SGN-6156, Lecture 8
Modeling biological regulatory networks

Harri Lähdesmäki, harri.lahdesmaki@tut.fi

Department of Signal Processing,

Tampere University of Technology

23.04.2008

Modeling of biological processes

- The problem of building a good mathematical model is to balance details versus higher-level generality, i.e., to capture essential biological features
- For example:
  - Diagrams + qualitative verbal descriptions
  - Simple (semi)quantitative models, such as deterministic or stochastic linear or discrete models
  - Coupled (stochastic) biochemical reactions
  - Systems of ordinary or partial differential equations, the chemical master equations

Simulation of biological processes

- Construct a mathematical model by combining the current knowledge of a particular biological system
  - Test the model against the current understanding (model validation)
  - Perform (simulation based) virtual experiments in different contexts/with different initial values or perturbations
  - Generate or redefine new biological experiments. Check the simulation based predictions in wet-lab
- Stochastic (and deterministic) simulations can be extremely useful if accurate model of a system is available
- The previous lecture introduced the essential ideas and algorithms for detailed simulation of a stochastic biological process: Chemical master equations, Gillespie algorithm and its variant, ODEs and SDEs

- Before simulations, a model needs to be defined.
  - Model selection: which variables $x_1, \ldots, x_n$ affect a variable y (y can be one of the x_is) and what is a specific type of function that describes stochastic relationship between $x_1, \ldots, x_n$ and y?
  - Parameter values: what are appropriate parameter values for a chosen model(s)?
- In some cases a model is known accurately, but more often we face a problem where we have no clue of the underlying biological model
- In the context of stochastic modeling (using previously introduced stochastic models)
  - Parameters can be learned from measurement data (this requires quite involved computation and is not discussed in this course)
  - Parameters (rate constants) can be "measured" in some cases
  - Model selection is difficult/practically impossible

Quantitative models of biochemical systems, recap

- CSB1 course introduced the essential concepts for quantitative models of biochemical systems, e.g.
  - Enzyme-catalyzed reactions, e.g., substrate S forms a product P (catalyzed by an enzyme E)

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + P,$$

    where ks are the rate constants, leads to a differential equation (similarly for other variables)

$$\frac{dP}{dt} = k_2 ES$$

- Michaelis-Menten equations assume a steady state condition has been reached

$$\frac{dP}{dt} = \frac{v_{\max} S}{K_m + S}$$

- A quantitative model can be specified (up to parameter values) starting from known chemical reactions

Quantitative models for transcriptional regulation

- Transcriptional regulation is a central regulatory control mechanism in cells and is a basis for many cellular processes
- A simplified example of eukaryotic transcription from (Wilkinson, 2006)
- Assume a case where two TFs, TF_1 and TF_2, regulate a gene g
- TF_1 binds the promoter of g (a specific location upstream of g)
- TF_2 binds the promoter of g (another specific location upstream of g) only if promoter is already bound by TF_1
- TF_1 cannot unbind DNA once TF_2 has bound
- TF_1 and TF_2 recruit RNA polymerase to bind the DNA and to initiate transcription
- All steps are reversible

- This can be modeled as (see Figure 1.5 in Wilkinson, 2006)

$$\begin{aligned} g + TF_1 &\rightleftharpoons g \cdot TF_1 \\ g \cdot TF_1 + TF_2 &\rightleftharpoons g \cdot TF_1 \cdot TF_2 \\ g \cdot TF_1 \cdot TF_2 + RNAP &\rightleftharpoons g \cdot TF_1 \cdot TF_2 \cdot RNAP \\ g \cdot TF_1 \cdot TF_2 \cdot RNAP &\rightarrow g \cdot TF_1 \cdot TF_2 + RNAP + r \end{aligned}$$

- The above model can be accurate enough for certain modeling purposes, but the transcription process is much more complex in reality (see additional material)
- A precise model (e.g. which TFs bind g) is most often unknown!

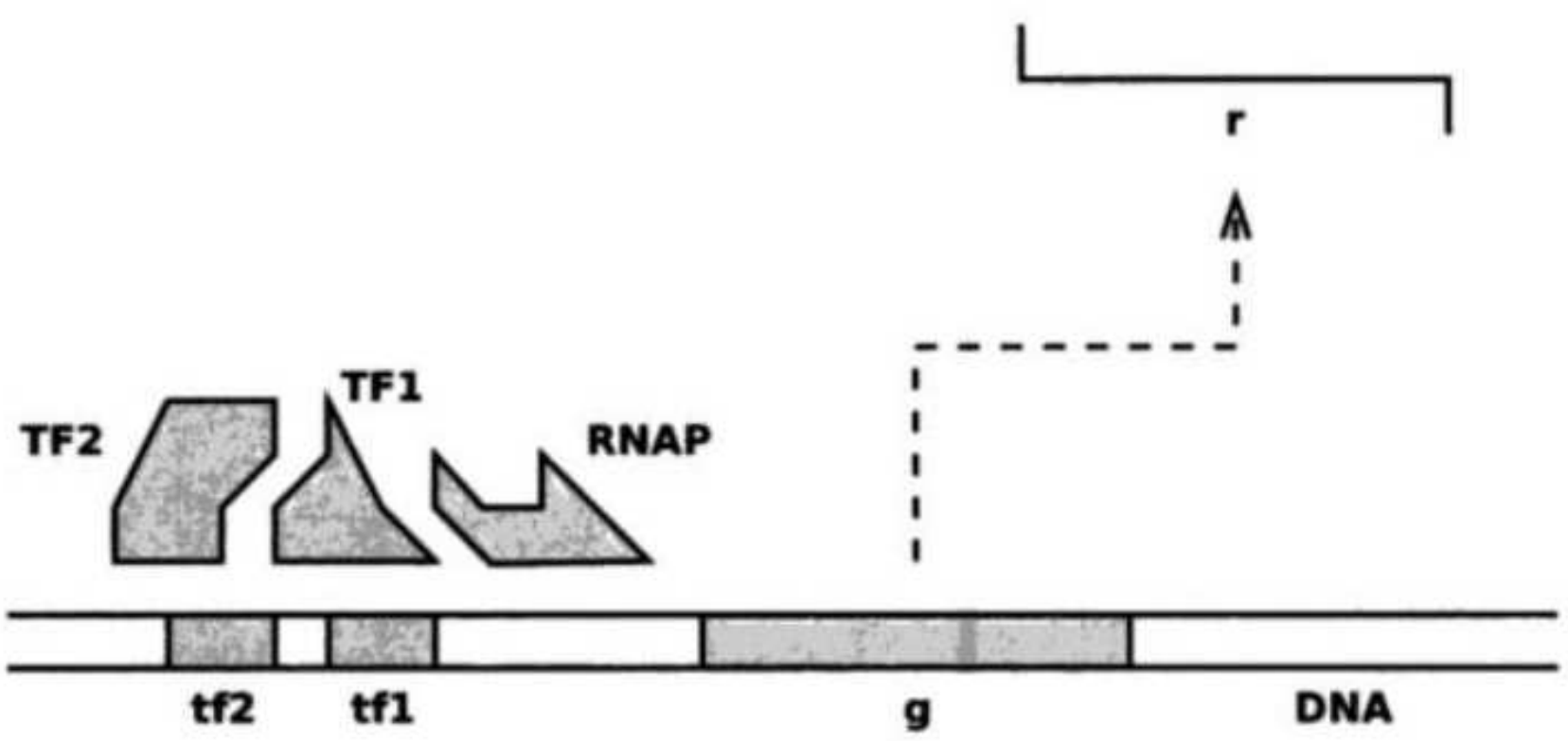

Figure 1.5 *A simple illustrative model of the transcription process in eukaryotic cells*

Figure from (Wilkinson, 2006)

Quantitative models for translation and degradation

- mRNA is translated into a protein with the help of ribosome and folded into a 3-D structure

$$r + Rib \rightleftharpoons r \cdot Rib$$
$$r \cdot Rib \rightarrow r + Rib + P_u$$
$$P_u \rightarrow P_f$$

- Degradation of mRNA by RNase

$$r + RNase \rightarrow r \cdot RNase$$
$$r \cdot RNase \rightarrow RNase$$

and degradation of folded protein (tagged by a signal molecule t)

$$P_f + t \rightarrow P_f \cdot t$$
$$P_f \cdot t \rightarrow t$$

Modeling transcriptional regulation

- The above models (or their more elaborated versions) can in principle give us a model for transcriptional regulation
  - The well-known lac operon model (see also Figure 1.8 in Wilkinson, 2006)
- As noted above, accurate models are rarely available in practice

$\rightarrow$ Learn the models from measurement data (recall also the binding site prediction problem from the last week)

Simulation of a model

- Recall the simplest possible numerical simulation method (Euler's method) for ODEs (more sophisticated methods exist)
- Variables $X = (X_1, \ldots, X_n)^T$ and an arbitrary function f of X with parameters θ

$$\begin{aligned} \frac{dX(t)}{dt} &= f(X(t)|\theta) \\ \lim_{\triangle t \to 0} \frac{X(t+\triangle t) - X(t)}{\triangle t} &= f(X(t)|\theta) \end{aligned}$$

- For small values of $\triangle t$ this is well approximated with the finite diffidence as

$$\frac{X(t+\triangle t) - X(t)}{\triangle t} \approx f(X(t)|\theta)$$

and by solving for $X(t + \triangle t)$ one gets

$$X(t + \triangle t) = X(t) + \triangle t f(X(t)|\theta)$$

- The above equation can be applied repeatedly to compute $X(t_0)$, $X(t_0 + \triangle t)$, $X(t_0 + 2\triangle t)$, ... which can be used to approximate the exact solution

$$X(t) = X(t_0) + \int_{t_0}^{t} f(X(t)|\theta)dt$$

Parameter estimation

- Assume first that we are given a model up to unknown parameter values
- Parameter estimation for θ given data $D = \{(Y_1, t_1), \ldots (Y_m, t_m)\}$
  1. Randomly choose θ
  2. Simulate model/numerically solve for $X(t)$
  3. Assess the goodness of the parameters, e.g.

$$e(\theta) = \sum_{i=1}^{m} (Y_i - X(t_i))^2$$

  4. Check for convergence of θ and stop if converged
  5. Update θ e.g. to the direction of negative gradient and go back to step 2
  6. Repeat the whole process with several different initial values

- Another commonly used but typically more crude approximation for $\frac{dX(t)}{dt} = f(X(t)|\theta)$ is

$$\frac{X(t_{i+1}) - X(t_i)}{t_{i+1} - t_i} \approx f(X(t_i)|\theta), \quad t_1 < t_2 \ldots < t_m$$

- Accuracy of the approximation depends on the measurement sampling times

- This can be interpreted as standard linear/nonlinear regression problem $y_i = f(X(t_i)|\theta)$ where

$$y_i = \frac{X(t_{i+1}) - X(t_i)}{t_{i+1} - t_i}$$

  and can be solved for θ by any standard methods

Model selection

- The most interesting case and the most often met in practice is the one where both the activation function f and the subset of variables that regulate y are unknown
- Without constraints, there are 2^n different combinations/subsets of $\{X_1, \ldots, X_n\}$
- There might also be a family of activation function f to consider, $f_1, \ldots, f_\ell$. In the most general case, there are infinitely many functions to consider...
- In that case, the use of the above simple search method for the identification of the best subset(s)/activation function(s), by minimizing squared error criterion on sample data D, is destined to fail for finite (i.e., in practice small) sample sizes

- This is due to the fact that the above error criterion is of the type of resubstitution, i.e., parameters of a model are fitted to the whole data without taking into consideration the model complexity
- Thus, more complex models/larger subsets will decrease the error although they are far away from the true model and do not generalize to unseen data points ((i.e., are overfitted to given data)
- A principled model selection method is needed
- Three different types of model selection methods
  - Assess predictive accuracy (cross-validation, bootstrap)
  - Bayesian model selection
  - Error-bound bounds

Cross-validation

- In k-fold cross-validation, the (training) data D is split into k non-overlapping parts D_i that have (approximately) the same size, i.e.: $D_i \cap D_j = \emptyset$, $i \neq j$, $|D_i| \approx |D_j|$, $i \neq j$ and $D = \cup_i D_i$
- Each set D_i is left out from the training data in turn and the model parameters are estimated from $D_1, \ldots, D_{i-1}, D_{i+1}, \ldots, D_k$. The accuracy of the model is tested on the left out set D_i
- This process is repeated for all k folds and the average prediction accuracy from the k repetitions is used as the error estimate
- The k-fold cross-validation can be repeated several times with randomly chosen D_is and again average
- If $K = m$ where m is the number of data points this corresponds to the leave-one-out cross-validation (LOOCV)

- Cross-validation gives an approximately unbiased prediction error estimate for data set size $m - m/K$
- Larger k gives a smaller bias but larger variance, and the other way round
- Computationally rather expensive at least for large values of k

An example

- Example from (Bonneau et al., 2006)
- Learn transcriptional regulatory networks from gene expression data using a model of the form

$$\frac{dY}{dt} = f(\beta_1 X_1 + \ldots + \beta_n X_n) - \tau Y$$

  where g is a sigmoidal type of function
- Model selection using cross-validation

Example (cont.)

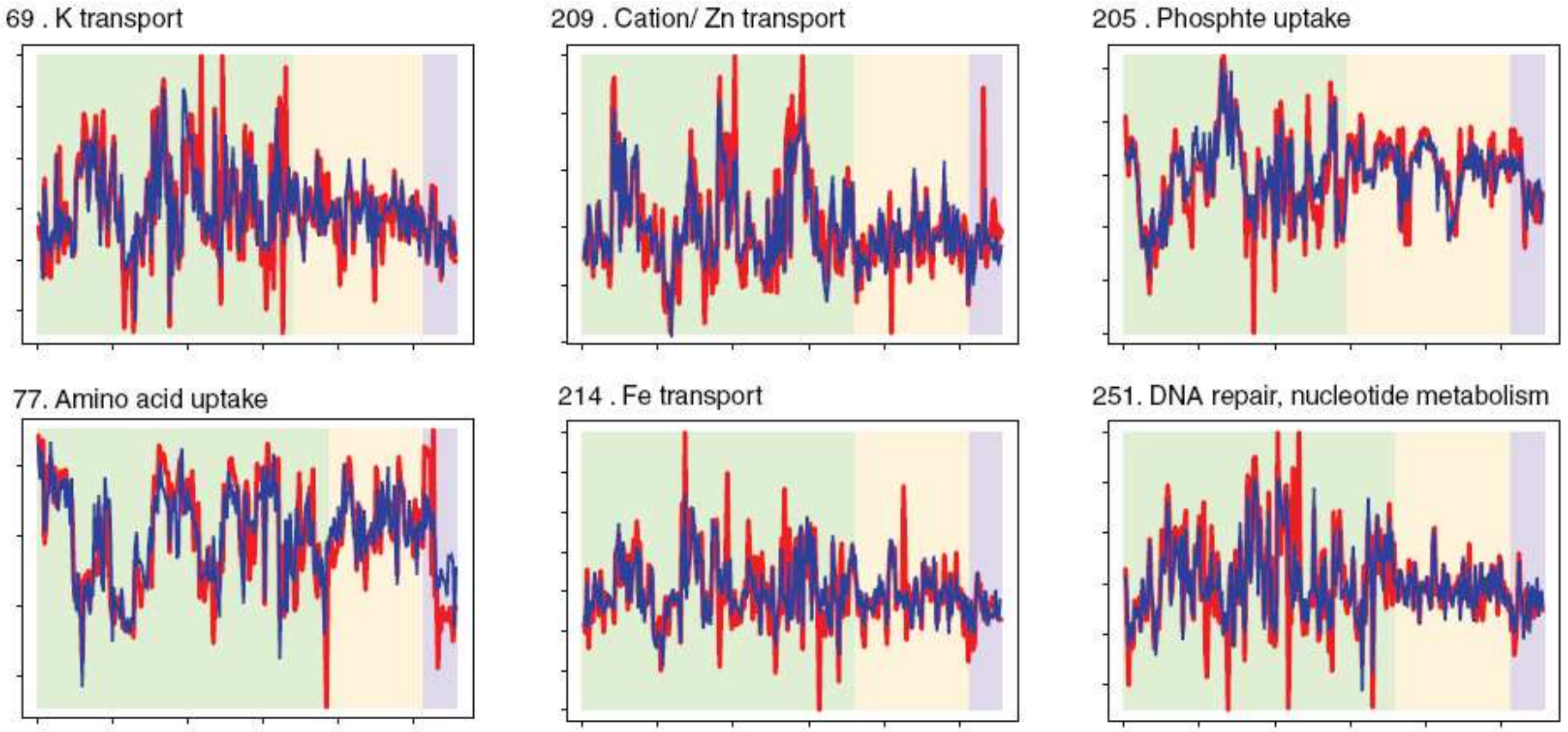

Figure adapted from (Bonneau et al., 2006)

An advertisement

- A new course will be taught next year: *Modeling Techniques for Stochastic Gene Regulatory Networks*, 3 cr, lectured by Dr. Andre S. Ribeiro

References

- Wilkinson D, *Stochastic Modelling for Systems Biology*, Chapman & Hall/CRC, 2006.
- Bonneau R et al., The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*, *Genome Biology*, 7:R36, 2007.