Inference of the statistics of a promoter process from population-snapshot gene expression data

Eugenio Cinquemani
INRIA Grenoble – Rhône-Alpes

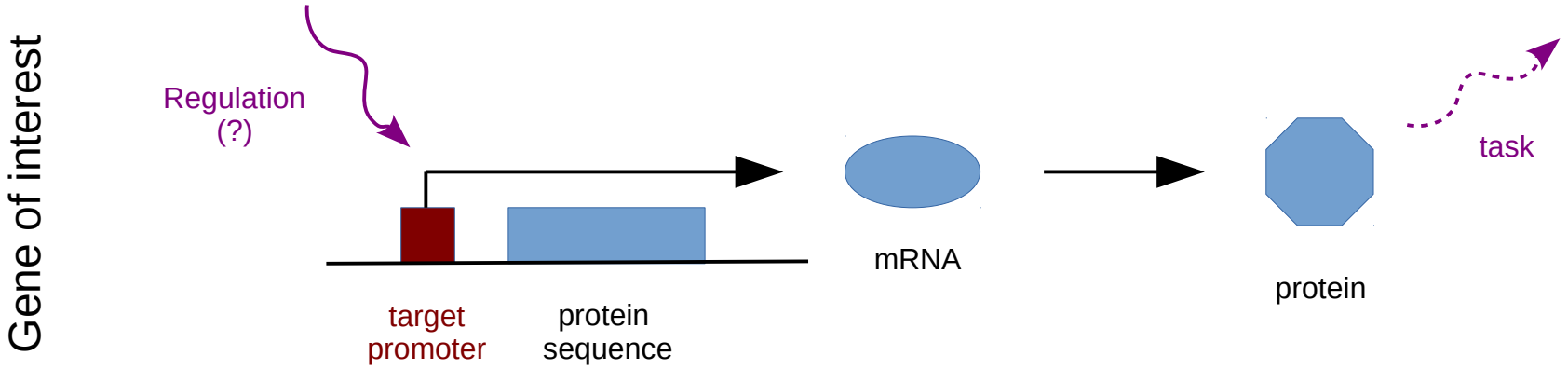
Biohasard 2021
June 10, 2021

Outline

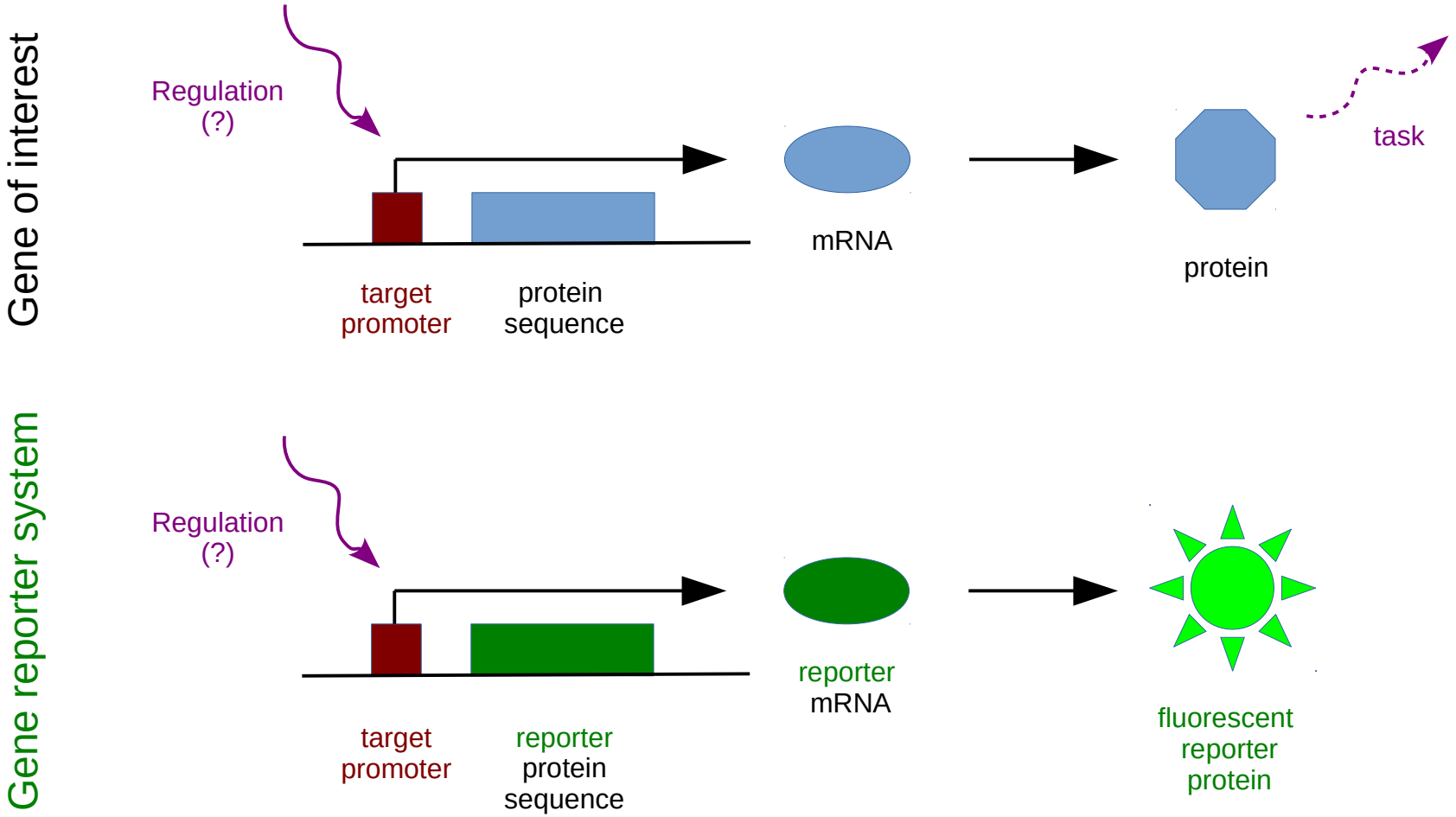
- Introduction : Gene expression, fluorescent reporters, snapshot data
- Single-cell gene expression modelling and moment equations
- Inference of the statistics of a stationary promoter process
- Inference of the statistics of a modulated promoter process
- Conclusions

Introduction

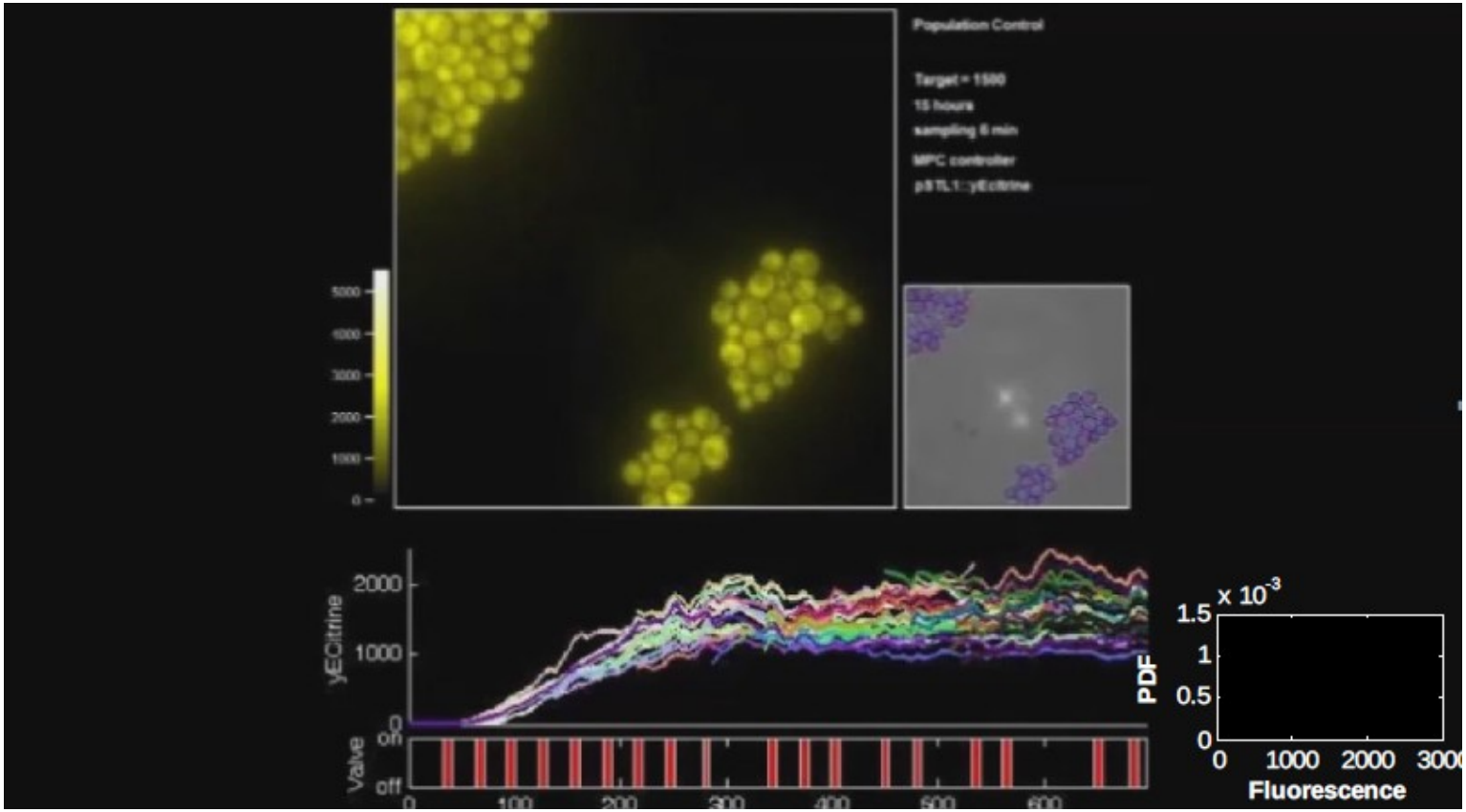
Gene expression



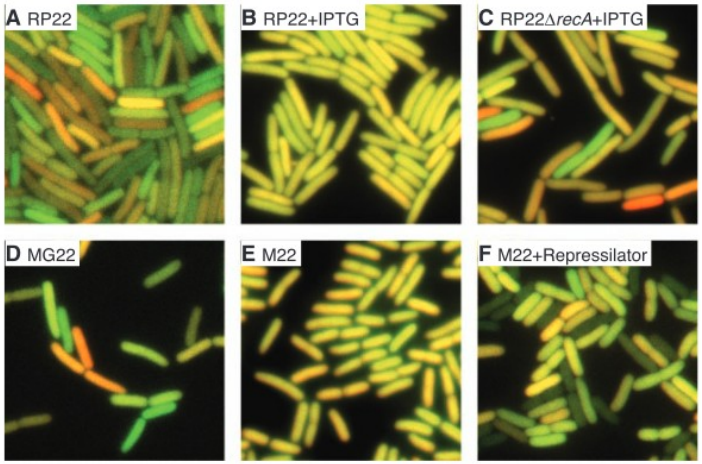
Gene expression and fluorescence reporting

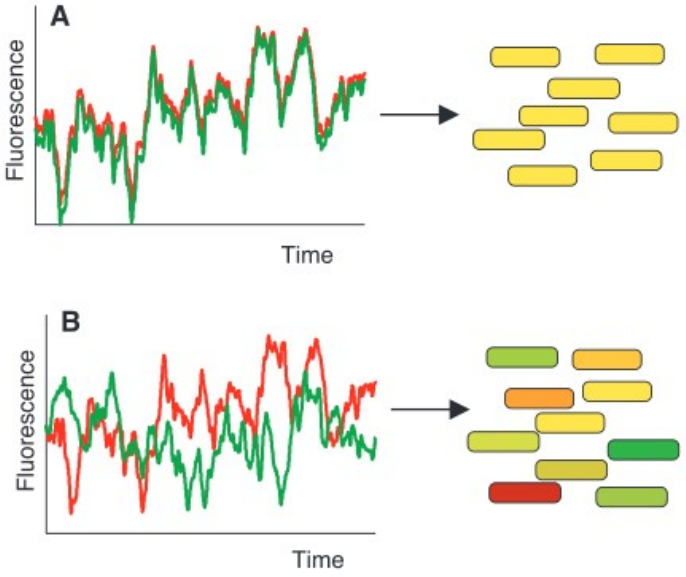


Gene expression variability



Intrinsic vs. extrinsic noise



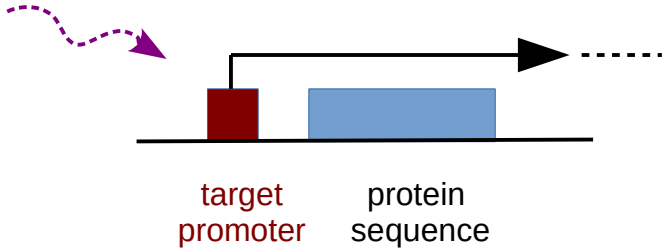


(Elowitz et al,Science 2002)

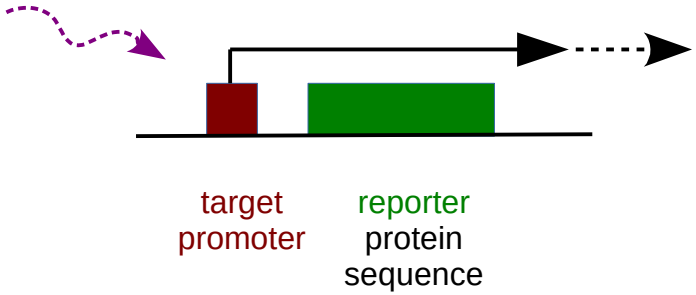
- Intrinsic : Random transcription and translation events
- Extrinsic : Other sources of variability (parameters, promoter activity, ...)

Population snapshot data

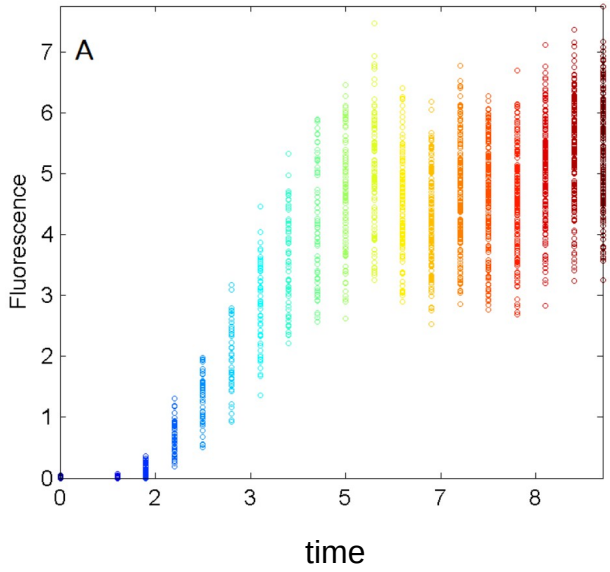
Gene of interest



Gene reporter system



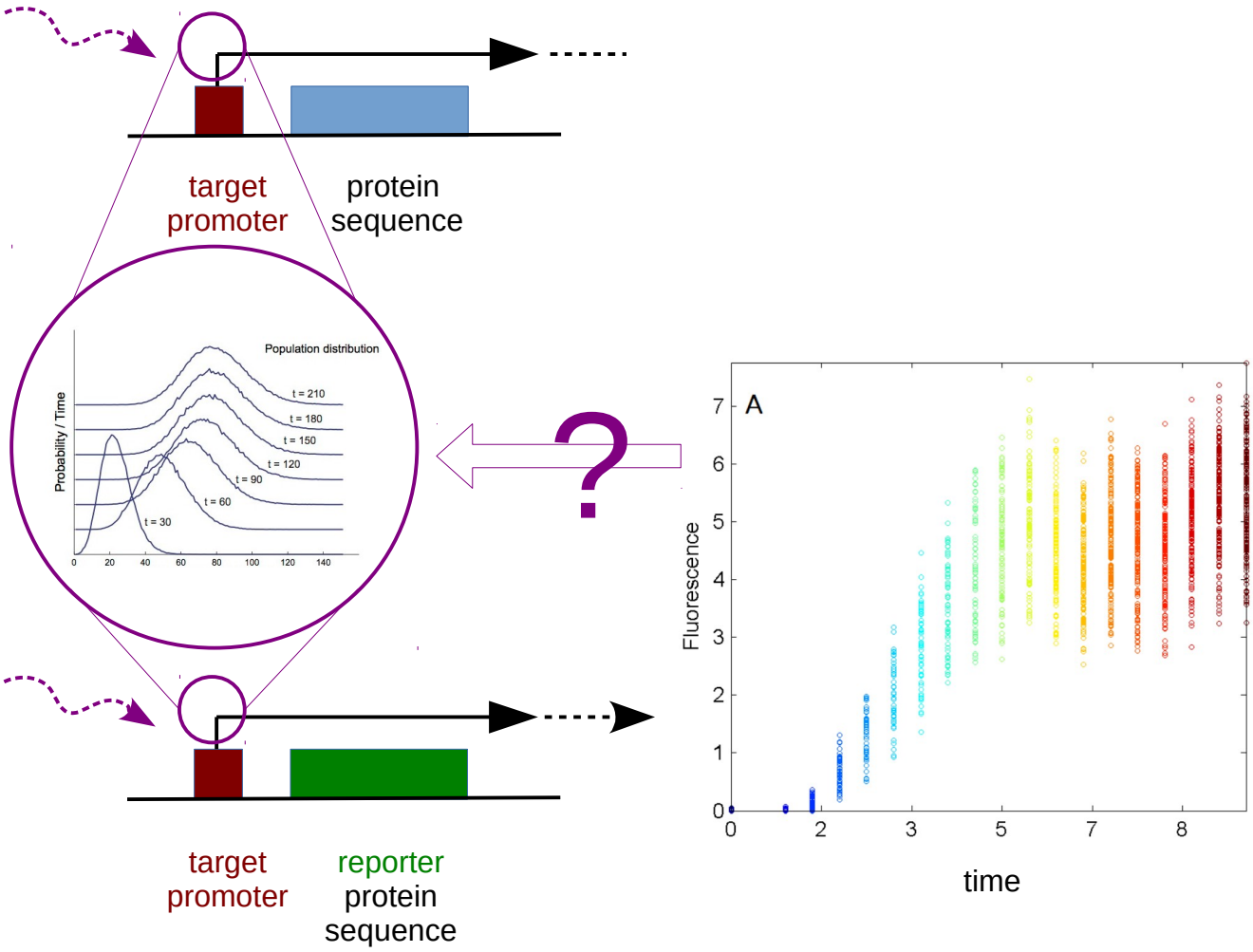
Fluorescence distribution in samples from a population of cells:



Goal : Inference of promoter activity statistics

Gene of interest

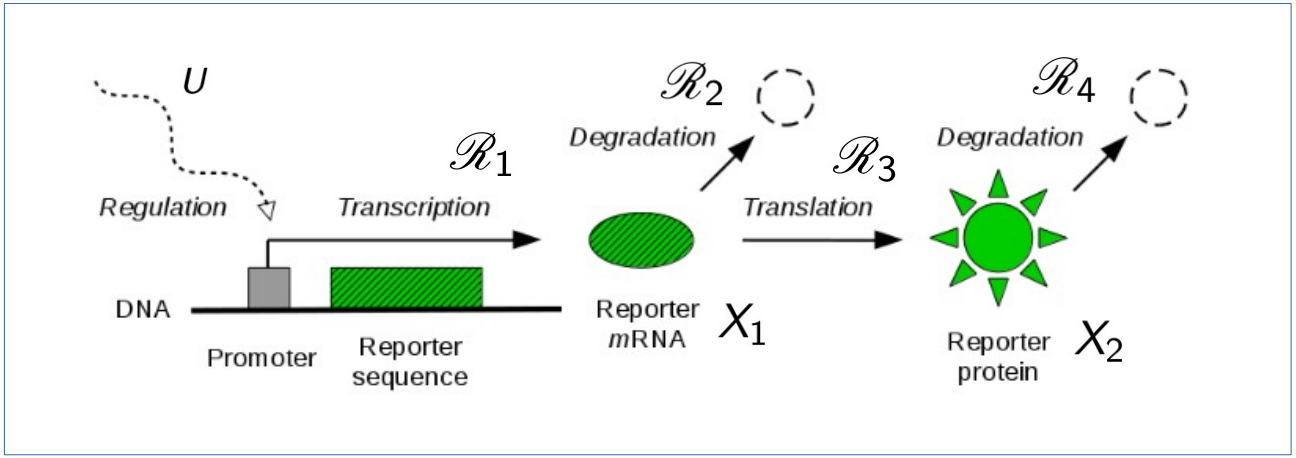
Gene reporter system



Gene expression modelling and analysis

(Cinquemani, *Automatica* 2019)

Random telegraph model



- Reactions :

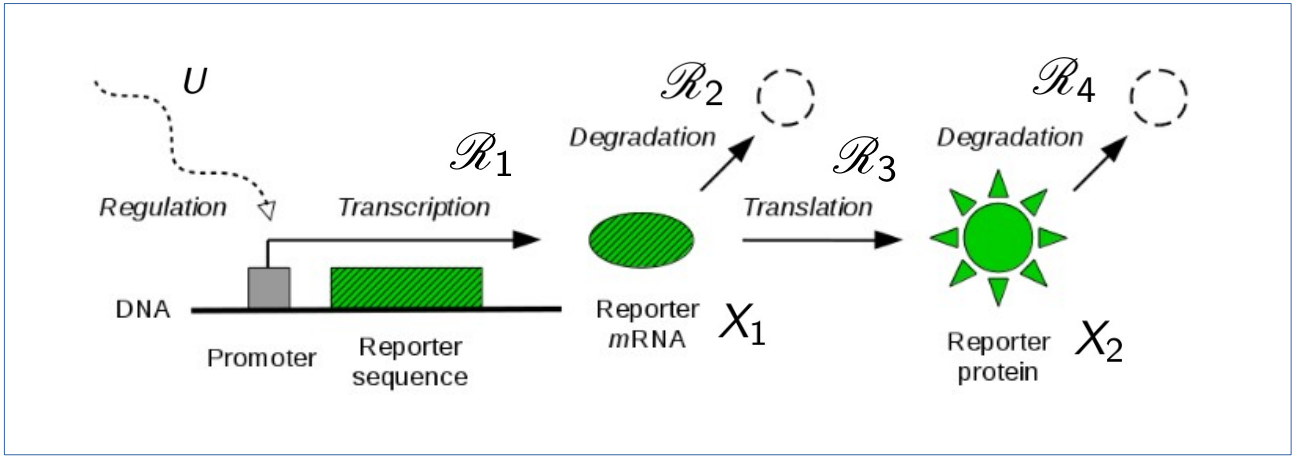
$$\mathscr{R}_1 : \emptyset \xrightarrow{k_M \cdot U} M \qquad \mathscr{R}_2 : M \xrightarrow{d_M} \emptyset$$

$$\mathscr{R}_3 : M \xrightarrow{k_P} M + P \qquad \mathscr{R}_4 : P \xrightarrow{d_P} \emptyset$$

- Stoichiometry matrix and reaction rates :

$$S = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \qquad w(t) = \begin{bmatrix} k_M U(t) \\ d_M X_1(t) \\ k_P X_1(t) \\ d_P X_2(t) \end{bmatrix}$$

Random telegraph model



- Reaction rates are affine in the state :

$$w(t) = WX(t) + F(t)$$

$$W = \begin{bmatrix} 0 & 0 \\ d_M & 0 \\ k_P & 0 \\ 0 & d_P \end{bmatrix} \qquad F(t) = \begin{bmatrix} k_M U(t) \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Moment Equations

- First- and second-order statistics of X :

$$\mu(t) = \mathbb{E}[X(t)], \quad \Sigma(t) = \mathrm{Var}\big(X(t)\big), \quad \rho(z,t) = \mathrm{Cov}\big(X(z), X(t)\big)$$

- If U (that is, F) is a deterministic function :

$$d\mu(t)/dt = SW\mu(t) + SF(t)$$
$$d\Sigma(t)/dt = SW\Sigma(t) + \Sigma(t)W^T S^T + S\mathrm{diag}\big(W\mu(t) + F(t)\big)S^T$$

$$\partial\rho(z,t)/\partial z = SW\rho(z,t)$$

Generalized Moment Equations

- Now assume F is a(ny) stochastic process :

$$\mu_F(t) = \mathbb{E}[F(t)], \quad \rho_F(z,t) = \mathrm{Cov}\big(F(z), F(t)\big), \quad \xi_F(t) = \mathrm{Cov}\big(X(0), F(t)\big)$$

- Assuming *absence of feedback* from X to F :

$$\begin{aligned} d\mu(t)/dt &= SW\mu(t) + S\mu_F(t) \\ d\Sigma(t)/dt &= SW\Sigma(t) + \Sigma(t)W^T S^T + S\mathrm{diag}\big(W\mu(t) + \mu_F(t)\big)S^T \\ &\quad + V_{\xi_F}(t,t) + V_{\xi_F}^T(t,t) + V_{\rho_F}(t,t) + V_{\rho_F}(t,t)^T \\ \partial\rho(z,t)/\partial z &= SW\rho(z,t) + V_{\xi_F}(z,t) + V_{\rho_F}(z,t) \end{aligned}$$

V_{ξ_F} linear (integral) functional of ξ_F

V_{ρ_F} linear (integral) functional of ρ_F

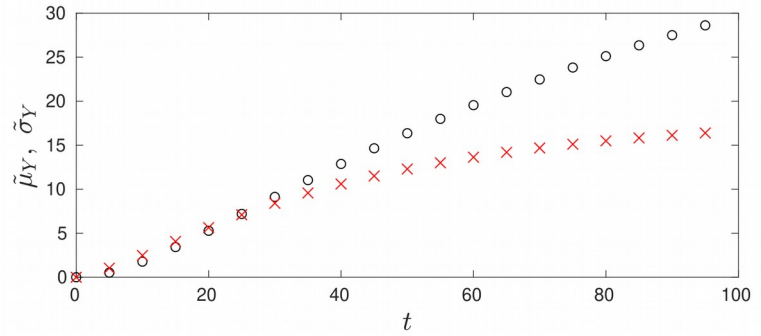
Inference in the case of a stationary process

(Cinquemani, *Automatica* 2019)

Inference of promoter activity statistics

- Given empirical statistics (population snapshot data)

$$\tilde{\mu}_2(t_k) = \mu_2^*(t_k) + e_k^{\mu}$$
$$\tilde{\sigma}_2^2(t_k) = \Sigma_{2,2}^*(t_k) + e_k^{\sigma}$$



with *k=1,...,K* and i.i.d. approx Gaussian noise

- Estimate unknown mean and autocovariance function of *U*

- Ill-posed but **linear** inversion : Efficient solutions possible using (Tikhonov) regularization

Special case : *U* stationary, $X_0=0$

$$F(t) = \begin{bmatrix} k_M U(t) \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

- Unknown stationary statistics : $\bar{\mu}_U = \mu_U(\cdot), \quad \bar{\rho}_U(\delta) = \rho_U(\cdot + \delta, \cdot)$
- Estimate constant mean by fitting mean data with the solution of

$$\frac{d}{dt}\mu(t) = SW\mu(t) + \begin{bmatrix} k_M \bar{\mu}_U \\ 0 \end{bmatrix}$$

- Estimate autocovariance as the solution of the convex optimization

$$\min_{\bar{\rho}_U \in \mathscr{C}} \sum_{k=1}^{K} \alpha_k^2 \left(\tilde{\sigma}_2^2(t_k) - \mathscr{L}(t_k | \bar{\rho}_U) \right)^2 + \gamma \mathscr{Q}(\bar{\rho}_U)$$

PSD functions (convex cone)

Inverse of meas. error variance

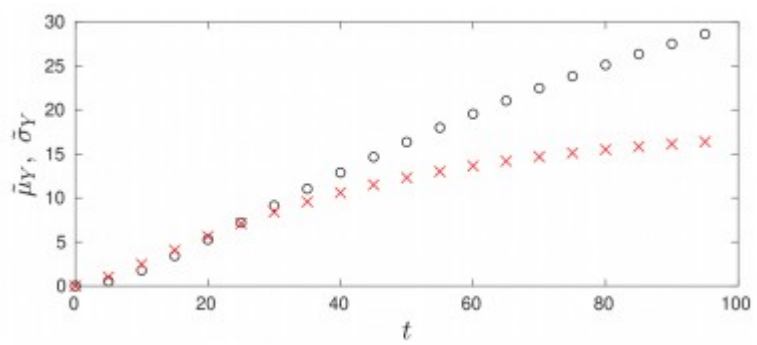
Solution at t_k of $d\Sigma(t)/dt = SW\Sigma(t) + \Sigma(t)W^T S^T + Q(t) + \Lambda(t|\bar{\rho}_U)$ with Q known function of the mean and Λ known functional

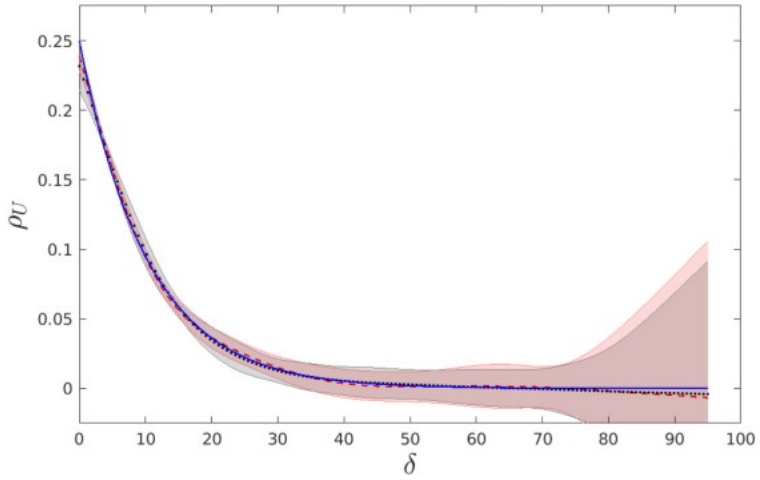
Convex penalization of irregular solutions

- Implementation : Finite autocovariance expansion, LQP problem

Results from numerical simulations

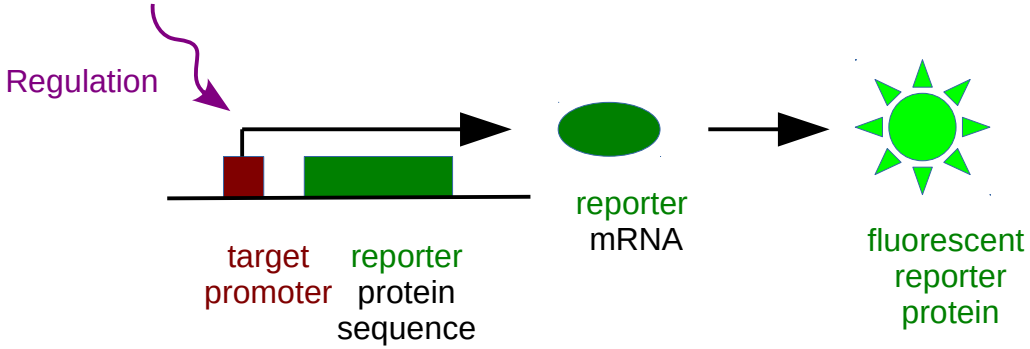
- Binary process *U* with stochastic switching rates
- Estimator mean ± 2 std for samples of 10^5 cells
- Automated vs. fixed choice of γ (vs. true autocovariance)





Considerations so far

- Regularized estimation method based on finite expansions
- Good results for reconstruction of stationary (1D) autocovariance function
- Stationarity assumption unsuitable for controlled gene expression



- Ideally, want to reconstruct nonstationary statistics
- Reconstructing 2D function from 1D data hard
- Relevant and tractable case : Modulated processes

The case of a modulated process

(Cinquemani, *Proc. of IFAC WC* 2020)

Modulated promoter process

- Known deterministic *G(t)* (control signal), stationary process *E(t)* :

$$F(t) = G(t)E(t)$$

- Nonstationary statistics for process *F(t)* : In particular, autocovariance is

$$\rho_F(t,\tau) = G(t)\bar{\rho}_E(t-\tau)G(\tau)^T$$

- Variance measurements relate with $\bar{\rho}_E$ via the integral

$$V_\rho(t) = \int_0^{+\infty} d\tau S G(t)\bar{\rho}_E(t-\tau)G(\tau)^T S^T \ell(t-\tau)^T \qquad \ell(t) = \exp(SWt)\mathbb{1}(t)$$

Contrary to stationary case, no simple expression as function of *t*

Autocovariance reconstruction

- Regularized linear inversion requires efficient evaluation of the integral as a function of *t* for at least a convenient class of functions *G(t)*
- For generic *r*, write integral as $V_r(t) = SG(t)H_{0,r}(t)$ where, for any *i*,

$$H_{i,r}(t) = \int_0^t d\delta r(\delta) G^{(i)}(t-\delta)^T S^T \ell(\delta)^T$$

- Consider (matrix) splines of degree *d*, such that, for suitable knots T_j,

$$G^{(d-1)}(\tau) = \begin{cases} G_0, & \tau < T_1, \\ G_j, & \tau \in [T_j, T_{j+1}),\ j = 1, \ldots, p-1, \\ G_p, & \tau \geq T_p. \end{cases}$$

Main result

Proposition : For G splines defined as above, it holds that

$$\dot{H}_i(t) = r(t) G^{(i)}(0)^T S^T \ell(t)^T + H_{i+1}(t), \qquad i = 0, 1, \ldots, d-2,$$

$$\dot{H}_{d-1}(t) = H_0^+(t) + \sum_{j=1}^{p-1} \left(H_j^+(t) - H_{j-1}^-(t) \right),$$

with $H_*(0) = 0$, where, for all relevant j,

$$H_j^+(t) = r(t - T_j) G_j^T S^T \ell(t - T_j)^T,$$
$$H_j^-(t) = r(t - T_{j+1}) G_j^T S^T \ell(t - T_{j+1})^T.$$

Estimation approach (sketch)

- Choose family of approximation ('basis') functions $\mathscr{R} = [r_1 \ \cdots \ r_N]$
- Compute their GME images $\mathscr{V}(t_k) = [v_1(t_k), \ldots, v_N(t_k)]$ at measurement times t_k via an augmented ODE system

$$
\begin{aligned}
\dot{\Sigma}(t) &= SW\Sigma(t) + \Sigma(t)W^T S^T + \\
&\qquad SG(t)H_{0,r_l}(t) + H_{0,r_l}(t)^T G(t)^T S^T, \\
\dot{H}_{0,r_l}(t) &= r_l(t)G^{(0)}(0)^T S^T \ell(t)^T + H_{1,r_l}(t), \\
&\vdots \\
\dot{H}_{d-2,r_l}(t) &= r_l(t)G^{(d-2)}(0)^T S^T \ell(t)^T + H_{d-1,r_l}(t), \\
\dot{H}_{d-1,r_l}(t) &= H^+_{0,r_l}(t) + \sum_{j=1}^{p-1} \left(H^+_{j,r_l}(t) - H^-_{j-1,r_l}(t) \right),
\end{aligned}
$$

- Compute estimates $\hat{\rho}_E(\cdot) = \mathscr{R}(\cdot)\hat{c}$ and $\hat{\rho}_F(z,t) = G(z)\hat{\rho}_E(z-t)G(t)^T$ from the solution of the *quadratic* optimization problem

$$
\hat{c} \in \arg\min_{c \in \mathbb{R}^N} \sum_{k=1}^{M} \alpha_k^2 \left(\tilde{\sigma}^2_Y(t_k) - v_0(t_k) - \mathscr{V}(t_k)c \right)^2 + \lambda \cdot c^T Qc
$$

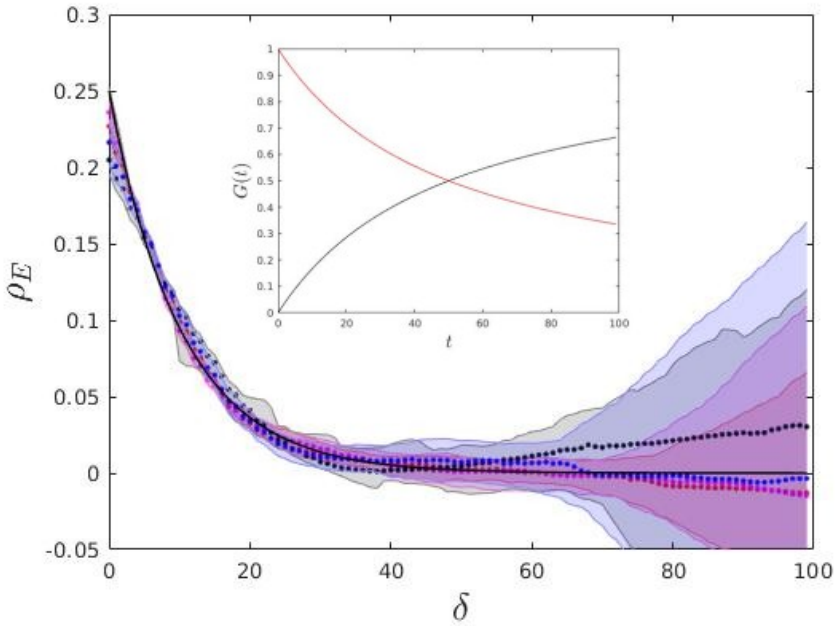
Measurements weighted by the inverse of their error variance

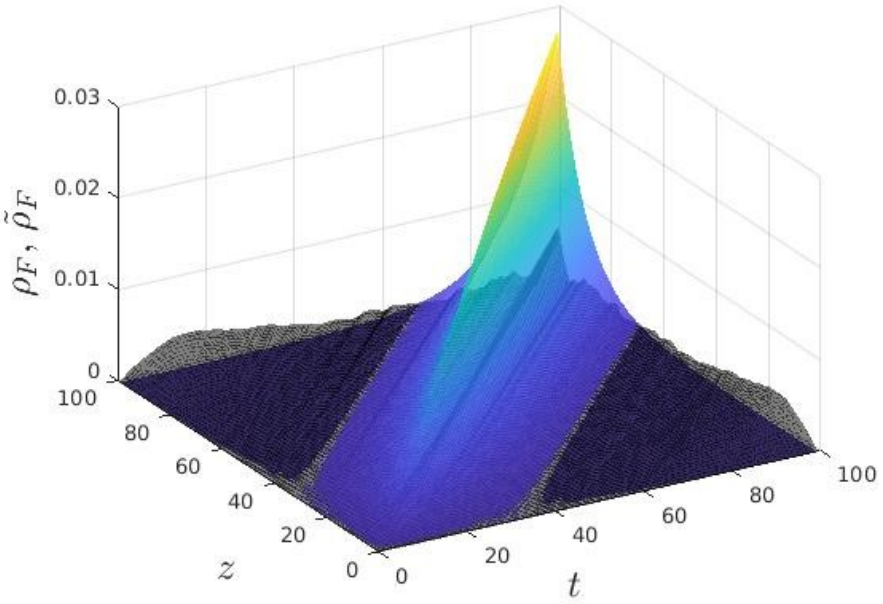
Homogeneous sol. of the GMEs

Convex penalization of irregular solutions

Numerical estimation results

- Estimation statistics for two different profiles of (control) signal *G(t)* :





- *Left :* Performance in estimation of $\bar{\rho}_E$ depends on *G(t)*
- *Right :* Small estimation error (black mash) of true ρ_F despite divergenent estimation accuracy in the tail of $\bar{\rho}_E$

Conclusions

Conclusions

- Generalized Moment Equations
- Efficient regularized inference algorithms based on finite autocovariance expansion
- Important role of modulating (control) signal
- Performance depends on whether source or modulated process is estimated
- Perspective applications to real data

… Thanks !

eugenio.cinquemani@inria.fr
team.inria.fr/ibis