

Computational Statistics for the Prediction of Gene
Regulatory Interactions
NDNS+ Workshop

Geert Geeven

April 14, 2010

Today's talk - Outline

The main aim of my project is to better understand the involvement of *transcription factors* (TFs) that govern *spatio-temporal* transcription of genes.

The outline for today's talk is as follows

- Biological background - basic mechanism of transcriptional gene regulation.
- Regression models for gene expression and DNA sequence data.
- GEMULA (Gene Expression Modeling Using Lasso)
- Application of GEMULA on data from a biological model of neuronal regeneration.

DNA → RNA → Protein

Production of a protein requires *transcription* of the corresponding gene, i.e. the production of a mRNA (*messenger RNA*) molecule which carries a "message" for the protein synthesizing apparatus of a cell.

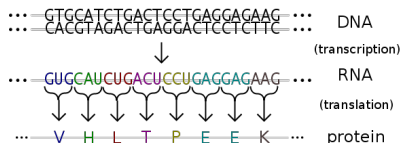
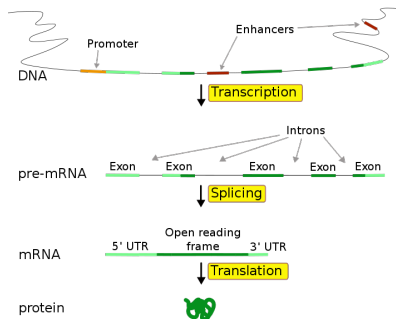


Figure: The biological processes of *transcription* and *translation*.

Regulation of (the rate of) *transcription* is a fundamental mechanism by which cells accomplish differential expression of proteins.

Typical Structure of an Eukaryotic Gene

The term *gene* is used to refer to the complete DNA sequence which is required for the production of a functional protein. Apart from *coding* sequences, genes contain regulatory DNA elements that are crucial for their proper function.



Gene Expression and Microarrays

Gene expression profiling experiments involve measuring the relative amount of mRNA expressed in two or more experimental conditions. DNA microarrays are widely used to quantify gene expression.

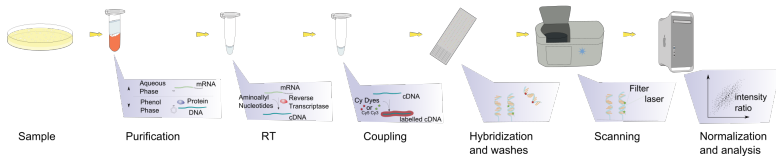


Figure: Graphical representation of the experimental steps in a typical microarray experiment.

Modeling DNA Sequences of TFBSs

Eac possible window of length 8 is scored according to a model and the score is compared to a threshold to predict putative TFBSs.

TGACATCA
TGACGTCA
TGACGTCA
TGACGTCC
TGACGTAG
TGACGTCA
TGACTGAT
.....
.....



	Binding site position							
	1	2	3	4	5	6	7	8
A	0.00	0.00	2900.00	0.00	0.00	0.00	499.41	417.87
C	0.00	0.00	0.00	2900.00	89.18	78.45	665.88	24.58
G	0.00	2497.04	0.00	0.00	2497.04	78.45	41.62	98.32
T	2900.00	89.18	0.00	0.00	0.00	2118.13	0.00	172.07

Position-specific scoring matrix

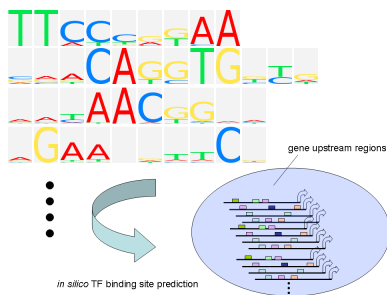


$S = \dots \text{TGTAGCTGACGTCAATGATGAAGGGTAGAATGACGTAAC} \dots$

In this case S contains 2 TFBSs for the TF CREB.

Construct Genome-wide Predictors of Gene Expression

Given the regulatory DNA sequences of all known genes (n) and a set of TFBS models (p), putative binding of TFs to gene sequences can be assessed genome-wide.



The result is a matrix $[X_1 \cdots X_p]$ of dimension $n \times p$ containing potential predictors that may explain observed variation in gene expression.

Modeling Gene Expression Using Regression

Suppose we observe a response vector $Y = (Y_1, \dots, Y_n)$ that represents gene expression for a set of n genes. Additionally, let a set of p predictor variables X_1, \dots, X_p which are *potentially* biologically related to Y be given.

We assume that Y and X_1, \dots, X_p are related through the following regression model

$$Y = \mathbf{X}\beta + \epsilon, \quad (1)$$

where

$$\mathbf{X} = [\mathbf{1} \quad f_1(X_1, \dots, X_s) \cdots f_d(X_1, \dots, X_s)],$$

is an unknown $n \times (d+1)$ design matrix, $\beta = (\beta_0, \dots, \beta_d)$ is an unknown vector of regression parameters and $\epsilon = (\epsilon_1, \dots, \epsilon_n) \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$.

Linear Models and Model Selection

- How can we model interactions between predictors using simple linear models ?
- And how to perform model selection ?

Solution: GEMULA (Gene Expression Modeling Using Lasso). GEMULA selects candidate models M_q , $q = 1, \dots, Q$ restricted to models sub-spaces \mathcal{M}_q constrained by parameters $\gamma_q = (\gamma_{q1}, \gamma_{q2}, \gamma_{q3})$, where γ_{q1} represents the maximum allowed order of interaction between terms, γ_{q2} is the maximum allowed power to which a candidate predictors is raised and γ_{q3} is the maximum number of terms allowed in the model. GEMULA uses the lasso for model selection.

The Lasso - An Example

Suppose we observe a response $Y = (Y_1, \dots, Y_n)$ and additionally let predictor variables X_1, \dots, X_{19} be given. We assume that $Y = \mathbf{X}_M \beta_M + \epsilon$, for $\mathbf{X}_M = [\mathbf{1} \ X_1 \cdots X_{19}]$. For a given shrinkage parameter $t \in \mathbb{R}^+$, lasso estimates of β_M minimize

$$\min_{\beta_M} \sum_{i=1}^n \left(Y_i - \beta_{M0} - \sum_{j=1}^{d_M} \beta_{Mj} X_{ij} \right)^2 \quad \text{subject to} \quad \sum_{j=1}^{d_M} |\beta_{Mj}| \leq t, \quad (2)$$

A fast algorithm called `lars` is available to solve this problem. It finds all solutions in a small number of steps.

The Lasso - An Example

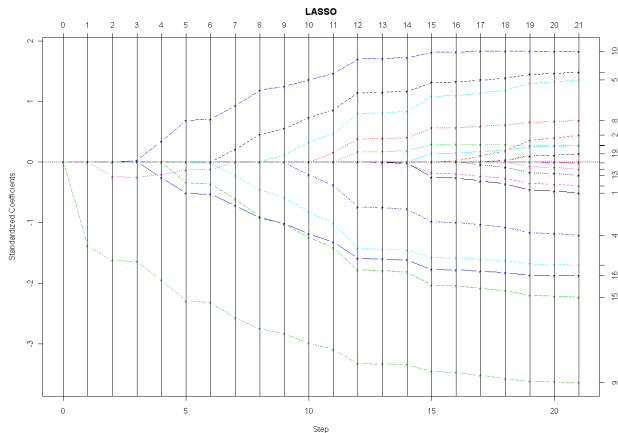


Figure: Example of a "lasso path". At each step of the `lars` algorithm, one predictor enters the "active set".

GEMULA: Step I

Let M_0 represent the model for which the design matrix satisfies $\mathbf{X}_{M_0} = [\mathbf{1} \ X_1 \cdots X_p]$. Since at each step k , exactly one predictor enters the "active set" $\mathcal{B}_M^k = \{j : \beta_{Mj}^k \neq 0\}$, GEMULA uses the mapping

$$r(j) = \min \{k : j \in \mathcal{B}_M^k\}, \quad j \in \{1, \dots, p\},$$

and its inverse r^{-1} defined by

$$r^{-1}(s) = j \quad \Leftrightarrow \quad r(j) = s \quad j \in \{1, \dots, p\}, s \in \{1, \dots, K\}$$

to define the order in which predictors enter the model. Now, e.g. when $\gamma_1 = (1, 1, 50)$, the sub-space \mathcal{M}_{γ_1} consists of all possible regression models that contain any subset of main effects for the first 50 predictors and GEMULA uses the lasso to select a model using the design matrix

$$\mathbf{X}_{\gamma_1} = [\mathbf{1} \ X_{r^{-1}(1)} \cdots X_{r^{-1}(50)}].$$

GEMULA: Step II

Within each sub-space \mathcal{M}_q determined by a parameter γ_q , GEMULA selects a model using lasso. When interactions between predictors are considered, the restrictions on the maximum number of allowed terms imposed by γ_{q3} force GEMULA to limit the number of predictors in the following way. Suppose we set $\gamma_2 = (2, 1, 150)$, then GEMULA first determines

$$s^* = \max\{s \in \{1, \dots, p\} : s + s(s-1)/2 \leq 150\},$$

and then \mathbf{X}_{γ_2} denotes the design matrix that contains all main effects and possible interactions between the predictors $X_{r-1(1)}, \dots, X_{r-1(s^*)}$. For each matrix \mathbf{X}_{γ_q} , we fit the entire path of lasso solutions and select the optimal lasso-parameter according to the AIC model selection criterion.

GEMULA: Step III

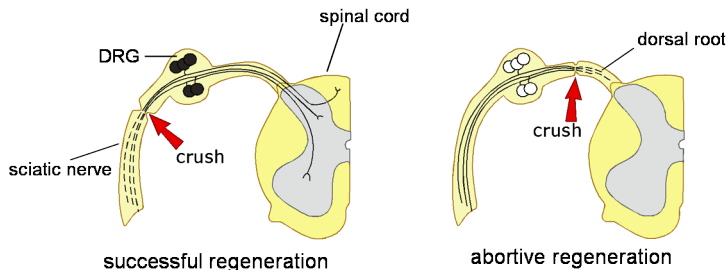
GEMULA uses V -fold cross-validation to evaluate the fit of the Q selected candidate models. The R^2 -statistics is used as *goodness-of-fit* measure, as its has an intuitive and biologically meaningful interpretation. For a model M_q with fitted response values \hat{Y}^{M_q} , it is given by

$$R^2(M_q) = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}^{M_q})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

Axonal Regeneration After Injury

- Neurons in the peripheral nervous system (PNS) successfully regenerate following axonal injury.
- Neurons in the central nervous system (CNS) generally do not.
- Combining different experimental and computational modeling approaches, we want to gain insight into the transcriptional network that underlies this difference.

The Dorsal Root Ganglion



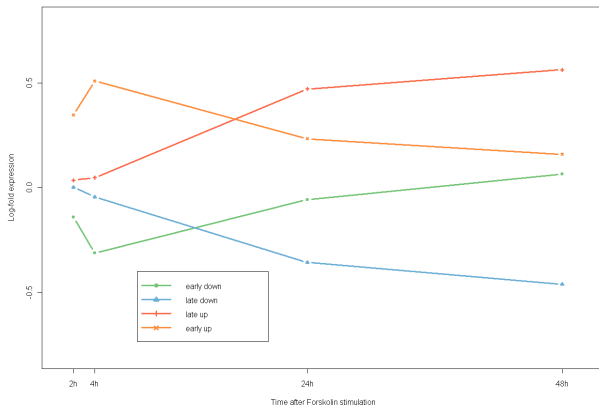
Neurons in the DRG have branches extending both into the PNS and CNS. This provides an excellent model to study the dramatic differences in regenerative capacity between PNS and CNS neurons.

Gene Expression Changes in F11 Cells in Response to Forskolin Stimulation

- The intrinsic potential of neurons to regrow damaged nerve fibers after an injury depends in part on their ability to initiate a growth promoting gene expression program. Coordinated expression of regeneration associated genes is believed to be governed by interactions between TFs and target genes.
- The F11 cell line is a fusion product of mouse neuroblastoma cells with embryonic rat DRG neurons. Upon stimulation with Forskolin, F11 cells acquire a neuronal phenotype which results in the outgrowth of neurites. F11 cells are easy to culture and transfect and provide a good *in vitro* model for neuronal regeneration *in vivo*.
- Since cultured F11 cells are a unicellular system, gene expression changes are more homogeneous and less complex than gene expression from *in vivo* samples of neuronal tissue where cells are in a complex and heterogeneous cellular environment.

Clustering Genes Based on Expression Profiles

We distinguish between genes whose expression profile show either early or late response to Forskolin stimulation.



Models Fitted With GEMULA

Time	Model type	Early responsive genes			Late responsive genes		
		Model P	Model T	\bar{R}_{cv}^2	Model P	Model T	\bar{R}_{cv}^2
2h	M1	30	30	0.14	8	8	-0.00
2h	M2	31	72	0.22	36	53	0.06
2h	M3	14	47	0.25	16	27	0.03
4h	M1	16	16	0.08	0	0	0
4h	M2	31	75	0.14	36	53	0.03
4h	M3	14	39	0.07	16	16	0.03
24h	M1	50	50	0.01	39	39	0.25
24h	M2	31	80	0.11	36	63	0.24
24h	M3	14	20	0.02	15	27	0.23
48h	M1	5	5	-0.01	44	44	0.25
48h	M2	31	85	0.11	35	52	0.27
48h	M3	14	60	0.04	16	37	0.24

Table: Comparison of models fitted using GEMULA for early and late Forskolin responsive genes in F11 cells at all four time-points. Columns 3-5 correspond to models fitted for the early responsive genes and columns 6-8 to models for the late responsive genes.

Predictors Associated to Gene Expression Changes












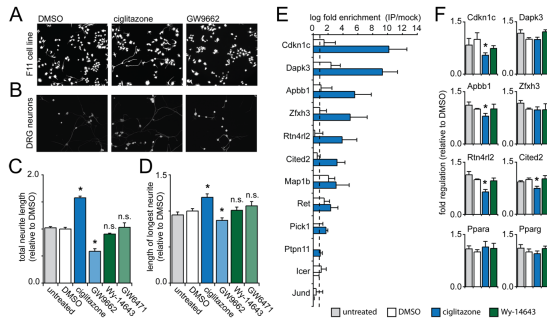
TFBS ID	Motif logo	Time	Activity
V.AP1.Q4.01		Early	Activator
V.AP1.Q4.01		Late	Activator
V.AREB6.02		Late	Activator
V.CREB.Q4.01		Early	Activator
V.CEBPDELTA.Q6		Early	Activator
V.CETS1P54.02		Early	Repressor
V.CETS1P54.02		Late	Repressor
V.E2F.Q6.01		Late	Repressor
V.EBF.Q6		Late	Activator
V.PPARA.01		Early	Repressor
V.PPARA.01		Late	Activator

Table: TFBS motif logos of predictors present in models selected by GEMULA.

Experimental Validation of Predictions

The functional role of PPAR α and PPAR γ was studied in F11 cultured cells *in vitro*. It turned out that PPAR γ , and *not* PPAR α promotes neurite outgrowth in F11 cells and that knock-down of PPAR γ significantly *decreases* neurite outgrowth.



Possible Future Improvements

In the future, as more and higher quality data becomes available and we better understand the different molecular mechanisms of gene regulation, possible improvements to the model may include

- (Even) better suitable TFBS predictors more focused at DNA elements that are most likely to be functional. For instance, how are absolute and relative location and conservation of DNA elements related to functionality?
- Inclusion of *in vivo* TF binding-assay data (ChIP-chip or ChIP-seq).
- Inclusion of predictors corresponding to other determinants of transcription rates such as DNA methylation status and nucleosome positions.
- Use of more accurate (absolute) measurements of gene transcription, e.g. RNA-seq data instead of DNA microarray data.

Acknowledgements

Mathisca de Gunst

*Department of Mathematics, Stochastics Section
VU University - Amsterdam*

Harold McGillavry and Ronald van Kesteren

*Department of Molecular and Cellular Neurobiology
VU University - Neuroscience Campus - Amsterdam*