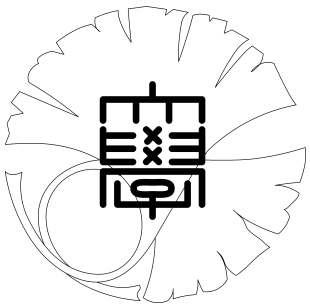
修士論文

動的微分ベイジアンネットワークによる遺伝子制御関係の推定



2005年1月31日

指導教官　伊庭 斉志 教授

東京大学大学院 新領域創成科学研究科

基盤情報学専攻
47-36311

杉本 直也

Abstract

We propose a dynamic differential Bayesian networks (DDBNs) and nonparametric regression model. This model is an extended model of traditional dynamic Bayesian networks (DBNs), which can incorporate temporal information in a natural way and directly handle real-valued data obtained from microarrays without any transformation. In addition, it can cope with differential information between gene expression levels, without any loss to the traditional advantage, i.e., the capability of estimating non-linear relationships between genes.

We apply DDBNs to analyze simulated data and real data, i.e., *Saccharomyces cerevisiae* cell cycle gene expression data. We have confirmed the effectiveness of our approach in the sense that some edges have been successfully detected only by DDBNs, not by DBNs.

内容概要

近年，ヒトが科学の対象として脚光を浴びており，生命の仕組みを解明しようとする研究が行われている．1980 年代終わりに，アメリカを中心としてイギリス・日本・フランス・ドイツからなる国際チームによる “ヒトゲノム計画” がスタートした．この計画によりヒトの DNA のドラフトシーケンス (大まかな配列) が 2000 年 6 月 15 日に決定し，高精度の配列も数年のうちに決定される見込みである．

遺伝子同士は，タンパク質や代謝産物を介して間接的ではあるが，お互いを誘導・抑制している．この相互作用をネットワークとして考え，遺伝子の発現を表すモデルとしてよく用いられている．遺伝子制御ネットワークによって，細胞機能を従来より大きな視点で解析することができる．

マイクロアレイデータから遺伝子同士の因果関係を表す遺伝子制御ネットワークを推定しようとすることはバイオインフォマティクスの主要テーマの一つであり，近年盛んに研究が行われている．遺伝子制御ネットワークのような化学反応に基づく制御関係を記述する場合，変数の変化量が重要な要素となる．しかし，微分方程式系の推定には大量のデータが必要であり，マイクロアレイで得られる時系列の時間ステップ数はそれに比べて少ない．さらに，微分方程式系の推定は誤差に敏感であり，マイクロアレイデータとの相性が悪いという欠点がある．

この欠点に対して，ベイズ統計に基づくグラフィカルモデルであるベイジアンネットワークを用いた研究が多くなされ，成果を挙げている．ベイジアンネットワークの問題点として循環構造を扱うことができないという点が挙げられるが，この問題を克服するために時系列データを扱うことで循環構造が表現可能な動的ベイジアンネットワークが研究されている．

ベイジアンネットワークでは普通，変数を離散として取り扱う．そのため，推定されるネットワークは離散化のための閾値に大きな影響を受ける．また，離散化による情報量の損失も問題となる．この問題を克服するために，発現量を離散化することなく実数値のまま取り扱い変数の非線形な関係を表現できるノンパラメトリック回帰モデルや，それを動的ベイジアンネットワークに拡張したモデルが提案されている．しかしながら，従来のモデルである動的ベイジアンネットワークでは時系列の変化量を扱うことはできなかった．

これらの問題を解決するために，本研究では微分方程式系モデルと動的ベイジアンネットワークモデルの長所を取り入れ，動的微分ベイジアンネットワークモデルを提案する．そして，遺伝子間関係を記述するモデルとして，実数値データを直接扱うことができ，遺伝子間の非線形な関係も取り扱うことができるノンパラメトリック回帰モデルを採用した．

本論文では，動的微分ベイジアンネットワークとノンパラメトリック回帰によるモデルを構築する．そして，そこから導出される評価規準によって遺伝子制御ネットワークの推定を行う．推定対象としては，人工的なネットワークから生成したデータと出芽酵母の細胞周期データを用いた．

ネットワーク探索アルゴリズムとしては，局所探索付き遺伝的アルゴリズムが最も良い成績を示した．比較の規準として，探索速度ではなく正答率が高いアルゴリズムに対して高い評価を与えた．

人工的なネットワークから生成したデータを対象とした推定実験については，S_n/S_p という指標によって評価を行った．その結果，提案手法は両指標に対して従来手法よりも良い成績を示した．

出芽酵母の細胞周期データと対象とした解析では，提案手法によって，従来手法では検出できなかった有用なエッジも検出することができた．ただし，提案手法で検出できなく，従来手法によって検出したエッジの中にも有用なエッジが存在していた．このことから，従来のモデルと本研究で提案するモデルを排他的に考えるのではなく，うまく融合する方法を模索していく必要がある．また，探索過程で有効と判定されたエッジの中に実際のデータベースに登録されているエッジが含まれていた．このことから，探索過程の情報もうまく取り入れることによって生物的に有用な結果が得られる可能性もある．さらに，よい評価を与えるネットワークが必ずしも生物学的に正しいとは限らなかった．

以上をまとめると，本論文では次の 4 つの知見が得られた．

1. 提案手法と従来手法とでは，検出能力を発揮できる変数間関係が異なる．
2. 探索過程にも，生物学的に有用な情報が含まれている．
3. ネットワーク構造の探索には確率的探索を用いる方がよい．
4. 生物学の知見を取り入れ，評価規準を修正する必要がある．

目次

図目次

表目次

第１章

序論

1.1 はじめに

近年，ヒトが科学の対象として脚光を浴びており，生命の仕組みを解明しようとする研究が行われている．1980 年代終わりに，アメリカを中心としてイギリス・日本・フランス・ドイツからなる国際チームによる “ヒトゲノム計画” がスタートした．この計画によりヒトの DNA のドラフトシーケンス (大まかな配列) が 2000 年 6 月 15 日に決定し，高精度の配列も数年のうちに決定される見込みである．DNA 配列が決定するということはその生物の設計図を手にするということに相当するものであるが，配列がわかったからといってその生物の仕組みが分かったことにはならない．つまり，ヒトの DNA 配列が解読されたことでヒトゲノム計画が終了したというわけではなく，得られた DNA 配列の意味を解読し，病気の原因や生命の原理を解明していくためのスタート地点に立ったということになるのである [41].

DNA 上の情報は 4 種類の文字の配列として表現され，これをもとにして生命活動のためのさまざまな部品が作られていく．こういった遺伝情報の解析は多分に情報学的なものとなる．また，DNA マイクロアレイなどに代表される生物学的実験技術の発達により，膨大な量のデータが日々蓄積されつつある．この大量のデータから遺伝情報を解析するためには，従来の生物学的手法のほかに，人工知能をはじめとする情報学的なアプローチが重要となってきている．このように生命科学と情報科学が融合した学問分野が注目されつつあり，バイオインフォマティクス (bioinformatics) や分子生物情報学と呼ばれている．

遺伝子同士は，タンパク質や代謝産物を介して，間接的ではあるがお互いを誘導・抑制している．この相互作用をネットワークとして考え，遺伝子の発現を表すモデルとしてよく用いられている．遺伝子制御ネットワークによって，従来よりも大きな視点で細胞機能を解析することができる．マイクロアレイデータから遺伝子間因果関係を表す遺伝子制御ネットワークを推定しようとすることはバイオインフォマティクスの主要テーマの一つであり，近年盛んに研究が行われている．

遺伝子制御ネットワークのような化学反応に基づく制御関係を記述する場合，変数の変化量が重要な要素となる．変数の変化量を扱う研究として微分方程式系 (Ordinary Differential Equations:ODEs) モデル [7,10] があるが，これらの研究は線形システムによるものであり，複雑な現象を表現するものとしては適していないだろう．非線形なモデルとして，一般質量作用則 (Generalized Mass Action:GMA) を簡単化した S-system と呼ばれる微分方程式系があり，盛んに研究が行われている [27,34,35,44,48,50,51].

微分方程式系の欠点として，推定に大量のデータが必要であることが挙げられるが，マイクロアレイで得られる時系列の時間ステップ数はそれに比べて少ない．さらに，微分方程式系の推定は誤差に敏感であるなど，マイクロアレイデータとの相性が悪い．

この欠点に対して，ベイズ統計に基づくグラフィカルモデルであるベイジアンネットワーク (Bayesian network:BN) [14,19,21,29] を用いた研究が多くなされており，成果を挙げている．

ベイジアンネットワークの問題点として，循環構造を扱うことができないという点が挙げられる．この問題を克服するために，時系列データを扱うことで循環構造が表現可能な動的ベイジアンネットワーク (Dynamic Bayesian Network:DBN) [5, 15, 28, 36] が研究されている．

ベイジアンネットワークでは普通，変数を離散として取り扱う．そのため，推定されるネットワークは離散化のための閾値に大きな影響を受ける．また，離散化による情報量の損失も問題となる．この問題を克服するために，発現量を離散化することなく実数値のまま取り扱い変数の非線形な関係を表現できるノンパラメトリック回帰モデル [19] や，それを動的ベイジアンネットワークに拡張したモデル [21] が提案されている．しかしながら，従来のモデルである動的ベイジアンネットワークでは，生体反応記述の重要な要素である，時系列の変化量を扱うことはできなかった．

マイクロアレイデータの特性から考えると，時間についての変化量を統計的枠組みによって扱うことができるようなモデルが，遺伝子制御ネットワークの推定には適していると考えられる．

1.2　本研究の目的

遺伝子の発現状態を計測するデバイスである DNA マイクロアレイの欠点として，大量の誤差が含まれている，得られる時間ステップが少ない，などが挙げられる．また，遺伝子制御ネットワークのような化学反応に基づく制御関係をモデル化する場合，微分方程式系で扱うような変数の変化量が重要な要素となる．さらに，遺伝子制御ネットワークにおいても存在すると考えられるフィードバックループなどの循環構造を表現する能力も必要である．

以上より，遺伝子制御ネットワークの推定においては次の条件を満たすモデルが有効であると考えられる．

- 誤差を扱うことができる
- 循環構造を表現できる
- 推定に多くの時間ステップを必要としない
- 時系列の変化量を扱うことができる．

本研究ではこれらの条件を満たすモデルの導出を目的とする．そして，推定実験を通じてそのモデルの評価を行う．

1.3　本論文の構成

本論文は全 6 章で構成されており，各章の内容は以下の通りである．

第 1 章

研究の背景を述べ，目的を明らかにする．

第 2 章

生物学的な背景から遺伝子制御ネットワークの定義についてふれ，その推定問題の定式化を行う．そして，遺伝子制御ネットワークの推定に関する研究例を紹介する．

第 3 章

提案する手法の導出に必要な関連研究について述べる．モデルの研究例としてノンパラメトリック回帰とベイジアンネットワークについて述べ，そのモデルを評価するための評価規準の研究例を紹介する．そして，ネットワーク構造の探索アルゴリズムについて説明する．

第 4 章

本論文で提案する遺伝子制御ネットワークの推定手法について述べる．まず本研究で提案するモデルである動的微分ベイジアンネットワークの導出を行い，動的微分ベイジアンネットワークとノンパラメトリック回帰に基づいてモデルを構築する．そして，モデルを評価するための評価規準の導出を行う．また，データの欠損値の取り扱いについてもふれる．

第 5 章

提案手法による遺伝子制御ネットワークの推定実験を行う．推定対象として，人工的なネットワークから生成したデータと出芽酵母の細胞周期データを用いた．実験内容は，ネットワーク構造探索手法の比較，人工データの解析，実データの解析である．

第 6 章

まとめを行い，本研究から得られた知見について述べる．そして，今後の課題を記す．

第２章

遺伝子制御ネットワーク

2.1　遺伝子制御ネットワークとは

すべての生命体において，各遺伝子はその発現量を調整しながら生命活動を維持している．さらに，各遺伝子は独立に発現しているのではなく，タンパク質や代謝産物によってその発現量が制御されている．つまりある遺伝子について考えると，他の遺伝子 (または自分自身) が発現して生成されたタンパク質や代謝産物によって，発現量が誘導されたり抑制されたりする．このように，遺伝子，タンパク質，代謝産物は複雑な相互関係を持っている．その関係をネットワークとしてとらえ，実験データからその構造やパラメータを推定する研究が行われている．

このような生命ネットワークを考える際，次のような表現方法が考案されている．

- 代謝ネットワーク．代謝産物間の化学反応を表す．
- タンパク質ネットワーク．情報伝達酵素によるタンパク質複合体形成などのタンパク質間の相互作用を表す．シグナル伝達ネットワークともいう．
- 遺伝子制御ネットワーク．遺伝子の発現量に関する相互作用を表す．

これらのネットワークを平面と考えれば，生命ネットワークは図 2.1 のように表される．細胞内では遺伝子間に直接的な相互作用は存在せず，遺伝子同士は，タンパク質や代謝産物によって間接的に作用している．実線は直接的な相互作用を示し，点線は遺伝子同士の仮想的な相互作用を表す．遺伝子制御ネットワークは，生命ネットワークにおける相互作用の遺伝子平面への写像であると考えられる [4].

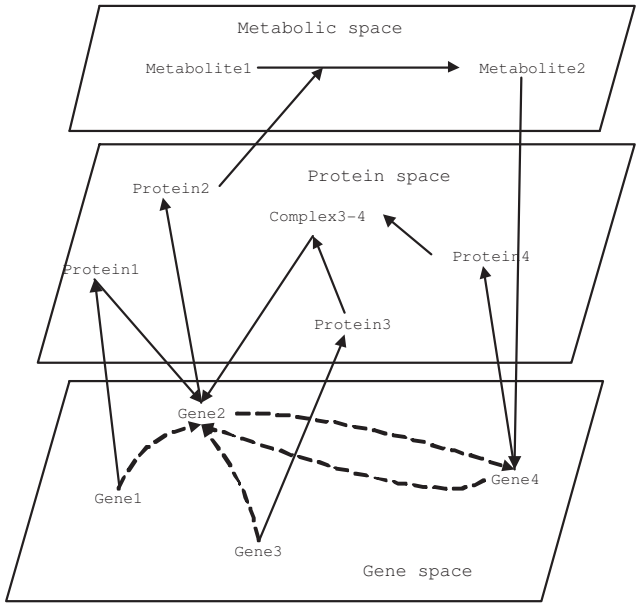


図 2.1: 3 層平面で表された全生化学ネットワークモデル

遺伝子制御ネットワークは他のネットワークに比べて大規模な視点から細胞機能を解析するこ

とができるため，遺伝子の発現を表すモデルとしてよく用いられている．遺伝子制御ネットワークを解明することは，遺伝子の機能や生命の設計図の解明を行ううえで非常に役に立つ．

2.2 DNA マイクロアレイ

遺伝子の発現量を測定する代表的なデバイスとして，DNA マイクロアレイがあげられる．このデバイスによって，遺伝子の発現量が時間や環境に従ってどのように変化していくのかを観測することができる．

DNA マイクロアレイとは，数千から数万種類の DNA 配列を，ガラスやシリコンでできた数 cm 四方のチップの上に格子状に配置したものである [11]．各 DNA 配列の長さは数百から数千塩基程度であり，ひとつのスポットの中には同じ配列の DNA が数多く含まれている．遺伝子の発現量はその遺伝子に対応する mRNA の量を測定することによって知ることができる．DNA マイクロアレイでは，対象となる mRNA に結合するような塩基配列を人工的に生成し (cDNA という)，その塩基配列に蛍光物質を結合させておく．各スポットに含まれる DNA 配列と結合する cDNA があれば各スポットが光り，発現している遺伝子を同定できることになる．また，蛍光の強度を測定することによって，発現量の測定を行うことができる．

また，ある遺伝子を破壊した状態や，ある遺伝子を強制的に発現させた状態での各遺伝子の発現量を測定するような実験も行うことができる．

DNA マイクロアレイによる計測には，非常に大きな誤差がつきまとう．時系列データを観測する際も，現段階では数ステップ ∼ 十数ステップしか計測することができない．

2.3 遺伝子制御ネットワークの推定とモデル化

2.3.1 問題概要

DNA マイクロアレイを用いることによって，多数の遺伝子の発現量を同時に測定し，どの遺伝子がどの程度発現しているのかを知ることができる．これらのデータは時間経過と共に観測され，その時間ステップ数はおよそ数十程度である．

また，DNA マイクロアレイによるデータはかなり多くの誤差を含んでおり，極めてよくコントロールされた状態でも 2 倍程度の誤差が生じる．さらに，遺伝子の発現量は発光の強度として観測されるため，その発光強度を数値化する段階でも誤差を含んでしまう．

このようなデータ，つまり，誤差を含む連続値の時系列データをもとにして，遺伝子のネットワークを推定していく．この問題を解くには，以下のようないくつかの課題を克服する必要がある [49]．

- 組み合わせ爆発：N 個の遺伝子から作りうるネットワークの数は，遺伝子の数が増えると幾何級数的に増大する．その中から，条件を満たすネットワークを探索するのは，非常に計算

時間がかかる．

- 複数の準最適解：元のデータと同じような発現プロファイルを作り出すネットワークは複数個存在し，その数も最悪の場合，非常に大きな数になる．
- パラメータ探索：同じ構造のネットワークでも，遺伝子間制御の強さなどのパラメータが変われば，違う発現プロファイルになる．元のデータを再現するパラメータを発見することも，構造の検討と同時に行われる必要があり，このこと自体でも膨大な計算量が必要となる．
- 非線形性：転写活性因子と転写産物の量の関係が非線形の入出力関係であり，その非線形性が大きい場合には，解析的に解を導き出すことができない上，その他の方法においても大きな誤差が生じる危険性がある．
- データのノイズと信頼性：データにノイズなどの誤差が多く含まれていること，また，実験の手法による信頼度の問題.

つまり，発現プロファイルからの遺伝子制御ネットワークの推定は，組合せ爆発，非線形最適化，確率プロセスという本質的に解決が難しい問題ばかりを集めたような課題であり，そのまま決定的な解放を見出すことは望めない．そのため，研究の初期においては，いくつかの仮定を置いて，厳密に問題を解くことができる，または，実用上問題がないように解くことができるより簡単な問題へと置き換える必要がある．

2.3.2 問題の定式化

生体内での真の遺伝子制御ネットワークを $GRN(G,\theta)$ とすると，DNA マイクロアレイによる発現データの観測は式 (2.1) のような写像として書ける．

$$f_{Observation} \quad : \quad GRN(G,\theta;Cnd_i) \mapsto X_{exp,i} \tag{2.1}$$

ただし，G は真の遺伝子間の制御 · 被制御関係を表すグラフ，θ は真のモデルに含まれる制御関係の条件や制御の強さなどを表すパラメータ，Cnd_i は観測を行ったときの実験条件，$X_{exp,i}$ は観測されたデータである．

n 種類の実験データ $X_{exp} = \{X_{exp,1}, \ldots, X_{exp,n}\}$ が得られたときの遺伝子制御ネットワークの推定問題は，式 (2.2) のように定式化できる．

$$f_{Inference} \quad : \quad X_{exp} \mapsto \hat{G}, \hat{\theta} \tag{2.2}$$

ただし，$f_{Inference}$ は推定アルゴリズム，$\hat{G}$ は推定された遺伝子制御関係，$\hat{\theta}$ は $f_{Inference}$ を構成するモデルに含まれるパラメータである．

$\hat{G}$ を真の遺伝子間制御関係 G に近づけることが遺伝子制御ネットワーク推定の目的である．そして，G にできるだけ近い $\hat{G}$ を出力するような推定アルゴリズム $f_{Inference}$ を構築することが研究の対象となる．

2.3.3 研究例

推定アルゴリズム $f_{Inference}$ を構成するには，遺伝子間制御関係をモデル化する必要がある．その研究例をいくつか紹介する．

閾値検定モデル

このモデルは，遺伝子 a の破壊実験や強制発現実験での遺伝子 b の発現量の変化を調べ，“遺伝子 a から遺伝子 b への影響” の有無を判定して “二項関係” として抽出する．そして，これらの二項関係が統計的に有意か否か判定する．時刻の変化は考慮していない．

ブーリアンネットワークモデル

このモデルは遺伝子がどの程度発現しているかは考慮せず，発現状態を on か off かのどちらかに丸めてしまうというものである [2].

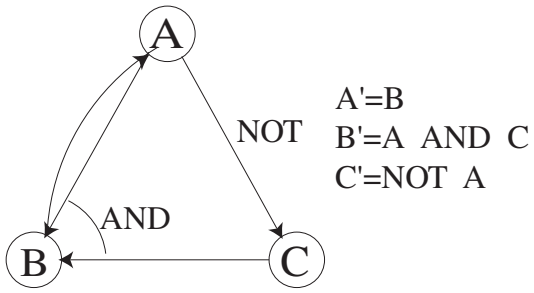


図 2.2: ブーリアンネットワークモデルの例

ブーリアンネットワークは図 2.2 のように，頂点集合 $V = \{v_1, v_2, \ldots, v_n\}$ と各頂点の値を決定するブール関数 (論理関数) の集合 $F = \{f_1, f_2, \ldots, f_n\}$ によって定められる．各頂点は 1 個の遺伝子に対応し，0(発現していないか) か 1(発現している) のどちらかの状態を取る．状態は離散的な時刻 $t = 1, 2, 3, \cdots$ ごとに同期して変化していき，時刻 $t + 1$ における遺伝子 v_i の状態 v_i' は

$$v_i{}' = f_i(v_{i_1}, \ldots, v_{i_K}) \tag{2.3}$$

によって定められる．ここで $v_{i_1}, \ldots, v_{i_K}$ は v_i に直接影響を与える頂点である (この入力頂点の個数 K を入次数という)．ブーリアンネットワークは，頂点集合の状態がどのように移り変わっていくかを表した状態遷移表によって表現することもできる．遺伝子発現量からネットワークを推定するということは，状態遷移表の一部が与えられたときにこれに矛盾しないブーリアンネットワークを求めるということである．

このモデルはシンプルにネットワークを表現しているが，発現レベルを 0,1 のみに限定してしまっている点，同期した変化を仮定している点など，設定に問題があるとの批判もある．

ニューラルネットワークモデル

このモデルでは有向グラフによってネットワークを表現する [42]．ノード，矢印，数値はそれぞれ遺伝子，遺伝子間制御関係，制御レベルを表しており (図 2.3)，これらのパラメータは行列の形で与えられる．重み行列 $W = \{w_{ij}\}_{n\times n}$ は，遺伝子 i と遺伝子 j の間の制御関係を表していて，この値が正のときは誘導，負のときは抑制の関係にあることを表す．このネットワークでの時刻 $t+1$ における遺伝子 i の発現量は，時刻 t におけるすべての遺伝子の発現量 $x_i(t)$ と重み付け行列から計算される値

$$s_i(t) = \sum_{j=1}^{n} w_{ji} x_j(t) \tag{2.4}$$

を用いて算出される．

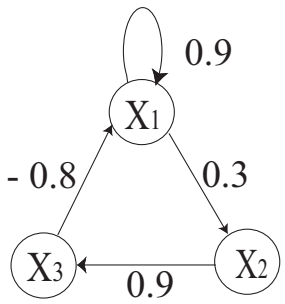


図 2.3: ニューラルネットワークモデルの例

ペトリネットモデル

ペトリネットを用いたモデルも提案されている．ペトリネットとは，プレース (place)，トランジション (transition) という二種類のノード (node) をもつ二部有向グラフ (bipartite digraph) である．プレースは円で表されるノードであり，条件を表す．トランジションは棒または箱で表されるノードであり，事象を表す．これらを結ぶアーク (arc) は条件，事象の間の関係を表す．プレース，トランジション，アークがシステムの構造を表現する．アークには正整数の重みがつけられる場合がある．重みはアークの本数で表すか，アークに重みを併記して表す．重みがつけられないアークは重みが 1 であるとみなす．プレースの上には，非負整数個のトークン (token) が置かれる．トークンは点で表され，条件の成立を表す．プレース上のトークンの配置をマーキング (marking) といい，システムの状態を表す．

形式的には，ペトリネットは $N = (P, T, F, W, M_0)$ であらわされる．

$$
\begin{array}{lcll}
P & = & \{p_1, p_2, \ldots, p_{|P|}\} & \text{プレースの有限集合} \\
T & = & \{t_1, t_2, \ldots, t_{|T|}\} & \text{トランジションの有限集合} \\
F & \subseteq & (P \times T) \cup (T \times P) & \text{アークの集合} \\
W & : & F \mapsto \{1, 2, \ldots\} & \text{アークの重み} \\
M_0 & : & P \mapsto \{0, 1, 2, \ldots\} & \text{初期マーキング}
\end{array}
$$

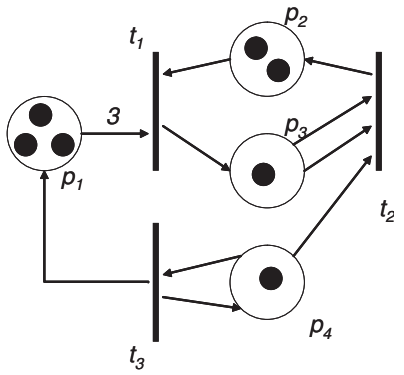


図 2.4: ペトリネットの例

このペトリネットモデルを用いて遺伝子制御ネットワークの推定を行い，よい結果を挙げたという研究も報告されている [23].

微分方程式系モデル

制御理論や統計などの分野でよく用いられるシステム同定のためのモデルとして微分方程式系があり，これには複雑な現象を記述できるという特徴がある．このモデルを用いることによって遺伝子制御ネットワークを推定しようという研究がある．しかし，どのような方程式系を用いれば遺伝子の制御関係をよく表現することができるのかについては現在のところあまりわかっておらず，実験データによく合う方程式系を見つけるアルゴリズムも提案されていない．なので，これまで提案されてきた研究例では，形の決まった微分方程式系をモデルとして用い，その式中のパラメータを最適化することを目的としている．

そのようなモデルのひとつとして S-system というものがあり，これは式 (2.5) で表される連立微分方程式である [48]．このモデルは一般質量作用則 (Generalized Mass Action:GMA) を簡単化したものであり，これまでにも遺伝子制御ネットワークに限らず内部の要素間相互作用が明らかでない系を記述するものとして用いられてきた．

$$
\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{n} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n} X_j^{h_{ij}} (i = 1, 2, \cdots, n) \tag{2.5}
$$

遺伝子制御ネットワークを表す場合，X_i は遺伝子の発現量にあたり，n はネットワーク内の遺伝子の個数に相当する．この式は遺伝子の誘導過程，抑制過程をそれぞれ一本のパスで近似したものとなっている．

また，式の形はあらかじめ固定せず，探索の過程で動的に決定していくという手法も提案されている [34,35,50,51]．この手法では，式構造の決定には遺伝的プログラミングを用い，パラメータの最適化には最小二乗法を用いている．

微分方程式系は複雑な現象を記述できるモデルではあるが，与えられたデータが少ない場合，過学習を起こしてしまうという問題がある．これについては，補完アルゴリズムによってデータ点を増やすという手法も提案されている [27]．また，時系列の振る舞いはパラメータに大きく影響を受けるため，データに誤差が含まれているような場合，ただしく微分方程式系が推定できないという問題が考えられる．

DNA マイクロアレイなどから得られる発現量データは時間ステップが少なく，誤差を多く含むため，微分方程式系で扱うには工夫が必要である．

ベイジアンネットワークモデル

ベイジアンネットワーク (Bayesian Network:BN) とは，不確かな出来事の連鎖について，確率の相互作用を集計する手法で，知能情報システム構築の有力な手段になっているモデルである．

このベイジアンネットワークによって，遺伝子制御ネットワークを推定しようという研究がいくつか報告されている [5,14,15,19,21,28,29,36]．DNA マイクロアレイから得られるデータは非常に多くの誤差を含んでいるため，このような確率モデルは遺伝子制御ネットワークの推定において強力なアプローチになると考えられる．

第3章

関連研究

概観

ベイジアンネットワークの学習を構成する要素は，大まかに次の 3 つである．

1. 変数の振る舞いを記述し，変数間の関係を表すためのモデル
2. 候補解を評価するための評価規準
3. 最適なネットワーク構造を探索するための探索アルゴリズム

本章では，提案する推定手法の導出に必要な関連研究について述べる．モデルとしてノンパラメトリック回帰とベイジアンネットワークについて述べ，評価規準の研究例を紹介する．そして，ネットワーク構造の探索アルゴリズムについて説明する．

3.1　ノンパラメトリック回帰

説明変数 $\boldsymbol{X}$ と目的変数 $\boldsymbol{Y}$ の関係が非線形であるとき，どのようなモデルを用いるかが重要となる．そしてこれまでも，観測されたデータから非線形構造をとらえるために様々なモデルが提案されてきた (例えば [16, 22, 30, 33] など)．変数間の関係が未知な場合，特定の分布 (例えば正規分布など) を仮定することなくモデル化を行う必要がある．このように，特定の分布を仮定することなく回帰を行う手法をノンパラメトリック回帰と呼び，非線形回帰においてしばしば用いられる．

非線形回帰においては，データの平均構造を記述するために移動平均，スプライン，核関数などが用いられるが，ここでは特に等間隔節点を持つ B-スプライン [9] を用いた方法を取り上げる．

本節ではまずノンパラメトリック回帰モデルについて述べ，説明変数が単数の場合，及び複数の場合について説明する．次に，B-スプラインによるノンパラメトリック回帰モデルについて述べる．

3.1.1　ノンパラメトリック回帰

単変量の場合

説明変数 $X(\in \Re)$ と目的変数 $Y(\in \Re)$ に関して大きさ n のデータ $\{(x_\alpha, y_\alpha); \alpha = 1, 2, \ldots, n\}$ が観測されたとする．ただし $\Re$ は実数の集合である．一般に回帰モデルは，各 x_α におけるデータ y_α の確率変動を表す成分と，その条件付き期待値 $E[Y_\alpha|x_\alpha] = \mu_\alpha (\alpha = 1, 2, \ldots, n)$ に対して仮定する系統的成分からなる．ここでは，データは

$$y_\alpha = \mu_\alpha + \varepsilon_\alpha, \quad \varepsilon_\alpha \sim N(0, \sigma^2), \quad \alpha = 1, 2, \ldots, n \tag{3.1}$$

に従って生成されたとする．ただし，$N(0, \sigma^2)$ は，平均 0，分散 σ^2 の正規分布を表す．

データ構造を説明変数の線形結合 $\mu_\alpha = \beta_0 + \beta_1 x_\alpha$ で表したものが，周知の正規線型モデルであり，また，十分滑らかな関数 $\omega(\cdot)$ を用いて $\mu_\alpha = \omega(x_\alpha)$ と仮定したものはノンパラメトリック回帰

モデルと呼ばれる．実用上，関数 $\omega(\cdot)$ としては，区分的多項式で与えられる自然 3 次スプラインを用いることが多い [16]．曲線 $\omega(x)$ の推定は，各データ x_α での値 $\omega(x_\alpha) = \omega_\alpha$ $(\alpha = 1, 2, \ldots, n)$ を推定する問題に帰着され，$\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)^T$ とパラメトライズする．

この他に B-スプラインによる研究 [20] も行われており，自然 3 次スプラインよりもパラメータ数が少なくてすむなどの利点がある．

多変量の場合

説明変数が複数ある場合を考える．つまり，p 個の変数からなる説明変数 $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ と目的変数 Y との関係をモデル化する．ただし，$\boldsymbol{X} \in \Re^p, Y \in \Re$ である．単変量の場合と同様に，$\boldsymbol{X}$ と Y に関して大きさ n のデータ $\{(\boldsymbol{x}_\alpha, y_\alpha); \alpha = 1, 2, \ldots, n\}$ が観測されたとする．ただし，$\boldsymbol{x}_\alpha = (x_{\alpha 1}, \ldots, x_{\alpha p})^T$ である．

このとき，式 (3.1) はノンパラメトリック回帰加法モデル [19] によって

$$\mu_\alpha = \sum_{k=1}^{p} m_k(x_{\alpha k}), \quad \alpha = 1, \ldots, n \tag{3.2}$$

と表される．ただし，$m_k (k = 1, \ldots, p)$ は $\Re$ から $\Re$ への滑らかな関数である．関数 m_k は

$$m_k(x_{\alpha k}) = \sum_{m=1}^{M_k} \gamma_{km} b_{km}(x_{\alpha k}), \quad \alpha = 1, \ldots, n;\ k = 1, \ldots, p \tag{3.3}$$

と分解され，$\boldsymbol{\gamma}_k = (\gamma_{k1}, \ldots, \gamma_{kM_k})^T$，$\boldsymbol{b}_k(x_{\alpha k}) = (b_{k1}(x_{\alpha k}), \ldots, b_{kM_k}(x_{\alpha k}))^T$ とすると式 (3.3) は

$$m_k(x_{\alpha k}) = \boldsymbol{\gamma}_k^T \boldsymbol{b}_k(x_{\alpha k}), \quad \alpha = 1, \ldots, n;\ k = 1, \ldots, p \tag{3.4}$$

となる．ただし，$\boldsymbol{b}_k(x_{\alpha k})$ は基底関数 (例えば Fourier 級数，多項式基底，回帰スプライン基底，B-スプライン基底，ウェーブレット基底など)，係数 $\boldsymbol{\gamma}_k$ は未知パラメータ，M_k は基底関数の数である．

式 (3.1)，(3.4) より，非線形回帰モデルは，期待値を式 (3.2)，誤差を正規分布とするモデルであるので，確率密度関数として式 (3.5) で表される．

$$f(y_\alpha | \boldsymbol{x}_\alpha; \boldsymbol{\gamma}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{\left(y_\alpha - \sum_{k=1}^{p} \boldsymbol{\gamma}_k^T \boldsymbol{b}_k(x_{\alpha k})\right)^2}{2\sigma^2} \right\} \tag{3.5}$$

ただし，$\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_p^T)^T$ はパラメータベクトルである．

3.1.2　B-スプラインによるノンパラメトリック回帰

非線形回帰モデルとして実用上多く用いられている自然 3 次スプラインには次のような欠点がある [47]．

1. モデルのパラメータ数が標本数を上回る

2. データ点 x_α が重複するとパラメータ推定の際に困難が生じる

これに対し，3 次の B-スプライン [9,12] を用いる研究も行われており，自然 3 次スプラインよりもパラメータ数が少なくてすむなどの利点がある [19]．ここでは B-スプラインを基底とするノンパラメトリック回帰モデルについて述べる．

B-スプライン

B-スプラインとは，基底関数の線形結合で滑らかな関数を表現する手法である．基底関数 B_j は節点と呼ばれる等間隔に配置された点 t_i において滑らかに連結した既知の区分的多項式で構成される．3 次 B-スプラインの例を図 3.1 に示す．例えば 1 番目の基底関数 B_1 は 5 つの節点 $t_1, \ldots, t_5$ において滑らかに連結した 4 つの 3 次多項式で構成されている．

基底関数とは，節点の幅が決まると一意に決定される関数であり，各節点の幅を h，$t_3 = 0$ としたとき，$B_1(x)$ は t_3 に関して対称であり，

$$B_1(x) = \begin{cases} \dfrac{1}{6h}\left\{(2-\dfrac{x}{h})^3 - 4(1-\dfrac{x}{h})^3\right\}, (t_3 = 0 < x < t_4 = h) \\ \dfrac{1}{6h}\left(2-\dfrac{x}{h}\right)^3, (t_4 = h < x < t_5 = 2h) \\ 0, (t_5 = 2h < x) \end{cases}$$

となる．他の基底関数は B_1 を幅 h ずつ平行移動すると得られる．

また基底関数は，データの点在する区間を等間隔に分割し，各小区間を 4 つの基底関数で覆うように構成する．つまり，基底関数の数を M_b とすると，節点の数 M_t，節点の値 t_i，幅 h は次のようになる．

$$M_t = M_b + 4, \tag{3.6}$$

$$t_4 = x_{min}, \tag{3.7}$$

$$t_{M+1} = x_{max}, \tag{3.8}$$

$$h = \frac{x_{max} - x_{min}}{M-3} \tag{3.9}$$

ただし，x_{max}，x_{min} はそれぞれデータの最大値，最小値である．

パラメータの推定

式 (3.5) より，これは複数の基底関数に基づく非線形関数

$$y_\alpha = \sum_{k=1}^{p} \boldsymbol{\gamma}_k^T \boldsymbol{b}_k(x_{\alpha k}) \tag{3.10}$$

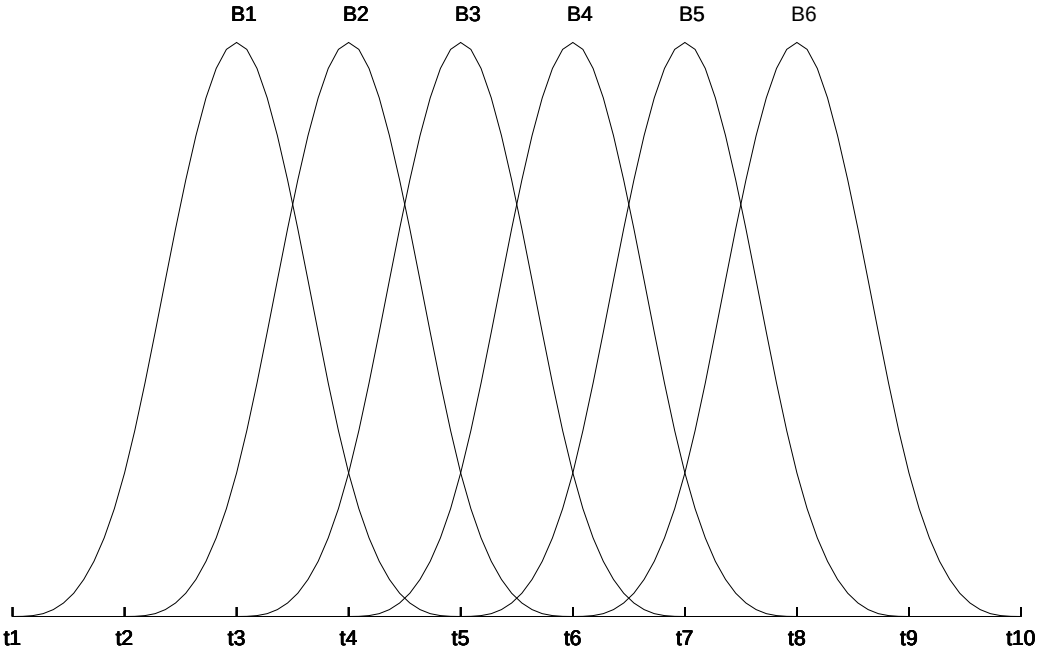


図 3.1: 3 次 B-スプラインの例

のデータへの当てはめと考えることができる．簡単のため $\boldsymbol{b}_k = \boldsymbol{b}_k(x_{\alpha k})$ と置くと，$\boldsymbol{b}_k$ は基底関数とデータから決定される既知のベクトルであるから，B-スプラインによる曲線推定問題は，$(m \times p)$ 次元の係数ベクトル $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_p^T)^T$ を推定することに帰着される．

$\boldsymbol{b}_k$ を用いて式 (3.5) を書き直すと

$$f(y_\alpha|\boldsymbol{x}_\alpha; \boldsymbol{\gamma}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{\left(y_\alpha - \sum_{k=1}^p \boldsymbol{\gamma}_k^T \boldsymbol{b}_k\right)^2}{2\sigma^2} \right\} \tag{3.11}$$

と表せる．ある程度多くの基底関数を用いたとき，モデルのパラメータ $\boldsymbol{\gamma}$ と σ^2 を最尤法によって推定すると，モデルの柔軟性の故にデータに強度に依存したモデルが推定される．そこで，パラメータは，対数尤度に曲線の局所変動の程度を考慮に入れた，次の罰則付き対数尤度関数

$$\begin{aligned} l_\lambda(\boldsymbol{\gamma}, \sigma^2) &= \sum_{\alpha=1}^n \log f(y_\alpha|\boldsymbol{x}_\alpha; \boldsymbol{\gamma}, \sigma^2) - \frac{n\lambda}{2} k(\boldsymbol{\gamma}) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{\alpha=1}^n \left(y_\alpha - \sum_{k=1}^p \boldsymbol{\gamma}_k^T \boldsymbol{b}_k \right)^2 - \frac{n\lambda}{2} k(\boldsymbol{\gamma}) \end{aligned} \tag{3.12}$$

の最大化に基づいて推定する．ここで，$\lambda(> 0)$ は平滑化パラメータと呼ばれ，推定曲線の局所変動の程度を制御するハイパーパラメータである．$k(\boldsymbol{\gamma})$ は回帰曲線の変動に対する罰則項であり，

B-スプライン非線形回帰モデルにおいては次の 2 次形式で与える．

$$\begin{aligned} k(\boldsymbol{\gamma}) &= \sum_{k=1}^{p} k(\boldsymbol{\gamma}_k) \\ &= \sum_{k=1}^{p} \sum_{j=d+1}^{M_k} (\Delta^d \gamma_{kj})^2 \\ &= \sum_{k=1}^{p} \boldsymbol{\gamma}_k^T D_{kd}^T D_{kd} \boldsymbol{\gamma}_k \end{aligned} \tag{3.13}$$

ただし，D_{kd} は，d 階差分を与える $(M_k - d) \times M_k$ 次の行列

$$D_{kd} = \begin{pmatrix} (-1)^0{}_dC_0 & \cdots & (-1)^d{}_dC_d & 0 & \cdots & 0 \\ 0 & (-1)^0{}_dC_0 & \cdots & (-1)^d{}_dC_d & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & (-1)^0{}_dC_0 & \cdots & (-1)^d{}_dC_d \end{pmatrix} \tag{3.14}$$

である．ただし，${}_dC_i$ は二項係数である．

3 次 B-スプライン基底関数を用いたときの 2 階差分の罰則 $\boldsymbol{\gamma}_k^T D_{k2}^T D_{k2} \boldsymbol{\gamma}_k$ に対して

$$\begin{aligned} \sum_{j=3}^{M_k} (\gamma_{k(j-2)} - 2\gamma_{k(j-1)} + \gamma_{kj}) &= \boldsymbol{\gamma}_k^T D_{k2}^T D_{k2} \boldsymbol{\gamma}_k \\ \approx \int \left\{ \frac{d^2 m_k(t)}{dt^2} \right\}^2 dt &= \int \left\{ \sum_{j=1}^{M_k} \gamma_{kj} \frac{d^2 b_{kj}(t)}{dt^2} \right\}^2 dt \end{aligned} \tag{3.15}$$

の近似が成り立つことが示されている [12].

式 (3.12) の罰則付き対数尤度関数の最大化に基づく推定値 $\hat{\boldsymbol{\gamma}}$，$\hat{\sigma}^2$ を式 (3.11) の確率密度関数 $f(y_\alpha | \boldsymbol{x}_\alpha; \boldsymbol{\gamma}, \sigma^2)$ に代入した

$$f(y_\alpha | \boldsymbol{x}_\alpha; \hat{\boldsymbol{\gamma}}, \hat{\sigma}^2) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp \left\{ - \frac{\left(y_\alpha - \sum_{k=1}^{p} \hat{\boldsymbol{\gamma}}_k^T \boldsymbol{b}_k \right)^2}{2\hat{\sigma}^2} \right\} \tag{3.16}$$

が一つの統計モデルである．

ここで，モデルのパラメータ $\boldsymbol{\gamma}$，σ^2 の推定法について述べる．B_k を $B_k = (\boldsymbol{b}_k(x_{1k}), \ldots, \boldsymbol{b}_k(x_{nk}))^T$ の $n \times M_k$ 行列とすると，パラメータの推定値 (モード) は，図 3.2 のようなバックフィッティングアルゴリズムによって求めることができる．ただし，$\beta_k = \sigma^2 \lambda_k (k = 1, 2, \ldots, p)$ はあらかじめ与えておくものとする．

推定値 $\hat{\boldsymbol{\gamma}}$，$\hat{\sigma}^2$ は，平滑化パラメータ λ_k と基底関数の個数 M_k に依存するため，適切な λ_k と M_k の選択を行う必要がある．

Algorithm *BackFitting*
begin
　for $k := 1$ to p **do** /*initialize*/
　　$\boldsymbol{\gamma}_{k(0)} := \mathbf{0}$;
　　$\sigma^2_{(0)} = \| \boldsymbol{y} \|^2 / n$
　end for
　$t := 0$;
　while $(\sigma^2_{(t)} - \sigma^2_{(t-1)})^2 > \delta$ **do**
　　for $k := 1$ to p **do** /*for each variable*/

$$\boldsymbol{\gamma}_{k(t+1)} := \left(B_k^T B_k + n\beta_k D_{kd}^T D_{kd}\right)^{-1} B_k^T \left(\boldsymbol{y} - \sum_{k' \neq k} B_{k'} \boldsymbol{\gamma}_{k'(t)}\right);$$

　　end for
　　$\sigma^2_{(t+1)} = \| \boldsymbol{y} - \sum_{k=1}^{p} B_k \boldsymbol{\gamma}_{k(t)} \|^2 / n$
　　$t := t + 1$;
　end while
　$\hat{\boldsymbol{\gamma}} := \boldsymbol{\gamma}_{(t)}$
　$\hat{\sigma}^2 := \sigma^2_{(t)}$
end

図 3.2: バックフィッティングアルゴリズム

3.2　ベイジアンネットワーク

確率変数をノードで表し，因果関係や相関関係といった依存する関係を持つ変数の間にリンクを張ったグラフ構造による確率モデルが確率ネットワーク (あるいはグラフィカルモデル) と呼ばれ，その中でとくにリンクが因果関係の方向に向きを持ち，このリンクをたどったパスが循環しない，非循環有向グラフで表されるモデルがベイジアンネットワークである [52].

ベイジアンネットワークの問題点として，循環構造を扱うことができないという点が挙げられる．この問題を克服するために，時系列データを扱うことで循環構造を表現可能な動的ベイジアンネットワーク [5, 15, 28, 36] が研究されている．

本節では，ベイジアンネットワーク，動的ベイジアンネットワークについて述べる．

3.2.1　ベイジアンネットワーク

ベイジアンネットワークは確率変数間の定性的な依存関係をグラフ構造によって表し，変数間の定量的な依存関係はその変数の間に定義される条件付き確率によって表すことで問題領域をモデル化する．二つの確率変数 X_i と X_j との間の条件付き依存性をベイジアンネットワークでは向

きのついたリンクによって $X_i \to X_J$ と表し，X_i を親ノード，X_j を子ノードと呼ぶ．親ノードが複数あるとき子ノード X_j の親ノードの集合を $\boldsymbol{P}_j = \{X_1, \cdots, X_i\}$ と書くことにする．この場合の変数に X_j 関する依存関係は条件付き確率，

$$P(X_j|\boldsymbol{P}_j) \tag{3.17}$$

で定義され，これは X_j を子ノード，$\boldsymbol{P}_j$ を親ノード群とする木構造になる．さらに n 個の確率変数 $X_1, \cdots, X_n$ があるとき，すべての確率変数の同時確率分布は式 (3.18) のようになり，各子ノードとその親ノード群からなる局所木を組み合わせたグラフ構造で表せる (表 3.3)．

$$P(X_1, \cdots, X_n) = \prod_j P(X_j|\boldsymbol{P}_j) \tag{3.18}$$

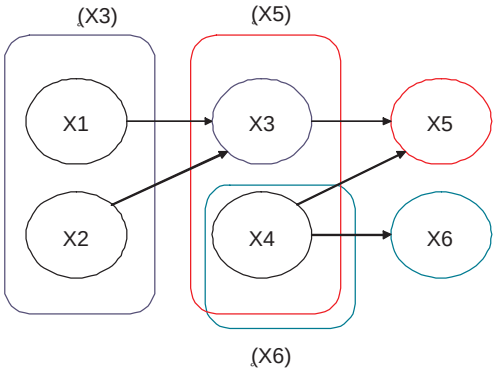


図 3.3: ベイジアンネットワークの例

つまり，式 (3.18) の左辺の同時確率分布は局所的な木構造に分割した右辺の各項の積として計算される．

このようなグラフ構造と，各ノードに割り当てた条件付き確率の集合によって，ベイジアンネットワークが構成される．変数間の依存関係は，因果律，状態遷移など，さまざまな相互作用に起因するが，それらを一括して条件付確率で表現することがベイジアンネットワークの大きな特徴である．

離散値の場合

離散変数の場合，子ノードの親ノードに関する条件付き確率はすべての状態における条件付き確率を並べた表，CPT(Conditional Probability Table) によって表す．例えば親ノードがある状態 $\boldsymbol{P}_j = \boldsymbol{y}$ ($\boldsymbol{y}$ は親ノード群の各値で構成したベクトル) のもとでの n 通りの離散状態を持つ変数 X_j の条件付き確率分布 $P(X_j|\boldsymbol{P}_j = \boldsymbol{y})$ を，

$$p(X_j = x_1|\boldsymbol{y}), \cdots, p(X_j = x_n|\boldsymbol{y}) \tag{3.19}$$

とする (ただし $\sum_{i=1}^{n} p(x_i|\boldsymbol{y}) = 1.0$).

これを行として，親ノードがとりうるすべての可能な状態 $\boldsymbol{P}_j = \boldsymbol{y}_1, \cdots, \boldsymbol{y}_m$ について列を構成した表 3.1 が X_j にとっての CPT，$P(X_j|\boldsymbol{P}_j)$ である．

表 3.1: X_j が持つ条件付確率表 (CPT)

$p(X_j = x_1\|\boldsymbol{P}_j = \boldsymbol{y}_1)$	$\cdots$	$p(X_j = x_n\|\boldsymbol{P}_j = \boldsymbol{y}_1)$
$\vdots$	$\ddots$	$\vdots$
$p(X_j = x_1\|\boldsymbol{P}_j = \boldsymbol{y}_m)$	$\cdots$	$p(X_j = x_n\|\boldsymbol{P}_j = \boldsymbol{y}_m)$

連続値の場合

$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$ を p 次元の確率変数，G を非循環有向グラフとし，ノード間にマルコフ性を仮定する．ベイジアンネットワークの枠組みにおいては同時確率を，

$$P(X_1, X_2, \ldots, X_p) = P(X_1|\boldsymbol{P}_1)P(X_2|\boldsymbol{P}_2) \times \cdots \times P(X_p|\boldsymbol{P}_p) \tag{3.20}$$

のように条件付き確率の積に分解する．ただし $\boldsymbol{P_j} = (P_1^{(j)}, P_2^{(j)}, \ldots, P_{q_j}^{(j)})^T$ はグラフ G における確率変数 X_j の親変数を表す q_j 次元ベクトルである．

p 個の確率変数からなる $\boldsymbol{X}$ の n 個の観測点を $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ とするとき，$\boldsymbol{P}_j$ の観測点を $\boldsymbol{p}_{1j}, \ldots, \boldsymbol{p}_{nj}$ と表記する．ただし，$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ であり，$\boldsymbol{p}_{ij}$ は $p_{ik}^{(j)}, k = 1, \ldots, q_j$ を k 番目の要素とする q_j 次元ベクトルである．

連続値を扱うために確率測度の代わりに確率密度を用いると式 (3.20) は，

$$f(x_{i1}, x_{i2}, \ldots, x_{ip}) = f_1(x_{i1}|\boldsymbol{p}_{i1})f_2(x_{i2}|\boldsymbol{p}_{i2}) \times \cdots \times f_p(x_{ip}|\boldsymbol{p}_{ip}) \tag{3.21}$$

となる．連続値を扱う場合，条件付き確率密度関数 $f_j(x_{ij}|\boldsymbol{p}_{ij})$ $(j = 1, \ldots, p)$ をどのように構築するかが重要な問題となる．

ベイジアンネットワークの学習

与えられたデータからそれに適するベイジアンネットワークを学習する研究は，いくつか報告されている．多くの学習手法の概略は同様なものであり，与えられたデータセットを D とすると，その手順は以下のようになる．

1. あるグラフ構造 G を仮定.
2. G おいて D よりモデルのパラメータを計算し，ベイジアンネットワークの解候補 B を得る.
3. 評価基準によって B を評価.
4. 探索アルゴリズムに新たなグラフ構造 G を生成し，2 へ.

つまり，ベイジアンネットワークの学習は次の三つの要素で構成されていると言える．

- ベイジアンネットワークの解候補 B を構築するためのモデル
- ベイジアンネットワークの解候補 B を評価するための評価基準
- 新たなグラフ構造 G を生成するための探索アルゴリズム

これらについては別章で詳しく説明する．

3.2.2 動的ベイジアンネットワーク

通常のベイジアンネットワークでは循環構造を表現することができない．この問題を克服するために，時系列データによって循環構造を表現できるように拡張したものが動的ベイジアンネットワークである．動的ベイジアンネットワークでは，通常の状態空間モデルの変数 (入力変数，隠れ変数，出力変数) に加えて，時間を表す変数 t を考慮することで，離散時間における確率過程が表現できる．

動的ベイジアンネットワークでは，連続系列データを取り扱うために，最も単純な因果モデルとして，一次マルコフ性を仮定する．つまり，各変数は直前の変数のみから直接に影響を及ぼされる．

時間 t における変数の集合を Z_t で表すと，ダイナミックベイジアンネットワークは，$P(Z_1)$ を与えるベイジアンネットワーク B_1 と，$P(Z_t|Z_{t-1})$ を与えるベイジアンネットワーク $B_{\rightarrow}$ のペア $(B_1, B_{\rightarrow})$ で一般に表される．

時系列データとして，$n \times p$ のデータ行列 $\boldsymbol{X}$ が与えられているとする．ただし，n はデータ点の数，p は確率変数の数とする．

時系列データを扱う動的ベイジアンネットワークにおいて時間に関する 1 次マルコフ性を仮定すると，各変数の同時確率は条件付確率の積として式 (3.22) のように分解できる．

$$P(X_{11}, \ldots, X_{np}) = P(\boldsymbol{X}_1)P(\boldsymbol{X}_2|\boldsymbol{X}_1) \times \cdots \times P(\boldsymbol{X}_n|\boldsymbol{X}_{n-1}) \tag{3.22}$$

ただし，$\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ は時刻 i における値を表す p 次元ベクトルである．

ベイジアンネットワークでは依存関係がない変数同士は独立として扱うため，条件付確率 $P(\boldsymbol{X}_i|\boldsymbol{X}_{i-1})$ も同様にして，式 (3.23) のように分解できる．

$$P(\boldsymbol{X}_i|\boldsymbol{X}_{i-1}) = P(X_{i1}|\boldsymbol{P}_{i-1,1}) \times \cdots \times P(X_{ip}|\boldsymbol{P}_{i-1,p}) \tag{3.23}$$

ただし，$\boldsymbol{P}_{i,j}$ は第 j 変数の親変数の，時刻 i における値を表す q_j 次元ベクトルである．

ここで，確率測度の代わりに確率密度を用いて式 (3.22) と式 (3.23) を書き直すと，

$$
\begin{aligned}
f(x_{11},\ldots,x_{np}) &= f_1(\boldsymbol{x}_1)f_2(\boldsymbol{x}_2|\boldsymbol{x}_1)\times\cdots\times f_n(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}) && (3.24)\\
&= f_1(\boldsymbol{x}_1)\prod_{i=2}^{n} g_1(x_{i1}|\boldsymbol{p}_{i-1,1})\times\cdots\times g_p(x_{ip}|\boldsymbol{p}_{i-1,p}) && (3.25)\\
&= f_1(\boldsymbol{x}_1)\prod_{j=1}^{p}\left\{\prod_{i=2}^{n} g_j(x_{ij}|\boldsymbol{p}_{i-1,j})\right\} && (3.26)
\end{aligned}
$$

となる．だたし，$\boldsymbol{x}_i=(x_{i1},\ldots,x_{ip})^T$ は時刻 i における各確率変数の観測値，$\boldsymbol{p}_{i,j}=(p_{i,1}^{(j)},\ldots,p_{i,q_j}^{(j)})^T$ はその親変数の観測値であり，

$$
f_i(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}) = g_1(x_{i1}|\boldsymbol{p}_{i-1,1})\times\cdots\times g_p(x_{ip}|\boldsymbol{p}_{i-1,p}) \tag{3.27}
$$

である．

このように時間についての因果関係を記述することで，変数間の循環構造を表現することができる．例えば図 3.4 では，左図の動的ベイジアンネットワークは右図のような循環構造を表現している．ただし，左図の点線は自己ループを表しており，右図では自己ループを省略してある．

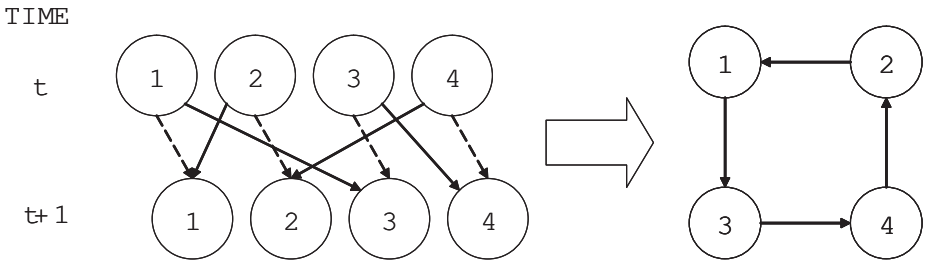


図 3.4: 動的ベイジアンネットワークによる循環構造の表現

3.3　ベイジアンネットワークの評価規準

あるグラフ構造を仮定すると，適当な方法によってベイジアンネットワークを構築・推定することができる．そして，対象とするシステムを最もよく近似するようなベイジアンネットワークを探すために，構築・推定されたベイジアンネットワークを評価する規準が必要となる．

本節では，ベイジアンネットワークの評価規準について述べる．まず研究例についてふれる．次に，ベイジアンネットワークとノンパラメトリック回帰モデルを評価するために提案されている BNRC(Bayesian Network and Nonparametric Regression Criterion) [19] について述べる．そして最後に，本研究で提案する動的微分ベイジアンネットワークとノンパラメトリック回帰モデルの評価規準，$\mathrm{BNRC}_{d-dynamic}$ の導出を行う．

3.3.1 研究例

ベイジアンネットワークの評価規準としてよく研究されているものに，Bayesian Dirichlet Metric(BDM) と，Bayesian Information Criteria(BIC) とがある．

BDM は対象問題に関する事前知識を考慮することができるという特徴を持つ．また，BIC はモデルが複雑になり過ぎない点が優れているが，BDM とは異なり，モデルに関する事前確率を考慮することができない．

Bayesian Dirichlet Metric

観測データ D が与えられたとき，ベイジアンネットワーク B がそのモデルである確率 $p(B|D)$ は，ベイズの定理により

$$p(B|D) = \frac{p(B)p(D|B)}{p(D)} \tag{3.28}$$

によって与えられる．この確率が大きいほど，モデル B が正しいと判断することができる．

ここで，$p(D)$ は定数とみなすことができるので BDM には影響を与えず，$p(B)$ によって対象問題に関する事前知識を考慮することができる．

BDM では，観測データに欠損値はないものとし，データの発生モデルを多項分布と仮定し，各変数に与えられる条件付き確率のパラメータはすべて独立であるとし，さらに，各変数に与えられるパラメータは Dirichlet 分布に従うものとする．

このとき，

$p(D|B) =$

$$\prod_{i=0}^{n-1} \prod_{\pi_i} \frac{\Gamma(m'(\pi_i))}{\Gamma(m'(\pi_i) + m(\pi_i)))} \prod_{x_i} \frac{\Gamma(m'(x_i, \pi_i) + m(x_i, \pi_i)))}{\Gamma(m'(x_i, \pi_i))} \tag{3.29}$$

となる．

ここで，$m(\pi_i)$ は，観測データ D の中で $\pi(X_i) = \pi_i$ となっているものの総数，同様に，$m(x_i, \pi_i)$ は，$X_i = x_i, \pi(X_i) = \pi_i$ となっている総数である．また，$m'(\pi_i)$ は，各変数の状態の事前頻度を表している．

Bayesian Information Criteria

BIC は MDL 原理 (最小記述長原理) に基づいており，モデルの記述長とデータの記述長の和 DL を評価規準とし，次のように表される．

$$DL = DL_{net} + DL_{data} \tag{3.30}$$

$$DL_{net} = DL_{graph} + DL_{param} \tag{3.31}$$

$$DL_{graph} = \sum_i \log n \cdot (1 + |\pi(X_i)|) \tag{3.32}$$

$$DL_{param} = \frac{\log N}{2} \sum_i |\pi(X_i)||X_i|(|X_i| - 1) \tag{3.33}$$

$$DL_{data} = -N \sum_i I(X_i; \pi(X_i)) \tag{3.34}$$

ただし，$I(X;Y)$ は相互情報量である．

BIC は，モデルが複雑になり過ぎない点が優れているが，BDM とは異なり，モデルに関する事前確率を考慮することができない．

3.3.2　BNRC

あるグラフ構造を仮定すると，適当な方法によって非線形ベイジアンネットワークモデルを構築・推定することができる．しかし，ここで問題となるのが，対象とするシステムを最もよく近似するようなグラフをいかにして見つけるかということである．尤度関数を用いると，モデルが複雑なほど尤度が大きくなるため，予測誤差，カルバック・ライブラー情報量，ベイズ法などの統計的アプローチが必要となる [1, 6, 24, 25].

ここでは，ベイジアンネットワークとノンパラメトリック回帰モデルにおけるグラフ評価規準をベイズ法にもとづいて導出した BNRC(Bayesian Network and Nonparametric Regression Criterion) [19] について述べる．

$\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$ を p 次元の確率変数, G を非循環有向グラフ, $\boldsymbol{P}_j = (P_1^{(j)}, P_2^{(j)}, \ldots, P_{q_j}^{(j)})^T$ をグラフ G における確率変数 X_j の親変数を表す q_j 次元ベクトルとし，ノード間にマルコフ性を仮定する．p 個の確率変数からなる $\boldsymbol{X}$ の n 個の観測点を $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}^T$ とするとき，$\boldsymbol{P}_j$ の観測点を $\boldsymbol{p}_{1j}, \ldots, \boldsymbol{p}_{nj}$ と表記する．ただし，$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ であり，$\boldsymbol{p}_{ij}$ は $p_{ik}^{(j)}$ $(k = 1, \ldots, q_j)$ を k 番目の要素とする q_j 次元ベクトルである．

変数間の一次マルコフ性を仮定するというベイジアンネットワークの枠組み (詳細は第 3.2 節参照) により，モデルは次のように表せる．

$$f(\boldsymbol{x}_i; \boldsymbol{\theta}_G) = \prod_{j=1}^{p} f_j(x_{ij}|\boldsymbol{p}_{ij}; \boldsymbol{\theta}_j), \quad (i = 1, \ldots, n) \tag{3.35}$$

ただし，$\boldsymbol{\theta}_G = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_p^T)^T$ はグラフ G にふくまれるパラメータベクトルである．

このとき，データ $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ に対する G の周辺尤度は

$$\int \prod_{i=1}^{n} f(\boldsymbol{x}_i; \boldsymbol{\theta}_G) \pi(\boldsymbol{\theta}_G | \boldsymbol{\lambda}) d\boldsymbol{\theta}_G \tag{3.36}$$

と表せる．データに対するグラフの事後確率 $\pi(G|\boldsymbol{X})$ は，グラフの事前確率 $\pi(G)$ とデータに対する周辺尤度の積を正規化定数で割った値として得られ，

$$\pi(G|\boldsymbol{X}) = \frac{\pi(G) \int \prod_{i=1}^{n} f(\boldsymbol{x}_i; \boldsymbol{\theta}_G) \pi(\boldsymbol{\theta}_G | \boldsymbol{\lambda}) d\boldsymbol{\theta}_G}{\sum_G \left\{ \pi(G) \int \prod_{i=1}^{n} f(\boldsymbol{x}_i; \boldsymbol{\theta}_G) \pi(\boldsymbol{\theta}_G | \boldsymbol{\lambda}) d\boldsymbol{\theta}_G \right\}} \tag{3.37}$$

と書ける．ただし，$\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})$ は $\log \pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda}) = O(n)$ を満たすパラメータ θ_G の事前分布であり，$\boldsymbol{\lambda}$ はハイパーパラメータベクトルである．ベイズ法においては，$\pi(G|\boldsymbol{X})$ が最大となるようなグラフを最適とする．式 (3.37) の分母は定数となるので，

$$\pi(G|\boldsymbol{X}) \propto \pi(G) \int \prod_{i=1}^{n} f(\boldsymbol{x}_i; \boldsymbol{\theta}_G) \pi(\boldsymbol{\theta}_G | \boldsymbol{\lambda}) d\boldsymbol{\theta}_G \tag{3.38}$$

となり，これが最大となるようなグラフを選択すればよい．

グラフの事後確率にもとづく評価規準を構築する際に重要な問題となるのが，式 (3.38) における高次の積分である．この積分にラプラス近似 [17] を用いる方法が報告されている [19]．データの周辺尤度に対するラプラス近似は，

$$\begin{aligned} \int \prod_{i=1}^{n} f(\boldsymbol{x}_i; \boldsymbol{\theta}_G) \pi(\boldsymbol{\theta}_G | \boldsymbol{\lambda}) \, d\boldsymbol{\theta}_G &= \int \exp\left\{ n l_\lambda(\boldsymbol{\theta}_G | \boldsymbol{X}_n) \right\} d\boldsymbol{\theta}_G \\ &= \frac{(2\pi/n)^{r/2}}{\left| J_\lambda(\hat{\boldsymbol{\theta}}_G) \right|^{1/2}} \exp\left\{ n l_\lambda(\hat{\boldsymbol{\theta}}_G | \boldsymbol{X}_n) \right\} \left\{ 1 + O_p(n^{-1}) \right\} \end{aligned} \tag{3.39}$$

で与えられる．ただし，r は $\boldsymbol{\theta}_G$ の次元，$\hat{\boldsymbol{\theta}}_G$ は $l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X}_n)$ のモードであり，

$$l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X}_n) = \frac{1}{n} \sum_{i=1}^{n} \log f(\boldsymbol{x}_i; \boldsymbol{\theta}_G) + \frac{1}{n} \log \pi(\boldsymbol{\theta}_G | \boldsymbol{\lambda}) \tag{3.40}$$

$$J_\lambda(\boldsymbol{\theta}_G) = -\frac{\partial^2 \left\{ l_\lambda(\boldsymbol{\theta}_G | \boldsymbol{X}_n) \right\}}{\partial \boldsymbol{\theta}_G \partial \boldsymbol{\theta}_G^T} \tag{3.41}$$

である．

以上より，グラフ選択のための評価規準，BNRC を得る．

$$\begin{aligned} BNRC(G) &= -2 \log \left\{ \pi(G) \int \prod_{i=1}^{n} f(\boldsymbol{x}_i; \boldsymbol{\theta}_G) \pi(\boldsymbol{\theta}_G | \boldsymbol{\lambda}) d\boldsymbol{\theta}_G \right\} \\ &= -2 \log \pi(G) - r \log \left(\frac{2\pi}{n} \right) + \log \left| J_\lambda(\hat{\boldsymbol{\theta}}_G) \right| - 2 n l_\lambda(\hat{\boldsymbol{\theta}}_G | \boldsymbol{X}_n) \end{aligned} \tag{3.42}$$

BNRC を最小とするようなグラフを最適とする．

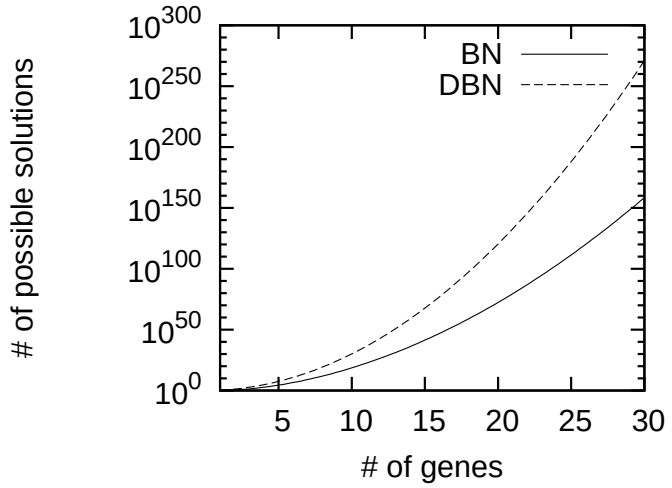


図 3.5: 遺伝子数と可能解数の関係

表 3.2: 代表的な n における $C_{BN}(n)$，$C_{DBN}(n)$ の値

	$n = 9$	$n = 20$	$n = 30$
$C_{BN}(n)$	1.21×10^{15}	2.34×10^{72}	2.71×10^{158}
$C_{DBN}(n)$	2.42×10^{24}	2.58×10^{120}	8.45×10^{270}

3.4　ネットワーク構造の探索

非循環有向グラフであるベイジアンネットワークの学習は NP 完全である [8]．ベイジアンネットワークの枠組みでは遺伝子は確率変数とみなすから，n 遺伝子からなる遺伝子制御ネットワークの推定問題は n 節点からなるグラフの最適化問題と見ることができる．n 節点から構築可能な非循環有向グラフの数 c_n は，

$$C_{BN}(n) = \frac{n! \cdot 2^{\frac{n}{2}(n-1)}}{r \cdot z^n}; \quad r \sim 0.57436; \quad z \sim 1.4881 \tag{3.43}$$

と概算される [32]．また，時系列データを用いる動的ベイジアンネットワークの場合，

$$C_{DBN}(n) = 2^{n^2} \tag{3.44}$$

と，さらに膨大な数となる．n に対する $C_{BN}(n)$，$C_{DBN}(n)$ のようすを図 3.5 に，$n = 9, 20, 30$ の場合でのこれらの値を表 3.2 に示す．ここからも分かるように，たとえ遺伝子数が少なくてもそれに対する可能解の数は爆発的に増大する．

このような広大な探索空間から最適なベイジアンネットワークを見つけ出すために，様々なヒューリスティクスが用いられてきた．

本節では探索空間を減少させる手法の一つである分割統治法と，ネットワーク構造学習手法の研究例について述べる．

3.4.1 分割統治法

問題を複数の部分に分解してそれぞれを個別に探索し，最後に統合して一つの解とする方法を分割統治法 (Divide and Conquer Approach) [3, 27, 38, 51] という．

局所構造に分解できるようなモデルを採用する場合，分割統治法を使うことができる．

動的ベイジアンネットワークの場合，分割統治法によって探索空間を 2^{n^2} から $n2^n$ に減らすことができる．この比較を図 3.6 に示す．また，動的ベイジアンネットワークに分割統治法を用いた場合の探索空間 $n2^n$ を図 3.7 に示す．

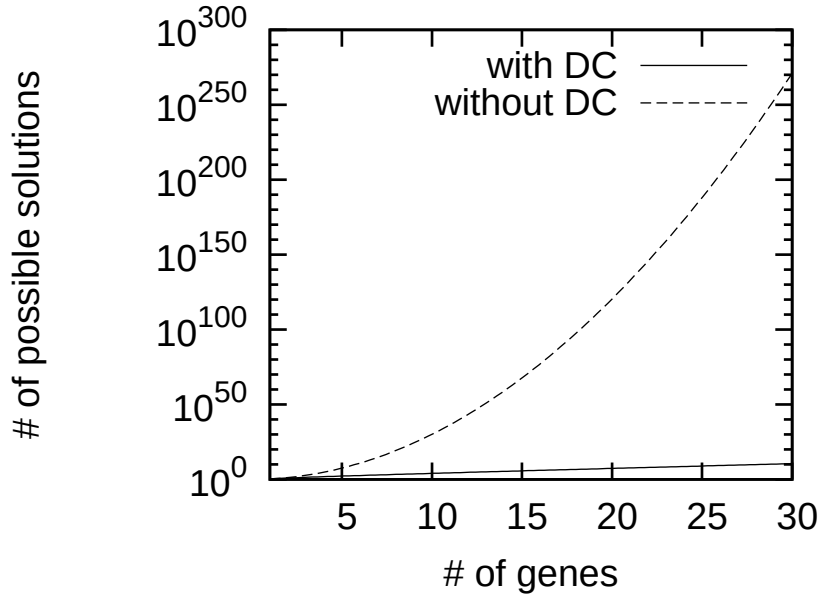


図 3.6: 分割統治法の有無による探索空間の比較

分割統治法で探索を行うことで，全体を一度に推定するよりも探索空間が小さいという利点がある．

3.4.2 研究例

欲張り山登り法 (GHC)

欲張り山登り法 (Greedy Hill Climbing:GHC) は山登り法の一種であり，評価を最も向上させるような更新を繰り返す．親の候補数をあらかじめ制限する GHC は K2 アルゴリズムとして知られている．GHC は次のような手順でグラフ構造の探索を行う．

1. 各エッジについて “追加”，“削除”，“反転” を行い，グラフを作る

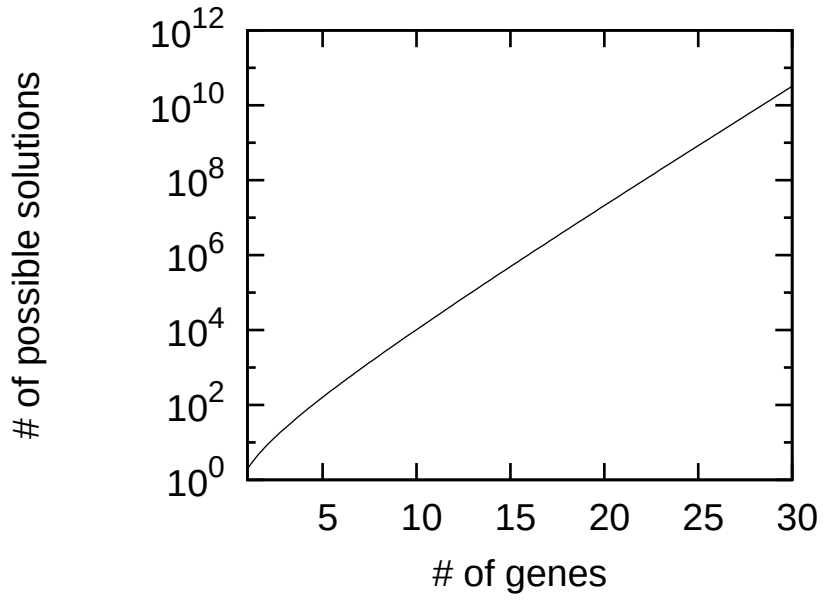


図 3.7: 分割統治法を用いた場合の動的ベイジアンネットワークの探索空間

2. そのグラフのもとでパラメータを決定し，評価する
3. 最も評価を向上させる更新を採用する
4. 評価が向上しなくなるまで繰り返す

GHC では効率的に評価を向上させることができるが，局所解に陥るという欠点もある．

ランダム山登り法 (RHC)

ランダム山登り法 (Random Hill Climbing:RHC) も山登り法の一種であるが，評価を最も向上させるような更新とは限らない．RHC は次のような手順でグラフ構造の探索を行う．

1. エッジをランダム選択し，“追加”，“削除”，“反転” のいずれかをランダムに行いグラフを作る
2. そのグラフのもとでパラメータを決定し，評価する
3. 評価を向上させるならその更新を採用する
4. 評価が向上しなくなるまで繰り返す

RHC では評価の収束が遅くなるが，試行によって異なる局所解に到達できるという利点もある．

ランダム探索

ランダム探索 (Random Search:RND) 各エッジを確率 r で生成し，そのネットワークの評価を行う．これを繰り返し，最も評価が良かったネットワークを採用する．

Random Search(RND) は確率的探索であるため局所解に陥ることはないが，候補解評価とその

Algorithm *LearningNetwork*
begin
  for $j := 1$ to p **do** /*for each variable*/
    $P_j := \emptyset$; $BestP_j := P_j$;
    $Score := Score(P_j)$;
    $BestScore := Score$;
    $T := T_0$; $M := M_0$; $Time := 0$;
    while $Time < MaxTime$ **do**
      Call *Metropolis*
      $T := decrease(T)$;
      $M := increase(M)$;
      $Time := Time + M$;
    end while
  end for
end

Algorithm *Metropolis*
begin
  for $i := 1$ to M **do**
    $NewP_j := Neighbor(P_j)$;
    $NewScore := Score(NewP_j)$;
    $\Delta Score := NewScore - Score$;
    if $\Delta Score < 0$ **then**
      $P_j := NewP_j$; $Score := NewScore$;
      if $Score < BestScore$ **then**
        $BestP_j := P_j$; $BestScore := Score$;
      end if
    else
      if $RANDOM < \exp(-\frac{\Delta Score}{T})$ **then**
        $P_j := NewP_j$; $Score := NewScore$;
      end if
    end if
  end for
end

図 3.8: 焼き鈍し法によるネットワーク構造学習手順の擬似コード

周辺を考慮しないため探索の効率が悪い.

焼き鈍し法 (SA)

焼き鈍し法 (Simulated Annearling:SA) [13, 18, 31, 43] は，物体の冷却過程にヒントを得た確率的探索手法である.

SA の特徴は，評価を向上させるような解は常に採用し，評価を悪くするような解も確率的に採用することである．図 3.8 に SA によるネットワーク推定の手順を示す．悪評価解を採用する確率を徐々に小さくしていくことで，一つの解に収束させる．SA では，温度の冷却，つまり確率の減少を十分ゆっくり行うことで大域解に達することが証明されている.

遺伝的アルゴリズム (GA)

遺伝的アルゴリズム (Genetic Algorithm:GA) [26, 45, 46] は生物の進化を模した確率的集団探索手法である．1 つの遺伝子はそれに対応する 1 つの解候補をコード化しており，これを個体と呼ぶ．この個体を評価し，その成績に応じて選択を行う.

さらに，交叉や突然変異という遺伝的操作を適用することで新たな個体を生成し，ネットワークを進化させる.

この過程をまとめると次のようになる.

1. 初期個体集団を生成する.

↓

2. 各個体の適合度を計算する (終了条件が満たされていれば終了).

↓

3. 適合度に基づき親を選択する.

↓

4. 交差・突然変異などを作用させ，次の世代を創出する.

↓

5. 2∼4 を繰り返す.

GA には，局所解に陥りにくい，並列計算と相性がよいなどの特徴がある.

局所探索付き遺伝的アルゴリズム (GA+LS)

局所探索付き遺伝的アルゴリズム (Genetic Algorithm with Local Search:GA+LS) は，遺伝的アルゴリズムと局所探索を組み合わせた手法である．各個体の評価時に周辺の解を探索し，評価が良ければそちらを採用する.

GA+LS では，GA の探索効率を向上させるという利点がある.

第4章

遺伝子制御ネットワークの推定手法

概観

本章では時系列遺伝子発現データから遺伝子制御ネットワークを推定する方法を述べる．まず動的ベイジアンネットワークを拡張し，動的微分ベイジアンネットワークの導出を行う．そして，動的微分ベイジアンネットワークとノンパラメトリック回帰モデルに基づいて本研究で用いるモデルを構築し，評価規準の導出を行う．さらに，本研究で用いたパラメータの設定について述べ，データの欠損値の取り扱いについてふれる．最後に遺伝子制御ネットワークの推定手順をまとめる．

4.1　動的微分ベイジアンネットワークの提案

本節では我々が提案する動的微分ベイジアンネットワークについて述べる．まず従来モデルの問題点を整理し，モデルに必要な条件を明らかにする．そして，その条件を満たすようなモデルの導出を行う．

4.1.1　問題点の整理

遺伝子の発現状態を計測するデバイスである DNA マイクロアレイの欠点として，大量の誤差が含まれている，得られる時間ステップが少ない，などが挙げられる．遺伝子制御ネットワークの推定を行うには，これらの欠点を克服する必要がある．

遺伝子制御ネットワークのような化学反応に基づく制御関係を記述する場合，化学反応に基づく制御関係をモデル化する場合，微分方程式系で扱うような変数の変化量が重要な要素となる．また，微分方程式系では循環構造も表現することができる．しかし，微分方程式系の推定には大量のデータが必要であり，マイクロアレイで得られる時系列の時間ステップ数はそれに比べて少ない．さらに，微分方程式系の推定は誤差に敏感であり，マイクロアレイデータとの相性が悪い．

通常のベイジアンネットワークは誤差を扱えるという点では有効であるが，循環構造を表現することができなかった．時系列データを用いることによってこの問題を解決したのが動的ベイジアンネットワークであり，循環構造を表現することができる．しかし，動的ベイジアンネットワークを用いても，時系列の変化量を扱うことはできない．

以上より，次の条件を満たすモデルが有効であると考えられる．

- 誤差を扱うことができる
- 循環構造を表現できる
- 推定に多くの時間ステップを必要としない
- 時系列の変化量を扱うことができる．

これらの条件を満たすモデルとして，本研究では動的微分ベイジアンネットワークを提案する．動的微分ベイジアンネットワークとは，変化量が扱えるように動的ベイジアンネットワークの拡張

したモデルである．

従来のモデルと動的微分ベイジアンネットワークモデルの比較を表 4.2 に示す．ただし，各モデルの名称はそれぞれ表 4.1 に示す通りである．

表 4.1: 各モデルの和名，英名，略称

和名	英名	略称
微分方程式系	Ordinary Differential Equations	ODEs
ベイジアンネットワーク	Bayesian Network	BN
動的ベイジアンネットワーク	Dynamic Bayesian Network	DBN
動的微分ベイジアンネットワーク	Dynamic Differential Bayesian Network	DDBN

表 4.2: 各モデルの比較

特徴	ODEs	BN	DBN	DDBN
誤差に強い	×	○	○	○
循環構造を表現可能	○	×	○	○
少ないデータで推定可能	×	○	○	○
変化量を扱える	○	×	×	○

4.1.2　動的微分ベイジアンネットワークへの拡張

本研究で提案する動的微分ベイジアンネットワークは，動的ベイジアンネットワークの拡張である．導出の準備として，まず動的ベイジアンネットワークモデルを示す．詳細は第 3.2.2 節にある．

時系列データとして，$n \times p$ のデータ行列 $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ が与えられているとする．ただし，n はデータ点の数，p は変数の数，$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ は時刻 i における各変数の値を表す p 次元ベクトルである．また，第 j 変数が q_j 個の親変数を持つとし，時刻 i における各親変数の値を表す q_j 次元ベクトルを $\boldsymbol{p}_{i,j} = (p_{i,1}^{(j)}, \ldots, p_{i,q_j}^{(j)})^T$ とする．

時間に関する 1 次マルコフ性を仮定し，依存関係がない変数同士は独立として扱うと，データの同時確率は

$$\begin{aligned} f(x_{11}, \ldots, x_{np}) &= f_1(\boldsymbol{x}_1) f_2(\boldsymbol{x}_2|\boldsymbol{x}_1) \times \cdots \times f_n(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}) \\ &= f_1(\boldsymbol{x}_1) \prod_{i=2}^{n} g_1(x_{i1}|\boldsymbol{p}_{i-1,1}) \times \cdots \times g_p(x_{ip}|\boldsymbol{p}_{i-1,p}) \\ &= f_1(\boldsymbol{x}_1) \prod_{j=1}^{p} \left\{ \prod_{i=2}^{n} g_j(x_{ij}|\boldsymbol{p}_{i-1,j}) \right\} \end{aligned} \tag{4.1}$$

のように条件付き確率の積として分解できる．ただし，

$$f_i(\boldsymbol{x}_i|\boldsymbol{x}_{i-1}) = g_1(x_{i1}|\boldsymbol{p}_{i-1,1}) \times \cdots \times g_p(x_{ip}|\boldsymbol{p}_{i-1,p}) \tag{4.2}$$

である．

式 (4.1) からも分かるように，動的ベイジアンネットワークでの関心の対象は $g_j(x_{ij}|\boldsymbol{p}_{i-1,j})$ である．つまり動的ベイジアンネットワークでは，時刻 $i-1$ と時刻 i における値の関係に注目し，これを条件付き確率密度関数によって表現している．

これに対して我々が提案する動的微分ベイジアンネットワークでは，時刻 $i-1$ における値とそこから時刻 i にいたるまでの変化分の関係に注目し，これを条件付き確率密度関数によって表現する．時刻 i から $i+1$ までの値の変化分を

$$\boldsymbol{d}_i = \boldsymbol{x}_{i+1} - \boldsymbol{x}_i \tag{4.3}$$

とし，$\boldsymbol{d}_i$ を第 i 行とする $(n-1) \times p$ のデータ行列を $\boldsymbol{D}$ とすると，

$$\begin{aligned}
f(\boldsymbol{X}, \boldsymbol{D}) &= f_1(\boldsymbol{x}_1) f_2(\boldsymbol{d}_1|\boldsymbol{x}_1) f_{2'}(\boldsymbol{x}_2|\boldsymbol{d}_1, \boldsymbol{x}_1) \times \cdots \times f_n(\boldsymbol{d}_{n-1}|\boldsymbol{x}_{n-1}) f_{n'}(\boldsymbol{x}_n|\boldsymbol{d}_{n-1}, \boldsymbol{x}_{n-1}) \\
&= f_1(\boldsymbol{x}_1) \prod_{i=2}^{n} \{ f_i(\boldsymbol{d}_{i-1}|\boldsymbol{x}_{i-1}) f_{i'}(\boldsymbol{x}_i|\boldsymbol{d}_{i-1}, \boldsymbol{x}_{i-1}) \} \\
&= f_1(\boldsymbol{x}_1) \prod_{i=1}^{n-1} \left\{ \prod_{j=1}^{p} g_j(d_{ij}|\boldsymbol{p}_{i,j}) \prod_{j=1}^{p} g_{j'}(x_{(i+1)j}|d_{ij}, x_{ij}) \right\} \\
&= f_1(\boldsymbol{x}_1) \prod_{j=1}^{p} \left[\prod_{i=1}^{n-1} \{ g_j(d_{ij}|\boldsymbol{p}_{i,j}) g_{j'}(x_{(i+1)j}|d_{ij}, x_{ij}) \} \right]
\end{aligned} \tag{4.4}$$

となる．ただし，

$$f_{i+1}(\boldsymbol{d}_i|\boldsymbol{x}_i) = \quad g_1(d_{i1}|\boldsymbol{p}_{i,1}) \times \cdots \times g_p(d_{ip}|\boldsymbol{p}_{i,p}), \tag{4.5}$$

$$f_{i+1'}(\boldsymbol{x}_{i+1}|\boldsymbol{d}_i, \boldsymbol{x}_i) = \quad g_{1'}(x_{(i+1)1}|d_{i1}, x_{i1}) \times \cdots \times g_{p'}(x_{(i+1)p}|d_{ip}, x_{ip}) \tag{4.6}$$

である．

式 (4.4) が，変数の変化量を扱うことができる動的微分ベイジアンネットワークの確率密度関数による表現である．

4.2 モデル

動的微分ベイジアンネットワークにおいて重要となることは，式 (4.4) における

$$f_1(\boldsymbol{x}_1), \quad g_j(d_{ij}|\boldsymbol{p}_{i,j}), \quad g_{j'}(x_{(i+1)j}|d_{ij}, x_{ij})$$

をどのようにモデル化するかということである．

$f_1(\boldsymbol{x}_1)$ はデータ初期値の分布を，$g_j(d_{ij}|\boldsymbol{p}_{i,j})$ は親変数と変化量との関係を，$g_{j'}(x_{(i+1)j}|d_{ij}, x_{ij})$ は時刻 i での値と変化量と，次の時刻 $i+1$ での値との関係を表している．本節ではこれらを表すモデルについて述べる．

4.2.1　データ初期値の分布

データ初期値の分布 $f_1(\boldsymbol{x}_1)$ をモデル化することを考える．本研究では，各変数の初期値は独立とし，それぞれを平均 μ_{1j}，分散 σ_{1j}^2 の正規分布でモデル化する．つまり，

$$f_1(\boldsymbol{x}_1; \boldsymbol{\theta}_f) = \prod_{j=1}^{p} f_{1j}(x_{1j}; \boldsymbol{\theta}_{fj}) \tag{4.7}$$

とする．ただし，

$$f_{1j}(x_{1j}; \boldsymbol{\theta}_{fj}) = \frac{1}{\sqrt{2\pi\sigma_{1j}^2}} \exp\left\{ -\frac{(x_{1j} - \mu_{1j})^2}{2\sigma_{1j}^2} \right\} \tag{4.8}$$

であり，$\boldsymbol{\theta}_{fj} = (\mu_{1j}, \sigma_{1j}^2)$ はパラメータベクトルである．

4.2.2　親変数と変化量との関係

親変数と変化量との関係 $g_j(d_{ij}|\boldsymbol{p}_{i,j})$ をモデル化することを考える．本研究では，d_{ij} と $\boldsymbol{p}_{i,j}$ との関係をモデル化するために B-スプラインによるノンパラメトリック回帰加法モデルを用いる．詳細は第 3.1.2 節にある．

モデルは，$\boldsymbol{p}_{i,j}$ における d_{ij} の確率変動を表す成分と，その条件付き期待値 $E[d_{ij}|\boldsymbol{p}_{i,j}] = \mu_{ij}$ に対して仮定する系統的成分からなる．ここでは，データは

$$d_{ij} = \mu_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2) \tag{4.9}$$

に従って生成されたとする．ただし，$(i = 1, 2, \ldots, n-1)$，$(j = 1, 2, \ldots, p)$ とし，ε_{ij} は互いに独立で平均 0，分散 σ_j^2 の正規分布 $N(0, \sigma_j^2)$ に従う．

μ_{ij} を，B-スプラインによるノンパラメトリック回帰加法モデルを用いて表すと，

$$\mu_{ij} = \sum_{k=1}^{q_j} m_k(p_{i,k}^{(j)}), \tag{4.10}$$

$$m_k(p_{i,k}^{(j)}) = \sum_{m=1}^{M_{jk}} \gamma_{mk}^{(j)} b_{mk}^{(j)}(p_{i,k}^{(j)}) \tag{4.11}$$

となり，$\boldsymbol{\gamma}_{jk} = (\gamma_{1k}^{(j)}, \ldots, \gamma_{M_{jk}k}^{(j)})^T$，$\boldsymbol{b}_{jk}(p_{i,k}^{(j)}) = (b_{1k}^{(j)}(p_{i,k}^{(j)}), \ldots, b_{M_{jk}k}^{(j)}(p_{i,k}^{(j)}))^T$ とすると式 (4.11) は

$$m_k(p_{i,k}^{(j)}) = \boldsymbol{\gamma}_{jk}^T \boldsymbol{b}_{jk}(p_{i,k}^{(j)}), \tag{4.12}$$

となる．ただし，$\boldsymbol{b}_{jk}(p_{i,k}^{(j)})$ は基底関数であり，係数 $\boldsymbol{\gamma}_{jk}$ は未知パラメータ，M_{jk} は基底関数の数である．本研究では基底関数として 3 次 B-スプラインを用いる．

式 (4.9)-(4.12) より，d_{ij} と $\boldsymbol{p}_{i,j}$ との関係は式 (4.13) のような条件付き確率密度関数として表される．

$$g_j(d_{ij}|\boldsymbol{p}_{i,j};\boldsymbol{\gamma}_j,\sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left\{-\frac{\left(d_{ij}-\mu(\boldsymbol{p}_{i,j})\right)^2}{2\sigma_j^2}\right\} \tag{4.13}$$

ただし，$\boldsymbol{\gamma}_j = (\boldsymbol{\gamma}_{j1}^T,\ldots,\boldsymbol{\gamma}_{jq_j}^T)^T$ はパラメータベクトルであり $\mu(\boldsymbol{p}_{i,j})$ は

$$\mu(\boldsymbol{p}_{i,j}) = \begin{cases} \displaystyle\sum_{k=1}^{q_j}\boldsymbol{\gamma}_{jk}^T\boldsymbol{b}_{jk}(p_{i,k}^{(j)}) & (\boldsymbol{p}_{i,j}\neq\emptyset) \\ \mu_j & (\boldsymbol{p}_{i,j}=\emptyset) \end{cases} \tag{4.14}$$

である．

4.2.3 値，変化量と次時刻での値との関係

時刻 i での値と変化量と，次の時刻 $i+1$ での値との関係 $g_{j'}(x_{(i+1)j}|d_{ij},x_{ij})$ をモデル化することを考える．ここでは (d_{ij},x_{ij}) における $x_{(i+1)j}$ の確率変動はないものとすると，モデルはその条件付き期待値 $E[x_{(i+1)j}|(d_{ij},x_{ij})] = \mu_{ij}$ のみからなる．ここでは，データは

$$x_{(i+1)j} = \mu_{ij} \tag{4.15}$$

に従って生成されたとする．式 (4.3) より，

$$\mu_{ij} = d_{ij} + x_{ij} \tag{4.16}$$

となるから，式 (4.15),(4.16) より $g_{j'}(x_{(i+1)j}|d_{ij},x_{ij})$ は

$$g_{j'}(x_{(i+1)j}|d_{ij},x_{ij}) = \begin{cases} 1 & (x_{(i+1)j} = d_{ij}+x_{ij}) \\ 0 & (x_{(i+1)j} \neq d_{ij}+x_{ij}) \end{cases} \tag{4.17}$$

となる．

本研究では $\boldsymbol{X}$ から $\boldsymbol{D}$ を求める際に確率変動を加えないため，式 (4.17) は結局，

$$g_{j'}(x_{(i+1)j}|d_{ij},x_{ij}) = 1 \tag{4.18}$$

となる．

4.2.4　動的微分ベイジアンネットワークとノンパラメトリック回帰モデル

式 (4.4)，(4.7)，(4.13)，(4.17) より，本研究で用いる “動的微分ベイジアンネットワークとノンパラメトリック回帰モデル” は確率密度関数として式 (4.19)-(4.22) で表される．

$$f(\boldsymbol{X},\boldsymbol{D};\boldsymbol{\theta}_G) = \prod_{j=1}^{p}\left\{ f_{1j}(x_{1j};\boldsymbol{\theta}_j)\prod_{i=1}^{n-1} g_j(d_{ij}|\boldsymbol{p}_{i,j};\boldsymbol{\theta}_j)\right\} \tag{4.19}$$

$$f_{1j}(x_{1j};\mu_{1j},\sigma_{1j}^2) = \frac{1}{\sqrt{2\pi\sigma_{1j}^2}}\exp\left\{-\frac{(x_{1j}-\mu_{1j})^2}{2\sigma_{1j}^2}\right\} \tag{4.20}$$

$$g_j(d_{ij}|\boldsymbol{p}_{i,j};\boldsymbol{\gamma}_j,\sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left\{-\frac{\big(d_{ij}-\mu(\boldsymbol{p}_{i,j})\big)^2}{2\sigma_j^2}\right\} \tag{4.21}$$

$$\mu(\boldsymbol{p}_{i,j}) = \begin{cases} \displaystyle\sum_{k=1}^{q_j}\boldsymbol{\gamma}_{jk}^T\boldsymbol{b}_{jk}(p_{i,k}^{(j)}) & (\boldsymbol{p}_{i,j}\neq\emptyset) \\ \mu_j & (\boldsymbol{p}_{i,j}=\emptyset) \end{cases} \tag{4.22}$$

ただし，$\boldsymbol{\theta}_G = (\boldsymbol{\theta}_1^T,\ldots,\boldsymbol{\theta}_p^T)^T$ はグラフ G にふくまれるパラメータベクトルであり，$\boldsymbol{\theta}_j$ はモデル f_1，モデル g_j にふくまれるパラメータベクトル，つまり，$\boldsymbol{\theta}_j = (\boldsymbol{\gamma_j^T},\sigma_j^2,\mu_{1j},\sigma_{1j}^2)^T$ または $\boldsymbol{\theta}_j = (\mu_j,\sigma_j^2,\mu_{1j},\sigma_{1j}^2)^T$ である．

4.3　ノンパラメトリック回帰モデルの設定

第 4.2.2 節で述べたように，本研究では滑らかな関数 $m_k(p_{i,k}^{(j)})$ の構築において基底関数を用いており，その基底関数として 3 次 B-スプライン [9] を用いている．本研究では [19, 21] と同様に，基底関数の数 M_{jk} を $M_{jk}=20$ とする．さらにこのとき，式 (3.6)-(3.9) より節点の数 $M_{jk,t}$，節点の値 $t_{jk,l}$，幅 h_{jk} は，

$$M_{jk,t} = M_{jk}+4 = 24, \tag{4.23}$$

$$t_{jk,4} = \min_i p_{i,k}^{(j)}, \tag{4.24}$$

$$t_{jk,M+1} = \max_i p_{i,k}^{(j)}, \tag{4.25}$$

$$h_{jk} = \frac{\max_i p_{i,k}^{(j)} - \min_i p_{i,k}^{(j)}}{M-3} \tag{4.26}$$

となる．

4.4　パラメータの事前確率分布

パラメータ $\boldsymbol{\theta}_G$ の事前確率分布においてパラメータベクトル $\boldsymbol{\theta}_j$ は互いに独立であるとすると，

$$\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda}) = \prod_{j=1}^{p}\pi_j(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j) \tag{4.27}$$

と分解でき，さらに，

$$\pi_j(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j) = \prod_{k=1}^{q_j} \pi_{jk}(\boldsymbol{\gamma}_{jk}|\lambda_{jk}) \tag{4.28}$$

と分解できるとする．ただし，λ_{jk} ハイパーパラメータである．$\boldsymbol{\gamma}_{jk}$ の事前分布としては非正則の M_{jk} 変量正規分布

$$\pi_{jk}(\boldsymbol{\gamma}_{jk}|\lambda_{jk}) = \left(\frac{2\pi}{n\lambda_{jk}}\right)^{(-M_{jk}-2)/2} |K_{jk}|_+^{1/2} \exp\left(-\frac{n\lambda_{jk}}{2}\boldsymbol{\gamma}_{jk}^T K_{jk}\boldsymbol{\gamma}_{jk}\right) \tag{4.29}$$

を用いる．ただし，K_{jk} は $M_{jk} \times M_{jk}$ の対称半正値行列であり，回帰曲線の 2 階差分に対する罰則

$$\boldsymbol{\gamma}_{jk}^T K_{jk}\boldsymbol{\gamma}_{jk} = \sum_{\alpha=3}^{M_{jk}} \left(\gamma_{\alpha,k}^{(j)} - 2\gamma_{\alpha-1,k}^{(j)} + \gamma_{\alpha-2,k}^{(j)}\right)^2 \tag{4.30}$$

を満たす．$|K_{jk}|_+$ は $(M_{jk}-2)$ 個の非負固有値の積である．

また，グラフ G の事前確率分布 $\pi(G)$ は

$$\begin{aligned}
\pi(G) &= \exp\{-(\text{ハイパーパラメータの数})\} && (4.31)\\
&= \prod_{j=1}^{p} \exp\{-(q_j+3)\} && (4.32)\\
&= \prod_{j=1}^{p} \pi_{L_j}(G) && (4.33)
\end{aligned}$$

を用いる．ただし，π_{L_j} は第 j 変数に関する局所構造の事前確率分布である．

4.5 評価規準

ベイジアンネットワークとノンパラメトリック回帰モデルによる推定 [19] では，第 3.3.2 節で述べた BNRC を用いている．本節では，本研究で用いる評価規準 $\text{BNRC}_{d-dynamic}$ を BNRC と同様な手順で導出する．

4.5.1 導出

式 (4.19)-(4.22) をモデルとするとき，データ $(\boldsymbol{X}, \boldsymbol{D})$ に対するグラフ G の周辺尤度は

$$\int f(\boldsymbol{X}, \boldsymbol{D}; \boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})d\boldsymbol{\theta}_G \tag{4.34}$$

と書ける．データに対するグラフの事後確率 $\pi(G|\boldsymbol{X}, \boldsymbol{D})$ は，グラフの事前確率 $\pi(G)$ とデータ $(\boldsymbol{X}, \boldsymbol{D})$ に対する周辺尤度の積を正規化定数で割った値として得られるから，

$$\pi(G|\boldsymbol{X}, \boldsymbol{D}) = \frac{\pi(G)\displaystyle\int f(\boldsymbol{X}, \boldsymbol{D}; \boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})d\boldsymbol{\theta}_G}{\displaystyle\sum_G \left\{\pi(G)\int f(\boldsymbol{X}, \boldsymbol{D}; \boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})d\boldsymbol{\theta}_G\right\}} \tag{4.35}$$

となる．これが最大となるようなグラフ G を最適とする．

式 (4.35) の分母はグラフ G に依存しないため定数となる．したがって，

$$\pi(G|\boldsymbol{X},\boldsymbol{D}) \propto \pi(G)\int f(\boldsymbol{X},\boldsymbol{D};\boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})d\boldsymbol{\theta}_G \tag{4.36}$$

となり，この右辺を考えればよい．式 (4.36) における高次の積分にラプラス近似を用い，評価規準 $\mathrm{BNRC}_{d-dynamic}$ を式 (4.37) によって与える．

$$\begin{aligned}\mathrm{BNRC}_{d-dynamic}(G) &= -2\log\left\{\pi(G)\int f(\boldsymbol{X},\boldsymbol{D};\boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda})d\boldsymbol{\theta}_G\right\}\\ &\approx -2\log\pi(G) - r\log\left(\frac{2\pi}{n-1}\right)\\ &\quad +\log\left|J_\lambda(\hat{\boldsymbol{\theta}}_G)\right| - 2(n-1)l_\lambda(\hat{\boldsymbol{\theta}}_G|\boldsymbol{X},\boldsymbol{D})\end{aligned} \tag{4.37}$$

で与えられる．ただし，

$$l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X},\boldsymbol{D}) = \frac{1}{n-1}\log f(\boldsymbol{X},\boldsymbol{D};\boldsymbol{\theta}_G) + \frac{1}{n-1}\log\pi(\boldsymbol{\theta}_G|\boldsymbol{\lambda}), \tag{4.38}$$

$$J_\lambda(\boldsymbol{\theta}_G) = -\frac{\partial^2\{l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X},\boldsymbol{D})\}}{\partial\boldsymbol{\theta}_G\partial\boldsymbol{\theta}_G^T}, \tag{4.39}$$

であり，r は $\boldsymbol{\theta}_G$，$\hat{\boldsymbol{\theta}}_G$ は $l_\lambda(\boldsymbol{\theta}_G|\boldsymbol{X}_n)$ のモードである．$\mathrm{BNRC}_{d-dynamic}$ を最小化するグラフ G を最適なグラフ G_{opt} として選択する．

第 4.4 節で行ったパラメータの分解によって，

$$l_\lambda(\hat{\boldsymbol{\theta}}_G|\boldsymbol{X},\boldsymbol{D}) = \sum_{j=1}^{p} l_{\lambda_j}^{(j)}(\hat{\boldsymbol{\theta}}_j|\boldsymbol{X},\boldsymbol{D}), \tag{4.40}$$

$$\log\left|J_\lambda(\hat{\boldsymbol{\theta}}_G)\right| = \sum_{j=1}^{p}\log\left|J_{\lambda_j}^{(j)}(\hat{\boldsymbol{\theta}}_j)\right| \tag{4.41}$$

となる．ただし，

$$\begin{aligned}l_{\lambda_j}^{(j)}(\boldsymbol{\theta}_j|\boldsymbol{X},\boldsymbol{D}) &= \frac{1}{n-1}\log f_{1j}(x_{1j};\boldsymbol{\theta}_j) + \frac{1}{n-1}\sum_{i=1}^{n-1}\log g_j(d_{ij}|\boldsymbol{p}_{ij};\boldsymbol{\theta}_j)\\ &\quad +\frac{1}{n-1}\log\pi_j(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j),\end{aligned} \tag{4.42}$$

$$J_{\lambda_j}^{(j)}(\boldsymbol{\theta}_j) = -\frac{\partial^2\{l_{\lambda_j}^{(j)}(\boldsymbol{\theta}_j|\boldsymbol{X},\boldsymbol{D})\}}{\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_j^T} \tag{4.43}$$

である．この分解によって，式 (4.44) のように $\mathrm{BNRC}_{d-dynamic}$ を第 j 変数についての局所評価規準に分解することができる．

$$\mathrm{BNRC}_{d-dynamic} = \sum_{j=1}^{p}\mathrm{BNRC}_{d-dynamic}^{(j)} \tag{4.44}$$

ただし，

$$\begin{aligned}\mathrm{BNRC}^{(j)}_{d-dynamic}(G) &= -2\log\left\{\pi_{L_j}(G)\int f_{1j}(x_{1j};\boldsymbol{\theta}_j)\prod_{i=1}^{n-1} g_j(d_{ij}|\boldsymbol{p}_{ij};\boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j|\boldsymbol{\lambda}_j)d\boldsymbol{\theta}_j\right\} \\ &\approx -2\log\pi_{L_j}(G) - r_j\log\left(\frac{2\pi}{n-1}\right) \\ &\quad +\log\left|J^{(j)}_{\lambda_j}(\hat{\boldsymbol{\theta}}_j)\right| - 2(n-1)l^{(j)}_{\lambda_j}(\hat{\boldsymbol{\theta}}_j|\boldsymbol{X},\boldsymbol{D})\end{aligned} \tag{4.45}$$

であり，r_j は $\boldsymbol{\theta}_j$ の次元である．

4.5.2　複数データセットの扱い

cDNA マイクロアレイで得られる時系列データのデータ点 n は通常，遺伝子の数よりもずっと少ない．そこでデータ点を増やすために，複数のデータセットを同時に扱うことを考える．

a 番目のデータセットに含まれているデータ点を n_a とすると，与えられるデータ行列 $\boldsymbol{X}^{(a)}$ の大きさは，$n_a \times p$ となる．また，発現量の変化分を与えるデータ行列 $\boldsymbol{D}^{(a)}$ の第 j 行を

$$\boldsymbol{D}^{(a)}_j = \boldsymbol{X}^{(a)}_{j+1} - \boldsymbol{X}^{(a)}_j,\ j = 1,\ldots,n_a - 1 \tag{4.46}$$

とする．

m 個のデータセットが与えられたとき，

$$\boldsymbol{X}' = (\boldsymbol{X}'^{(1)T},\ldots,\boldsymbol{X}'^{(m)T})^T, \tag{4.47}$$

$$\boldsymbol{D}' = (\boldsymbol{D}^{(1)T},\ldots,\boldsymbol{D}^{(m)T})^T \tag{4.48}$$

をデータ行列として用いる．ただし，$\boldsymbol{X}'^{(a)} = (\boldsymbol{X}^{(a)T}_1,\ldots,\boldsymbol{X}^{(a)T}_{n-1})^T$ である．$\boldsymbol{X}'$ と $\boldsymbol{D}'$ はともに $N \times p$ の行列となり，

$$N = \sum_{a=1}^{m}(n_a - 1) = \sum_{a=1}^{m} n_a - m \tag{4.49}$$

である．

これを用いると，

$$f(\boldsymbol{X}',\boldsymbol{D}';\boldsymbol{\theta}_G) = \prod_{a=1}^{m} f(\boldsymbol{X}^{(a)},\boldsymbol{D}^{(a)};\boldsymbol{\theta}_G) \tag{4.50}$$

となり，

$$\begin{aligned}\mathrm{BNRC}^{(j)}_{d-dynamic}(G) &\approx -2\log\pi_{L_j}(G) - r_j\log\left(\frac{2\pi}{N}\right) \\ &\quad +\log\left|J^{(j)}_{\lambda_j}(\hat{\boldsymbol{\theta}}_j)\right| - 2Nl^{(j)}_{\lambda_j}(\hat{\boldsymbol{\theta}}_j|\boldsymbol{X}',\boldsymbol{D}')\end{aligned} \tag{4.51}$$

となる．ただし，

$$
\begin{aligned}
l_{\lambda_j}^{(j)}(\boldsymbol{\theta}_G|\boldsymbol{X}',\boldsymbol{D}') &= \frac{1}{N}\sum_{a=1}^{m}\log f_{1j}(x_{1j}^{(a)};\boldsymbol{\theta}_j)+\frac{1}{N}\sum_{i=1}^{N}\log g_j(d_{ij}^{(a)}|\boldsymbol{p}_{ij}^{(a)};\boldsymbol{\theta}_j) \\
&\quad +\frac{1}{N}\log\pi_j(\boldsymbol{\theta}_j|\boldsymbol{\lambda_j}), \qquad (4.52)\\
J_{\lambda_j}^{(j)}(\boldsymbol{\theta}_G) &= -\frac{\partial^2\{l_{\lambda_j}^{(j)}(\boldsymbol{\theta}_j|\boldsymbol{X}',\boldsymbol{D}')\}}{\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_j^T} \qquad (4.53)
\end{aligned}
$$

である．

4.5.3　評価規準の計算

以上で複数のデータセットが与えられたときの動的微分ベイジアンネットワークの評価規準が導出された．ここではこの評価規準をさらに，ノンパラメトリック回帰モデルの基底関数に B-スプラインを用いた際の評価規準として具体化し，実際に計算できる形に式を変形していく．

まず，式 (4.43) の Hessian 行列式の対数に対して近似を適用して，

$$
\begin{aligned}
\log\left|-\frac{\partial^2\{l_{\lambda_j}^{(j)}(\boldsymbol{\theta}_j|\boldsymbol{X}',\boldsymbol{D}')\}}{\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_j^T}\right| &\approx \sum_{k=1}^{q_j}\log\left|-\frac{\partial^2\{l_{\lambda_j}^{(j)}(\boldsymbol{\theta}_j|\boldsymbol{X}',\boldsymbol{D}')\}}{\partial\boldsymbol{\gamma}_{jk}\partial\boldsymbol{\gamma}_{jk}^T}\right|+\log\left|-\frac{\partial^2\{l_{\lambda_j}^{(j)}(\boldsymbol{\theta}_j|\boldsymbol{X}',\boldsymbol{D}')\}}{\partial(\sigma_j^2)^2}\right| \\
&\quad +\log\left|-\frac{\partial^2\{l_{\lambda_j}^{(j)}(\boldsymbol{\theta}_j|\boldsymbol{X}',\boldsymbol{D}')\}}{\partial{\mu_{1j}}^2}\right|+\log\left|-\frac{\partial^2\{l_{\lambda_j}^{(j)}(\boldsymbol{\theta}_j|\boldsymbol{X}',\boldsymbol{D}')\}}{\partial(\sigma_{1j}^2)^2}\right| \qquad (4.54)
\end{aligned}
$$

とする．さらに B_{jk} を $B_{jk}=(\boldsymbol{b}_{jk}(p_{1k}^{(j)}),\ldots,\boldsymbol{b}_{jk}(p_{nk}^{(j)}))^T$ の $N\times M_{jk}$ 行列としてそれぞれの項を計算すると，

$$
\begin{aligned}
\log\left|-\frac{\partial^2\{l_{\lambda}^{(j)}(\hat{\boldsymbol{\theta}}_j|\boldsymbol{X}',\boldsymbol{D}')\}}{\partial\boldsymbol{\theta}_j\partial\boldsymbol{\theta}_j^T}\right| &\approx \sum_{k=1}^{q_j}\left\{\log\left|B_{jk}^TB_{jk}+N{\hat{\sigma_j}}^2\lambda_{jk}K_{jk}\right|-M_{jk}\log(N{\hat{\sigma_j}}^2)\right\} \\
&\quad +\left\{-\log(2{\hat{\sigma_j}}^4)\right\}+0+\left\{\log(m)-\log(2N\hat{\sigma}_{1j}^4)\right\} \qquad (4.55)
\end{aligned}
$$

となる．

式 (4.51)，(4.52)，(4.55) より，$\mathrm{BNRC}_{d-dynamic}^{(j)}(G)$ は次のようになる．

$$
\begin{aligned}
\mathrm{BNRC}_{d-dynamic}^{(j)}(G) &= 2\,(q_j+3)-\left(\sum_{k=1}^{q_j}M_{jk}+3\right)\log\left(\frac{2\pi}{N}\right) \\
&\quad +\sum_{k=1}^{q_j}\left\{\log\left|B_{jk}^TB_{jk}+N\hat{\sigma}_j^2\lambda_{jk}K_{jk}\right|-M_{jk}\log(N{\hat{\sigma_j}}^2)\right\} \\
&\quad -\log(2{\hat{\sigma_j}}^4)+\log(m)-\log(2N\hat{\sigma}_{1j}^4) \\
&\quad +\sum_{k=1}^{q_j}\left\{(M_{jk}+2)\log\left(\frac{2\pi}{N\lambda_{jk}}\right)-\log|K_{jk}|_+-\frac{N\lambda_{jk}}{2}\hat{\boldsymbol{\gamma}}_{jk}^TK_{jk}\hat{\boldsymbol{\gamma}}_{jk}\right\} \\
&\quad +N\log\left(2\pi{\hat{\sigma_j}}^2\right)+N+m\log\left(2\pi\hat{\sigma}_{1j}^2\right)+m \qquad (4.56)
\end{aligned}
$$

式 (4.56) を計算するためにはモード $\hat{\boldsymbol{\theta}}_j$ を求める必要がある．$(\hat{\gamma}_{jk}, {\hat{\sigma}_j}^2)$ は，$\beta_{jk} = \lambda_{jk}\sigma_j^2$ を与えると，図 3.2 に示したバックフィッティングアルゴリズムによって求めることができる．また，

$$\hat{\mu}_{1j} = \frac{1}{m}\sum_{a=1}^{m} x_{1j}^{(a)} \tag{4.57}$$

$$\hat{\sigma}_{1j}^2 = \frac{1}{m}\sum_{a=1}^{m}\left(x_{1j}^{(a)} - \hat{\mu}_{1j}\right)^2 \tag{4.58}$$

によって与える．

4.6 ハイパーパラメータ

${\hat{\sigma}_j}^2$ と $\hat{\gamma}_{jk}$ は β_{jk} の値に依存しており，β_{jk} を最適化することが重要な問題となる．本研究では，$\mathrm{BNRC}_{d-dynamic}^{(j)}(G)$ を最小にするような β_{jk} を選択する．

4.7 ネットワーク構造の探索

式 (4.44) から分かるように，評価規準は各変数で独立に計算できる．つまり，各変数についての局所構造を独立に推定し，最後に併合すればよい．ここでの局所構造とは，対象変数とその親変数からなる有向グラフ G_j である．問題を小さな部分に分割できることから，分割統治法 (詳細は第 3.4.1 節参照) を用いることができる．全体を一度に推定するよりも探索空間が小さいという利点がある．

ネットワーク構造の探索は NP 完全 [8] であり，探索アルゴリズムの工夫が必要である．本研究では第 3.4 章で述べた次の 6 つの手法による探索を試みた．

- 欲張り山登り法 (GHC)
- ランダム山登り法 (RHC)
- ランダム探索 (Random)
- 焼き鈍し法 (SA)
- 遺伝的アルゴリズム (GA)
- 局所探索付き遺伝的アルゴリズム (GA+LS)

なお，用いたパラメータは表 4.3 の通りである．

4.8 欠損値補完

動的微分ベイジアンネットワークの学習は，データに欠損値がないことを前提としている．しかし，実際のマイクロアレイデータには欠損値が存在し，そのままでは扱うことができない．そこで前処理として，欠損値を推定し，補完を行う．

表 4.3: ネットワーク探索アルゴリズムのパラメータ

GHC,RHC,SA 共通	最大評価回数			
	3,000			
GA 共通	エリート率	交叉率	突然変異率	トーナメントサイズ
	0.01	0.8(二点交叉)	0.1	5
GA	個体数		世代数	
	100		30	
GA+LOCAL	個体数	世代数	局所探索手法	局所探索の深さ
	50	10	GHC	5

欠損値推定については KNNimpute(weighted K-nearest neighbors)，SVDimpute(特異値分解:Singular Value Decomposition (SVD) basedmethod)，データの平均値の 3 つの手法を比較した研究も行われており，KNN アルゴリズムが有効であるという報告がなされている [40].

データに欠損値がある場合，本研究でも KNN を用いて欠損値の補完を行った.

4.9　遺伝子制御ネットワーク推定の手順

式 (4.56) を整理し，次のように書き直す.

$$\mathrm{BNRC}^{(j)}_{d-dynamic}(G_j) = C_j + \Gamma_j(\hat{\boldsymbol{\beta}}_j) \tag{4.59}$$

ただし，

$$\begin{aligned} C_j &= 2\log\left(\frac{N}{2}\right) + (2q_j + N + m - 3)\log(2\pi) + (2q_j + 6 + n + m) \\ &\quad + \log(m) - \sum_{k=1}^{q_j} |K_{jk}|_+ + (m-2)\log\left(\hat{\sigma}^2_{1j}\right), \\ \Gamma_j(\boldsymbol{\beta}_j) &= (2q_j - 2 + N)\log\left(\hat{\sigma}^2_j\right), \\ &\quad + \sum_{k=1}^{q_j}\left\{\log|\Lambda_{jk}| - (M_{jk}+2)\log(N\beta_{jk}) - \frac{N\beta_{jk}}{2\hat{\sigma}^2_j}\hat{\boldsymbol{\gamma}}^T_{jk}K_{jk}\hat{\boldsymbol{\gamma}}_{jk}\right\}, \\ \Lambda_{jk} &= B^T_{jk}B_{jk} + N\beta_{jk}K_{jk} \end{aligned} \tag{4.60}$$

であり，G_j は第 j 変数に対する局所構造，$\hat{\boldsymbol{\beta}}_j = (\boldsymbol{\beta}_{j1}, \ldots, \boldsymbol{\beta}_{jk})^T$ は $\Gamma_j(\boldsymbol{\beta}_j)$ を最小とする β_{jk} の組

$$\hat{\boldsymbol{\beta}}_j = \arg\min_{\boldsymbol{\beta}_j}\left\{\Gamma_j(\boldsymbol{\beta}_j)\right\} \tag{4.61}$$

である.

以上より，遺伝子制御ネットワークの推定手順 (図 4.1) を得る.

Algorithm $Minimeze\mathrm{BNRC}_{d-dynamic}$
begin
　begin for $j := 1$ to p **do** /* for each gene */
　　$G_j = G_j(0)$;
　　while $!convergence(\mathrm{BNRC}^{(j)}_{d-dynamic})$ **do** /* search best G_j */
　　　$\boldsymbol{\beta}_j := \boldsymbol{\beta}_j(0)$;
　　　while $!convergence(\Gamma_j)$ **do** /* optimize $\boldsymbol{\beta}_j$ */
　　　　$\hat{\boldsymbol{\theta}}_j := BackFitting(\boldsymbol{\beta}_j)$;
　　　　calculate $\Gamma_j(\boldsymbol{\beta}_j)$;
　　　　$\boldsymbol{\beta}_j := NextCandidate(\boldsymbol{\beta}_j)$;
　　　end while
　　　calculate C_j;
　　　$\mathrm{BNRC}^{(j)}_{d-dynamic}(G_j) := C_j + \Gamma_j$;
　　　$G_j := NextCandidate(G_j)$;
　　end while
　end for
end

図 4.1: 遺伝子制御ネットワークの推定手順

第5章

推定実験

概要

提案手法によって，遺伝子制御ネットワークの推定実験を行う．比較のため，提案手法である “動的微分ベイジアンネットワークとノンパラメトリック回帰モデル (Dynamic Differential Bayesian Network and Nonparametric Regression model:DDBN)” と，先行研究である “Dynamic Bayesian Network and Nonparametric Regression model:DBN” [21] の両手法によって推定実験を行った．提案手法，従来手法をそれぞれ DDBN，DBN と表記する．

実験には人工的なネットワークから生成したデータと出芽酵母の細胞周期データを用いた．実験は

- ネットワーク構造探索手法の比較
- 人工データの解析
- 実データの解析 1
- 実データの解析 2

の 4 つを行った．

5.1　実験に用いたデータ

5.1.1　人工データ

5 つの遺伝子からなる図 5.1 のような遺伝子制御ネットワークを考える．人工的にデータを生成するために，このネットワークを数理モデルで表現する．数理モデルとして式 (5.1) で示した S-system [39] を採用する．具体的にパラメータを与えた式 5.1 から人工的な時系列データを生成し，ターゲットデータとして実験を行った．なお，図 5.2 は図 5.1 を簡素化したものである．

一つの初期値 $\boldsymbol{x}(0) = (x_1(0), \ldots, x_5(0))$ に対して一つのデータセットが生成できる．実験では，3 つのデータセット用意した．

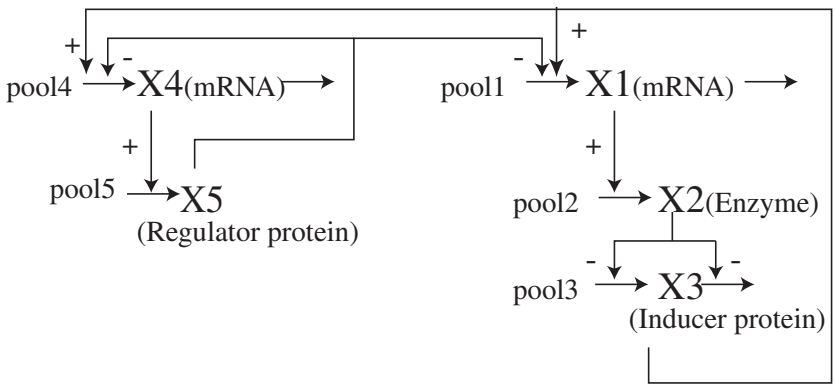


図 5.1: ターゲットとした遺伝子制御ネットワーク (人工データ)

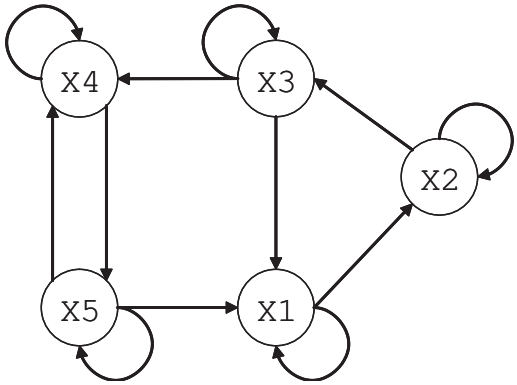


図 5.2: 簡素化したターゲットネットワーク (人工データ)

$$\begin{cases} \dot{X}_1 = & 15.0X_3X_5^{-0.1} - 10.0X_1^{2.0} \\ \dot{X}_2 = & 10.0X_1^{2.0} - 10.0X_2^{2.0} \\ \dot{X}_3 = & 10.0X_2^{-0.1} - 10.0X_2^{-0.1}X_3^{2.0} \\ \dot{X}_4 = & 8.0X_3^{2.0}X_5^{-1.0} - 10.0X_4^{2.0} \\ \dot{X}_5 = & 10.0X_4^{2.0} - 10.0X_5^{2.0} \end{cases} \tag{5.1}$$

5.1.2　実データ

実データとして，出芽酵母の細胞周期を計測したマイクロアレイデータ (*Saccharomyces cerevisiae* cell cycle gene expression data) を用いた．このデータは Spellman *et al.* [37] によって収集されたものであり，6,178 個の遺伝子についての 6 つの時系列データからなる．それぞれの実験名と，データ点数を表 5.1 に示す.

表 5.1: マイクロアレイの実験名と各実験のデータ点数

実験名	cln3	clb2	alpha	cdc15	cdc28	elu
データ点数	2	2	18	24	17	14

本研究では，これらの全ての遺伝子についてではなく KEGG(Kyoto Encyclopedia of Genes and Genomes) データベース[1]にある細胞周期パスウェイの 1 部に注目し，そのパスウェイを構成する遺伝子のみを扱った．KEGG データベースにある出芽酵母の細胞周期パスウェイを図 5.3 に，推定対象とする部分パスウェイ (2 カ所) を図 5.4，5.5 に示す.

[1] http://www.genome.ad.jp/kegg/

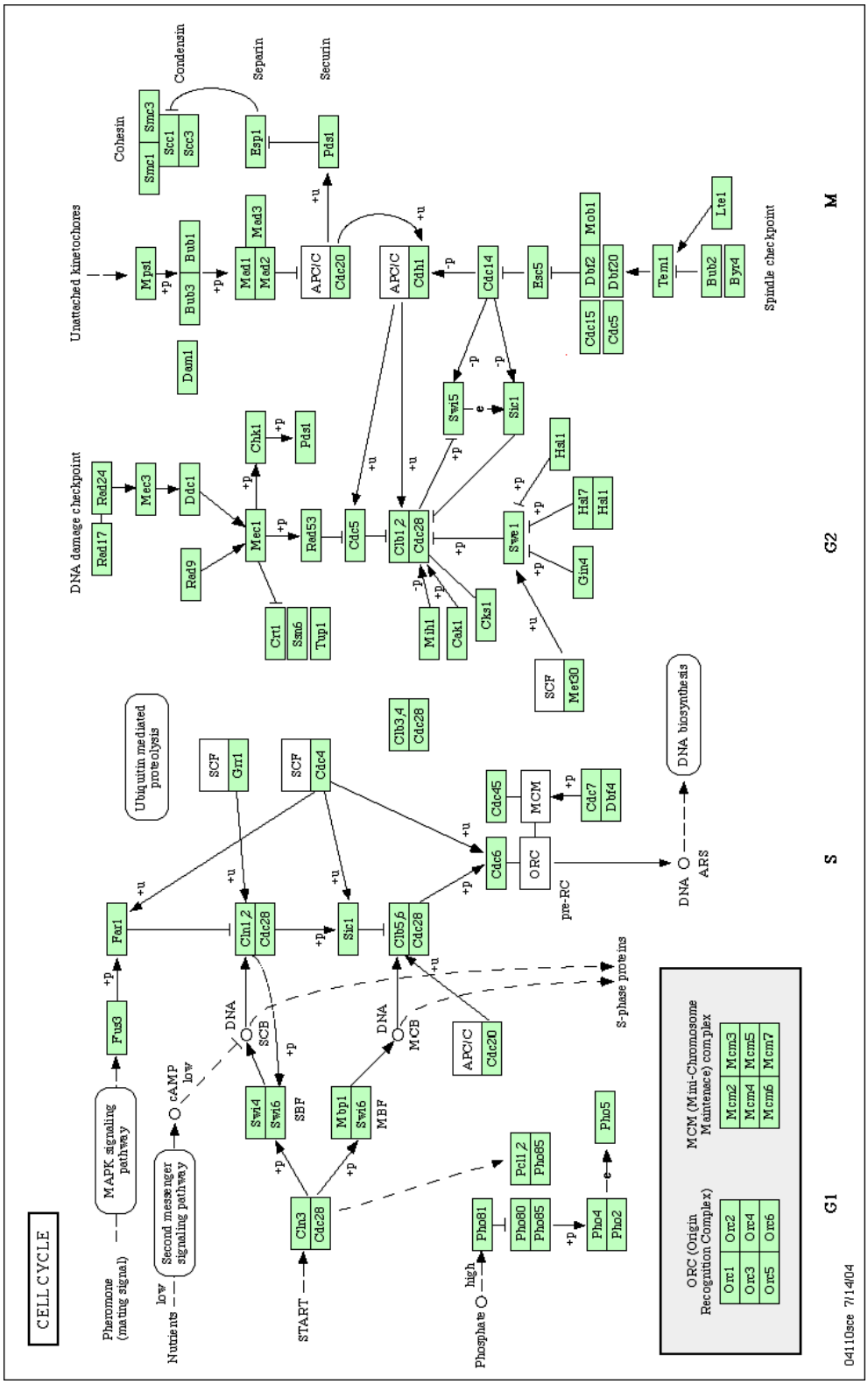


図 5.3: KEGG データベース中の出芽酵母の細胞周期パスウェイ

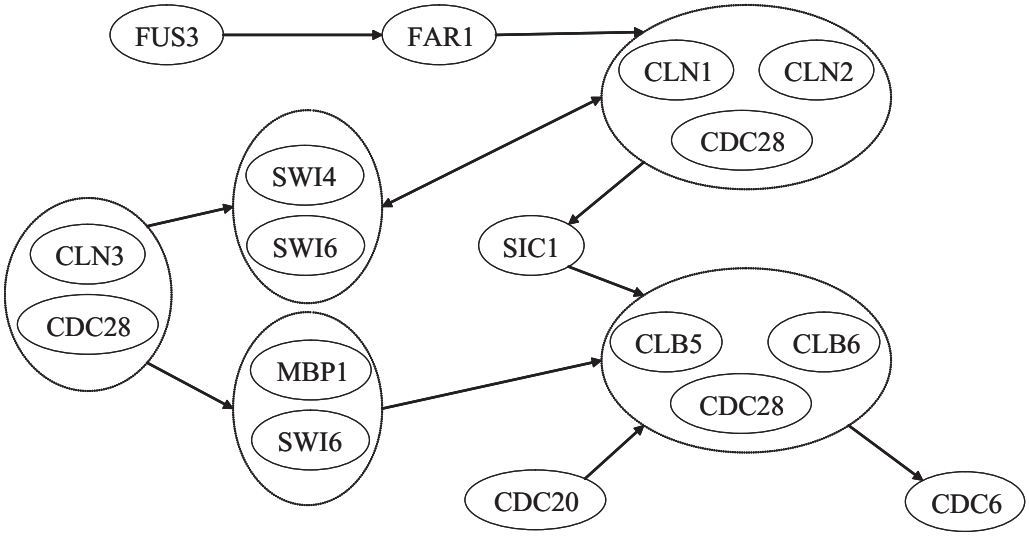


図 5.4: ターゲットネットワーク #1(実データ)

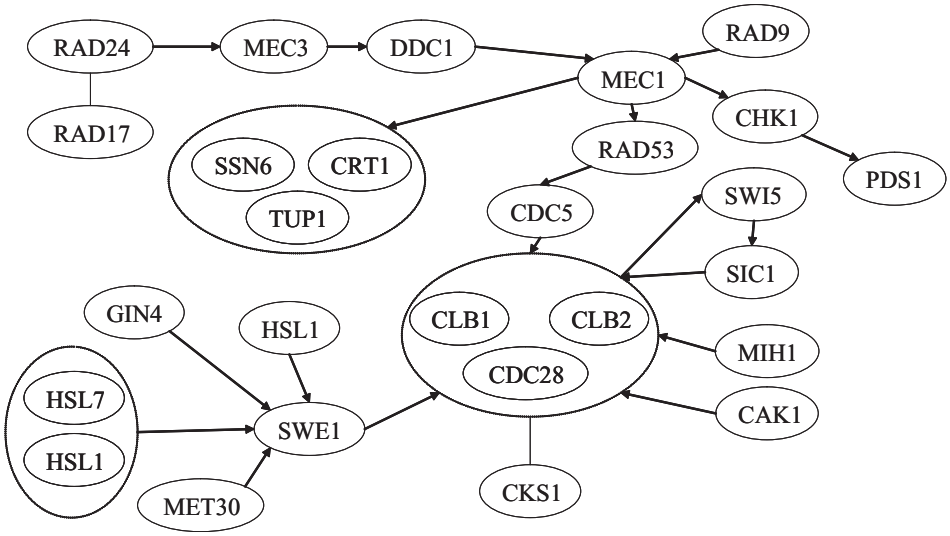


図 5.5: ターゲットネットワーク #2(実データ)

5.2　ネットワーク構造探索手法の比較

ネットワーク構造の探索手法として，次の 6 つの手法を比較した．これらの手法の詳細は第 3.4 節にある．

- 欲張り山登り法 (GHC)
- ランダム山登り法 (RHC)
- ランダム探索 (Random)
- 焼き鈍し法 (SA)
- 遺伝的アルゴリズム (GA)
- 局所探索付き遺伝的アルゴリズム (GA+LS)

エッジの初期生成率 r を $r = (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6)$ として，DDBN，DBN それぞれを用いてそれぞれ 20 回ずつの実験を行った．実際の生物データの特性に合った手法ほど評価が高いと考えられるため，推定対象とするデータは実データ #1 を用いた．

DDBN による推定結果を表 5.2，5.3，5.4 に示す．ただし，太字はスコアの最小値 (探索成功) を表す．

表 5.2: 分割統治法と DDBN によって得られた最良スコア

algorithm	$r = 0$	$r = 0.1$	$r = 0.2$	$r = 0.3$	$r = 0.4$	$r = 0.5$	$r = 0.6$
GHC	1476.35	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**
RHC	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**
RND	1951.81	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**	1507.46
SA	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**
GA	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**
GA+LS	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**

表 5.3: 分割統治法と DDBN によって得られた 20 回のスコアの平均値

algorithm	$r = 0$	$r = 0.1$	$r = 0.2$	$r = 0.3$	$r = 0.4$	$r = 0.5$	$r = 0.6$
GHC	1476.35	1480.76	1481.77	1482.59	1484.08	1483.24	1481.72
RHC	1488.84	1486.91	1484.88	1487.33	1486.74	1486.97	1486.58
RND	1951.81	1475.74	1473.13	1473.71	1486.69	1537.34	1619.78
SA	1471.61	1471.73	1471.86	1471.7	1471.58	1471.61	1471.63
GA	1471.47	**1471.42**	1471.47	**1471.42**	**1471.42**	**1471.42**	**1471.42**
GA+LS	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**

表 5.4: 分割統治法と DDBN によって得られた 20 回のスコアの最悪値

algorithm	$r = 0$	$r = 0.1$	$r = 0.2$	$r = 0.3$	$r = 0.4$	$r = 0.5$	$r = 0.6$
GHC	1476.35	1495.32	1491.93	1494.16	1506.5	1500.23	1490.2
RHC	1508.66	1503.36	1499.31	1500.03	1501.86	1500.97	1498.15
RND	1951.81	1480.4	1477.47	1478.22	1504.4	1560.14	1648.61
SA	1472.48	1473.52	1473.61	1474.35	1472.48	1472.98	1473.28
GA	1472.48	**1471.42**	1472.48	**1471.42**	**1471.42**	**1471.42**	**1471.42**
GA+LS	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**	**1471.42**

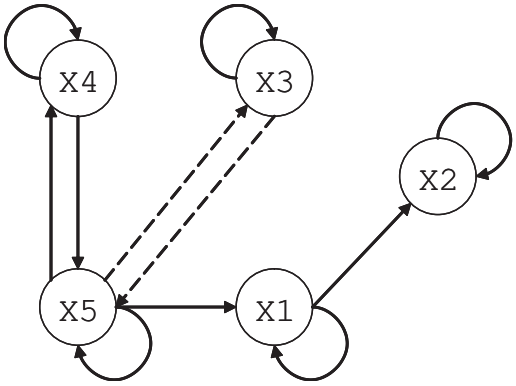


図 5.6: DDBN によって推定されたネットワーク (人工データ)

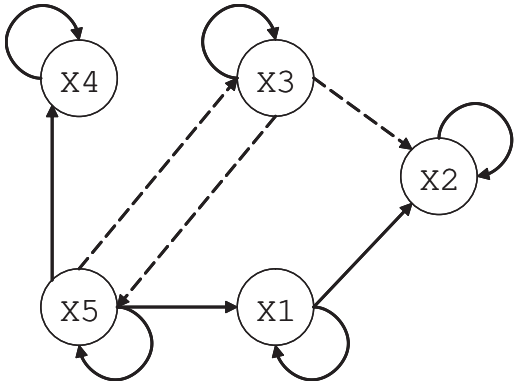


図 5.7: DBN によって推定されたネットワーク (人工データ)

GA+LS のみが全ての条件において正答率 100% を獲得した．これは DBN のときにも同様であった．GA+LS では $r = 0.2$ のときに平均として最も早く解の探索に成功した．

5.3　人工データの解析

DDBN，DBN によって推定されたネットワークをそれぞれ図 5.6，5.7 に示す．ターゲットと同じ制御関係は実線で，ターゲットにはない制御関係は点線で表した．

第 5.2 節の結果をふまえ，ネットワークの探索アルゴリズムには GA+LS($r = 0.2$) を用いた．

DDBN と DBN の性能評価のために，Sensitivity/Specificity という指標を用いる．ターゲットネットワークに対するモデルの Sensitivity(S_N) と Specificity(S_P) は次のように定義する．

$$S_N = \frac{\text{獲得エッジの正解数}}{\text{ターゲットのエッジ数}} = \frac{TP}{TP + FN}, \tag{5.2}$$

$$S_P = \frac{\text{獲得非エッジの正解数}}{\text{ターゲットの非エッジ数}} = \frac{TN}{TN + FP} \tag{5.3}$$

ただし，TP，TN，FP，FN はそれぞれ True Positive，True Negative，False Positive，False

Negative であり，各エッジに対して表 5.5 のように定義する．

表 5.5: TP，TN，FP，FN の定義

		ターゲット ○	ターゲット ×
獲得	○	TP	FP
	×	FN	TN

両手法によって獲得したネットワークに対する S_N/S_P を，表 5.6 に示す．S_N，S_P ともに，DDBN の方がよい性能を示した．

表 5.6: DBN と DDBN での sensitivity/specificity

	S_N	S_P
DDBN	**0.75**	**0.85**
DBN	0.67	0.77

5.4　実データの解析

5.4.1　#1

DDBN，DBN によって推定されたネットワークをそれぞれ図 5.8，5.9 に示す．また，両手法によって推定されたエッジ，DDBN のみによって推定されたエッジ，DBN のみによって推定されたエッジをそれぞれ図 5.10，5.11，5.12 に示す．さらに各エッジについての詳細をそれぞれ表 5.7，5.8，5.9 に示す．表中での類似度とは推定されたエッジを評価する値であり，推定されたエッジが KEGG データベースに登録されているパスウェイを通った数を表す．例えば，KEGG データベースに登録されているパスウェイと同じエッジを推定できた場合は 1, KEGG のパスウェイで一つの遺伝子をバイパスしたエッジを推定した場合は 2 となる．また，KEGG のパスウェイと逆方向のエッジを推定した場合は -1，それ以外を unknown とした．

第 5.2 節の結果をふまえ，ネットワークの探索アルゴリズムには GA+LS($r = 0.2$) を用いた．

これらの結果には自己ループは省略してある．DDBN ではすべての遺伝子に対して，DBN では CDC20，CLB6，CLN2，CLN3，FAR1，SWI4 に対して自己ループを検出した．

5.4.2　#2

DDBN，DBN によって推定されたネットワークをそれぞれ図 5.13，5.14 に示す．また，両手法によって推定されたエッジ，DDBN のみによって推定されたエッジ，DBN のみによって推定

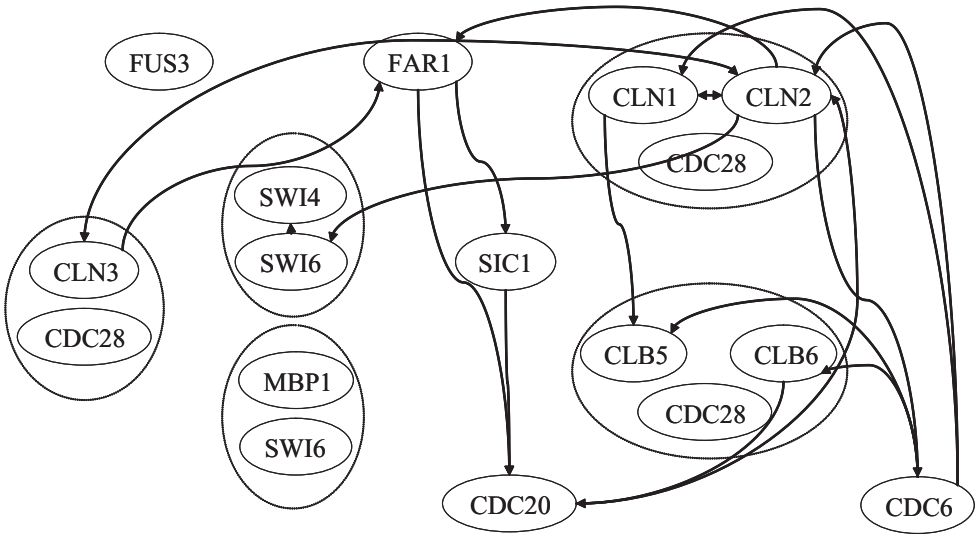


図 5.8: DDBN で推定されたネットワーク (実データ #1)

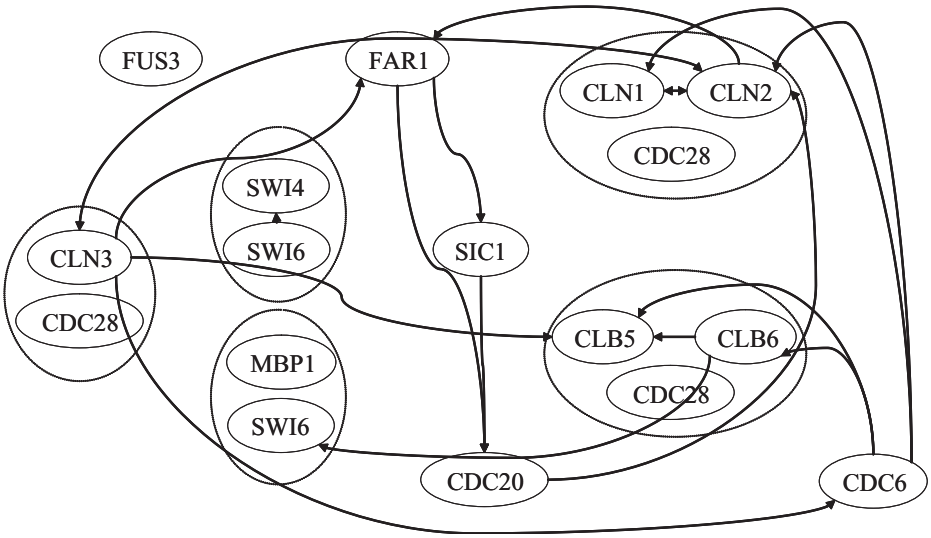


図 5.9: DBN によって推定されたネットワーク (実データ #1)

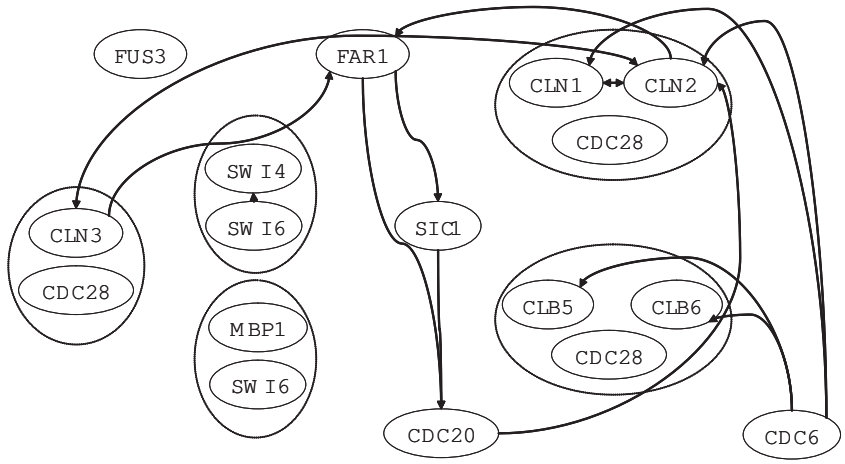


図 5.10: 両手法で推定されたエッジ (実データ #1)

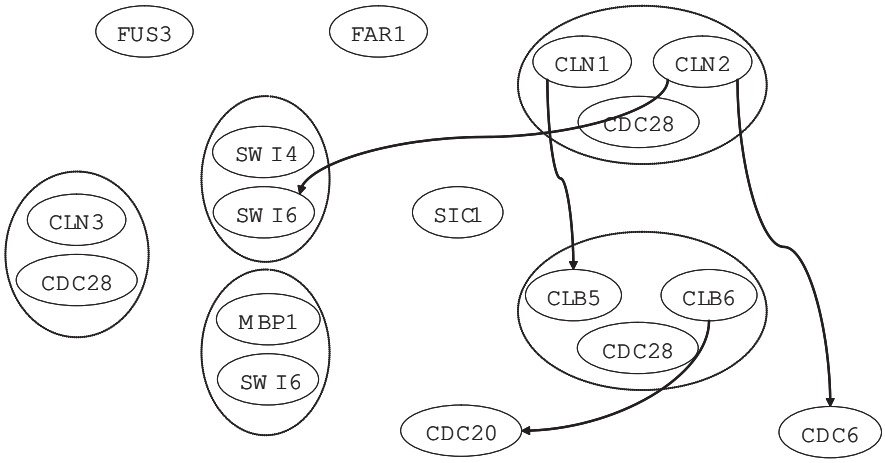


図 5.11: DDBN のみで推定エッジ (実データ #1)

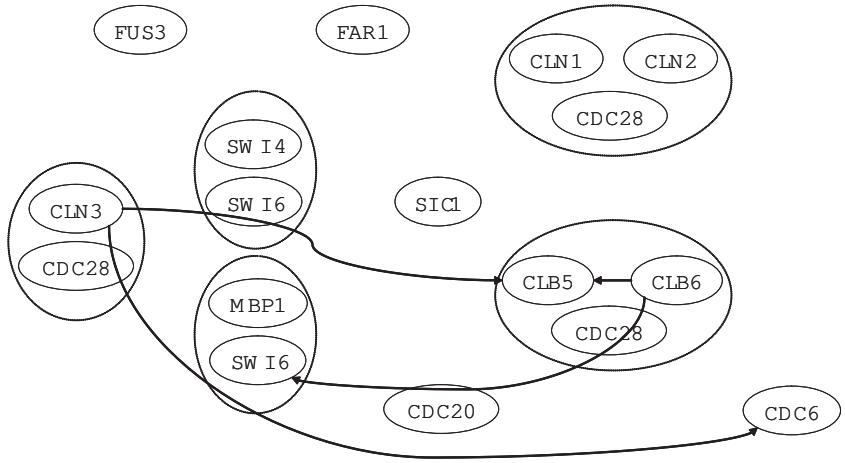


図 5.12: DBN のみで推定されたエッジ (実データ #1)

表 5.7: 両手法で推定されたエッジ (実データ #1)

エッジ	類似度
FAR1 → CDC20	unknown
SIC1 → CDC20	unknown
CDC6 → CLB5	-1
CDC6 → CLB6	-1
CDC6 → CLN1	unknown
CLN2 → CLN1	1
CDC20 → CLN2	unknown
CDC6 → CLN2	unknown
CLN1 → CLN2	1
CLN3 → CLN2	2
CLN2 → CLN3	unknown
CLN2 → FAR1	-1
CLN3 → FAR1	unknown
FAR1 → SIC1	2
SWI6 → SWI4	1

表 5.8: DDBN のみで推定されたエッジ (実データ #1)

エッジ	類似度
CLB6 → CDC20	-1
CLN2 → CDC6	3
CLN1 → CLB5	2
CLN2 → SWI6	1

表 5.9: DBN のみで推定されたエッジ (実データ #1)

エッジ	類似度
CLN3 → CDC6	3
CLB6 → CLB5	1
CLN3 → CLB5	2
CLB6 → SWI6	-1

されたエッジをそれぞれ図 5.15，5.16，5.17 に示す．さらに各エッジについての詳細をそれぞれ表 5.10，5.11，5.12 に示す．表中での類似度は #1 と同様である．ただし，“−1or2” は，ループ構造によるものであり，2.5% は KEGG データベース中のパスウェイのうち無向エッジがあるためである．

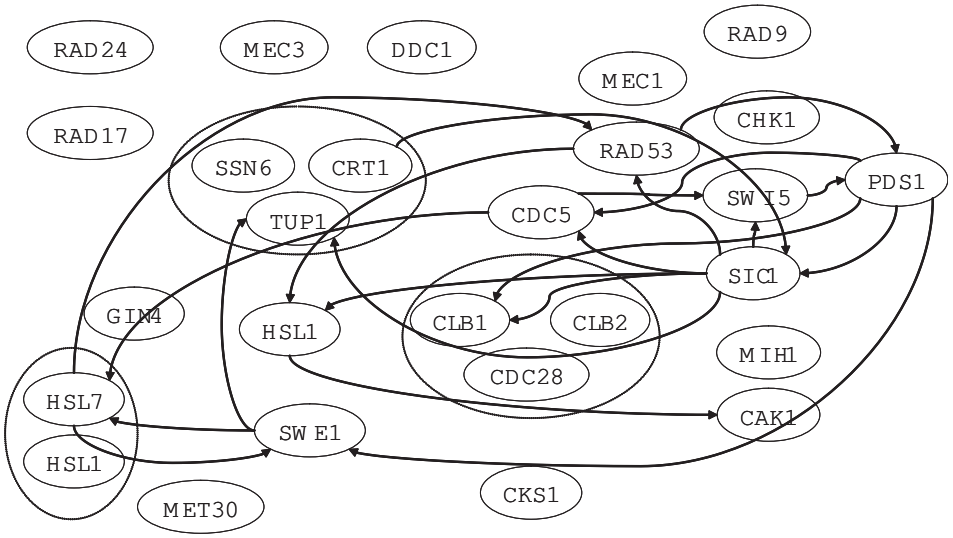


図 5.13: DDBN で推定されたネットワーク (実データ #2)

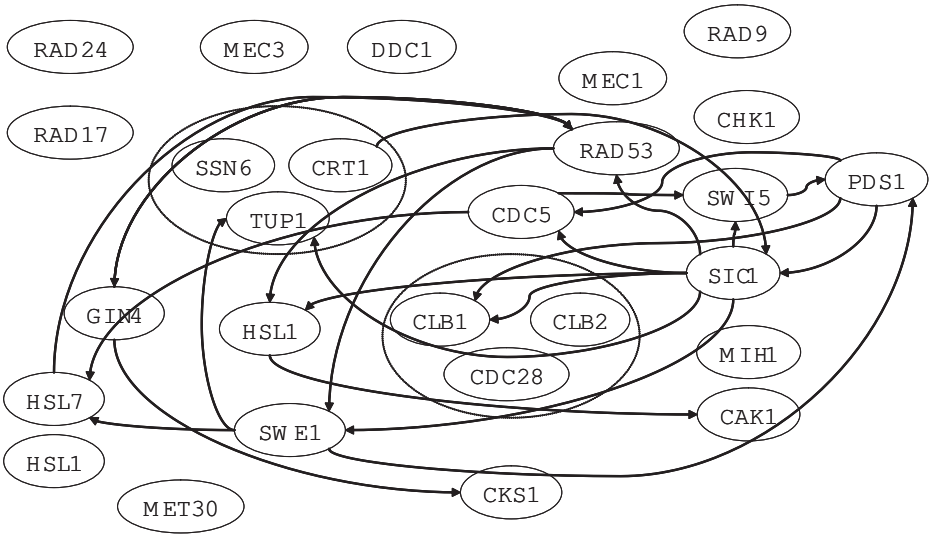


図 5.14: DBN によって推定されたネットワーク (実データ #2)

これらの結果には自己ループは省略してある．DDBN ではすべての遺伝子に対して，DBN では CDC5，CLB1，HSL1，MIH1，SIC1 に対して自己ループを検出した．

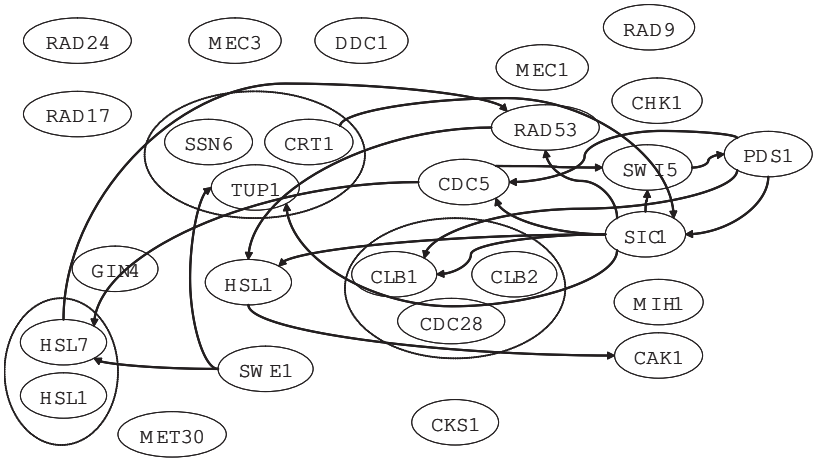


図 5.15: 両手法で推定されたエッジ (実データ #2)

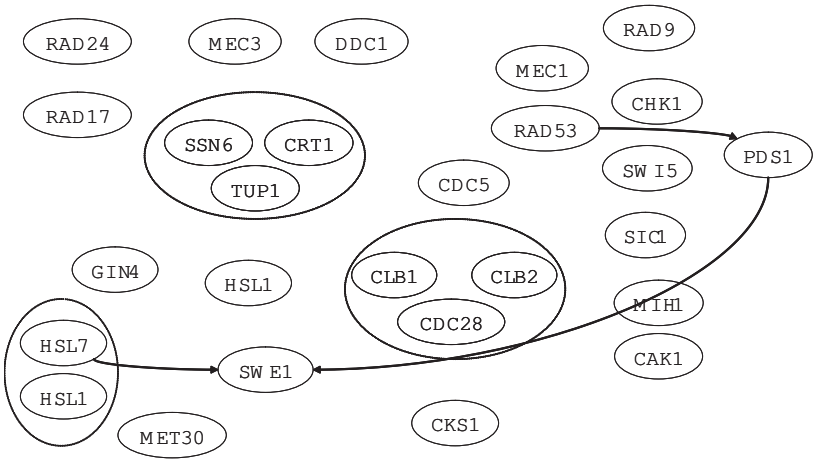


図 5.16: DDBN のみで推定エッジ (実データ #2)

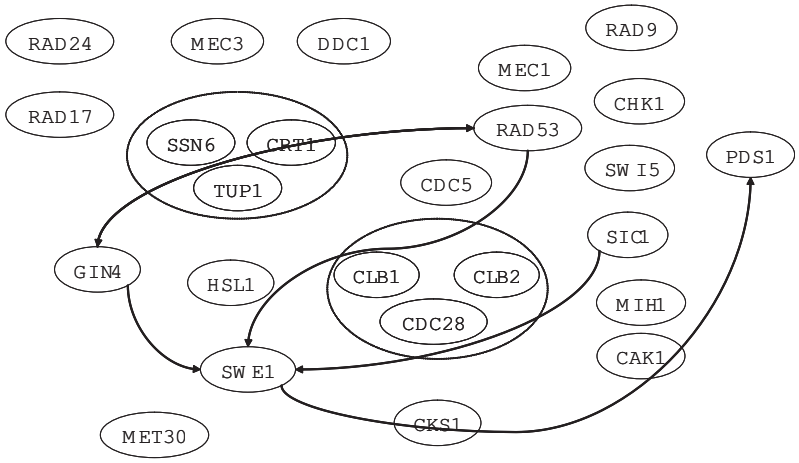


図 5.17: DBN のみで推定されたエッジ (実データ #2)

表 5.10: 両手法で推定されたエッジ (実データ #2)

エッジ	類似度
HSL1 → CAK1	unknown
PDS1 → CDC5	unknown
SIC1 → CDC5	unknown
PDS1 → CLB1	unknown
SIC1 → CLB1	1
RAD53 → HSL1	unknown
SIC1 → HSL1	unknown
CDC5 → HSL7	unknown
SWE1 → HSL7	-1
SWI5 → PDS1	unknown
HSL7 → RAD53	unknown
SIC1 → RAD53	unknown
PDS1 → SIC1	unknown
CRT1 → SIC1	unknown
CDC5 → SWI5	unknown
SIC1 → SWI5	-1 or 2
SIC1 → TUP1	unknown
SWE1 → TUP1	unknown

表 5.11: DDBN のみで推定されたエッジ (実データ #2)

エッジ	類似度
RAD53 → PDS1	unknown
HSL7 → SWE1	1
PDS1 → SWE1	unknown

表 5.12: DBN のみで推定されたエッジ (実データ #2)

エッジ	類似度
GIN4 → CKS1	2.5
RAD53 → GIN4	unknown
SWE1 → PDS1	unknown
GIN4 → RAD53	unknown
RAD53 → SWE1	unknown
SIC1 → SWE1	unknown

5.5 考察

ネットワーク探索アルゴリズムに山登り法を用いた場合局所解に陥ってしまい，最良スコアを得ることができなかった．実験の結果では GA+LS($r = 0.2$) が最も良い成績を示し，確率的探索の有効性，局所探索との組み合わせによる効率向上が認められた．

推定されたネットワークについては，DBN では検出できなかったエッジのうち，DDBN では検出することができたものがいくつかあった．しかし逆に，DDBN では検出できていないが，DBN では検出しているエッジも存在した．また，最終結果ではエッジとして検出されていないが，探索過程で有用とされ，かつ，KEEG データベースに登録されているパスがいくつか見られた．さらに，高い評価を与えるネットワークが必ずしも生物学的に正しいとは限らなかった．

以上より，遺伝子制御ネットワークの推定における次のような改善案が挙げられる．

- DDBN と DBN を排他的に用いるのではなく，お互いを補助的な立場とする．
- 探索過程の情報も用いる．
- 並列性の高い GA の特徴を活かし，複数の解候補を提示する．
- 生物学者からのフィードバックにより，評価規準を修正する．

これらのより詳しい内容については第 6 章で述べる．

第6章

結論

6.1 まとめ

本研究の目的は，次の 4 つの条件を満たすモデルの導出と，推定実験によるモデルの評価であった．

- 誤差を扱うことができる
- 循環構造を表現できる
- 推定に多くの時間ステップを必要としない
- 時系列の変化量を扱うことができる．

これについて，本論文では以下のような構成で論じてきた．

第 1 章では，研究の背景を述べ，本研究で達成すべき目的を明らかにした．

第 2 章では，生物学的な背景から遺伝子制御ネットワークの定義についてふれ，遺伝子発現状態を測定する装置である DNA マイクロアレイを説明した．そして，遺伝子制御ネットワークの推定問題の定式化を行い，研究例を紹介した．

第 3 章では，提案する推定手法の導出に必要な関連研究，つまり，ベイジアンネットワークの学習の構成要素であるモデル，評価規準，ネットワーク構造の探索アルゴリズムについて述べた．モデルとして，ノンパラメトリック回帰とベイジアンネットワークについて述べ，評価規準の研究例を紹介した．そして，ネットワーク構造の探索アルゴリズムについて説明を行った．

第 4 章では，提案手法によって時系列遺伝子発現データから遺伝子制御ネットワークを推定する方法を述べた．まず本研究で提案する動的微分ベイジアンネットワークの導出を行った．導出に際して問題点の整理を行い，上に記した 4 つの条件を明らかにした．動的微分ベイジアンネットワークの導出を行った後，変数間関係を記述するモデルについて述べた．非線形な関係を連続値として扱うために，ノンパラメトリック回帰を採用し，動的微分ベイジアンネットワークとノンパラメトリック回帰に基づいてモデルを構築した．そして，モデルを評価するための評価規準を導出した．

第 5 章では，提案手法と従来手法による遺伝子制御ネットワークの推定実験を行った．推定対象として人工的なネットワークから生成したデータと，出芽酵母細胞周期データの解析を行った．ネットワーク探索アルゴリズムの比較，人工データ解析による従来手法との比較，そして実データの解析を行った．

そして，この第 6 章で本論文を振り返った．次節で本研究から得られた知見についてまとめ，最後に今後の課題を記し，本論文の結論とする．

6.2 結論

遺伝子制御ネットワークのような化学反応に基づく制御関係を記述する場合，微分方程式系で扱うような変数の変化量が重要な要素となる．しかし，微分方程式系の推定には大量のデータが必

要であり，マイクロアレイで得られる時系列の時間ステップ数はそれに比べて少ない．さらに，微分方程式系の推定は誤差に敏感であり，マイクロアレイデータとの相性が悪いという欠点がある．

この欠点に対して，ベイズ統計に基づくグラフィカルモデルであるベイジアンネットワークを用いた研究が多くなされ，成果を挙げている．しかしながら，従来のモデルである動的ベイジアンネットワークでは時系列の変化量を扱うことはできなかった．

これらの問題を解決するために，本研究では微分方程式系モデルと動的ベイジアンネットワークモデルの長所を取り入れ，動的微分ベイジアンネットワークモデルを提案した．そして，遺伝子間関係を記述するモデルとして，実数値データを直接扱うことができ，遺伝子間の非線形な関係も取り扱うことができるノンパラメトリック回帰モデルを採用した．

ネットワーク探索アルゴリズムの比較では，局所探索付き遺伝的アルゴリズムによって最も良い正答率を獲得した．

人工的なネットワークから生成したデータを対象とした推定実験については，S_n/S_p という指標によって評価を行った．その結果，提案手法は両指標において従来手法よりも良い成績を示した．

出芽酵母の細胞周期データと対象とした解析では，提案手法によって，従来手法では検出できなかった有用なエッジも検出することができた．ただし，提案手法で検出できなく，従来手法によって検出したエッジの中にも有用なエッジが存在していた．このことから，従来のモデルと本研究で提案するモデルを排他的に考えるのではなく，うまく融合する方法を模索していく必要がある．また，探索過程で有効と判定されたエッジの中に実際のデータベースに登録されているエッジが含まれていた．このことから，探索過程の情報もうまく取り入れることによって生物的に有用な結果が得られる可能性もある．さらに，よい評価を与えるネットワークが必ずしも生物学的に正しいとは限らなかった．

以上をまとめると，本研究について次の 4 つの知見が得られる．

1. 提案手法と従来手法とでは，検出能力を発揮できる変数間関係が異なる．
2. 探索過程にも，生物学的に有用な情報が含まれている．
3. ネットワーク構造の探索には確率的探索を用いる方がよい．
4. 生物学の知見を取り入れ，評価規準を修正する必要がある．

6.3　今後の課題

本研究によって得られた 4 つの知見をもとに，今後行うべき課題について述べる．

提案手法と従来手法とでは，検出能力を発揮できる変数間関係が異なる．

検出可能 · 不可能な変数間関係を具体的に知る必要がある．この変数間関係を特定するためには，実際のデータ生成メカニズムと推定結果を比較しなくてはならない．そのための手法としては，例えば人工データを生成する関数にゆらぎを与えて推定結果に及ぼす影響を見る，というこ

とが考えられる．ただし，ここでのゆらぎとは，仮定するモデルや式の構造，パラメータに対して与える微少変動である．

探索過程にも，生物学的に有用な情報が含まれている．

探索過程に出現した有効な情報を活かすためには，探索過程の情報を何らかの形で最終結果に反映させる必要がある．

例えば，各エッジに対して探索過程での出現頻度測定を行ったり，スコアによる重み付けを行ったりする方法が考えられる．

ネットワーク構造の探索には確率的探索を用いる方がよい．

大規模なネットワークの推定を行うためには，ネットワーク構造探索の方法を工夫する必要がある．例えば，探索アルゴリズムの性能を向上させるというだけでなく，生物学の事前知識を制約として加える，遺伝子の挙動によってあらかじめクラスタリングを施す，などが考えられる．

生物学の知見を取り入れ，評価規準を修正する必要がある．

いくら最良の評価を得たネットワークであったとしても，評価規準が適切でなければ正しい知見を得ることはできない．そこで，推定された結果と既に生物学の知見として蓄えられているデータベースとを比較することで，評価規準を修正していく必要がある．修正の方法として，グラフの事前確率分布として与えるというのが最も自然な方法に思われる．

その他，改善の余地がある点

変数間関係を記述するモデルについて

動的微分ベイジアンネットワークは，次の 3 つのモデルからなる．

1. データ初期値の分布
2. 親変数と変化量との関係
3. 値，変化量と次時刻での値との関係

これらのモデルを再検討し，修正を加える．

例えば，データ初期値の分布に生物的知見を加える，ノンパラメトリック回帰モデルでなく他のモデルを用いる，値，変化量と次時刻での値との関係に，確率変動を加える，などが考えられる．

ノンパラメトリック回帰モデルについて

本研究で採用したノンパラメトリック回帰モデルにも改良の余地は多分にある．

例えば，B-スプライン以外の基底関数を検討する，基底関数の数を固定にしない，パラメータの事前確率分布を検討する，などが考えられる．特に，グラフの事前分布には生物学的知見を導入しやすく，修正の価値は高いと思われる．

マルコフ性の拡張

本研究では時間に対する 1 次マルコフ性を仮定している．しかし，実際の遺伝子制御関係には時間遅れが存在する可能性がある．これを実現するためには，例えば時間遅れや n 次マルコフ過程への拡張が考えられる．

モデルの構築

現段階では生物学的にも未解明の部分が多いため，本研究ではさまざまな変数間関係を表すことができるノンパラメトリック回帰モデルを採用した．しかし，より詳細なメカニズムを解明するためには生物学的な知見に基づくモデルを構築する必要がある．

欠損値の扱い

動的微分ベイジアンネットワークの学習では欠損値がのという仮定を用いている．しかし，実際に観測されるデータには大量の欠損値が含まれている．

そこで，欠損値の取り扱いを工夫する必要がある．例えば，欠損値補完を行う際に生物学的知見を取り入れる，動的微分ベイジアンネットワークの拡張を行って欠損値を取り扱えるようにする，などが考えられる．

謝辞

本研究を進めるにあたり，多くの方からご指導，ご鞭撻を賜りました．

指導教官である伊庭斉志教授には，大変熱心にご指導していただきました．充実した研究室環境を提供していただいたばかりでなく，非常に多忙であるにも関わらず，研究に関する私の相談に熱心に耳を傾けていただき，親切なアドバイスをしていただきました．また論文や資料の作成にも特別のご配慮をいただきました．今，このように私の研究をまとめることができたのも，ひとえに伊庭教授のご指導の賜物であります．厚く御礼申し上げます．

秘書の島津美和氏には様々な事務手続きをしていただき，私の研究活動がスムースに行えるようにしていただきました．深く感謝いたします．

研究室の先輩方に感謝いたします．井上豊氏には，研究面だけでなく私生活においてのさまざまなことを教えていただきました．神尾正太郎氏には，研究室の計算機，ネットワークの管理者としてお世話になりました．

同じ修士２年のメンバーに感謝します．三橋秀行氏，青木勝洋氏，長谷川禎彦氏とは多くの時間を共有し，さまざまな刺激を受けました．

その他，ここには書ききれない先輩方，同期，後輩達に深く感謝します．

また，わたしの学生生活をさまざまな面で支えてくださった家族，友人たちにも深く感謝いたします．

最後に，わたしの健康面，精神面をはじめ生活の全てにおいて支えになってくれた入江尚子さんに心から感謝します．

参考文献

[1] H. Akaike. Information theory and an extension of the maximum likelihood principli. In *2nd Internatoinal Symposium on InformationTheory*, pp. 267–281, 1973.

[2] T. Akutsu, S. Miyano, and S. Kuhara. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In *Proc. of Pacific Symp. Biocomputing'99*, 1999.

[3] S. Ando and H. Iba. Estimation of gene regulatory network by genetic algorithm and pairwise correlation analysis. In *Congress on Evolutionary Computing(CEC)*, 2003.

[4] Paul B, Alberto de la Fuente, and Pedro M. Gene network: how to put the function in genomics. *TRENDS in Biotechnology*, Vol. 20, No. 11, Novenber 2002.

[5] J. Bilmes. Dynamic bayesian multinets. In *Proc. Conf. on Uncertainty in Artificial Intelligence*, pp. 38–45, 1985.

[6] K. P. Burnham and D. R. Anderson. Model selection and inference. In *a Practical Information-Theoretical Approach*. Springer-Verlag, 1998.

[7] T. Chen, H. L. He, and G.M. Church. Modeling gene expression with differential equations. In *Proc. Pacific Symposium on Biocomputing*, pp. 29–40, 1994.

[8] D. M. Chickering. Learning bayesian networks is np-complete. In D. Fisher and H.-J. Lenz, editors, *Learning from Data*, Artificial Intelligence and Statistics V. Springer-Verlag, 1996.

[9] C. de Boor. *A Practical Guide to Splines*. Springer, Berlin, 1978.

[10] M. J. L. De Hoon, S. Imoto, K. Kobayashi, N. Ogasawara, and S. Miyano. Inferring gene regulatory networks from time-ordered gene expression data of *Bacillussubtilis* using differential equations. In *Proc. Pacific Symposium on Biocomputing*, pp. 17–28, 2003.

[11] J. L. DeRisi, V. R. lyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, Vol. 278, , October 1997.

[12] P. Eilers and B. Marx. Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, Vol. 11, pp. 89–121, 1994.

[13] N. Metropolis et al. Equation of state calculations by fast computing machines. *Journal of Chem. Physics*, Vol. 21, pp. 1087–1092, 1953.

[14] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian network to analyze expression data. *J.Comp.biol*, Vol. 7, pp. 601–620, 2000.

[15] N. Friedman, K. Murphy, and S. Russell. Learning the structure of dynamic probabilistic networks. In *Proc. Conf. on Uncertainty in Artificial Intelligence*, pp. 139–147, 1998.

[16] P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall, 1994.

[17] D. Heckerman. A tutorial on learning with bayesian networks. In *Learning and Inference in Graphical Models*. Kluwer Academic Publisher, 1998.

[18] M. D. Huang, F. Romeo, and A. L. Sangiovanni-Vincentelli. An efficient general cooling schedule for simulated annearing. In *IEEE International Conference on Computer-Aided Design*, pp. 381–329, 1986.

[19] S. Imoto, T. Goto, and S. Miyano. Estimation of genetic networks and functional structure between genes by using bayesian networks and nonparametric regression. In *Pacific Symposium on Biocomputing(PSB2002)*, Vol. 7, pp. 175–186, 2002.

[20] S. Imoto and S. Konishi. Nonlinear regression models using B-spline and information criteria. *the Insutitute of Statistical Mathematics*, Vol. 2, No. 47, pp. 359–373, 1999.

[21] S. Kim, S. Imoto, and S. Miyano. Dynamic bayesian network and nonparamtric regression for nonlinear modeling of gene networks from time series gene expression data. *Biosystems*, Vol. 75, pp. 57–65, July 2004.

[22] G. Kitagawa and W. Gersch. *Smoothness Priors Analysis of Time Series*. Springer, 1996.

[23] J. Kitagawa and H. Iba. Identifying gene regulatory networks as a petri net by genetic algorithms. In Gary Fogel and David Corne, editors, *Evolutionary Computation and Bioinformatics*. Morgan Kaufmann, 2002.

[24] S. Konishi. Statistical model evaluation and information criteria. In S.Ghosh(ed.), editor, *Multivariate Analysis, Design of Experiments and Surbey Sampling*. Marcel Dekker, 1999.

[25] S. Konishi and G. Kitagawa. Generalised information criteria in model selection. *Biometrika*, Vol. 83, pp. 875–890, 1996.

[26] J. Koza. On the programming of computers by means of natural selection. In *Genetic Programming*. MIT Press, 1992.

[27] A. Mimura and H. iba. Inference of a gene regulatory network by means of interactive evolutionary computing. In *Proc. of Fourth Conference on Computational Biology and Genome Informatics*, 2002.

[28] I. M. Ong, J. D. Glasner, and D. Page. Modelling regulatory pathways in *e. coli* from time series expression profiles. *Bioinformatics*, Vol. 18, pp. S241–S248, 2002.

[29] D. Peér, A. Regev, G. Elidan, and N. Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, Vol. 17, pp. S215–S224, 2001.

[30] J. O. Ramsay. Monotone regression splines in action (with discussion). *Statistical Science*, Vol. 3, pp. 425–461, 1988.

[31] F. Reif. *Statistical and Thermal Physics*. McGraw–Hill, New York, 1965.

[32] R. W. Robinson. Counting labeled acyclic digraphs. *New Directions in the theory of Graphs*, pp. 239–273, 1973.

[33] Simonoff J. S. *Smoothing Method in Statistics*. Springer, 1996.

[34] E. Sakamoto and H. Iba. Identifying gene regulatory network as differential equation by genetic programming. In *Proc. of Genome Informatics Workshop*, 2000.

[35] E. Sakamoto and H. Iba. Inferring a system of differential equations for a gene regulatory network by using genetic programming. In *Congress on Evolutionary Computation*, 2001.

[36] E. V. Someren, L. Wessels, and M. Reinders. Linear modeling of genetic networks from experimental data. *Bioinformatics*, Vol. 18, pp. 355–366, ISMB2002.

[37] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comperhensive identification of cell cycleregulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*, Vol. 9, pp. 3273–3297, 1998.

[38] Naoya Sugimoto and Hitoshi Iba. Inference of gene regulatory networks by means of dynamic differential bayesian networks and nonparametric regression. In *The 15th International Conference on Genome Informatics(GIW 2004)*, Vol. 15, pp. 121–130, 2004.

[39] D. Tomonaga, N. Koga, and M. Okamoto. Efficient numerical optimization algorithm based on genetic algorithm for inverse problem. In *Proc. of Genetic and Evolutionary Computation Conference*, 2000.

[40] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, Vol. 17, pp. 520–525, 2001.

[41] Ognjenka Goka Vukmirovic and Shirley M. Tilghman. Exploring genome space. *Nature*, Vol. 405, , June 2000.

[42] D. C. Weaver, C. T. Workman, and G. D. Storm. Modeling regulatory networks with weighted matrices. In *Proc. on Pasific Symposium on Bioinformatics 4*, 1999.

[43] S. R. White. Concept of scale in simulated annealing. In *IEEE International Conference of Computer Design*, pp. 646–651, 1984.

[44] 安藤晋, 伊庭斉志. 遺伝的アルゴリズムを用いたネットワーク構造の学習. 人工知能学会論文誌, Vol. 18, No. 5, 2003.

[45] 伊庭斉志. 遺伝的アルゴリズムの基礎. オーム社, 1994.

[46] 伊庭斉志. 遺伝的プログラミング入門. 東京大学出版会, 2001.

[47] 井元清哉, 小西貞則. 情報量規準に基づく B-スプライン非線形回帰モデルの推定. 応用統計学, 1999.

[48] 岡本正宏. 第 10 章 シミュレーション技術. バイオプロセスシステム工学. IPC, 1994.

[49] 北野宏明. 遺伝子回路のリバースエンジニアリング. システムバイオロジー, 第 3 章. 秀潤社, 2001.

[50] 杉本直也, 伊庭斉志. 遺伝的プログラミングによる超越関数を含む微分方程式系の推定. 情報処理学会第 65 回全国大会講演論文集, 2003.

[51] 杉本直也, 坂本栄里奈, 伊庭斉志. 遺伝的プログラミングによる微分方程式系の推定. 人工知能学会論文誌, Vol. 19, No. 6, pp. 450–459, 9 2004.

[52] 本村陽一. Bayesian network による大規模データのモデル化について. ベイジアンネットセミナー 2002.

発表文献

学術雑誌等

[J1] 杉本直也, 坂本栄里奈, 伊庭斉志. 遺伝的プログラミングによる微分方程式系の推定. 人工知能学会論文誌, Vol. 19, No. 6, pp. 450–459, 9 2004.

国際会議発表論文

[C1] Naoya Sugimoto and Hitoshi Iba. Inference of gene regulatory networks by means of dynamic differential bayesian networks and nonparametric regression. In *The 15th International Conference on Genome Informatics(GIW 2004)*, Vol. 15, pp. 121–130, 2004.

全国大会発表論文

[P1] 杉本直也, 伊庭斉志. 遺伝的プログラミングによる超越関数を含む微分方程式系の推定. 情報処理学会第 65 回全国大会, 2003. 学生奨励賞受賞.

研究会 · シンポジウム等発表論文

[M1] 杉本直也, 伊庭斉志. Dynamic differential bayesian networks and nonparametric regression による遺伝子ネットワークの推定. “大規模遺伝子ネットワークの相互作用推定” “遺伝子 · タンパク質系ダイナミクスの非線形システム的理解”2004 年度合同会議, 2004.

[M2] 杉本直也, 伊庭斉志. ベイジアンネットワークによる遺伝子ネットワークの推定. “大規模遺伝子ネットワークの相互作用推定” “遺伝子 · タンパク質系ダイナミクスの非線形システム的理解”2003 年度合同会議, 2003.

著書等

[B1] 伊庭, その他訳. Evolutionary Computation in Bioinformatics (David Fogel 他編).