線形ダイナミカルシステムモデルの変分ベイズ推定による遺伝子発現時系列のシステム同定

行縄直人[†] 吉本潤一郎[††,†]
大羽成征[†] 石井信[†]

遺伝子発現ダイナミクスの解析のために，状態空間モデルに基づく解析法が提案されている．従来の解析法では，状態変数のダイナミクスを仮定せず，また，システムノイズと観測ノイズを無視したモデルを仮定していたため，状態空間に含まれるノイズ成分を状態変数として誤検出する可能性がある．本研究では，ノイズプロセスに白色ガウシアンを仮定した線形ダイナミカルシステムモデルを考え，変分ベイズ法による推定とモデル選択を行う．本手法を出芽酵母細胞周期に関する公開データセットに適用したところ，従来手法で選択されたモデルと比較し，よりシンプルでもっともらしいモデルが選択された．また，この結果得られたモデルパラメータは，生物学的な考察とよく一致した．人工データへの適用も行い，ノイズを含む時系列データに対する有効性が示された．

System Identification of Gene Expression Time-series Based on a Linear Dynamical System Model with Variational Bayesian Estimation

Naoto Yukinawa,[†] Jun-ichiro Yoshimoto,[††,†] Shigeyuki Oba[†] and Shin Ishii[†]

Several methods based on state space models have been proposed for analyzing dynamics of gene expression. Existing analysis methods can detect false noisy internal variables which seem to have no dynamics in state space because the methods do not assume any dynamics with system noise and observation noise. In this study, we propose a linear dynamical system model in which state variables and observation variables are generated by Gaussian white noise process and provide a variational Bayes inference for the model. We first show effectiveness of our method when applied to a synthesized noisy time-series data set. We also applied our method to a published yeast cell-cycle gene expression data set, then our method could select a simpler and more plausible model than existing method did. In addition, the resultant model parameters well matched the biological considerations.

1. 序　　論

細胞の機能を分子レベルで理解するためには，どの遺伝子が，いつ，どんな条件で，どの細胞内小器官で，どれだけ発現しているのかを詳細に知る必要がある．しかし，生物における遺伝子の発現制御は，核酸や酵素から低分子にいたるまでの多くの構成要素の複雑な相互作用により形成される制御ネットワークを通して実現されている．そのため，たとえ細胞内の生体分子に関する濃度情報が詳細に得られたとしても，それらのダイナミクスを直接解釈するのは困難である．現在，この問題に対して mRNA 定量化技術を用いて細胞の状態と包括的な遺伝子発現量の関係を遺伝子発現プロファイルとして蓄積し，これらに対して統計学的，情報科学的な手法に基づく解析を行うアプローチがとられている[1)]．

本研究では，遺伝子発現プロファイルをもとに，多数の遺伝子の発現をコントロールする遺伝子発現制御因子の数を推定する問題を扱う．細胞には数千から数万の遺伝子が含まれており，その個別の発現挙動は複雑な制御ネットワークによる．一方で，遺伝子発現の大域的な挙動については，転写制御因子や外的環境といったわずかな数の要因に支配されており，そのことが生物の恒常性維持のために重要である．こうした仮説に基づき，細胞状態の時間変化を観測した遺伝子発現プロファイルからのダイナミクスの解析のために，

† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
†† 独立行政法人科学技術振興機構沖縄新大学院大学先行的研究事業
Initial Research Project, Okinawa Institute of Science and Technology, JST

線形状態空間モデルをベースにした解析法が提案されている[2)~4)]．

遺伝子発現プロセスの状態空間モデルでは，観測系列は遺伝子発現量の時間変化に対応し，非観測な内部状態空間における状態変数および遷移行列は，細胞における潜在的な上位システムあるいは外部環境，すなわち，遺伝子発現を制御する因子を仮想している．遺伝子発現のダイナミクスの複雑さを知ることは，システム同定という工学的逆問題において重要なだけでなく，生物の恒常性維持と環境適応との競合のメカニズムを知るうえで手がかりとなる．このため，モデルの複雑さを規定する状態空間の次元数の最適決定が問題となる．従来手法では，状態空間モデルにおいて，しばしば状態空間でのダイナミクスと，システムノイズと観測ノイズを無視したモデルを仮定している．特に文献 2) では，観測行列と内部状態変数の推定を因子分析の問題として定式化している．しかし，遺伝子発現プロファイルは高次元でありノイズが多く含まれるため，こうした簡略なモデルでは，ノイズを含みデータの生成過程にダイナミクスが想定されるデータを扱うには十分とはいえない．

本研究では，遺伝子制御系のモデルとして線形ダイナミカルシステム（Linear Dynamical System; LDS）モデルを仮定したシステム同定法を提案し，遺伝子発現レベルの時系列データから，生きた細胞内で動的に変化する発現制御因子の挙動と，個々の遺伝子の特徴を同時に解析する手法を提案する．LDS モデルは，白色ガウスノイズをともなうガウス過程モデルの 1 つであり，時間変化する内部状態変数の系列から観測系列が生成されるものとする．

因子分析モデルや LDS の生成モデルは，内部状態変数とパラメータを持つ指数族に属するため，EM アルゴリズムによる最尤（maximum likelihood; ML）推定法を用いて推定することができる[5),6)]．ここで内部状態変数の次元数の同定が問題となるが，ML 推定法では複雑なモデルほど選択されやすいため，次元同定が困難である．この問題に対する 1 つの解決法としては，情報量基準を用いてモデル選択を行う方法があげられる．たとえば，因子分析モデル[2)] では次式で定義される Bayesian information criterion（BIC）[7)] によるモデル選択を行っている．

$$BIC \equiv -2L + F\log_2 n$$

ここで，L は推定されたモデルの対数尤度，F はモデルに含まれるパラメータ数，n は標本数である．BIC はモデルの自由度に対する罰則付きの負の対数尤度を表すため，この値が最小のモデルが，データを表現するのに適切なモデルとして選択される．これに対し，ベイズ推定法の一手法である，変分ベイズ（variational Bayes; VB）法[8)] は，有効なパラメータ推定アルゴリズムであるとともに，特にデータ数が不十分なときなどで情報量基準よりも有効なモデル選択法として用いることもできる．

本研究では，LDS モデルの変分ベイズ推定法[9),10)] を用いて，ロバストなパラメータ推定と，システムの複雑さに関わる状態空間の次元の同定を行う．また，学習により得られた LDS モデルのパラメータのうち，特に観測行列に着目した．これは遺伝子の特徴を表す特徴ベクトルの集合と解釈できることから，遺伝子に関する既知の生物学的研究との比較を行うことで，手法の有効性を検討した．適用実験では，まず人工データからのモデルパラメータの推定を行い，ダイナミクスを持つ内部状態変数系列の次元を正しくとらえられることを示した．次に，出芽酵母の細胞周期の各位相における遺伝子発現についての公開データ[11)] を用いて，モデルパラメータ推定を試み，本モデルとデータとの適合性，データ生成の内部状態に関して検討した．また，得られた観測行列と公開データに対する生物学的知識との関連付けを行い，遺伝子を特徴付ける情報が観測行列に抽出されている可能性があることを示した．

2. 線形ダイナミカルシステムモデル

2.1 遺伝子発現プロファイル

遺伝子発現プロファイルとは，さまざまな実験条件下での細胞サンプルにおける遺伝子の発現レベルを網羅的に測定したデータである．通常，各遺伝子の発現レベルとして，測定対象サンプルとコントロールサンプルの対数発現比が用いられる．

測定時点 t における発現プロファイルベクトル $\boldsymbol{y}_t$ を，

$$\boldsymbol{y}_t = (y_{t1}, \ldots, y_{tD})'; \quad t = 1, \ldots, T \tag{1}$$

で表す．ここで y_{tj} は測定時点 t，遺伝子 j の発現レベル，D は測定対象となる遺伝子数，また T は測定時点数である．

2.2 線形ダイナミカルシステムの確率モデル

本研究で用いる LDS モデルは，離散時間で遷移する N 次元の非観測な内部状態変数の系列 $\boldsymbol{x}$ と，その線形変換により生成される D 次元の可観測な観測状態変数の系列 $\boldsymbol{y}$ の 2 つの状態系列について，以下のシステム方程式として定式化される．

$$\boldsymbol{x}_t = \boldsymbol{W}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t; \quad t = 2, \ldots, T, \tag{2}$$

$$\boldsymbol{y}_t = \boldsymbol{V}\boldsymbol{x}_t + \boldsymbol{\eta}_t; \quad t = 1, \ldots, T, \tag{3}$$
$$\boldsymbol{x}_1 \sim \mathcal{N}_N(\boldsymbol{x}_1|\boldsymbol{\mu}_1, \sigma_1^2\boldsymbol{I}_N), \tag{4}$$
$$\boldsymbol{\epsilon}_t \sim \mathcal{N}_N(\boldsymbol{\epsilon}_t|\boldsymbol{0}_N, \sigma_\epsilon^2\boldsymbol{I}_N), \tag{5}$$
$$\boldsymbol{\eta}_t \sim \mathcal{N}_D(\boldsymbol{\eta}_t|\boldsymbol{0}_D, \sigma_\eta^2\boldsymbol{I}_D) \tag{6}$$

$\boldsymbol{x}_1$ は $\boldsymbol{x}$ の初期値である．$\boldsymbol{\epsilon}_t \in \mathcal{R}^N$ および $\boldsymbol{\eta}_t \in \mathcal{R}^D$ はそれぞれ観測ノイズとシステムノイズである．これらのノイズは正規分布に従うものと仮定する．ここで，観測ノイズとシステムノイズをともに一般的な正規分布でモデル化するのは，ノイズモデルに関して先験的な知識がないこと，推定の対象となるパラメータ数を抑えて推定をロバストにすること，および，状態変数の推定がロバストに行えるようにするためである．なお，

$$\mathcal{N}_p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{S}) \equiv (2\pi)^{\frac{-p}{2}}|\boldsymbol{S}|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

は，平均 $\boldsymbol{\mu}$，共分散行列 $\boldsymbol{S}$ の p 次元正規分布の確率密度関数である．

$\boldsymbol{\mu}_1 \in \mathcal{R}^N$ は状態変数の初期値の平均値，$\boldsymbol{W} \in \mathcal{R}^{N\times N}$ は内部状態遷移行列（遷移行列），$\boldsymbol{V} \in \mathcal{R}^{D\times N}$ は観測状態生成行列（観測行列）である．σ_1^2，σ_ϵ^2，σ_η^2 はそれぞれ $\boldsymbol{x}_1$，$\boldsymbol{\epsilon}_t$，$\boldsymbol{\eta}_t$ の分散である．$\boldsymbol{\theta} \equiv \{\boldsymbol{\mu}_1, \sigma_1^2, \boldsymbol{W}, \sigma_\epsilon^2, \boldsymbol{V}, \sigma_\eta^2\}$ が，モデルパラメータのセットとなる．

状態変数 $\boldsymbol{x}_t$ と観測変数 $\boldsymbol{y}_t$ に関するシステム方程式 (2)，(3) と白色ガウシアンノイズの仮定 (4)，(5)，(6) より，以下の確率モデルが導かれる．

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta}) = \begin{cases} \mathcal{N}_N(\boldsymbol{x}_1|\boldsymbol{\mu}_1, \sigma_1^2\boldsymbol{I}_N) & t = 1, \\ \mathcal{N}_N(\boldsymbol{x}_t|\boldsymbol{W}\boldsymbol{x}_{t-1}, \sigma_\epsilon^2\boldsymbol{I}_N) & t = 2, \ldots, T, \end{cases}$$

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta}) = \mathcal{N}_D\left(\boldsymbol{y}_t|\boldsymbol{V}\boldsymbol{x}_t, \sigma_\eta^2\boldsymbol{I}_D\right)$$

以上をまとめると，完全データ $\boldsymbol{X}_{1:T} \equiv \{\boldsymbol{x}_t\}$，$\boldsymbol{Y}_{1:T} \equiv \{\boldsymbol{y}_t\}$ に対するモデルパラメータ $\boldsymbol{\theta}$ の尤度関数

$$p(\boldsymbol{Y}_{1:T}, \boldsymbol{X}_{1:T}|\boldsymbol{\theta}) = \prod_{t=1}^{T} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta})p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta}) \tag{7}$$

が得られる．

本研究では，モデルパラメータ $\boldsymbol{\theta}$ の事前分布として以下で与えられる共役事前分布を仮定した．

$$p(\boldsymbol{\mu}) = \mathcal{N}_N\left(\boldsymbol{\mu}|\boldsymbol{0}, \gamma_0^{-1}\boldsymbol{I}_N\right), \tag{8}$$
$$p(\sigma_1^2) = \mathcal{G}\left(\sigma_1^{-2}|\gamma_0, \gamma_0\tau_{\mu 0}\right), \tag{9}$$
$$p(\boldsymbol{W}) = \prod_{i=1}^{N}\mathcal{N}_N\left(\boldsymbol{w}_i|\boldsymbol{0}_N, \gamma_0^{-1}\boldsymbol{I}_N\right), \tag{10}$$
$$p(\sigma_\epsilon^2) = \mathcal{G}\left(\sigma_\epsilon^{-2}|\gamma_\epsilon, \gamma_\epsilon\tau_{\mu_\epsilon}\right), \tag{11}$$
$$p(\boldsymbol{V}) = \prod_{j=1}^{D}\mathcal{N}_D\left(\boldsymbol{v}_j|\boldsymbol{0}_D, \gamma_0^{-1}\boldsymbol{I}_D\right), \tag{12}$$
$$p(\sigma_\tau^2) = \mathcal{G}\left(\sigma_\tau^{-2}|\gamma_\tau, \gamma_\tau\tau_{\mu_\tau}\right) \tag{13}$$

ただし，$\mathcal{G}(\sigma^{-2}|\gamma, \gamma\tau)$ は，

$$\mathcal{G}\left(\sigma^{-2}|\gamma, \gamma\tau\right) \equiv \frac{(\gamma\tau)^\gamma(\sigma^{-2})^{\gamma-1}}{\Gamma(\gamma)}\exp[-\gamma\tau\sigma^{-2}]$$

で定義されるガンマ分布である．$\boldsymbol{w}_i$ と $\boldsymbol{v}_i$ はそれぞれ，$\boldsymbol{W}$ の第 i 行ベクトルと $\boldsymbol{V}$ の第 j 行ベクトルを示す．ほぼ無情報な事前分布を実現するために，$\gamma_0 = 0.0001$，$\gamma_{\epsilon 0} = \gamma_{\eta 0} = 0.01$，$\tau_{\epsilon 0} = \tau_{\epsilon 0} = 0.01$ を用いた．また後の適用実験では，$\tau_{\mu 0}$ を遺伝子ごとの分散の 1 遺伝子あたりの平均値とした．

2.3 観測行列の性質

観測行列 $\boldsymbol{V}$ は $D \times N$ 行列である．各行ベクトル $\boldsymbol{v}_i \in \mathcal{R}^{1\times N}, i = 1, \ldots, D$ は内部状態変数 $\boldsymbol{x}_t$ から観測変数 y_{ti} への写像を規定し，大域的因子に対する遺伝子 i の応答特性を示すものである．この性質から，$\boldsymbol{v}_i$ を遺伝子 i に対する特徴量と見なすことができ，観測ベクトルと呼ぶ．

2.4 変分ベイズ法

観測変数の系列 $\boldsymbol{Y}$ が与えられたとき，未知変量に関する事後分布 $p(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{Y})$ を求めることがベイズ推定の目的である．この事後分布はベイズの定理により，以下で与えられる．

$$p(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{Y}) = \frac{p(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{Y})}{p(\boldsymbol{Y})}, \tag{14}$$
$$p(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{Y}) = p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{15}$$
$$p(\boldsymbol{Y}) = \int p(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\theta})p(\theta)d\boldsymbol{\theta}d\boldsymbol{X} \tag{16}$$

正規化項 $P(\boldsymbol{Y})$ は，周辺化尤度と呼ばれ，LDS モデルにおける内部状態変数の次元 N に対する尤度を表すため，モデル選択のための指標として用いることができる[12]．LDS モデルは内部状態変数についてはガウス過程モデルの一種であるが，パラメータと内部状態変数の事後分布および周辺化尤度を解析的に求めることは困難である．このため，本研究では変分ベイズ法を用い，事後分布および周辺化尤度の近似計算を行う．

変分ベイズ推定では，内部状態変数 $\boldsymbol{X}$ およびパラメータ $\boldsymbol{\theta}$ の事後分布 $p(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{Y})$ を近似するための試

験事後分布 $q(\boldsymbol{X},\boldsymbol{\theta}) \approx p(\boldsymbol{\theta},\boldsymbol{X}|\boldsymbol{Y})$ を用意し，以下で定義される対数周辺化尤度 $\ln p(\boldsymbol{Y})$ の下界である自由エネルギー（variational free energy）$\mathcal{F}[q(\boldsymbol{\theta},\boldsymbol{X})]$ を，試験事後分布に関して変分法的に最大化することでベイズ推定を実現する．

$$\begin{aligned}\ln p(\boldsymbol{Y}) &\equiv \ln \int p(\boldsymbol{Y},\boldsymbol{X}|\boldsymbol{\theta})p(\theta)d\boldsymbol{\theta}d\boldsymbol{X} \\ &\geq \int q(\boldsymbol{\theta},\boldsymbol{X}) \ln \frac{p(\boldsymbol{Y},\boldsymbol{X}|\theta)p(\boldsymbol{\theta})}{q(\boldsymbol{\theta},\boldsymbol{X})} d\boldsymbol{\theta}d\boldsymbol{X} \\ &= \log p(\boldsymbol{Y}) - \mathrm{KL}(q(\boldsymbol{X},\boldsymbol{\theta})||p(\boldsymbol{X},\boldsymbol{\theta}|\boldsymbol{Y})) \\ &\equiv \mathcal{F}[q(\boldsymbol{\theta},\boldsymbol{X})] \end{aligned} \tag{17}$$

ここで，$\mathrm{KL}(\cdot||\cdot)$ は 2 つの分布間の Kullback-Leibler 情報量であり，$q(\boldsymbol{\theta},\boldsymbol{X}) = p(\boldsymbol{X},\boldsymbol{\theta}|\boldsymbol{Y})$ で最小値 0 となる．

$\mathcal{F}[q(\boldsymbol{\theta},\boldsymbol{X})]$ の最大化は，独立分解近似

$$\begin{aligned} q(\boldsymbol{\theta},\boldsymbol{X}) &= q(\boldsymbol{\theta})q(\boldsymbol{X}), \\ q(\boldsymbol{X}) &= \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \\ q(\boldsymbol{x}_1|\boldsymbol{x}_0) &\equiv q(\boldsymbol{x}_1) \end{aligned} \tag{18}$$

のもとで，$q(\boldsymbol{X})$ に関する最大化，$q(\boldsymbol{\theta})$ に関する最大化を交互に繰り返す変分法的 EM（VB-EM）アルゴリズムによって行うことができ，収束性が保証されている．また，自由エネルギーの最大値は対数周辺化尤度の近似値となっているため，パラメータ数の異なるモデル間での，モデル選択基準となりうる[12]．

自由エネルギーは，前述の BIC と比較した場合，指数分布族で与えられる確率モデルに対して，より有効なモデル選択基準である．LDS モデルでは事後分布が単峰に近い可能性があるため，変分ベイズ法はサンプリング手法に迫る性能を，大幅に少ない計算コストで実現できると考えられる[13]．

3. 関 連 研 究

3.1 状態空間モデルの先行研究

状態空間モデルを用いた発現プロファイルからのシステム同定を目指した先行研究では，内部状態変数のダイナミクスとノイズの過程を無視した簡略なモデルを想定し，特異値分解[3] や因子分析に対する EM アルゴリズム[2] により状態変数とパラメータを求める手法が提案されているが，現状で発現制御因子（内部状態）数の推定まで踏み込んでいるのは Wu らの研究[2] のみである．Wu らは，因子分析モデルにおける自由度を，最尤推定の結果から得られる BIC を用いて決定することにより，発現制御因子数の推定を試みた．

3.1.1 因子分析モデル

Wu らの因子分析モデルは，

$$\boldsymbol{Y} = \boldsymbol{V}\boldsymbol{X} \tag{19}$$

で定義される．ここで，$\boldsymbol{Y}$ は T 点の発現プロファイルをまとめた $D \times T$ の遺伝子発現行列，$\boldsymbol{V}$ は LDS モデルと同様の $D \times N$ の観測行列，$\boldsymbol{X}$ は T 点の状態変数ベクトルからなる $N \times T$ の状態変数行列であり，これらがモデルパラメータとなる．

パラメータ推定では，与えられたデータ $\boldsymbol{Y}$ に対し，因子分析モデルの EM アルゴリズムによる推定[14] により，因子得点行列と因子負荷行列に対応する $\boldsymbol{V}$ と $\boldsymbol{X}$ を求める．

3.2 その他の関連研究

状態空間モデル以外の遺伝子発現時系列データの解析法では，S-systems による非線形微分方程式モデル や[15),16)] ブーリアンネットワークモデル[17),18)] に基づいた手法が代表的なものとしてあげられる．S-systems によるモデル化では，mRNA やタンパク質など生体分子の濃度変化のダイナミクスをそれぞれ微分方程式により記述し，連立微分方程式を構成する．そして，各種数理最適化手法によりデータから係数を同定する．ブーリアンネットワークモデルでは，各遺伝子の発現に応じて状態を二値化し，その制御規則をブール関数で表す．データからその状態遷移規則を学習することで，遺伝子制御ネットワークの構造推定を行うことができる．ブーリアンネットワークモデルと線形計画法を組み合わせて，微分方程式モデルの係数を最適化する手法も提案されている[19]．

我々が提案する LDS モデルを含む状態空間モデルと，これらのモデルが大きく異なる部分は，前者が遺伝子間の相互作用を陽に仮定せず，全 D 個の遺伝子に共通する N 個の遺伝子制御因子により発現が駆動されるとするのに対し，後者では，個々の遺伝子が個別に相互作用を持つことで発現が制御されるとする点にある．このため，状態空間モデルでは $N = D$ という特殊なケースを除くと，遺伝子間の相互作用を直接扱うことができない．だが，その代わりに，観測ベクトル v_i を，内部状態変数のダイナミクスのモデルに応じて低次元に射影された遺伝子発現ベクトルと見なし，遺伝子間の類似度を評価できる．これが状態空間モデルならではの特徴である．特に，LDS モデルでは，連続的に時間変化する遺伝子制御因子のみを抽出するために，従来の因子分析では規定されていなかった線形の内部状態変数のダイナミクスを考えているため，大域的な遺伝子を制御する滑らかな基底と，それに対応した遺伝子の情報の両者の抽出が期待できる．

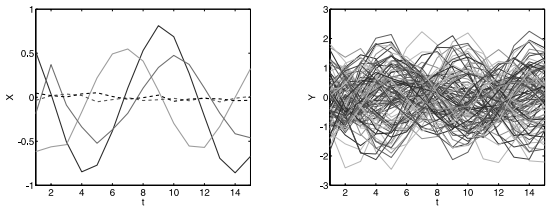


図 1 人工データ．左図は 5 つの内部状態変数の時系列，右図は内部状態変数と観測行列から生成された 100 サンプルの観測系列を示す

Fig. 1 Synthesized data. The left figure shows time-series of three-dimensional internal state, and the right one shows the observation time-series of 100 samples generated by the internal-state time-series and an observation matrix.

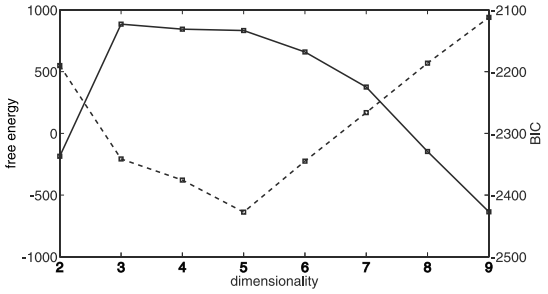


図 2 人工データに関するモデルの内部状態空間の次元に対する提案モデルでの自由エネルギーと因子分析モデルでの BIC．実線が自由エネルギー，破線が BIC を示す

Fig. 2 Plot of the free energy of LDS models (solid) and BIC of factor analysis models (dash) versus the dimensionality of the internal state space for the synthesized data set.

4. 適 用 実 験

4.1 人工データによる評価

状態空間に含まれるダイナミクスを持つ成分を抽出できるかどうかを評価するために，人工データを用いて提案手法の性能評価を行った．

4.1.1 人工データ

まず，状態遷移行列

$$\boldsymbol{W} = \begin{bmatrix} 0.9071 & 0.7655 & -0.2499 \\ -0.3238 & 0.7116 & -0.2128 \\ 0.6780 & 0.0002 & -0.2133 \end{bmatrix}$$

を持つ LDS モデル（$N = 3$）を用いて，15 時点（$T = 15$）の内部状態変数系列を生成した．これらの内部状態変数系列に加え，ダイナミクスを仮定しない無情報な 2 つの内部状態変数系列を，区間 $[-0.053, 0.053]$ の一様乱数より生成することで，合計 5 つの内部状態変数系列を得た（**図 1** 左）．ここで，システムノイズの標準偏差 σ_ϵ は 0.02 とした．次に，得られた内部状態変数系列に対し，$\mathcal{N}_{100}(0, \boldsymbol{I}_{100})$ に従って生成した観測行列 $\boldsymbol{V}$ を用いて，100 サンプルの観測状態系列 $\boldsymbol{Y}_{1:15}$ を生成した（図 1 右）．観測ノイズの標準偏差 σ_η は 0.05 とした．

4.1.2 システム同定

生成したデータ $\boldsymbol{Y}_{1:15}$ を用いて，LDS モデルを用いたパラメータ推定，およびモデル選択を行った．データに対し，$N = 1$ から $N = 10$ までの 10 個の LDS モデルを用意し，推定の結果最大の自由エネルギーが得られたモデルを最適なモデルと決定した．比較のため，従来手法である，因子分析と BIC によるモデル選択も同様に行った．

図 2 は，LDS モデルと因子分析モデルでの自由エネルギーおよび BIC を示す．自由エネルギー最大化の観点からモデル選択を行うと，データ $\boldsymbol{Y}_{1:15}$ について，LDS モデルでは状態空間の次元 $N = 3$ が選択された．また，因子分析モデルでは，BIC より $N = 5$ のモデルが選択された．

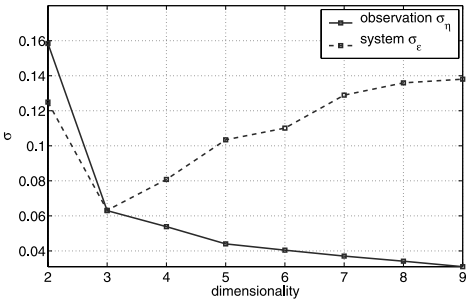


図 3 人工データに関する LDS モデル（$N = 2, \ldots, 9$）のシステムノイズと観測ノイズの標準偏差の推定値．実線と破線はそれぞれ，観測ノイズの標準偏差 σ_ϵ とシステムノイズの標準偏差 σ_η を示す

Fig. 3 Standard deviation of noise on estimated LDS models for the synthesized data set ($N = 2, \ldots, 9$). Solid and dash lines represent standard deviation of observation noise and system noise, respectively.

次に，LDS モデルに関するシステムノイズ分散と観測ノイズ分散の推定値に関して評価を行った（**図 3**）．モデルの複雑さが増加するほど，データに適合しやすくなるため，観測ノイズ分散は内部状態変数の次元に対して単調減少を示す．一方，システムノイズ分散は，$N = 3$ のモデルで最小値をとる形となっている．これは，LDS で選択されたモデルの内部状態変数の次元と一致しており，本 LDS モデルと変分ベイズによる推定では，システムノイズを最小化する方向でモデル選択が行われたと考えられる．

以上より，LDS モデルでは，そのダイナミクスに従

う内部状態変数の成分の数 $N = 3$ を自動的に検出できることが示された．これに対し，因子分析モデルでは，ダイナミクスを持つ成分だけでなく，無情報な成分も検出してしまう結果となった．

4.2 酵母遺伝子発現プロファイルに対する適用

4.2.1 データセット

提案手法の実データに対する適用性を評価するため，公開遺伝子発現プロファイルデータに対する適用実験を行った．用いるデータは，Spellman らが文献 11) の実験において，出芽酵母 *cdc15-2* の変異株の細胞周期における 6177 遺伝子の発現量の 24 時点にわたる時間変化を cDNA マイクロアレイを用いて観測して得た対数発現比である．本データは，http://cellcycle-www.stanford.edu/☆から入手可能である．

本データは，1) 包括的な遺伝子発現の観点から細胞内の現象を明らかにする目的で計測されたものである，2) 実験結果に基づいた 800 個の遺伝子の機能分類情報が提供されているため，LDS モデルによる特徴抽出結果との対応付けが可能である，3) 時系列データの形式をとり時点数も十分である，4) 従来手法の適用実験[2] でも用いられた，といった性質を持ち，LDS モデルの評価に適していると考え，評価に用いることにした．

まず，前処理としてデータセットから，Spellman らによって同定された 800 個の遺伝子と，等間隔（10 分間隔）で測定された 19 時点のサンプルのデータ（800 遺伝子 ×19 時点）を選択した．ついで，このデータに含まれる 1,023 個（5.3%）の欠測値について，Bayesian PCA アルゴリズム[20] により補完を行った．さらに，800 個の遺伝子から，ランダムに 200 遺伝子を抽出し学習データを構成した．

4.2.2 システム同定

内部状態変数の次元を $N = 1, \ldots, 10$ に設定した 10 個の LDS モデルを用意し，VB-EM アルゴリズムによるパラメータ推定を行った．事後分布は単峰に近いものと予想されるが，アルゴリズムが局所最適解に収束する可能性もあるため，推定の初期値を変えつつ 20 試行繰り返し，20 試行中で最大の自由エネルギーを実現したモデルを採用した．また，因子分析モデルに対する EM アルゴリズム（最尤推定）によるシステム同定を行い比較した．因子分析モデルでは解が一意に求まるため，各モデルに対し 1 試行のみ推定を行った．

☆ Stanford 大学 Yeast Cell Cycle Analysis Project.

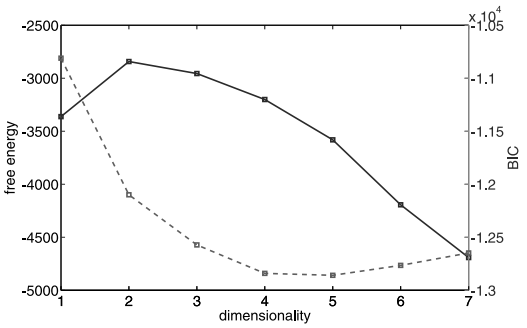


図 4 酵母データに関するモデルの内部状態空間の次元に対する提案モデルでの自由エネルギーと因子分析モデルでの BIC．実線が自由エネルギー，破線が BIC を示す

Fig. 4 Plot of the free energy of LDS models (solid) and BIC of factor analysis models (dash) versus the dimensionality of the internal state space for the yeast data set.

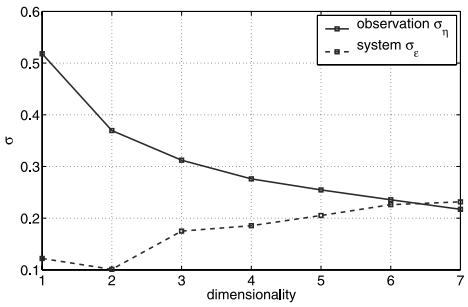


図 5 酵母データに関する LDS モデル（$N = 1, \ldots, 7$）のシステムノイズと観測ノイズの標準偏差の推定値．実線と破線はそれぞれ，観測ノイズの標準偏差 σ_ϵ とシステムノイズの標準偏差 σ_η を示す

Fig. 5 Standard deviation of noise on LDS estimated models for the yeast data set. ($N = 1, \ldots, 7$). Solid and dash lines represent standard deviation of observation noise and system noise, respectively.

図 4 は，内部状態変数の次元 $N = 1, \ldots, 7$ に対する，LDS モデルにおける自由エネルギーの最大値と，因子分析モデルにおける BIC を示すプロットである．これより，自由エネルギー最大化の観点から評価すると，LDS モデルでは最適な状態空間の次元数は $N = 2$ であるといえる．図には示していないが，状態変数の次元が 7 より大きいモデルでも，自由エネルギーは単調減少の傾向が見られた．一方，因子分析モデルでは $N = 5$ のモデルが選択されている．

選択された $N = 1, \ldots, 7$ の LDS モデルに関する，システムノイズおよび観測ノイズの標準偏差の推定値を**図 5** に示す．やはり $N = 2$ のモデルにおいてシステムノイズが最小となっていることが分かる．

図 6 は，自由エネルギーが最大となった $N =$

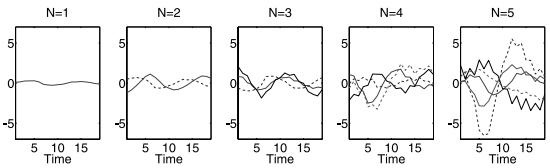


図 6 自由エネルギーが最大となったモデルでの内部状態変数 $\boldsymbol{x}$ の時系列．各列はモデルの内部状態変数の次元に対応する

Fig. 6 Time-series of the internal variable $\boldsymbol{x}$ in the model with the largest free energy. Each panel corresponds to a single internal dimensionality.

$1,\ldots,5$ のモデルの内部状態変数の変動を，推定したパラメータから再現したものである．$N=1$ のモデルでは発現プロファイルの変動を表すには十分ではないと考えられる．また，$N=4$ や $N=5$ のモデルでは，ある状態変数の変動が，他の状態変数のものの定数倍，もしくは，状態変数の変動どうしの重ね合わせで表現されるような，冗長性が観察される．このことは，図 5 において，システムノイズ分散が $N=4$ や $N=5$ で $N=2$ よりも大きくなっているということに対応すると考えられる．あらゆるモデル中で自由エネルギーが最大となった $N=2$ では，ちょうどフーリエ基底に対応するような，位相が異なりながら周期的挙動を示す 2 種類の変動が抽出されている．

図 7 は，自由エネルギーが最大となった $N=2$ のモデルにおける $\boldsymbol{V}$ の推定値における観測ベクトル $\boldsymbol{v}_i, i=1,\ldots,D$ を，二次元の要素空間にプロットしたものである．図中の各点が 1 遺伝子に対応する．Spellman らは，細胞周期における機能が既知である 93 個の遺伝子の発現プロファイルをもとに，細胞周期に関与すると考えられる 800 個の遺伝子を同定し，各遺伝子に対して，G_1/S，S，S/G_2，G_2/M，M/G_1 の 5 つに分割できる細胞周期の中でいつ活性されるのかを，既知遺伝子との時系列の類似性に基づき分類を行った．図中のシンボルは，この分類結果に対応している．観測ベクトルの空間である v_1–v_2 空間において，時計回りの回転方向に 5 つの細胞周期フェーズに分類された遺伝子が並んでいることから，LDS モデルが Spellman らが遺伝子を分類した際の特徴空間を自動的に構成していることが分かる．

5. 議　　論

体細胞分裂は，多細胞生物にみられる最も基本的な周期的かつ自律的な生理現象である．細胞が分裂し遺伝情報を複製した後，再び分裂するまでの過程を細胞周期というが，これは巨視的な観点から明確に 4 つの段階に分けることができる．分子レベルで見た場合

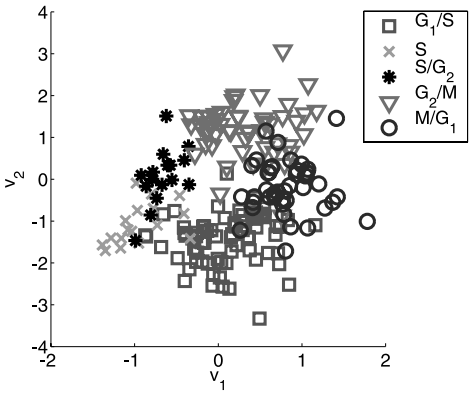


図 7 $N=2$ の LDS モデルの推定結果から得られた $\boldsymbol{V}$ の横ベクトルの散布図．各シンボルは，Spellman らによって同定された，体細胞分裂の過程において遺伝子が高いレベルで発現するフェーズを示す

Fig. 7 Scatter plot of the observation vectors of $\boldsymbol{V}$ in the LDS model with $N=2$. Each symbol represents the known phase information identified by Spellman and others.

も，各段階で発現する遺伝子は特異的であり，各遺伝子の発現量は細胞周期の中でつねに動的に変動している．Spellman ら[11] は，細胞周期における遺伝子発現変化の周期性を仮定し，解析のための遺伝子発現のダイナミクスのモデルとして，周期変動する 2 種類の基底，サイン波とコサイン波の線形和すなわちフーリエ基底を採用している．このモデルでは，位相と振動数がシステムを規定するパラメータであり，それらは LDS モデルにおける状態遷移行列と状態変数の初期値に対応する．また，2 つの線形和の重みパラメータは，LDS モデルの観測ベクトルに対応する．今回，我々は LDS モデルと変分ベイズ推定を組み合わせることにより，Spellman らの解析モデルにおいて仮定されているものと等価な $N=2$ の基底（内部状態時系列）を自動的に構成した．

一方，ノイズと状態変数のダイナミクスを陽に仮定しない因子分析モデルを，我々が用いたものと同様のデータへ適用した結果では，我々の結論とは異なる状態空間次元 $N=5$ のモデルが選ばれた．人工データの解析結果をふまえると，これは，彼らのモデルが，状態空間に含まれる意味のないノイズ成分を別の因子としてとらえてしまったことが一因であると考えられる．

我々の LDS モデルは，状態変数のダイナミクスとシステムノイズと観測ノイズを組み込んだ確率モデルとなっているため，データのノイズに対して比較的ロバストに，時間変化する状態変数の基底を抽出できると考えられる．マイクロアレイ実験などから得られた

ノイズのある時系列データを解析するためには，この性質は大きな利点としてはたらくと思われる．

6. 結　　論

我々の手法の一番の強みは，定常的な過程にあると考えられる現象から観測されたダイナミクスを持つ時系列データに対して，最適な基底を自動的に求めることができることにある．これは，生物のような自律的に恒常的活動を刻むシステムの背後にある要因を探ることを可能にする道具となりうる．

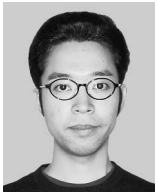
一方で，本手法の欠点としては，一般的に構成要素間の因果関係が非線形であると考えられている生物のシステムに線形性の仮定を行っていること，システムに定常性を要求すること，また，内部状態変数に対する外部因子の入力を省いた状態空間モデルとなっていることの 3 つが主に考えられる．特に第 3 の簡略化は，生物の環境への適応性を議論するためには問題であり，将来的には，これらの簡略化を除去した手法を提案したい．また，より広範なデータに対して本手法を適用し，その有効性を検討する予定である．

参 考 文 献

1) de Jong, H.: Modeling and simulation of genetic regulatory system: a literature review, *Journal of Computational Biology*, Vol.9, pp.67–103 (2002).
2) Wu, F.X., Zhang, W.J. and Kusalik, A.J.: Modeling gene expression from microarray expression data with state-space equations, *Pacific Symposium on Biocomputing*, Vol.9, pp.581–592 (2004).
3) Dewey, T.G. and Galas, D.J.: Dynamic models of gene expression and classification, *Functional & Integrative Genomics*, Vol.1, pp.269–278 (2001).
4) Holter, N.S., Maritan, A., Cieplak, M., Fedoroff, N. and Banavar, J.R.: Dynamic modeling of gene expression data, *Proc. National Academy of Sciences of the United States of America*, Vol.98, pp.1693–1698 (2001).
5) Dempster, A.P., Laird, N.M. and Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society B*, Vol.39, pp.1–38 (1977).
6) Roweis, S. and Gharahmani, Z.: A unifying review of Linear Gaussian models, *Neural Computation*, Vol.11, pp.305–345 (1999).
7) Schwarz, G.: Estimating the dimension of a model, *Annals of Statistics*, Vol.6, pp.461–464 (1978).
8) Attias, H.: Inferring parameters and structure of latent variable models by variational Bayes, *Proc. 15th Conference on Uncertainty in Artificial Intelligence*, pp.21–30 (1999).
9) Gharahmani, Z. and Beal, M.J.: Propagation algorithms for variational Bayesian learning, *Advances in Neural Information Processing Systems 13*, pp.507–513 (2001).
10) Yoshimoto, J., Ishii, S. and Sato, M.: System identification based on on-line variational Bayes method and its application to reinforcement learning, *Artificial Neural Networks and Neural Information Processing* (*ICANN/ICONIP 2003*), Lecture Notes in Computer Science 2714, pp.123–131 (2003).
11) Spellman, P.T., Sherlock, G., Zang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Molecular Biology of the Cell*, Vol.9, pp.3273–3297 (1998).
12) Yoshimoto, J., Ishii, S. and Sato, M.: Hierarchical model selection for NGnet based on variational Bayes inference, *Artificial Neural Networks* (*ICANN 2002*), Lecture Notes in Computer Science 2415, pp.661–666 (2002).
13) Beal, M.T. and Ghahramani, Z.: The variational Bayesian EM algorithm for imcomplete data: with application to scoring graphical model structures, *Bayesian Statistics*, Vol.7, pp.453–464 (2003).
14) Rubin, D. and Thayer, D.: EM algorithms for ML factor analysis, *Psychometrika*, Vol.47, pp.69–76 (1982).
15) Arvine, D.H. and Savageau, M.A.: Efficient solution of nonlinear ordinary differential equations expressed in S-System canonical form, *SIAM Journal on Numerical Analysis*, Vol.27, pp.704–735 (1998).
16) Tominaga, D. and Okamoto, M.: Design of canonical model describing complex nonlinear dynamics, *Proc. IFAC International Conference*, CAB7, pp.85–90 (1998).
17) Liang, S., Fuhrman, S. and Somogyi, R.: REVEAL, a general reverse engineering algorithm for inference of genetic network architectures, *Pacific Symposium on Biocomputing*, Vol.3, pp.18–29 (1998).
18) Akutsu, T., Miyano, S. and Kuhara, S.: Identification of genetic networks from a small number of gene expression patterns under the Boolean network model, *Pacific Symposium on Biocomputing*, Vol.4, pp.17–28 (1999).

19) Akutsu, T., Miyano, S. and Kuhara, S.: Inferring qualitative relations in genetic networks and metabolic pathways, *Bioinformatics*, Vol.16, pp.727–734 (2000).
20) Oba, S., Sato, M., Takemasa, I., Monden, M., Matsubara, K. and Ishii, S.: A Bayesian missing value estimation method, *Bioinformatics*, Vol.19, pp.2088–2096 (2003).

(平成 16 年 8 月 16 日受付)
(平成 16 年 10 月 5 日再受付)
(平成 16 年 10 月 18 日採録)



行縄 直人（学生会員）

平成 13 年東京工業大学生命理工学部卒業．平成 15 年奈良先端科学技術大学院大学情報科学研究科博士前期課程修了．現在，同大学院博士後期課程に在学中．遺伝子発現解析の研究に従事．

吉本潤一郎

平成 10 年関西大学総合情報学部卒業．平成 14 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．科学技術振興機構（JST）CREST 研究員を経て，平成 16 年より JST 沖縄新大学院大学先行的研究事業研究員および奈良先端科学技術大学院大学非常勤講師となり現在に至る．博士（工学）．強化学習，ニューラルネットワーク，統計的学習理論，システム同定の研究に従事．

大羽 成征

平成 14 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了．平成 15 年より同研究科助手．遺伝子発現量データの統計的解析手法開発の研究に従事．博士（工学）．

石井 信

昭和 63 年東京大学大学院工学系研究科修士課程修了．昭和 63 年（株）リコー中央研究所研究員．平成 6 年（株）ATR 人間情報通信研究所研究員．非線形力学系と最適化の研究に従事．平成 9 年奈良先端科学技術大学院大学情報科学研究科助教授．脳型情報処理と統計的学習の研究に従事．平成 13 年より奈良先端科学技術大学院大学情報科学研究科教授．バイオインフォマティクスとシステム神経生物学の研究に従事．博士（工学）．