad S_{11} > sThr, P_{11} < pThr$$

$$\text{Equivalent} \quad S_{01} > sThr, P_{01} < pThr, S_{10} > sThr, P_{10} < pThr$$

$$\text{Opposite} \quad S_{00} > sThr, P_{00} < pThr, S_{11} > sThr, P_{11} < pThr$$

FDR correction

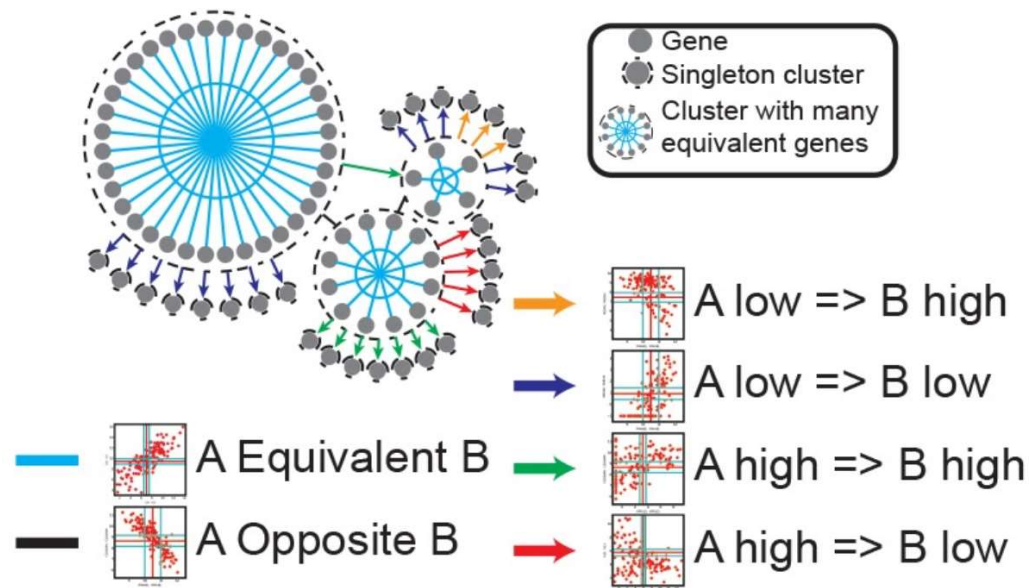
Given the large number of genes and potential relationships, it is necessary to evaluate the significance of the discovered relationships.

A FDR is computed by randomly permuting the expression values for each gene independently several times.

The FDR is the ratio between the average number of boolean relationships in the randomized data and the observed number of relationships.

Boolean implication network

Labeled directed graph where vertices are genes and edges are implications, labeled with the implication type.



Implication vs causality

A boolean implication does not necessarily imply causality.

For example, $A \text{ high} \Rightarrow B \text{ high}$ means that the set of samples where A is high is a subset of the set of samples where B is high.

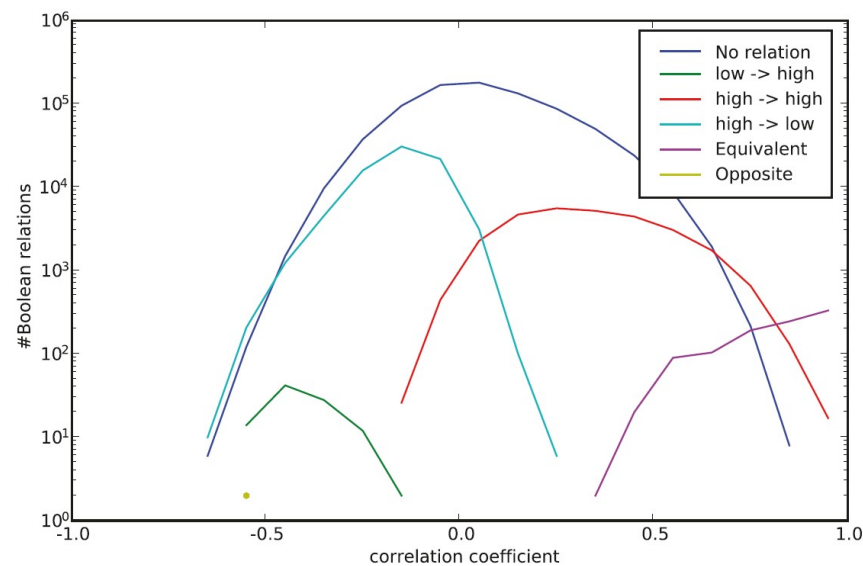
Possible interpretations:

- B is specific to a particular cell type or tissue type and A is specific to a subclass of it;
- A is *one of several* transcription factors that increases expression of B or vice versa.

Implication vs correlation

Boolean implications capture many more relationships than simple correlation.

There may be a highly significant implication between genes with weakly correlated expressions.



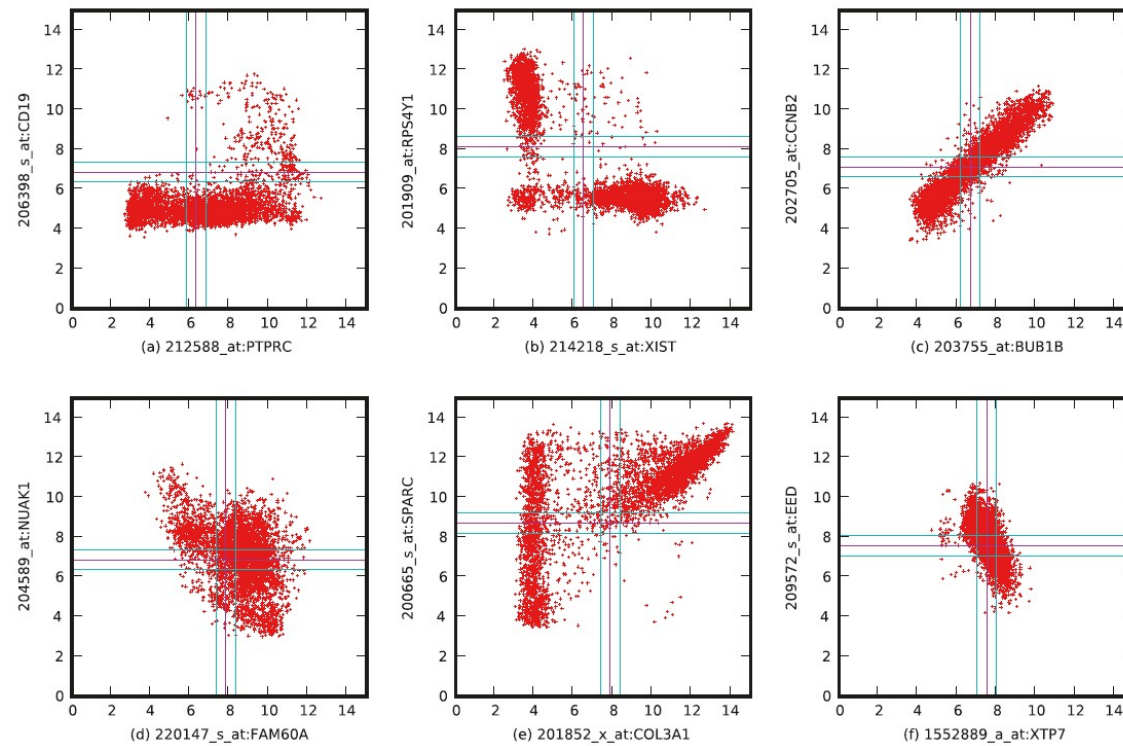
Experimental results

Collection of Affymetrix arrays from three different species:

- Human: 4,787 Affymetrix U133Plus 2.0 arrays;
- Mouse: 2,154 Affymetrix 430 2.0 arrays;
- Fly: 450 Affymetrix Genome 1.0 Drosophila arrays.

Data are normalized using RMA and log2-transformed before using BooleanNet.

Boolean relationships in human



Biological interpretation of implications

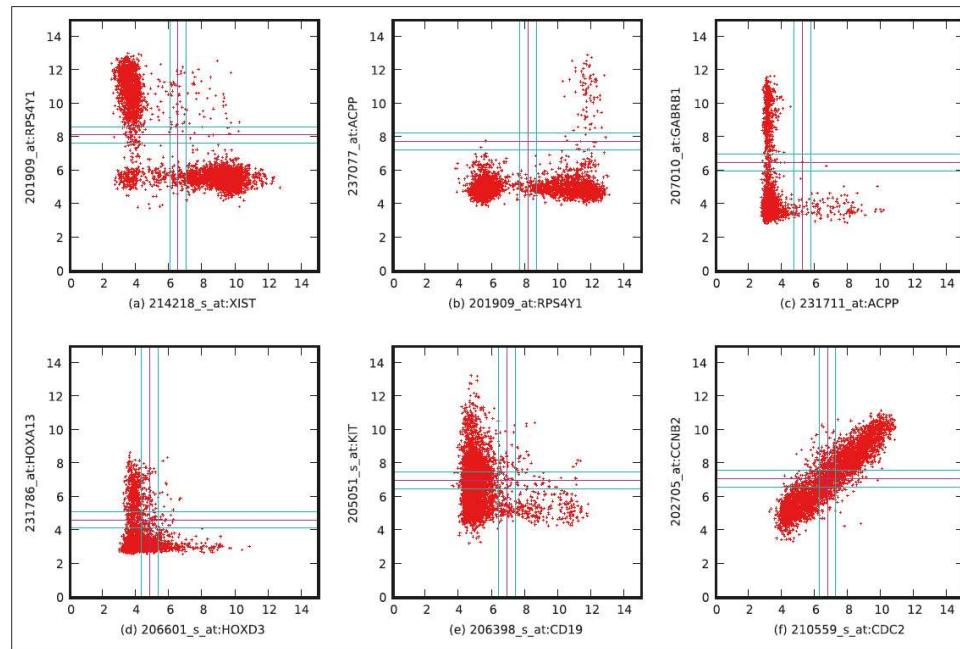


Figure 2

Boolean relationships follow known biology. (a) Gender difference, XIST high \Rightarrow RPS4Y1 low, male and female genes are not expressed in the same sample. (b) Gender tissue specific, RPS4Y1 low \Rightarrow ACPP low, prostate cells are from males. (c) Tissue difference, ACPP high \Rightarrow GABRB1 low, prostate and brain genes are not expressed in the same samples. (d) Development, HOXD3 high \Rightarrow HOXA13 low, anterior is different from posterior. (e) Differentiation, KIT high \Rightarrow CD19 low, differentiated B cell is different from hematopoietic stem cell. (f) Co-expression, CDC2 versus CCNB2.

Number of boolean implications found

Number (in millions) of Boolean relationships in human, mouse and fruit fly datasets

Dataset	Total	Low implies high	High implies low	Low implies high	High implies low	Equivalent	Opposite
Human	208	2	128	38	38	1.6	0.4
Mouse	336	8	208	57.6	57.6	4.1	0.7
Fruit fly	17	0.3	7.3	3.7	3.7	1.9	0.1

Conserved implications

