A new statistical framework to infer gene regulatory networks with hidden transcription factors

Ivan Vujačić, Javier González, Ernst Wit

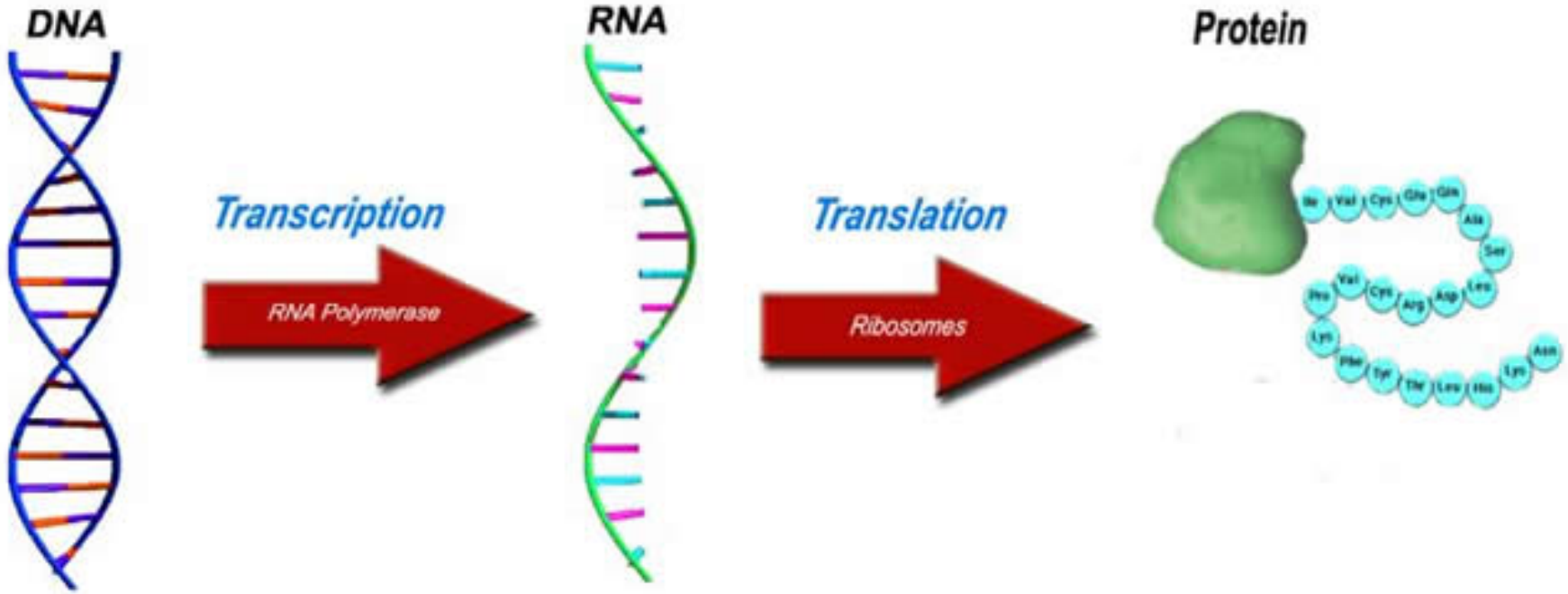
University of Groningen
Johann Bernoulli Institute

PEDS II, June 4-6, Eindhoven

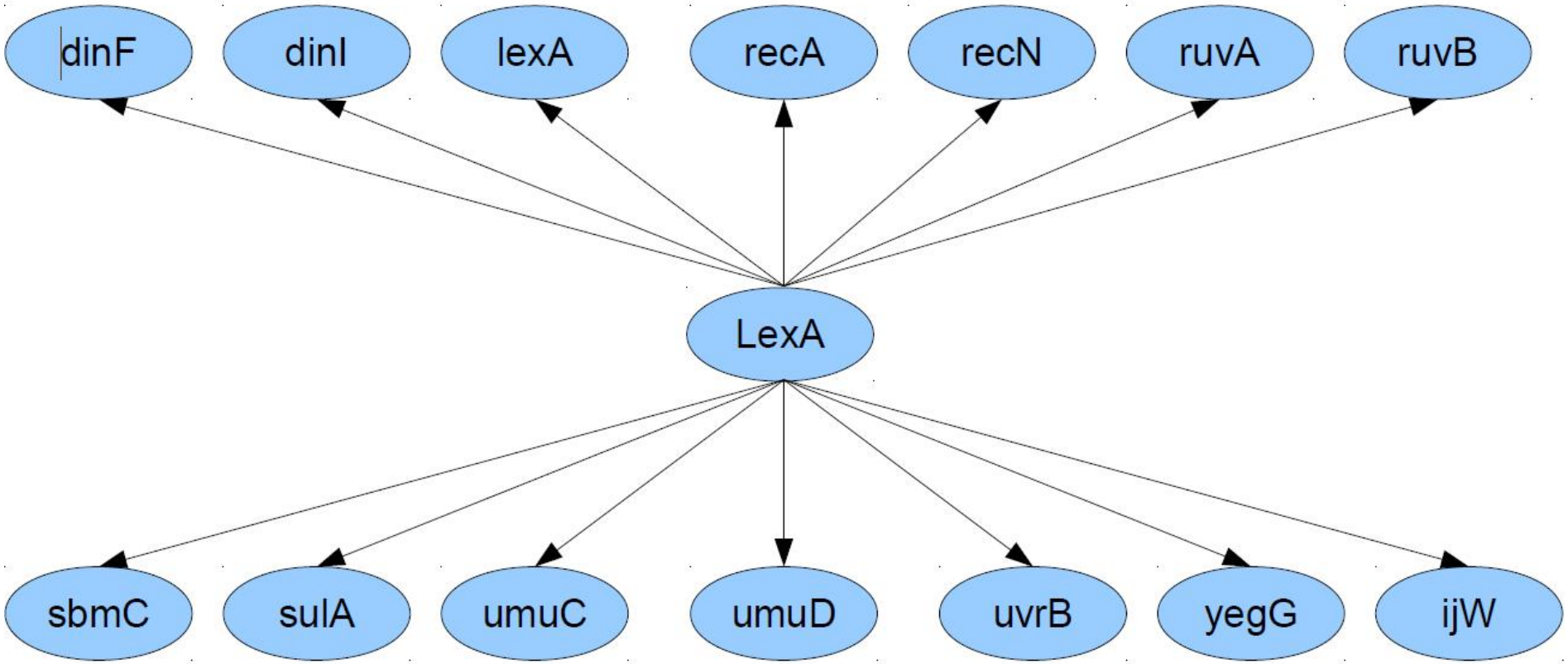
university of groningen

faculty of mathematics and natural sciences

Central dogma of molecular biology



SOS system in Escherichia coli



The goal

The goal of the analysis is to

- Reconstruct the activity level of the repressor LexA.
- Identify kinetic parameters of the system.
- Order the genes in terms of speed they are repressed by LexA protein.

ODE model

- $\eta(t)$ - the abundance of the protein (TF) LexA at time $t \in T$.
- $x_1(t), \ldots, x_m(t)$ - mRNA concentrations of genes.

ODE model for expressions of the genes

$$\dot{x}_k(t) = p(t; \boldsymbol{\theta}_k, \eta(t)) - \delta_k x_k(t), \qquad k = 1, \ldots, m.$$

Michaelis-Menten form for p, i.e.

$$p(t; \boldsymbol{\theta}_k, \eta(t)) = \beta_k \frac{1}{\gamma_k + \eta(t)} + \varphi_k.$$

Here $\boldsymbol{\theta}_k = (\beta_k, \gamma_k, \varphi_k)$, where

- β_k - production rate,
- γ_k - half-saturation rate,
- φ_k - basal level of gene transcription.

Modelling the protein abundance

We assume that η is cubic spline on T:

$$\eta(t) = \sum_{j=1}^{d} \mu_j \phi_j(t),$$

$\mu_j \in \mathbb{R}$ and $\{\phi_1, \ldots, \phi_d\}$ is truncated-power basis set for cubic spline. This induces a new set of parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^T$.

Noise model

- $\mathbf{t} = \{t_1, \ldots, t_n\}$
- y_{ki} - the measured expression of gene k at t_i.
- $S_k = \{(y_{ki}, t_i) \in \mathbb{R} \times T\}_{i=1}^n$

Gene expression measurements of genes are independent and

$$y_{ki} \sim \mathcal{N}(x_k(t_i), \sigma_k^2).$$

The log-likelihood of a single observation is up to an additive constant

$$l(x_k(t_i), \sigma_k^2 | y_{ki}) \quad = \quad -\frac{1}{2}\left(\frac{y_{ki} - x_k(t_i)}{\sigma_k}\right)^2 - \frac{1}{2}\log(\sigma_k^2).$$

The contribution of each gene k to the log-likelihood of the network is

$$l_k(S_k; \delta_k, \boldsymbol{\theta}_k, \sigma_k^2, \boldsymbol{\mu}) = \sum_{i=1}^n l(x_k(t_i), \sigma_k^2 | y_{ki}),$$

where it is assumed that each function x_k satisfies the corresponding ODE.

The likelihood is up to an additive constant

$$l = -\frac{1}{2}\sum_{k=1}^{m}\frac{1}{\sigma_k^2}\|\mathbf{y}_k - x_k(\mathbf{t})\|^2 - \frac{n}{2}\sum_{k=1}^{m}\log(\sigma_k^2),$$

where $x_k(\mathbf{t}) = (x_k(t_1), \ldots, x_k(t_n))^T$, $\mathbf{y}_k = (y_{k1}, \ldots, y_{kn})^T$.

Discretization

$p(\mathbf{t}; \boldsymbol{\theta}_k, \eta(\mathbf{t})) = (p(t_1; \boldsymbol{\theta}_k, \eta(t_1)), \ldots, p(t_n; \boldsymbol{\theta}_k, \eta(t_n)))^T$.
The difference operator

$$\mathbf{D} = \Delta^{-1} \begin{pmatrix} -1 & 1 & & & & \\ -1 & 0 & 1 & & & \\ & & & \ddots & & \\ & & & -1 & 0 & 1 \\ & & & & -1 & 1 \end{pmatrix},$$

for $\Delta = diag(t_2 - t_1, t_3 - t_1, \ldots, t_n - t_{n-1})$.
$\mathbf{P}_{\delta_k} = \mathbf{D} + \delta_k \mathbf{I}$, where $\mathbf{I}$ is the identity matrix.
The discretization of the ODE for the gene k can be written as

$$\mathbf{P}_{\delta_k} x_k(\mathbf{t}) = p(\mathbf{t}; \boldsymbol{\theta}, \eta(\mathbf{t})).$$

Penalizing the likelihood

We want to penalize the likelihood with the term

$$\|\mathbf{P}_{\delta_k} x_k(\mathbf{t}) - p(\mathbf{t}; \boldsymbol{\theta}_k, \eta(\mathbf{t}))\|^2.$$

To use RKHS theory, penalty needs to be equal to $\|\mathbf{P}_{\delta_k}\tilde{\mathbf{x}}_k\|^2$ for some $\tilde{\mathbf{x}}_k$. This leads to

$$\tilde{\mathbf{x}}_k = x_k(\mathbf{t}) - \mathbf{P}_{\delta_k}^{-1} p(\mathbf{t}; \boldsymbol{\theta}_k, \eta(\mathbf{t})),$$

and to transformation of the observations

$$\tilde{\mathbf{y}}_k = \mathbf{y}_k - \mathbf{P}_{\delta_k}^{-1} p(\mathbf{t}; \boldsymbol{\theta}_k, \eta(\mathbf{t})).$$

Penalized log-likelihood is defined as

$$l_\lambda = -\frac{1}{2}\sum_{k=1}^{m}\frac{1}{\sigma_k^2}\|\tilde{\mathbf{y}}_k - \tilde{\mathbf{x}}_k\|^2 - \frac{n}{2}\sum_{k=1}^{m}\log(\sigma_k^2) - \lambda\sum_{k=1}^{m}\|\mathbf{P}_{\delta_k}\tilde{\mathbf{x}}_k\|^2.$$

RKHS theory guarantees that for some $\boldsymbol{\alpha}_k \in \mathbb{R}^n$ and $\mathbf{K}_{\delta_k} = (\mathbf{P}_{\delta_k}^T \mathbf{P}_{\delta_k})^{-1}$

$$\Omega(\tilde{x}_k) = \|\mathbf{P}_{\delta_k}\tilde{x}_k\|^2 = \boldsymbol{\alpha}_k^T \mathbf{K}_{\delta_k}\boldsymbol{\alpha}_k, \qquad \tilde{\mathbf{x}}_k = \mathbf{K}_{\delta_k}\boldsymbol{\alpha}_k.$$

$\tilde{S}_k$ is transformed set of expression measurements for the gene k.
The penalized log-likelihood is

$$\begin{aligned} l_\lambda(\Delta, \Theta, \Sigma, \boldsymbol{\mu}|\tilde{S}) &= -\frac{1}{2}\sum_{k=1}^{m}\frac{1}{\sigma_k^2}\|\tilde{\mathbf{y}}_k - \mathbf{K}_{\delta_k}\boldsymbol{\alpha}_k\|^2 - \frac{n}{2}\sum_{k=1}^{m}\log(\sigma_k^2) \\ &- \lambda\sum_{k=1}^{m}\boldsymbol{\alpha}_k^T\mathbf{K}_{\delta_k}\boldsymbol{\alpha}_k, \end{aligned}$$

$\tilde{S} = \{\tilde{S}_1, \ldots, \tilde{S}_m\}$, $\Delta = \{\delta_1, \ldots, \delta_m\}$, $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$,
$\Sigma = \{\sigma_1^2, \ldots, \sigma_m^2\}$.

Estimation

$A = \{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_m\}$ the set of parameters characterizing $\tilde{x}$.
The penalized maximum likelihood estimators of $\Delta, \Theta, \Sigma, \mu$ and A are given by

$$(\hat{\Delta}_\lambda, \hat{\Theta}_\lambda, \hat{\Sigma}_\lambda, \hat{\boldsymbol{\mu}}_\lambda, \hat{A}_\lambda) = \arg \max_{\Delta, \Theta, \Sigma, \boldsymbol{\mu}, A} l_\lambda(\Delta, \Theta, \Sigma, \boldsymbol{\mu}, A | \tilde{S}).$$

Choosing the tunning parameter

To choose λ we use AIC type of criteria. The smoother matrix is

$$\mathbf{S}_{\lambda,k} = \mathbf{K}_{\delta_k}(\mathbf{K}_{\delta_k} + 2\sigma_k^2\lambda\mathbf{I})^{-1}.$$

The complexity of the model is defined as $df_k = Tr(\mathbf{S}_{\lambda,k})$. λ is chosen as

$$\lambda_{opt} = \arg\min_{\lambda} \sum_{k=1}^{m} [-2 \cdot l_k(S_k; \hat{\delta}_k, \hat{\boldsymbol{\theta}}_k, \hat{\sigma}_k^2, \hat{\boldsymbol{\mu}}) + 2df_k].$$

Data

- $m = 14$ expression genes of the E-coli SOS system.
- mRNA measurements in $n = 6$ time points: 0,5,10,20, 40 and 50 min.

RESULTS

LexA activity

Figure: Reconstruction of the activity of the master repressor LexA scaled between 0 an 1. The smoothed LexA profile is obtained using a cubic spline. Time is given in minutes.

Gene Profiles

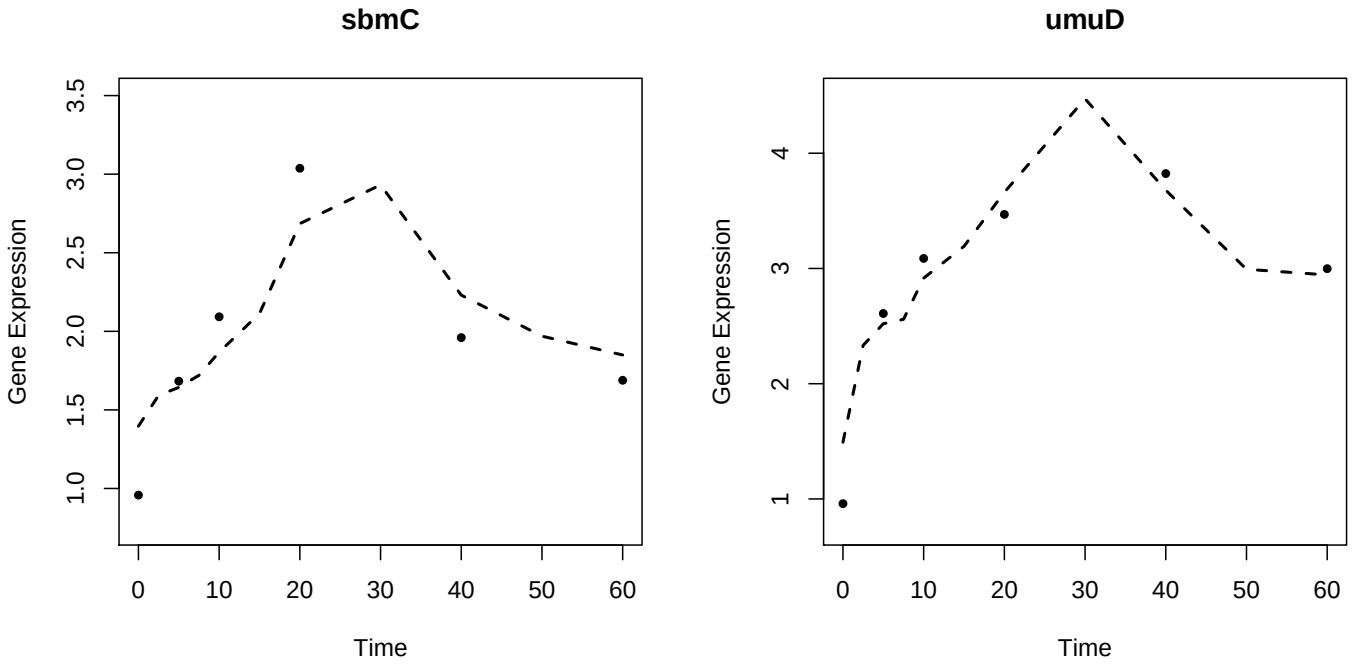


Figure: Reconstruction of gene profiles sbmC and umuDB. Points represent the data values and lines the predicted profiles. The rest of the genes are fitted in a similar way. Time is given in minutes.

Kinetics

To rank the genes in terms of their production rate we use the values of

$$r_k = \frac{\beta_k}{\gamma_k + \bar{\eta}}$$

for $i = 1, \ldots, 14$ and $\bar{\eta}$ the averaged TF level,

- recN ($r = 15.51$), sulA ($r = 13.62$), unuC ($r = 16.95$) are the fastest
- dinF ($r = 0.97$) is the slowest
- recN and umuC show the largest values for δ_k
- φ_k were found to be negligible for genes sulA and umuC and very small for recN.