

Transcription Factor Identification: a Bayesian Approach

José Miguel Hernández-Lobato, Tom Heskes and Tjeerd Dijkstra

Department of Information and Knowledge Systems (IRIS).
Radboud University. Nijmegen.

December 14, 2006

Outline

- 1 Introduction
- 2 Problem description
- 3 A simpler problem
- 4 A Bayesian model for transcription
- 5 Looking for transcription factors
- 6 Final notes

Outline

- 1 Introduction
- 2 Problem description
- 3 A simpler problem
- 4 A Bayesian model for transcription
- 5 Looking for transcription factors
- 6 Final notes

Microarray chips

- 1 Allow to simultaneously measure the level of expression of many genes (**RNA transcripts**) within a cell.
 - 2 RNA transcripts are reverse transcribed to **dyed** cDNA.
 - 3 Chips have **spots** with the complementary strands for the dyed cDNA of the genes.
 - 4 The **amount of dye** on each spot indicates the **level of expression** for each gene.
- Images of microarrays are analyzed by computer to get a final measurement of expression level.

Example of microarray

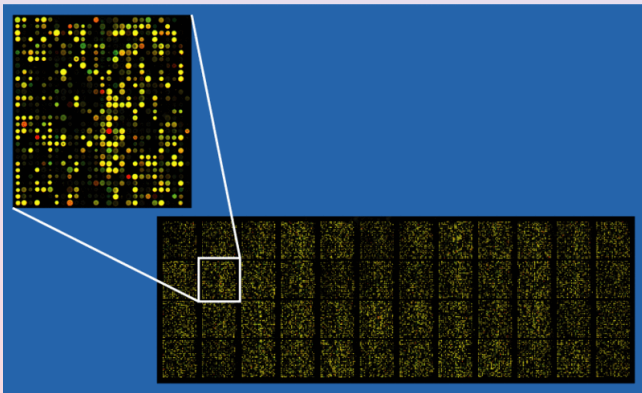
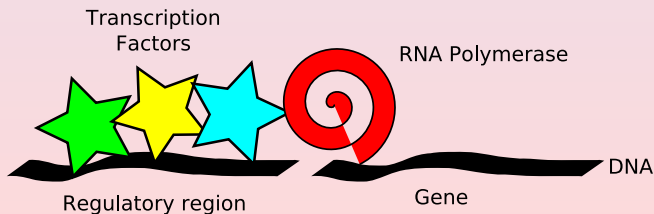


Figure: An approximately 40,000 probe spotted microarray.

Transcription factors (TF)

- 1 Are proteins that control the expression level of other genes (RNA transcripts).
- 2 The expression of a TF is **correlated** in time with the expression of the genes it regulates.
- 3 Correlations can be appreciated in consecutive microarray experiments.



Consecutive microarray experiment

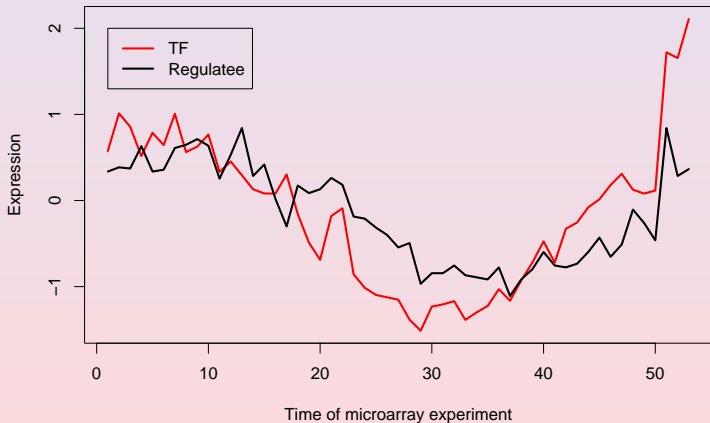


Figure: How a transcription factor could influence another gene.

Outline

- 1 Introduction
- 2 Problem description**
- 3 A simpler problem
- 4 A Bayesian model for transcription
- 5 Looking for transcription factors
- 6 Final notes

Problem description

Objective

- Given the results of a consecutive microarray experiment we want to **identify** the genes that are **transcription factors**.

Main difficulties

- Microarray experiments contain only **a few measurements** for each gene.
- There are **thousands of genes** and most of them are **correlated** with each other.
- Measurement **error** is big.

Example

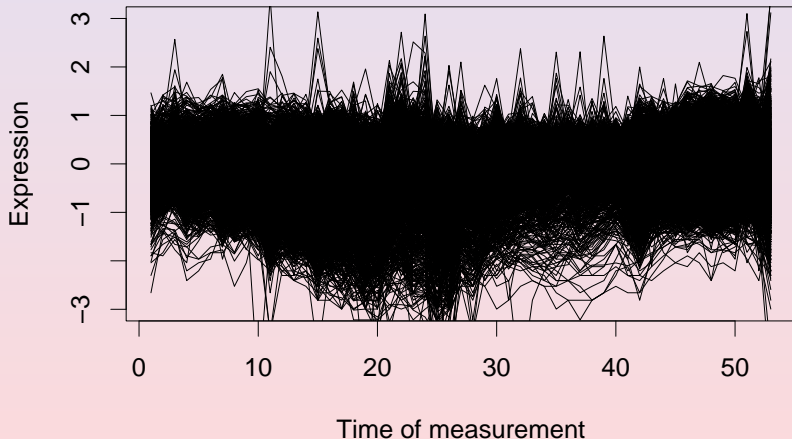


Figure: Expression level for 4199 genes of *Plasmodium falciparum* 3D7.

Proposed solution

We apply Bayesian inference

- We propose a probabilistic model \mathcal{M} for transcription.
- We define a variable t_i which takes value 1 if gene i is a transcription factor and 0 otherwise.
- Given the data \mathcal{D} of a microarray experiment, the probability of each gene to be a TF is

$$\mathcal{P}(\mathbf{t}|\mathcal{D}, \mathcal{M}) = \frac{\mathcal{P}(\mathcal{D}|\mathbf{t}, \mathcal{M})\mathcal{P}(\mathbf{t})}{\mathcal{P}(\mathcal{D}|\mathcal{M})} \quad (1)$$

- \mathcal{M} should be **simple** if we want to be able to compute (1).

Outline

- 1 Introduction
- 2 Problem description
- 3 A simpler problem**
- 4 A Bayesian model for transcription
- 5 Looking for transcription factors
- 6 Final notes

Bayesian variable selection for the linear model

Problem

- Vector \mathbf{y} contains the expression of one gene.
- Vectors $\mathbf{x}_1, \dots, \mathbf{x}_p$ contain the expressions of the candidates to be TF (all the other genes).
- Which of $\mathbf{x}_1, \dots, \mathbf{x}_p$ regress \mathbf{y} ?

Example

The transcription factors that regulate gene \mathbf{y} are genes \mathbf{x}_1 and \mathbf{x}_2 :

$$\mathbf{y} = \mathbf{x}_1 - \frac{1}{2}\mathbf{x}_2 + 0\mathbf{x}_3 + 0\mathbf{x}_4 + \dots + 0\mathbf{x}_p$$

Solution to the variable selection problem

Bayesian solution based on sampling: George and McCulloch 1994.

- 1 r_i indicates when x_i is a regressor of y ($r_i = 1$) or not ($r_i = 0$).
- 2 c_i is the regression coefficient between y and x_i .
- 3 If $r_i = 1$ then c_i is different from 0, otherwise $c_i \simeq 0$:

$$\mathcal{P}(c_i|r_i) = r_i\mathcal{N}(c_i; 0; v_1) + (1 - r_i)\mathcal{N}(c_i; 0; v_0),$$

where $v_0 \simeq 0$ and v_1 is big.

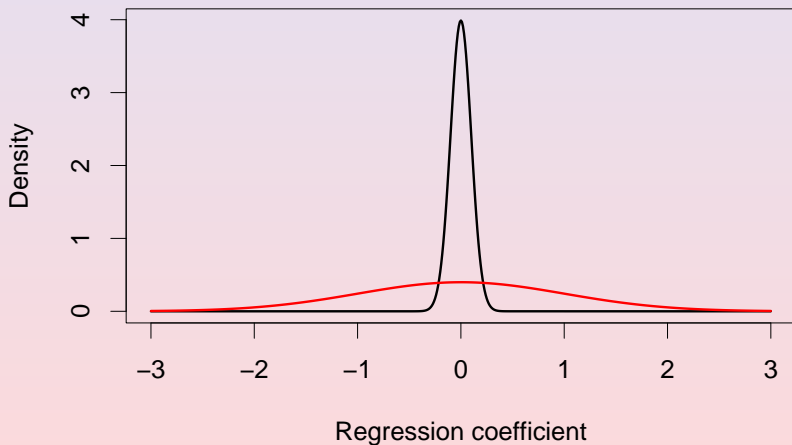
Densities $\mathcal{P}(c_i|r_i)$ 

Figure: In red $\mathcal{P}(c_i|r_i = 1)$ and in black $\mathcal{P}(c_i|r_i = 0)$.

Solution to the variable selection problem

- We assume a **Gaussian error** with variance $\frac{\sigma^2}{2}$ in the measurements for \mathbf{y} and $\mathbf{x}_1, \dots, \mathbf{x}_p$.
- If \mathbf{X} is the matrix with $\mathbf{x}_1, \dots, \mathbf{x}_p$ as columns, we have that

$$\mathcal{P}(\mathbf{r}, \mathbf{c}, \sigma^2 | \mathbf{y}, \mathbf{X}) \propto \mathcal{N}(\mathbf{y}; \mathbf{X}\mathbf{c}; \sigma^2 I) \mathcal{P}(\mathbf{c} | \mathbf{r}) \mathcal{P}(\mathbf{r}) \mathcal{P}(\sigma^2) \quad (2)$$

$$\mathcal{P}(\mathbf{c} | \mathbf{r}) = \prod_i \mathcal{P}(c_i | r_i)$$

$$\mathcal{P}(c_i | r_i) = r_i \mathcal{N}(c_i; 0; v_1) + (1 - r_i) \mathcal{N}(c_i; 0; v_0)$$

- We can approximate the left part of (2) by **expectation propagation** (much faster than sampling).

Outline

- 1 Introduction
- 2 Problem description
- 3 A simpler problem
- 4 A Bayesian model for transcription**
- 5 Looking for transcription factors
- 6 Final notes

Intuition

- We can extend the variable selection method for regression to a Bayesian model for transcription very easily.
- 1 We just have to perform the regression of the expression of **each gene** delayed some time **against all the others**.
 - 2 If a gene is a transcription factor it should appear as a regressor many times.

Probabilistic formulation I

- $\mathbf{x}_i^{(-1)}$ represents the expression of gene i delayed one unit in time and $\mathbf{x}_i^{(0)}$ the expression without any delay.
- $r_{i,j} = 1$ when $\mathbf{x}_j^{(0)}$ is a regressor of $\mathbf{x}_i^{(-1)}$ and $r_{i,j} = 0$ otherwise.
- Then, $\mathcal{P}(r_{i,j} = 1 | t_j = 1) = w_1$ and $\mathcal{P}(r_{i,j} = 1 | t_j = 0) = w_0$ where $w_1 > w_0$

Probabilistic formulation II

- If \mathbf{X}_{-i} is the matrix with $\mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_p$ as columns and \mathbf{c}_{-i} is the vector of coefficients $c_{i,j \neq i}$, we have that

$$\mathcal{P}(\mathbf{R}, \mathbf{C}, \mathbf{t}, \sigma^2 | \mathbf{X}) \propto \prod_{i=1}^p \mathcal{N}(\mathbf{x}_i^{(-1)}; \mathbf{X}_{-i} \mathbf{c}_{-i}; \sigma^2 I)$$

$$\mathcal{P}(\mathbf{C} | \mathbf{R}) \mathcal{P}(\mathbf{R} | \mathbf{t}) \mathcal{P}(\mathbf{r}) \mathcal{P}(\sigma^2)$$

- Again we can approximate the posterior distribution by expectation propagation. This time the sampling methods are not feasible.

Example 1

- We generated the expression for a TF as $z \sim \mathcal{N}(0, 3I)$.
 - We generated the expression for 49 genes as $x_j = z^{(1)}$.
 - We stored 50 observations of x_1, \dots, x_{49} and z in a dataset adding a measurement error of $\mathcal{N}(0, 3I)$.
-
- We ran the algorithm for TF identification with $w_1 = 0.9$, $w_0 = 0.1$ and the prior for a gene to be a TF is set to 0.02.

Results

- The TF is identified with the highest probability.

Dataset used

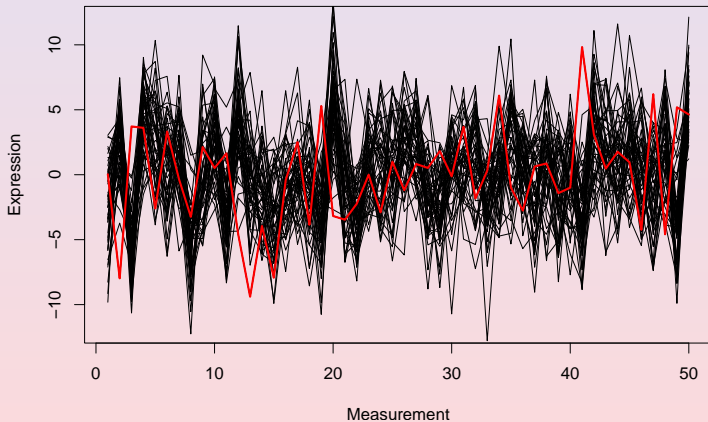


Figure: Dataset used in example 1. In red the TF.

Example 2

- This time the TF z follows a **smoothed curve**.
 - We generated the expression for 49 genes as $x_i = z^{(1)}$.
 - We stored 50 observations of x_1, \dots, x_{49} and z in a dataset adding a Gaussian error with $sd = \frac{1}{3}sd(z)$.
-
- We ran the algorithm for TF identification with $w_1 = 0.9$, $w_0 = 0.1$ and the prior for a gene to be a TF is set to 0.02.

Results

- Again the TF obtained the highest probability among all the other genes.

Dataset used

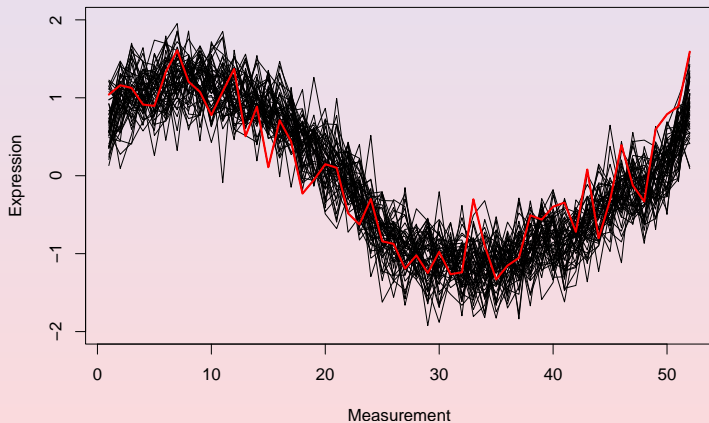


Figure: Dataset used in example 2. In red the TF.

Outline

- 1 Introduction
- 2 Problem description
- 3 A simpler problem
- 4 A Bayesian model for transcription
- 5 Looking for transcription factors**
- 6 Final notes

Data acquisition and preprocessing

- 1 We took the expression dataset for the IDC of *Plasmodium falciparum* 3D7 (<http://malaria.ucsf.edu>).
- 2 We estimated missing values with *impute.knn* (R cran package).
- 3 We centered at 0 the expression time series for each gene and performed a *K-means* clustering with $k = 6$.

Cluster 1: 902 genes.

Cluster 2: 150 genes.

Cluster 3: 693 genes.

Cluster 4: 1178 genes. Almost constant expression.

Cluster 5: 976 genes.

Cluster 6: 299 genes.

Clusters

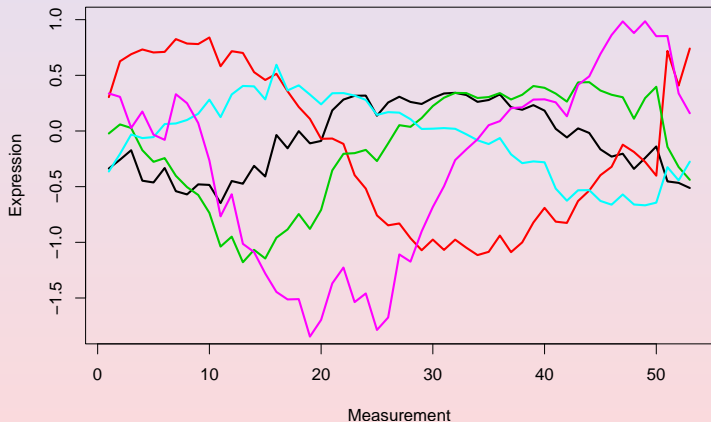


Figure: Means of all the clusters except cluster 4 which has an almost constant mean around 0.

Elements of cluster 2

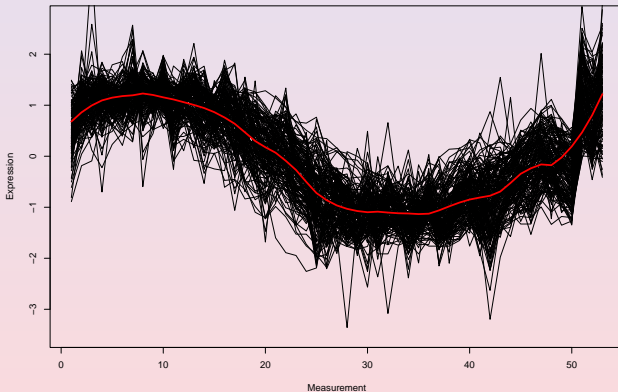


Figure: Standardized expressions for the elements of cluster 2 and loess estimated mean in red.

Running the algorithm for Bayesian TF discovery

- We ran the algorithm for TF identification with $w_1 = 0.9$, $w_0 = 0.1$ and the prior for a gene to be a TF is set to $1/150$.

Results

- The algorithm assigned gene PFC0240c the highest probability of being a transcription factor.
- We looked PFC0240c up at the PlasmoDB site.
- It appeared in the BLASTP section to be similar in a 28% to a transcription factor of *Dictyostelium discoideum*.

PFC0240c

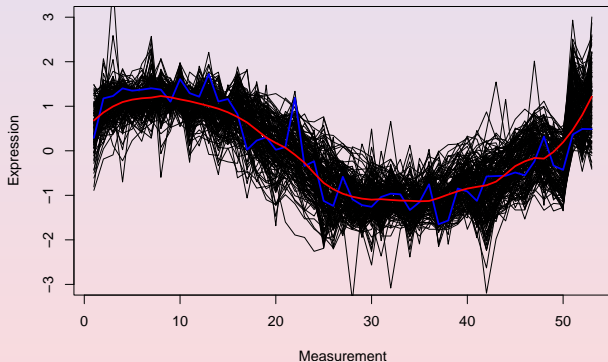


Figure: Standardized expressions for the elements of cluster 2, loess estimated mean in red and expression for gene PFC0240c in blue.

Elements of cluster 6

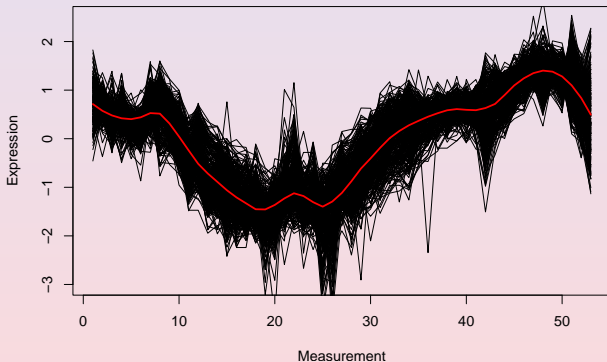


Figure: Standardized expressions for the elements of cluster 6 and loess estimated mean in red.

Running the algorithm for Bayesian TF discovery

- We ran the algorithm for TF identification with $w_1 = 0.9$, $w_0 = 0.1$ and the prior for a gene to be a TF is set to $1/299$.

Results

- The algorithm assigned gene PFD0800c the highest probability of being a transcription factor.
- We looked PFD0800c up at the PlasmoDB site.
- It appeared in the BLASTP section to be similar in a 33% to a transcription factor of *Dictyostelium discoideum*.

PFD0800c

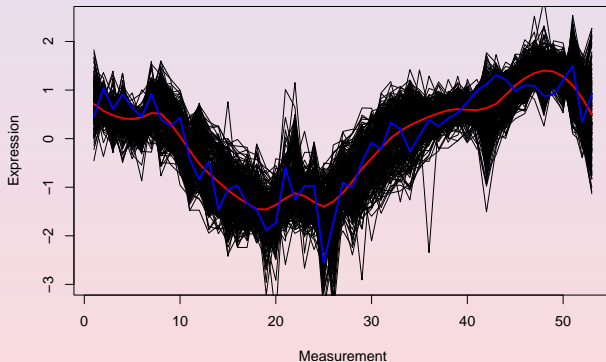


Figure: Standardized expressions for the elements of cluster 6, loess estimated mean in red and expression for gene PFD0800c in blue.

Outline

- 1 Introduction
- 2 Problem description
- 3 A simpler problem
- 4 A Bayesian model for transcription
- 5 Looking for transcription factors
- 6 Final notes**

Conclusion

- 1 The implemented algorithm seems to produce coherent results.
- 2 We expect to identify more possible TFs after running the algorithm on a bigger dataset (possibly the whole dataset).

Questions...

QUESTIONS?