2008－BIO－13 (3)
2008／6／26

# ペトリネットによる転写制御ネットワークのモデリングと統計的推測

吉田 亮[1]，長崎 正郎[2]，山口 類[2]，井元 清哉[2]，宮野 悟[2]，樋口知之[1]
[1] 情報・システム研究機構 統計数理研究所
[2] 東京大学 医科学研究所 ヒトゲノム解析センター

## 概要

遺伝子発現情報を利用して生体内分子ネットワークを推測するための統計解析理論について議論する．生体内分子ネットワークとは，細胞内で起こる様々な化学反応の連鎖を表す多義的な用語である．具体例としては，転写制御 (mRNA の合成)，タンパク質相互作用ネットワーク，代謝ネットワークなどが挙げられる．ネットワークの統計的推測の過程は次のように要約される: (1) 生体内分子ネットワークの (インシリコ) モデリング (2) 観測データ（例えば遺伝子発現データ）にもとづくモデルパラメータの推定 (4) 推定モデルの性能評価 (5) リモデリング．本研究の目標は，これら一連の過程に関わる方法論を整備することである．ここでは，時系列遺伝子発現データを利用した転写制御ネットワークの推測を例に，提案手法について概説する．

# Graphical modeling and statistical inferences of transcription regulatory networks using hybrid functional Petri net

Ryo Yoshida[1], Masao Nagasaki[2], Rui Yamaguchi[2], Seiya Imoto[2],
Satoru Miyano[2], Tomoyuki Higuchi[1]
[1]Institute of Statistical Mathematics, Research Organization of Information and Systems
[2]Human Genome Center, Institute of Medical Science, University of Tokyo

## Abstract

Building *in silico* simulation models of genetic regulatory networks provides a rigorous tool for unraveling complex machinery of biological pathways. To proceed to *in silico* simulations, it is an essential first step to find the effective values of kinetic rate constants, which are difficult to measure directly from *in vivo* and *in vitro* experiments. The aim of this research is to present a new statistical technology, called Genomic Data Assimilation, for the data-driven construction of *in silico* simulation models.

## モデリングと推測の方法

$n$ 時点に観測された遺伝子発現ベクトルを $\boldsymbol{x}_n \in \mathbb{R}_+^p$ と表す．$\boldsymbol{x}_n$ の第 $i$ 要素 $x_{in}$ は遺伝子 $i$ の発現レベル (mRNA の転写レベル) に対する計測値である．ここで，データの生成機構に対して，次のような状態空間表現を行う．

$$\boldsymbol{x}_n = \boldsymbol{H}\boldsymbol{z}_n + \boldsymbol{w}_n, \quad n \in \mathcal{N}_{\mathrm{obs}}, \tag{1}$$

$$\boldsymbol{z}_n = \boldsymbol{f}(\boldsymbol{z}_{n-1}; \boldsymbol{\theta}) + \boldsymbol{v}_n, \quad n \in \mathcal{N}. \tag{2}$$

$\boldsymbol{z}_n \in \mathbb{R}_+^m$ は $n$ 時点での制御タンパク質の濃度や mRNA の転写レベルを要素にもつ長さ $m$ の正値ベクトルである．これら $m$ 個の変数は，$p$ 個の遺伝子発現 ($\boldsymbol{x}_n$) の制御に関与するもので，非観測量 (潜在確率変数) として取り扱われる．システムモデル (2) は，これら潜在的確率変数の時間発展を記述するマルコフ性を保持するインシリコシミュレーションモデルである．$\boldsymbol{v}_n$ と $\boldsymbol{\theta}$ はそれぞれシステムノイズ，未知のモデルパラメータを表す．観測モデル (1) は，非観測量 $\boldsymbol{z}_n$ と観測量 $\boldsymbol{x}_n$ の関係を記述する．観測行列 $\boldsymbol{H}$ の各要素は，0 か 1 の値をとる．例えば，$\boldsymbol{z}_n$ の第 $j$ 番目の変数が $x_{in}$ によって観測されれば，$(\boldsymbol{H})_{ij} = 1$，そうでなければ $(\boldsymbol{H})_{ij} = 0$ となる．$\boldsymbol{w}_n$ は観測ノイズを表

す．システムモデルと観測モデルの時間スケール $\mathcal{N}$ と $\mathcal{N}_{\rm obs}$ は非同一で，一般に $\mathcal{N}_{\rm obs}$ は $\mathcal{N}$ の部分集合となる．

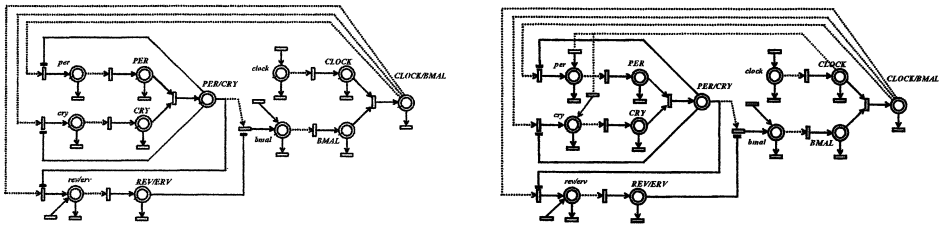


図 1: 哺乳動物の体内時計遺伝子の転写制御ネットワーク．左図は文献（Fujii *et al.* (2004)）に報告されている制御関係を HFPN 表現によって整理したものである (Model 1)．図中に示されている 12 個のノードは，5 種類の mRNA 分子とその翻訳タンパク質，および 2 種類のタンパク複合体によって構成される．各エッジはノード間の分子相互作用を表わす．ここでは示していないが，それぞれのエッジには，ノード間の反応速度を規定する方程式が割り当てられている．右図は，CLOCK タンパクから per, cry への制御を新たに追加したモデル (Model 2)．

本研究では，Hybrid functional Petri net (HFPN) というグラフィカルモデル (差分方程式) にもとづき，仮説的知見を積極的に取り込むことでシステムモデル $\boldsymbol{f}$ を定める．HFPN は，システム生物学の分野で，生体内ネットワークの表現モデルとして比較的広範に用いられているシミュレーションモデル作成エンジンである (Matusno *et al.* (2000))．これを統計モデルとして利用する．モデリングの詳細については，当日の発表内で述べるが，ミカエリス・メンテン式など酵素反応速度理論を模倣して，$m$ 変数間の制御関係を記述する．パラメータ $\boldsymbol{\theta}$ は主に反応速度 (転写速度やタンパクのリン酸化，結合速度) を規定するものである．図 1 は，文献情報にもとづき哺乳動物の体内時計遺伝子の転写制御ネットワークを構築した後，それを HFPN 表現したものである．

モデル $\boldsymbol{f}$ は不完全情報にもとづくものであり，統計的モデル選択の過程でより良いモデルの探索を行う必要がある．そのために，まずなすべきことは，パラメータ $\boldsymbol{\theta}$ の推定である．しかしながら，遺伝子発現データの観測時点数は通常 10 から 20 と極端に少ないため，パラメータの推定は極めて困難である．したがって，ベイズ的な生化学的事前情報の活用や正則化パラメータ推定の設計が鍵となる．本研究では，パラメータ推定問題を次の制約付き事後確率最大化によって定式化する．

$$\begin{aligned} \min_{z,\theta} \quad & \sum_{n\in\mathcal{N}_{obs}} \frac{1}{\epsilon}||\boldsymbol{x}_n - \boldsymbol{H}\boldsymbol{z}_n||^2 + \sum_{n\in\mathcal{N}} \frac{1}{\lambda}||\boldsymbol{z}_n - \boldsymbol{f}(\boldsymbol{z}_{n-1};\boldsymbol{\theta})||^2 + \frac{1}{\eta}||\boldsymbol{z}_0 - \boldsymbol{\mu}_0||^2 + \frac{1}{\gamma}||\boldsymbol{\theta} - \boldsymbol{\omega}||^2 \\ \text{s.t.} \quad & \boldsymbol{z}_n \in \mathbb{R}_+^m,\ \boldsymbol{\theta} \geq \boldsymbol{0},\ \epsilon,\lambda,\gamma,\eta \geq 0 \end{aligned}$$

第 1 項は観測モデルによって規定される $\boldsymbol{z}_n$ のデータ $\boldsymbol{x}_n$ への適合度，第 2 項はモデル $\boldsymbol{f}$ の滑らかさを規定する正則化項である．残りの 2 項は，初期変数とパラメータに関する事前分布を表す．$(\epsilon,\lambda,\eta,\gamma)$ は正則化パラメータである．$\boldsymbol{z}_n$ はタンパク質の濃度や mRNA の転写レベル，$\boldsymbol{\theta}$ は速度パラメータであり，正の符号制約が課される．速度パラメータや初期変数の事前平均 $\boldsymbol{\omega}$ や $\boldsymbol{\mu}$ については，文献情報や生物実験の結果を利用して定める．一般にモデル式 $\boldsymbol{f}$ は非線形な関数であり，この最適化問題を直接解くことは困難である．そこで，概線形モデルという HFPN のサブクラスを定義し，その中でネットワークモデリングを行う戦略を採用する．概線形モデルの場合，上述の最適化問題は，部分的に凸最適化問題に帰着するため，最適化が容易になる利点がある．概線形モデルはミカエリス・メンテン式やその拡張型など，標準的な反応速度モデルを包含するものであり，実際面において過度な制約にはならない．

次に，一旦モデルが推定された後，何らかの統計的規準にもとづきモデルの性能評価を行う．ここでは詳述しないが，本研究では，ベイズ原理からモデルの感度分析のための規準を導く．提案するベイズ規準は，構築したモデルに対して擬似的なノイズを混入させ，システムの揺らぎを測るための自然な尺度となっている．当日の報告では，Ueda *et al.* (2002) の時系列プロファイルを用いた解析例にもとづき，(1) HFPN によるモデリング (2) 事前分布のデザイン (3) パラメータ推定 (4) モデル選択について紹介する．

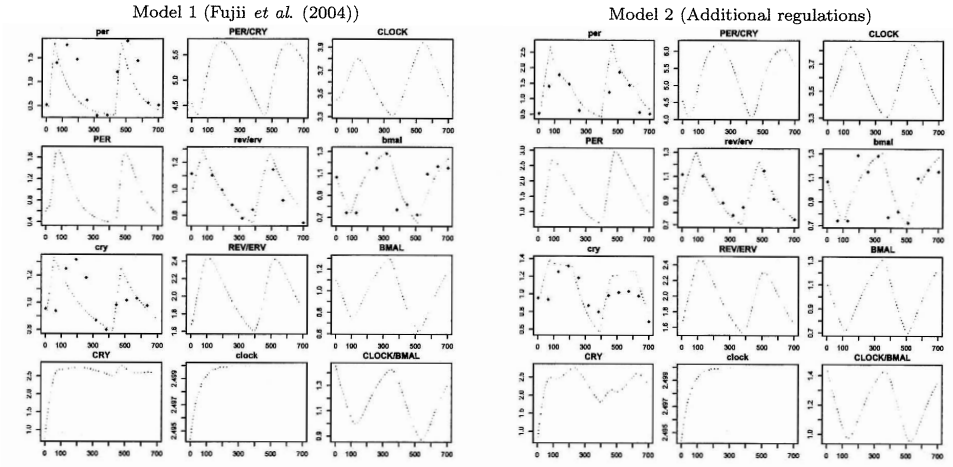


図 2: Ueda *et al.* (2002) の時系列遺伝子発現プロファイル (Affymetrix, mouse, GeneChip) からえられたパラメータ推定の結果．データセットは遺伝子発現量の時間変化を約二日間に渡り計測したものである（体内時計周期 2 に相当）．観測時点は 12 時点からなる．各図は推定したパラメータと初期濃度を用いた 12 変数のシミュレーション結果 (左は Model 1, 右は Model 2)．観測データは黒点によって示されている．

## 参考文献


[1] Fujii,Y. *et al.* (2004) A new regulatory interactions suggested by simulations for circadian genetic control mechanism in mammals, *Genome Inform.*, available at `http://www.jsbi.org/journal/GIW04/GIW04P003.pdf`.

[2] Matsuno,H. *et al.* (2000) Hybrid Petri net representation of gene regulatory network, *Pac Symp Biocomput.*, 341-352.

[3] Nagasaki,M., Yamaguchi,R., Yoshida,R., Imoto,S., Doi,A., Tamada,Y., Matsuno,H., Miyano,S., Higuchi,T. (2006) Genomic data assimilation for estimating Hybrid Functional Petri Net from time-course gene expression data, *Genome Inform.*, **17**, 46-61.

[4] Ueda *et al.* (2002) A transcription factor response element for gene expression during circadian night, *Nature*, **418**, 534-539.

[5] Tasaki,S., Nagasaki,M., Oyama,M., Hata,H., Ueno,K., Yoshida,R., Higuchi,T., Sugano,S., Miyano,S. (2006) Modeling and estimation of dynamic EGFR pathway by data assimilation approach using time series proteomic data, *Genome Inform.*, **17**, 226-238.

[6] Yoshida,R., Nagasaki,M., Yamaguchi,R., Imoto,S., Miyano,S., Higuchi,T. (2008) Bayesian learning of biological pathways on genomic data assimilation (under submission).