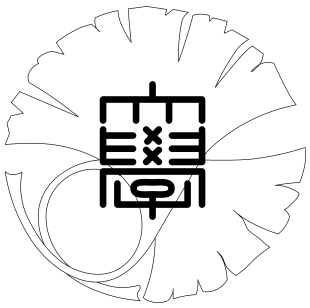
修士論文

AdaBoostを用いた
遺伝子制御ネットワークの統合的推定



2007年2月2日

指導教官　　伊庭 斉志 教授

東京大学大学院 工学系研究科

電子工学専攻
37-56446

生田目 慎也

Abstract

In order to estimate Gene Regulatory Networks (GRNs) from gene expression time series data, various recurrence or differential equation based models have been proposed, such as S-system, Linear model etc. Generally, it is assumed that a specific recurrence or differential equation model is sufficient to estimate the network from the expression profile. However, with so many different models available, it is not easy to recognize the model that will be most suitable for a particular network inference problem. To deal with the problem, integrative estimation with multiple recurrence or differential equation based models seems promising. In this paper, we propose the integration of multiple estimation methods by means of AdaBoost. Empirical studies show the effectiveness of our proposal.

内容概要

本論文では複数の漸化式・微分方程式モデルによる遺伝子制御ネットワークの推定結果を AdaBoost により統合する方法を提案する．

発現量時系列データから遺伝子制御ネットワークを推定するために S-system や線形モデルなどの様々な漸化式・微分方程式モデルが提案されている．現在は特定のモデルがターゲットとするネットワークの推定に適していると仮定して推定が行われている．しかしながら，複数のモデルが提案されている状況から分かるように，対象とする時系列データに対してどのモデルが推定に適しているか不明であるという問題がある．この問題に対して複数の漸化式・微分方程式モデルによる統合的な推定方法は有効であると考えられる．本論文では AdaBoost を用いた複数の推定手法（モデル）の統合を提案し，さらに推定実験を行い提案手法の有効性を示す．

目次

第 1 章　序論 **9**

1.1　研究背景 10

1.2　本論文の目的 11

1.3　本論文の構成 11

第 2 章　遺伝子制御ネットワーク **13**

2.1　はじめに 14

2.2　遺伝子制御ネットワーク 14

2.2.1　ゲノム・遺伝子・DNA 14

2.2.2　遺伝子の発現過程 14

2.2.3　遺伝子制御ネットワーク 14

2.2.4　DNA マイクロアレイ 17

2.2.5　Pathway データベース 18

2.3　問題の定義：遺伝子制御ネットワークの推定 18

2.4　ネットワーク推定・解明の意義 19

第 3 章　関連研究 **20**

3.1　はじめに 21

3.2　提案されている漸化式・微分方程式モデル 21

3.2.1　線形モデル (Linear モデル) 21

3.2.2　重み行列モデル (Weaver モデル) 21

3.2.3　確率微分方程式モデル (SDE モデル) 22

3.3　漸化式・微分方程式による GRN 推定 22

3.4　AdaBoost 22

3.4.1　2 値判別問題の AdaBoost 23

3.4.2　回帰問題の AdaBoost.R 23

3.5　その他の推定手法 24

3.5.1　S-system モデル 24

3.5.2　遺伝的プログラミングによる構造の推定 24

3.5.3 閾値検定モデル . . . 25
3.5.4 ブーリアンネットワークモデル . . . 25
3.5.5 ベイジアンネットワーク モデル . . . 26

第 4 章 提案手法 27
4.1 はじめに . . . 28
4.2 GRN 推定研究における問題点 . . . 28
4.3 提案手法・AdaBoost を利用した GRN 推定 . . . 29
4.3.1 仮説器の推定方法 . . . 29
4.3.2 情報量基準 . . . 31
4.3.3 GRN 推定のための AdaBoost.R アルゴリズム . . . 31

第 5 章 実験 34
5.1 はじめに . . . 35
5.2 人工データ実験 1：ノイズ無し . . . 35
5.2.1 データの作成方法 . . . 35
5.2.2 評価方法 . . . 36
5.2.3 実験結果の観察と考察 . . . 37
5.3 人工データ実験 2：ノイズ有り . . . 43
5.3.1 ノイズ有りデータの作成方法 . . . 43
5.3.2 実験結果の観察と考察 . . . 43
5.4 実データを用いた実験 . . . 46
5.4.1 データの適用方法 . . . 46
5.4.2 実験結果の観察と考察 . . . 47

第 6 章 考察 55
6.1 ターゲットネットワークの構造について . . . 56
6.2 ノイズについて . . . 56
6.3 計算量について . . . 56
6.4 モデルの種類について . . . 57

第 7 章 結論 58

付録 60
A Linear モデルによる推定方法 . . . 61
B Weaver モデルによる推定方法 . . . 62
C SDE モデルによる推定方法 . . . 62

D SDE のヒューリスティクス 63

参考文献 **65**

発表文献 **67**

図目次

2.1 遺伝子の発現過程 15
2.2 生命ネットワーク 15
2.3 大腸菌の DNA 修復機構 16
2.4 大腸菌 DNA 修復機構の遺伝子制御ネットワーク 17
2.5 Ronen らにより得られた SOS ネットワークの発現量時系列データ 17
2.6 Stanford 大学のマイクロアレイデータベース 18
2.7 Kegg データベース：出芽酵母菌の細胞周期機構 19

3.1 漸化式・微分方程式とネットワークの関係 23
3.2 訓練データの使用方法 24
3.3 木構造による方程式の表現 25
3.4 Boolean network and Transition table 26

4.1 提案手法の流れ 29

5.1 人工ネットワーク 35
5.2 実験 1 の結果（ノイズ無し） 38
5.3 実験 2 の結果（ノイズ無し） 38
5.4 実験 3 の結果（ノイズ無し） 39
5.5 実験 4 の結果（ノイズ無し） 39
5.6 学習状況，採用仮説器 No 40
5.7 学習状況, 信頼度 41
5.8 実験 1 の結果（ノイズ有り） 44
5.9 実験 2 の結果（ノイズ有り） 44
5.10 実験 3 の結果（ノイズ有り） 45
5.11 実験 4 の結果（ノイズ有り） 45
5.12 大腸菌 DNA 修復機構の遺伝子制御ネットワーク (再掲) 46
5.13 各手法の精度，(正確度順の上位 8 手法) 47
5.14 各手法の精度 (正確度順の下位 8 手法) 47
5.15 SDE(BIC) によって推定されたネットワーク 48

5.16 SDE(AIC) によって推定されたネットワーク 49
5.17 AdaBoost によって推定されたネットワーク 49
5.18 SDE(R) によって推定されたネットワーク 50
5.19 SDE(Cp) によって推定されたネットワーク 50
5.20 uvrD の制御関係推定時の学習状況 51
5.21 lexA の制御関係推定時の学習状況 51
5.22 umuD の制御関係推定時の学習状況 52
5.23 recA の制御関係推定時の学習状況 52
5.24 uvrA の制御関係推定時の学習状況 53
5.25 uvrY の制御関係推定時の学習状況 53
5.26 ruvA の制御関係推定時の学習状況 54
5.27 polB の制御関係推定時の学習状況 54

表目次

5.1 実験条件 36
5.2 各パターンの割合（ノイズ無し） 42
5.3 各パターンの割合（ノイズ有り） 43

第１章

序論

1.1 研究背景

遺伝子の発現量を測定する装置である DNA マイクロアレイなどから得られた発現量時系列データから遺伝子の制御関係を表す GRN を推定する研究が盛んに行われている．これまでに漸化式・微分方程式モデルだけでなく，Boolean Network や Dynamic Bayesian Network，遺伝的プログラミングなど様々な推定方法など提案されてきた [1, 10].

推定手法の 1 つである漸化式・微分方程式モデルでは，現在までに線形モデル (以下，Linear モデル)，重み付き行列モデル (以下，Weaver モデル)，確率微分方程式モデル (以下，SDE モデル) や S-system モデルなどが提案されている [3, 8, 20]．これら従来の研究では遺伝子の発現量時系列データが特定の微分方程式モデルに従うと仮定していた．例えば SDE モデルに従っているとして[1]，そのモデルのパラメータを求めることによりネットワークの推定が行われてきた.

現在までに，主に以下の 2 つの問題に対する研究が行われてきた.

- 時系列データにノイズがのっている状態でのパラメータ推定のための研究
- 大規模なネットワークの推定をするための計算時間短縮のための研究

現在の DNA マイクロアレイは測定誤差が大きいという観点から，前者の研究には意義がある．また後者では，特に S-system パラメータを求める研究で遺伝子の破壊実験データを用いることによりネットワークを分割してパラメータを求める方法 [12] や，数値積分をスプライン補間法で代用して問題を分割する方法が提案されるなど [11]，計算時間を短縮する上で重要な研究がなされてきた.

しかしながら，現在あまり考慮されていない問題が 1 つある．この問題は現在様々な漸化式・微分方程式モデルが提案され，さらに 1 つの GRN に対して様々なモデルが推定に用いられている状況から分かる．例えば大腸菌の DNA 修復機構である SOS ネットワークやラットの中枢神経系に対して，S-system や 2 階の常微分方程式モデル，Weaver モデルや Linear モデル等で推定実験が行われている [2, 8, 14, 17]．つまり，対象とする発現量時系列データに対してどのモデルが推定に適しているか不明であるという問題がある．そしてこの問題については現在までにあまり議論されていない.

一方で，機械学習手法の 1 つである AdaBoost が近年注目を浴びている．AdaBoost とは複数の推定手法による仮説（以下，推定手法を仮説器と呼ぶ）を統合し，全体としての推定精度を向上させるアルゴリズムである．バイオインフォマティクスでは疾患の分類において精度向上に貢

[1] ここでは，SDE モデルに従っている=SDE モデルが GRN 推定に適しているという意味で用いている

献しており，その効果が確認されている [7]．また，複数の仮説器を必要とすることから，多数の解候補 (推定手法) を提示することができる遺伝的プログラミング (Genetic Programming，GP) と相性が良く，AdaBoost と GP を組み合わせた研究なども行われている [9]．

GRN 推定研究の現状，つまり複数の推定手法 (モデル) が混在する状況を鑑みると AdaBoost の GRN 推定への適用は有効であると考えられる．しかしながら，現在のところ GRN 推定の分野に適用した研究は殆ど無い．これは学習に用いる例題をどう扱うかという課題が有るからである．

[4, 6] では時系列データの各サンプル点を学習データとして利用し，回帰の精度を向上させるアルゴリズム・AdaBoost.R を提案している．我々は AdaBoos.R を用いて，その一部を変更することにより GRN 推定に適したアルゴリズムを考案する．具体的なアルゴリズムについては 4 の提案手法で述べる．

1.2　本論文の目的

本論文の目的は，複数の漸化式・微分方程式モデルによる遺伝子制御ネットワーク (Gene Regulatory Network, 以下 GRN) の推定結果を AdaBoost により統合する方法を提案することである．提案手法の性能を評価するために，人工ネットワークと実データを用いた推定実験を行った．

本論文は以下の 3 点において GRN 推定研究に貢献していると考えられる．

- 複数の漸化式・微分方程式モデルによる推定方法の AdaBoost による統合
- 生成したモデルの種類に依らない高精度の推定結果の達成
- 実際の生物データでの有効性の検証

1.3　本論文の構成

また，本論文は以下の様な構成をとっている．

第 1 章

遺伝子制御ネットワークの推定研究についての背景の概略を述べ，現在の問題点，本研究で対象とする問題点，本論分の目的，構成の順に述べる．

第 2 章

遺伝子制御ネットワークの定義を述べるために，ゲノム・遺伝子・DNA などについて説明する．さらに，大腸菌の DNA 修復機構である SOS ネットワークを例として，遺伝子制御ネットワーク

について説明する．最後に本研究のネットワーク推定問題の定義や推定・解明の意義について述べる．

第 3 章

現在までに提案されている推定方法述べる．最初に本研究で用いる漸化式・微分方程式モデルについて，提案されているモデルや推定方法，ネットワークの解釈について述べた後，統合方法として用いる AdaBoost.R について説明する．最後に，S-sytem，遺伝的プログラミング，ベイジアンネットワークによる推定などの手法について述べる．

第 4 章

遺伝子制御ネットワークの現状についての問題点を述べる．問題点に有効であると考えられる手法・AdaBoost を用いた統合的な推定方法を提案し，そのアルゴリズムを示す．

第 5 章

提案手法を評価するために人工ネットワーク，実ネットワークを用いて推定実験を行う．前者では推定に適したモデルが異なる状況を想定しており，人工ネットワークの構造は一定で 4 種の条件で推定を行う．後者の実ネットワークは大腸菌の修復機構である SOS ネットワークを用いる．また，それぞれの実験で得られた結果より考察を行う．

第 6 章

提案手法の利点や今後解決すべき課題等を，ノイズや計算量，モデルの種類といった観点から述べる．

第 7 章

本論文のまとめを行う．

第２章

遺伝子制御ネットワーク

2.1 はじめに

本章では，ゲノムや遺伝子，遺伝子制御ネットワークについて述べた後，推定問題の定義とその意義について述べる．

2.2 遺伝子制御ネットワーク

2.2.1 ゲノム・遺伝子・DNA

ゲノム，DNA，遺伝子といった言葉はよく使われるが，どう違うのか紛らわしいのでここで説明する．

ゲノムとは，各生物のもつ一揃いの遺伝情報の総体を表す抽象的な概念である．DNA は，A(アデニン)，C(シトシン)，G(グアニン)，T(チミン) の 4 種類の塩基という分子が糖やリン酸を介して連なってできた高分子であり，ゲノムの実体である．この塩基の連なりは鎖とよばれ，TCTTGAGAA というように 4 種類の文字の配列として表現される．

DNA に書かれている情報は生物の設計図に過ぎず，生命活動のための機能的な部品は，DNA の情報を元にしてタンパク質を合成することによって作られる．しかし，DNA 上の情報全てが部品を定義しているわけではなく，定義箇所はごく一部である．

遺伝子とは，ゲノムの中で部品を定義している箇所のことであり，RNA 分子やタンパク質に翻訳される部分を指す．ヒトの場合，遺伝子領域は DNA 配列全体のわずか 5%にすぎない．ヒトのゲノムサイズ (ゲノムの文字列の長さ) は 30 億，遺伝子の数は 10 万程度である (遺伝子数はまだ完全にはわかっていないので概算値である)．

2.2.2 遺伝子の発現過程

DNA 上の遺伝子は，タンパク質に翻訳されてその機能を発現する．その過程を図 2.1 に示す．まず，DNA 配列情報の一部が mRNA にコピーされる．次にこの mRNA 上の文字列が，3 文字分を単位 (コドンと呼ぶ) としてアミノ酸の 1 つに翻訳され，これによってアミノ酸の連なりが生成される．そしてこれが折りたたまれてタンパク質ができる (mRNA への転写の際に T は U(ウラシル) に置換される)．なお，DNA は設計図，タンパク質は部品にあたるので，1 個の遺伝子から多数の mRNA，タンパク質が作られる．つまり，ある遺伝子の発現レベルが高いときは，その対応する mRNA，タンパク質が数多く作られるということになる．

2.2.3 遺伝子制御ネットワーク

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行っている．また，各遺伝子は独立に発現しているのではなく，タンパク質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

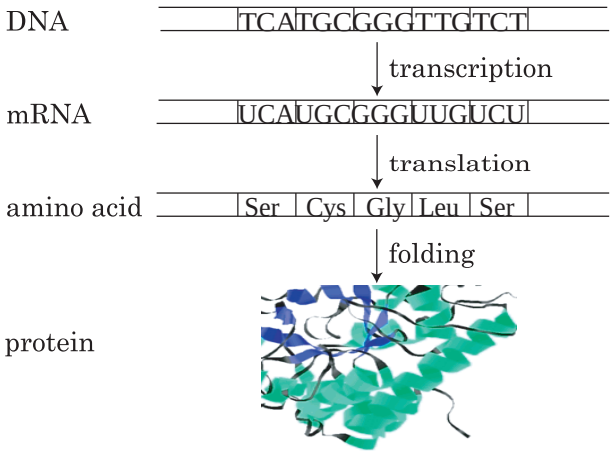


図 2.1: 遺伝子の発現過程

ク質によって，発現が活性されたり抑制されたりする．このように，遺伝子，タンパク質は複雑な相互作用の関係を持っている．これらの関係をネットワークとして考えると，生命ネットワークは図 2.2 のように表される．図 2.2 内の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．線の数字は対応する相互作用を示している．生命ネットワークにおける遺伝子・タンパク質間の相互作用を遺伝子にのみ着目したネットワークを遺伝子制御ネットワークという．

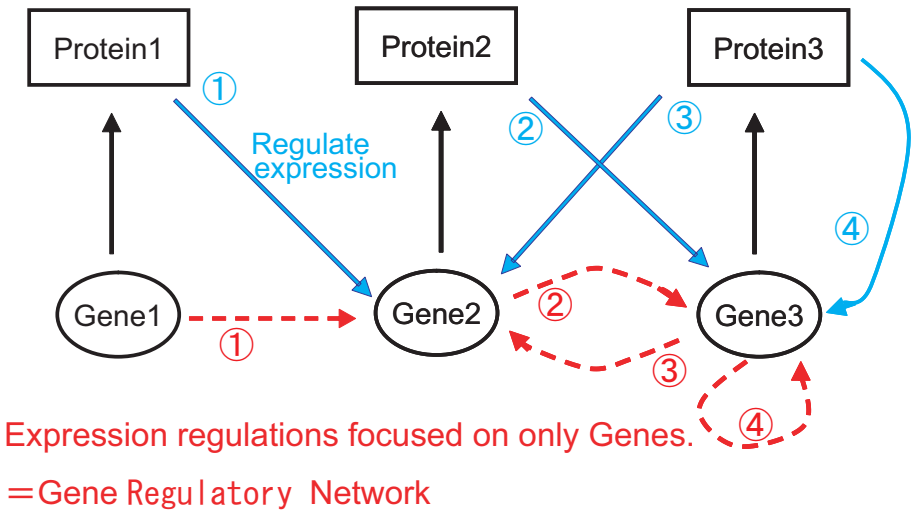


図 2.2: 生命ネットワーク

生命ネットワークの例として大腸菌の DNA 修復機構である SOS ネットワークを図 2.3 に示す．遺伝子は楕円で示し，その名前は小文字で記述している．また，その遺伝子によって生成される

タンパク質は方形で示し，名前の最初を大文字で記述している．

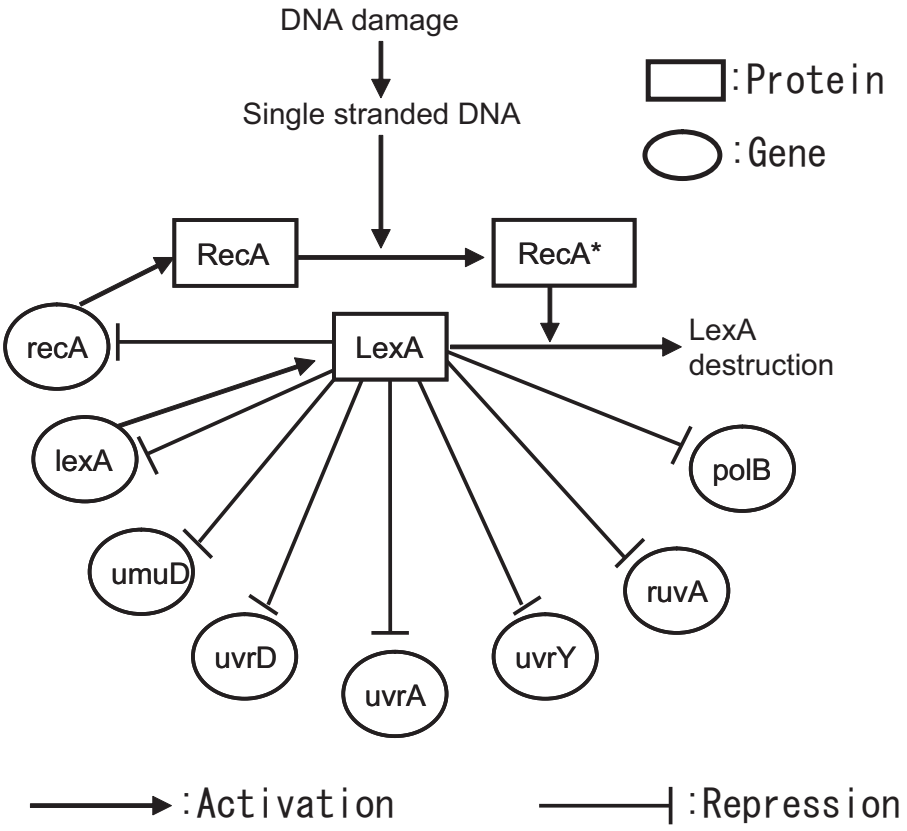


図 2.3: 大腸菌の DNA 修復機構

図 2.3 は紫外線などによって生じる DNA の損傷を修復する機構を生命ネットワークとして表した図である．これを SOS ネットワークといい，各遺伝子は SOS 遺伝子と呼ばれる．以下 SOS ネットワークについて簡単に説明する．

通常は SOS 遺伝子の発現は LexA によって調整されている．しかし，紫外線などにより DNA が損傷を受けるとネットワーク内のセンサーとしての役割を果たしている RecA が一本鎖 DNA(single stranded DNA) に結合し RecA*酵素となり，LexA の抑制を破壊して修復を開始する．LexA が破壊されると抑制が無くなった各 SOS 遺伝子の mRNA 合成が起こる．lexA 自身も発現し LexA が合成され，時間が経つとともに SOS 遺伝子は再び抑制される．

図 2.3 の生命ネットワークを遺伝子のみに着目すると図 2.4 を遺伝子制御ネットワークと見ることが出来る．

Ronen らは大腸菌の DNA 修復機構に紫外線を照射することにより，DNA に損傷を与え SOS

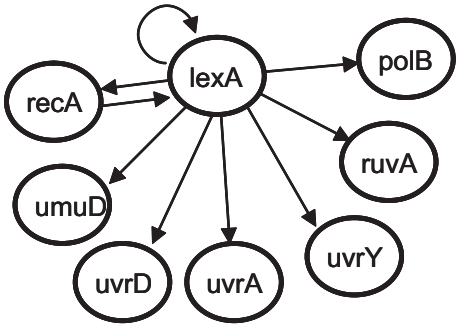


図 2.4: 大腸菌 DNA 修復機構の遺伝子制御ネットワーク

ネットワーク内の遺伝子の発現データの時系列を獲得している [18]．そのデータを図 2.5 に示す．

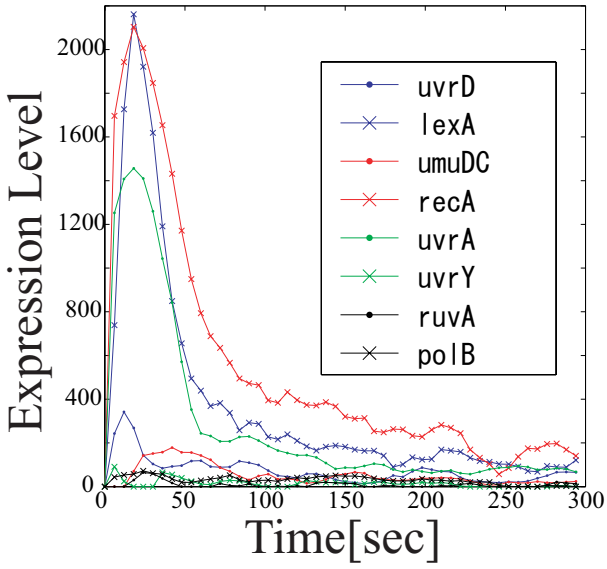


図 2.5: Ronen らにより得られた SOS ネットワークの発現量時系列データ

2.2.4　DNA マイクロアレイ

近年の測定装置の発達によって，生物から大量のデータを得ることを可能にした。その中の 1 つとして複数の遺伝子の発現量を同時に測定する装置である DNA マイクロアレイがある．測定されたデータはインターネット上に公開され自由にダウンロードできるようになっている．世界中の研究者がこれらのデータをダウンロードして研究に用いている．2.6 はデータベースの一つである Stanford 大学のマイクら例データベースである．

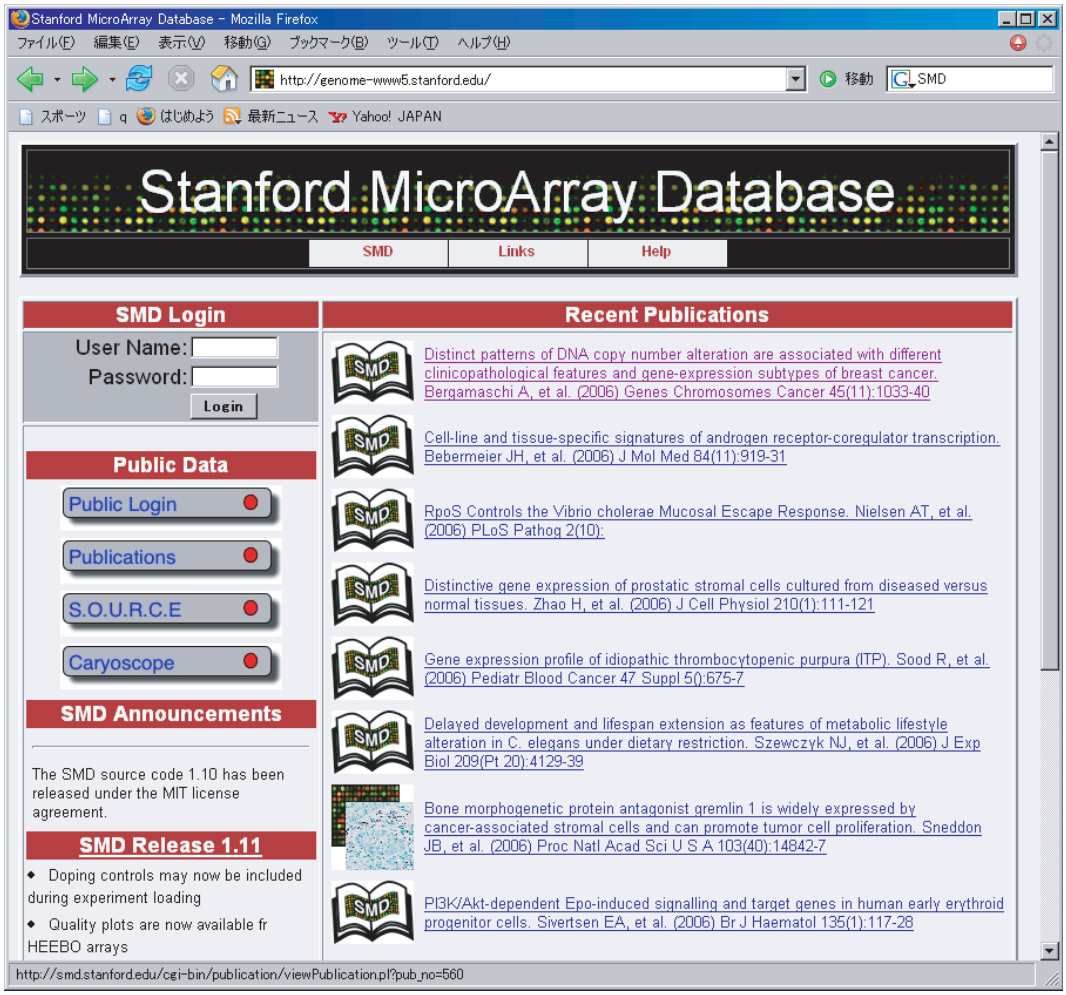


図 2.6: Stanford 大学のマイクロアレイデータベース

2.2.5　Pathway データベース

解明されたネットワークも WEB 上のデータベースに登録され，誰でも閲覧できるようになっている．図 2.7 は出芽酵母菌の細胞周期機構のネットワークの一部である．

2.3　問題の定義：遺伝子制御ネットワークの推定

遺伝子制御ネットワークの推定とは，各遺伝子間の制御関係の有無を推定することを意味する．本研究における遺伝子制御ネットワークの推定問題とは，図 2.5 のような各遺伝子の発現量を示す時系列データから図 2.4 のような各遺伝子間の制御関係を推定する問題を意味する．

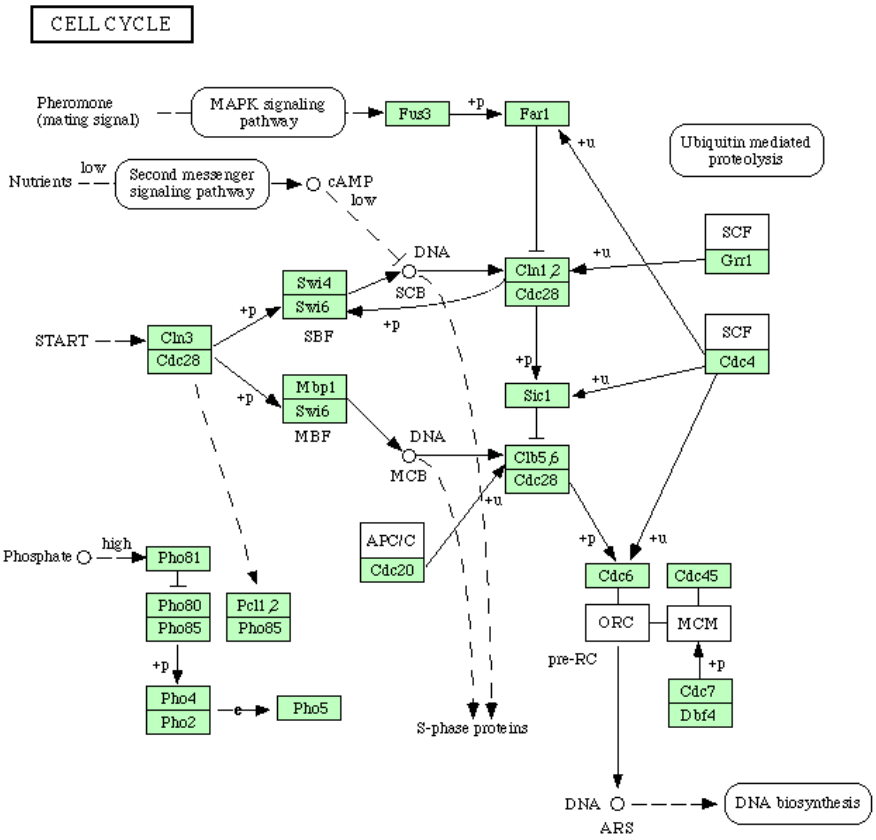


図 2.7: Kegg データベース：出芽酵母菌の細胞周期機構

2.4　ネットワーク推定・解明の意義

推定された遺伝子制御ネットワークは，実際に生物学的実験により解明する際の手がかりとして役に立つ．何十万とある遺伝子の相互作用を効率的に解明する上で，手がかりとなる推定結果は重要な情報であると考えられる．さらに，解明された遺伝子制御ネットワークは薬の開発などにおいて役に立つと考えられている．例えば遺伝子レベルで作用する薬を開発する際に，遺伝子レベルの副作用の効果を考慮に入れる必要がある．そのため遺伝子間の相互作用に関する知識，つまり遺伝子制御ネットワークに関する知識は必須であると考えられる．

第3章

関連研究

3.1 はじめに

本章では，本研究に関係する漸化式・微分方程式モデル，提案手法に関係する AdaBoost，その他の推定手法の順に述べる．

3.2 提案されている漸化式・微分方程式モデル

本研究で用いる線形モデル，重み行列モデル，確率微分方程式モデルについて紹介する．

3.2.1 線形モデル (Linear モデル)

線形モデルは遺伝子の発現レベルの変化量がネットワーク内の遺伝子の発現レベルの重み付き総和に依存すると仮定したモデルであり，以下の式 3.1 で表される [8].

$$X_i(t+\Delta t) = C_{i,0} + \sum_{j=1}^{n} C_{i,j} X_j \tag{3.1}$$

ここで,
$X_i(t)$：時間 t における遺伝子 i の発現レベル
$C_{i,0}$：外部からの遺伝子 i への寄与を示す係数
$C_{i,j}$：遺伝子 j から遺伝子 i への寄与を示す係数
n：ネットワーク内の遺伝子の数
である．

3.2.2 重み行列モデル (Weaver モデル)

重み行列モデルは，線形モデルと同様に発現レベルの変化量がネットワーク内の遺伝子の重み付総和に依存すると仮定した漸化式モデルである [20]．ただしシグモイド関数に重み付総和を代入する点が異なる．重み行列モデルは

$$X_i(t+1) = \frac{M_i}{1 + exp(-S_i(t))} \tag{3.2}$$

$$M_i = \max_t X_i(t) \tag{3.3}$$

$$S_i(t) = C_{i,0} + \sum_{j=1}^{n} C_{i,j} X_j(t) \tag{3.4}$$

で表される．$X_i(t), C_{i,0}, C_{i,j}, n$ は Linear モデルと同様の意味を持つ．

3.2.3 確率微分方程式モデル (SDE モデル)

SDE モデルは遺伝子の転写過程を確率微分方程式で表したモデルである．確率微分を行うことにより式 3.5 で表される [3]（導出過程は論文を参照).

$$dX_i = C_{i,0} + \sum_{j=1}^{n} C_{i,j} f(X_j(t)) \tag{3.5}$$

$$f(X_j(t)) = \frac{1}{1 + e^{\{-(X_j(t)-\mu_j)/\omega_j\}}} \tag{3.6}$$

$X_i(t), C_{i,0}, C_{i,j}, n$ は Linear モデルと同様の意味を持つ．μ_j, ω_j は遺伝子 j の時系列データの平均, 標準偏差であり，以下の式で表される．

$$\mu_j = \frac{1}{m} \sum_{k=1}^{m} X_j(k) \tag{3.7}$$

$$\omega_j = \sum_{k=1}^{m} (X_j(k) - \mu_j)/(m-1) \tag{3.8}$$

ここで，m は時系列サンプル数である．

3.3 漸化式・微分方程式による GRN 推定

線形モデルを例に挙げて説明する．漸化式・微分方程式モデルによる GRN 推定では，得られた発現量時系列データによく「フィットする」，つまり回帰してパラメータ $C_{i,j}$ を求め遺伝子間の制御関係を推定する．例えば遺伝子 X_3 について回帰を行い，以下のようなパラメータが得られたとする．

$$X_3(t + \Delta t) = 1.2 + 0.9X_2 - 1.2X_3 \tag{3.9}$$

$$= F(X_2, X_3) \tag{3.10}$$

式 3.10 は，遺伝子 X_3 の発現量は X_2, X_3 に依存していることを意味している．これを，X_3 は X_2, X_3 に制御されていると解釈し，$X_2 \to X_3$，$X_3 \to X_3$ の制御関係があると推定する．他の遺伝子についても同様にパラメータを求め制御関係を求める．漸化式・微分方程式と 5 個の遺伝子からなる GRN の関係を図 3.1 に示す.

3.4 AdaBoost

本研究で用いる AdaBoost について簡単に説明を行う.

$$\frac{dX_1}{dt} = F_1(X_1(t), X_3(t), X_5(t))$$
$$\frac{dX_2}{dt} = F_2(X_1(t), X_2(t))$$
$$\frac{dX_3}{dt} = F_3(X_2(t), X_3(t))$$
$$\frac{dX_4}{dt} = F_4(X_3(t), X_4(t), X_5(t))$$
$$\frac{dX_5}{dt} = F_5(X_4(t), X_5(t))$$

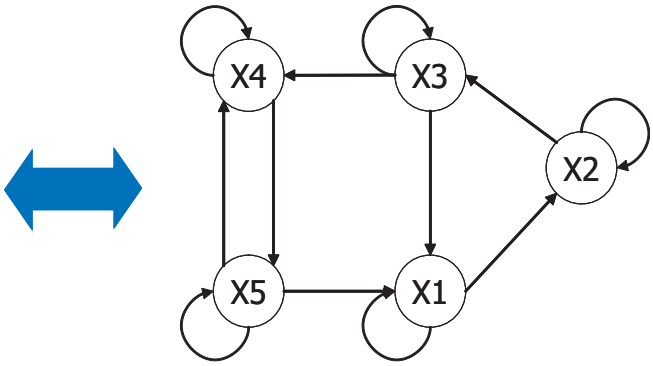


図 3.1: 漸化式・微分方程式とネットワークの関係

3.4.1　2 値判別問題の AdaBoost

機械学習手法の 1 つである AdaBoost が近年注目を浴びている．AdaBoost とは複数の推定手法による仮説（以下，推定手法を仮説器と呼ぶ）を統合し，全体としての推定精度を向上させるアルゴリズムである．バイオインフォマティクスでは疾患の分類において精度向上に貢献しており，その効果が確認されている [7]．また，複数の仮説器を必要とすることから，多数の解候補 (推定手法) を提示することができる遺伝的プログラミング (Genetic Programming，GP) と相性が良く，AdaBoost と GP を組み合わせた研究なども行われている [9].

GRN 推定研究の現状，つまり複数の推定手法 (モデル) が混在する状況を鑑みると AdaBoost の GRN 推定への適用は有効であると考えられる．しかしながら，現在のところ GRN 推定の分野に適用した研究は殆ど無い．これは学習に用いる例題をどう扱うかという課題が有るからである．

3.4.2　回帰問題の AdaBoost.R

[4, 6] では時系列データの各サンプル点を学習データとして利用し，回帰の精度を向上させるアルゴリズム・AdaBoost.R を提案している．これは，図 3.2 に示すように，各仮説器の回帰結果の各サンプル点を訓練データとして用いることにより損失関数や信頼度等を計算し，全体としての回帰精度を向上させるのに有効なアルゴリズムである．

我々は AdaBoos.R を用いて，その一部を変更することにより GRN 推定に適したアルゴリズムを考案する．具体的なアルゴリズムについては 4 の提案手法で述べる．

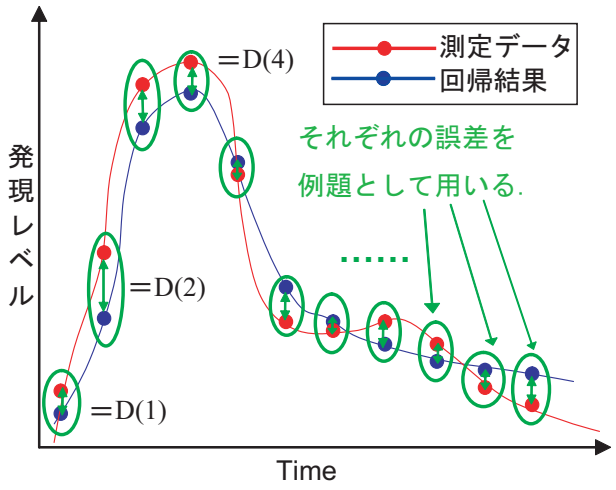


図 3.2: 訓練データの使用方法

3.5　その他の推定手法

3.5.1　S-system モデル

S-system は一般の生化学反応を近似したモデルであり，以下の式 3.11 によって記述される非線形の微分方程式である [19].

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{n} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n} X_j^{h_{ij}}$$

パラメータ数が膨大であり，さらに非線形であるためパラメータを解析的に求めることが出来ない．そのため進化論的計算手法を用いた様々なパラメータ探索方法が提案されている [11, 16].

3.5.2　遺伝的プログラミングによる構造の推定

遺伝的プログラミング (Genetic Programming, GP）によって，方程式の形を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 [5, 21]．GP による推定は，これまで紹介してきたアプローチである線形モデルや SDE モデルなどの方程式の構造をあらかじめ仮定してそのパラメータを求めることにより GRN を推定する方法とは異なる．GP では，方程式の構造を木構造を用いて表現する．例えば以下の式 (3.11),(3.12) は図 3.3 ように表現できる．これを個体とみなして，発現量時系列データに上手くフィットすることのできる木構造を探す.

$$\frac{dX_1}{dt} = 0.3X_1X_2 + X_2 \tag{3.11}$$

$$\frac{dX_2}{dt} = 0.5X_1X_2 \tag{3.12}$$

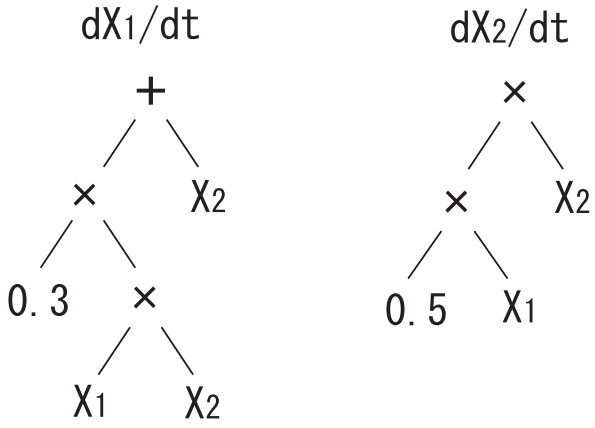


図 3.3: 木構造による方程式の表現

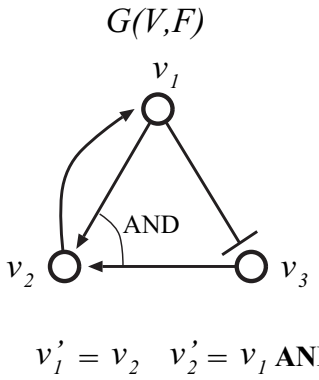
パラメータは最小二乗法で求める方法が提案されている [5]．GP による推定手法は関数の構造を構成する要素に制約が無く，方程式の次数に対する制限も無いために大きな探索空間を柔軟に探索することが可能である．しかしその一方で目的とするモデル構造の探索が困難であり，オーバーフィッティングした局所解が得られやすいというジレンマを持つ.

3.5.3　閾値検定モデル

遺伝子 a を破壊又は強制発現させたときの遺伝子 b の発現量の変化により，遺伝子 a から遺伝子 b への影響の有無を判定し，を「二項関係」として抽出する．これらの二項関係を統計的に有意か否か判定する．時間による変化は考慮に入れていれていない.

3.5.4　ブーリアンネットワークモデル

ブーリアンネットワークは，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 [1]．ブーリアンネットワークは，頂点集合 $V = v_1, v_2, ..., v_n$ と各頂点の値を決定するブール関数 (論理関数) の集合 $F = f_1, f_2, ..., f_n$ 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 $t = 1, 2, 3, ...$ ごとに同期して変化していく．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．発現量時系列データ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おり，また推定のための計算量が小さい．しかし，発現量を 2 値に簡単化している点，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.



INPUT			OUTPUT		
v_1	v_2	v_3	v_1'	v_2'	v_3'
0	0	0	0	0	1
0	0	1	0	0	1
0	1	0	1	0	1
0	1	1	1	0	1
1	0	0	0	0	0
1	0	1	0	1	0
1	1	0	1	0	0
1	1	1	1	1	0

図 3.4: Boolean network and Transition table

3.5.5 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確率モデルである．このモデルは確率変数をノードで表し, 因果関係などの依存の関係を持つ変数間にリンクを張ったグラフ構造である．また，リンクが依存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．

$\pi(X_j)$ を X_j の親ノードとしたとき，全ての確率変数の同時確率分布は以下の式 (3.13) で表現される．

$$P(X_1, X_2, ..., X_n) = \prod_j P(X_j|\pi(X_j)) \tag{3.13}$$

BN を時系列データを用いて循環構造を表現できるように拡張したものが Dynamic Bayesian Network(DBN) である．DBN では時間を表す変数 t を考慮することで, 循環構造を表し，離散時間における確率過程を表現している．

DBN を用いて遺伝子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 [15]. DBN による推定ではグラフの構造の探索空間はグラフのノード数に対して指数オーダーとなっている．そのため探索には工夫が必要となる．

第4章

提案手法

4.1 はじめに

本章では，遺伝子制御ネットワーク研究に対する現状の問題点を挙げ，それに対処するための提案手法を述べる．

4.2 GRN 推定研究における問題点

推定手法の 1 つである漸化式・微分方程式モデルでは，現在までに線形モデル (以下，Linear)，重み付き行列モデル (以下，Weaver)，確率微分方程式モデル (以下，SDE) や S-system モデルなどが提案されている [3,8,20]．これら従来の研究では遺伝子の発現量時系列データが特定の微分方程式モデルに従うと仮定していた．例えば SDE モデルに従っているとして[1]，そのモデルのパラメータを求めることによりネットワークの推定が行われてきた．

現在までに，主に以下の 2 つの問題に対する研究が行われてきた．

- 時系列データにノイズがのっている状態でのパラメータ推定の研究
- 大規模なネットワークの推定をするための計算時間短縮研究

現在の DNA マイクロアレイは測定誤差が大きいという観点から，前者の研究には意義がある．また後者では，特に S-system パラメータを求める研究で遺伝子の破壊実験データを用いることによりネットワークを分割してパラメータを求める方法 [12] や，数値積分をスプライン補間法で代用して問題を分割する方法が提案されるなど [11]，計算時間を短縮する上で重要な研究がなされてきた．

しかし，もう 1 つ問題がある．この問題は現在様々な漸化式・微分方程式モデルが提案され，さらに 1 つの GRN に対して様々なモデルが推定に用いられている状況から分かる．例えば大腸菌の DNA 修復機構である SOS ネットワークやラットの中枢神経系に対して，S-system や 2 階の常微分方程式モデル，Weaver モデルや Linear モデル等で推定実験が行われている [2,8,14,17]．**つまり，対象とする発現量時系列データに対してどのモデルが推定に適しているか不明であるという問題がある**．しかしこの問題については現在までにあまり議論されていない．

GRN 推定研究の現状，つまり複数の推定手法 (モデル) が混在する状況を鑑みると AdaBoost の GRN 推定への適用は有効であると考えられる．しかしながら，現在のところ GRN 推定の分野に適用した研究は殆ど無い．これは学習に用いる例題をどう扱うかという課題が有るからである．3.4.2 で述べたように，[4,6] では時系列データの各サンプル点を学習データとして利用し，回

[1] ここでは，SDE モデルに従っている = SDE モデルが GRN 推定に適しているという意味で用いている

帰の精度を向上させるアルゴリズム・AdaBoost.R を提案している．本研究では AdaBoos.R を用いて，その一部を変更することにより GRN 推定に適したアルゴリズムを考案する．

4.3　提案手法・AdaBoost を利用した GRN 推定

提案手法の概要図を図 4.1 に示す．SDE+AIC，Linear+BIC などの複数の推定手法（以下，仮説器と呼ぶ）により得られた推定結果を AdaBoost により統合する．本節では提案手法について，仮説器の具体的な推定方法，AdaBoost による推定結果の統合アルゴリズムの順に説明する．

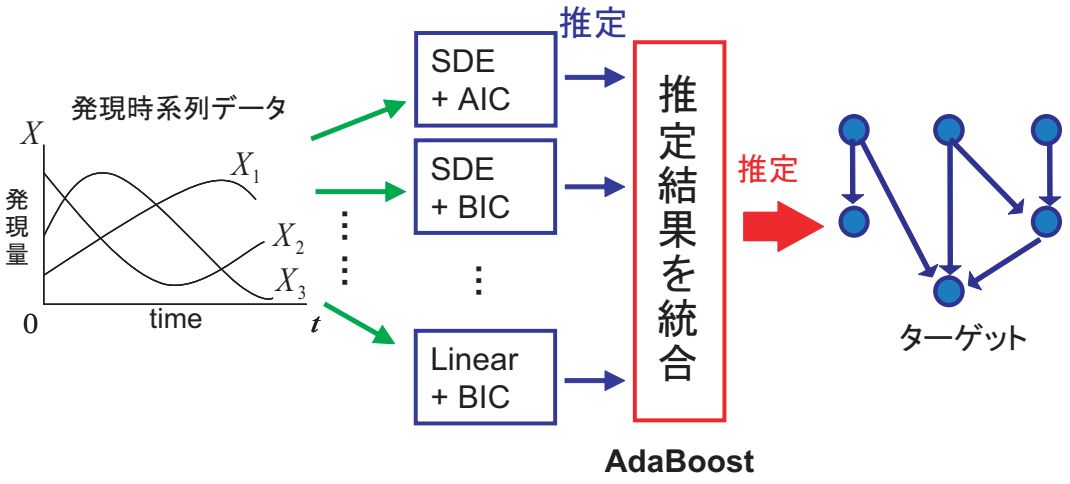


図 4.1: 提案手法の流れ

4.3.1　仮説器の推定方法

本研究では仮説器は SDE,Linear, Weaver の 3 モデルを用いる．これら 3 モデルでは最尤推定によりパラメータを求めることができる．これは発現量時系列データが与えられると GRN を一意に推定でき，AdaBoost による統合の効果が分かりやすいからである．また，今回はパラメータ推定に実数値遺伝的アルゴリズムなどの確率的手法を必要とする S-system モデルは使用しない．

パラメータ決定のさいにはオーバーフィッティングを防ぐために尤度ではなく，赤池情報量基準 (AIC) やベイズ情報量基準 (BIC) などを用いる．以下に仮説器の推定方法を示す．

仮説器の推定アルゴリズム

For i = 1 To n　　(n：GRN 内の遺伝子数)

Step1　　回帰する遺伝子 (被説明変数) の選択

遺伝子 X_i を選択する．

Step2　情報量基準の計算

n 個の遺伝子のうち任意の遺伝子を任意の個数だけ選択する．この組合せが遺伝子 X_i を制御している (つまり説明変数) と仮定する．例として，ここでは遺伝子を 3 個選んだとする．これを X_j, X_k, X_l とする．

X_j, X_k, X_l の時系列データを用いて遺伝子 i の時系列データにフィッティングするように最尤法でパラメータを求める．さらにこの組合せの情報量基準を計算する．これを全ての組合せについて行う．

Step3　推定

情報量基準が最小となる遺伝子の組合せが遺伝子 i を制御していると推定する．

end For

ここでの仮説器とは，例えば SDE モデルのパラメータを AIC または BIC などを基準として求める手法を指す．今後はこれらの仮説器を SDE(AIC) や SDE(BIC) などと表記する．

推定例

5 個の遺伝子の発現時系列データから遺伝子ネットワークを推定する場合を例に挙げる．Linear(AIC) を用いて遺伝子 4 について回帰した結果，AIC が最小になる組合せが X_2, X_4, X_5 となり以下のようにパラメータが得られたとする．

$$\begin{aligned} X_4(t+1) &= 1.1 - 1.2X_2(t) - 0.3X_4(t) + 0.5X_5(t) \\ &= F(X_2, X_4, X_5) \end{aligned} \tag{4.1}$$

式 4.1 より，発現量 X_4 は X_2, X_4, X_5 に依存しているということができる．よって，$X_2 \to X_4$，$X_4 \to X_4$，$X_5 \to X_4$ の制御関係があると推定する．このとき Linear(AIC) による遺伝子 4 の制御関係についての仮説 $\mathbf{h}_{\mathbf{Lin(AIC)}} \in \Re^{1\times n}$ を以下のように定義する．

$$\mathbf{h}_{\mathbf{Lin(AIC)}} = [-1, 1, -1, 1, 1] \in \Re^{1\times n} \tag{4.2}$$

ここで，$\mathbf{h}$ ベクトルの

- 要素 j が　$1 \Rightarrow X_j$ は X_4 を制御している
- 要素 j が $-1 \Rightarrow X_j$ は X_4 を制御していない

を意味する[2]．$\mathbf{h}$ は 4.3.3 の AdaBoost アルゴリズムの式 4.17 で使用する．パラメータ，情報量基準の計算方法は各モデル毎に付録 A，B，C に示す．

[2]制御関係が活性か抑制かまで推定するのが理想であるが，本研究ではそこまでの精度を求めない．制御関係の有無のみを推定する．有無の情報のみでも実際に生物学実験に GRN を調べる際に十分役立つと考えられる．

4.3.2 情報量基準

仮説器の数を増やして多様な仮説を得るために，上記の AIC,BIC に加えて 3 個の基準を用いる．つまり 1 モデルあたり 5 個の仮説器を用いる．以下，n' は制御している遺伝子の数 (説明変数の数)，m は時系列サンプル数を表す.

- 赤池情報量基準 (以下 AIC)

$$AIC = -2logL + 2(n' + 1) \tag{4.3}$$

- ベイズ情報量基準 (以下 BIC)

$$BIC = -2logL + mlogn' \tag{4.4}$$

- Hanann-Quinn の情報量基準 (以下 HQ)

$$HQ = -2logL + c \times mlog(logn') \tag{4.5}$$

本研究では $c = 3$ として用いている.

- 自由度調整済み決定係数 (以下 R)

$$R = 1 - \frac{S_E/(n' - m - 1)}{S_T/(n' - 1)} \tag{4.6}$$

ここで,
残差平方和 $S_E = \sum_{t=1}^{m}(X(t) - \hat{X}(t))^2$
総平方和 $S_T = \sum_{t=1}^{m}(X(t) - \bar{X}(t))^2$
$\hat{X}(t)$ は回帰して得られた発現レベル,
$\bar{X}(t)$ は X(t) の $t = 1$ から m を通しての平均である [22].

- Mallows の Cp 統計量 (以下 Cp)

$$Cp = \frac{S_E}{\hat{\omega}^2} + 2(m + 1) - n' \tag{4.7}$$

$\hat{\omega}^2$ は全ての説明変数を用いたときの誤差分散の不偏推定量である.

4.3.3 GRN 推定のための AdaBoost.R アルゴリズム

以下に GRN 推定のために AdaBoost.R の一部を変更したアルゴリズムを示す.
For i = 1 To n　　(n：GRN 内の遺伝子数)

Step 1　全ての仮説器について回帰誤差を計算

$$D_k(t) = \left| X_i(t) - \hat{X}_{k,i}(t) \right| \tag{4.8}$$

例題に対する重みの初期化

$$w(t) = \frac{1}{m}, (t = 1, ..., m) \tag{4.9}$$

Step 2　学習

For $\ell = 1$ To Iteration

1. 損失関数の計算

$$L_k(t) = \frac{D_k(t)}{D_{k,max}} \tag{4.10}$$

$$D_{k,max} = \max_t D_k(t) \tag{4.11}$$

$$\overline{L}_k = \sum_{i=1}^{m} w(t) L_k(t) \tag{4.12}$$

2. 損失関数が最小となる仮説器を選択

$$\overline{L} = \min_k \overline{L}_k \tag{4.13}$$

$$k'(\ell) = arg \min_k \overline{L}_k \tag{4.14}$$

3. 信頼度の計算，重みの更新

$$\beta_\ell = \frac{\overline{L}}{1 - \overline{L}} \tag{4.15}$$

$$w(t) \leftarrow w(t) \beta_\ell^{1 - L_{k'(\ell)}(t)} \tag{4.16}$$

4. 終了条件のチェック

if $\sum_{t=1}^{m} w(t) < 0.001$ ，then 学習終了

end For

Step 3　重み付多数決による推定

式 4.14 で選択された各仮説器の推定結果 **h** と信頼度を用いて遺伝子 **i** の制御関係についての仮説 $\mathbf{H_i} \in \Re^{1 \times n}$ を生成する．

$$\mathbf{H_i} = sign\Big(\sum_{\ell=1}^{Iteration} \mathbf{h}_{k'(\ell)} \times log \frac{1}{\beta_\ell} \Big) \tag{4.17}$$

Step4　信頼度の小さい制御関係を削除するヒューリスティクス

$$\mathbf{H_i}' = \sum_{\ell=1}^{Iteration} \mathbf{h}_{k'(\ell)} \times log \frac{1}{\beta_\ell} \tag{4.18}$$

とする．

$\mathbf{H_i'}$ の各要素で正の信頼度の総和が小さい制御関係を消去する．

$if \quad H_i'(j) > 0 \quad and \quad H_i'(j) < r \times \bar{H_i'}$

Then　$H_i(j) = -1$

ここで，

$H_i(j)$：$\mathbf{H_i}$ の j 番目の要素，$\bar{H_i} = \sum_{j=1}^{n} |H_i'(j)|$

r：閾値，実験では $r = 0.5$ としている．

end For

ここで

k：仮説器の番号[3]

$\hat{X}_{k,i}(t)$：仮説器 k が遺伝子 i の時系列データを回帰して推定した時刻 t での発現レベル

m：時系列サンプル数

$L_k(t)$：仮説器 k の回帰結果における例題 t の損失．

$k'(\ell)$：学習 ℓ 回目で採用された仮説器の番号

$\mathbf{h}_k$：仮説器 k の推定結果

AdaBoost.R は回帰の精度を高めることを目的として提案されているが，これを一部を変更して 2 値の判別問題とすることにより GRN 推定に適用する．つまり，**Step3** では本来は重み付メディアンにより回帰を行うが，これを制御している・してないの 2 値の判別問題として重み付多数決を行う．重み付多数決は学習で得られた信頼度を利用する．

[3]仮説器の番号を以下の様に定義する．
SDE:　AIC$\cdots$1，BIC$\cdots$2，HQ$\cdots$3，R$\cdots$4，Cp$\cdots$5　Linear：同様の情報量基準の順番で 6〜10，Weaver：11〜15.

第5章

実験

5.1　はじめに

本研究では提案手法の有効性を調べるために人工データ・実データを用いた推定実験を行い，実験結果の観察・考察を行う．

人工データではノイズ無し・有りの各ケースについて実験を行う．

5.2　人工データ実験 1：ノイズ無し

人工データの作成方法，評価結果，推定結果，観察・考察の順に述べる．

5.2.1　データの作成方法

図 5.1 に示すように，遺伝子 9 個からなる GRN をターゲットとして人工データを作成する．GRN の特徴として制御関係が少ないという点，つまりネットワークは疎であるという考えに基いてターゲットを作成した．

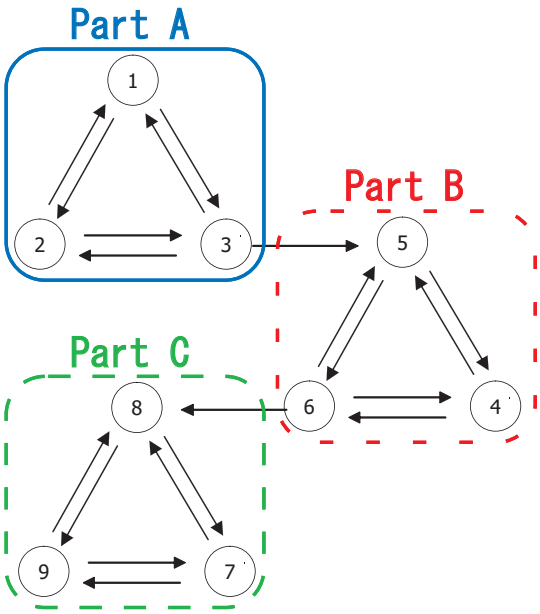


図 5.1: 人工ネットワーク

図 5.1 に示すようにネットワークを Part A, Part B, PartC に分ける．それぞれの部分で異なるモデルにより発現時系列データを生成する．モデルの組合せは表 5.1 に示す 4 通りを行う．実験 1，2，3 はそれぞれ，SDE，Linear，Weaver モデルが推定に適している GRN を想定している．実験 4 は単一の推定手法では完璧に推定できないケースを想定している．

表 5.1: 実験条件

	Part A	Part B	Part C
実験 1	SDE	SDE	SDE
実験 2	Linaer	Linear	Linear
実験 3	Weaver	Weaver	Weaver
実験 4	SDE	Linear	Weaver

時系列データは離散形で作成する．例えば，実験 4 では遺伝子 X_6 は Part B に属しているので Linear モデルである．さらに X_4, X_5 に制御されているので以下の式 5.1 より作成する．

$$X_6(t+1) = C_{6,0} + C_{6,4}X_4(t) + C_{6,5}X_5(t) \tag{5.1}$$

ここで，パラメータ $C_{6,0}, C_{6,4}, C_{6,5}$ をランダムに決定する．具体的には $C_{i,i} \sim -U(0,1)$[1]，$C_{i,j}(i \neq j)$ はそれぞれ 1/2 の確率で $U(0,1)$ 又は $-U(0,1)$ の乱数を取る．初期値 $X_i(0)$ は $U(0,1)$ で決定する．他の遺伝子の場合も同様にランダムに決定する．時系列サンプル数 m は 20 とする．

5.2.2　評価方法

推定手法の性能評価指標として，以下の式で示す正確度 A，感度 Sn，特異度 Sp，的中度 P を用いる．

- 正確度 A

$$A = \frac{\sharp ofTP + \sharp ofTN}{\sharp ofTP + \sharp ofTN + \sharp ofFP + \sharp ofFN} \tag{5.2}$$

ここで，TP，TN，FP，FN は TruePositive，TrueNegative，FalsePositive，FalseNegative の略であり，
TP：正しく検出した制御関係
TN：正しく検出しなかった制御関係
FP：誤って検出した制御関係
FN：誤って検出しなかった制御関係
である．
正確度 = 正解数/(正解数+不正解数) であり，総合的な精度を示す．

[1] $U(a,b)$ は a〜b の値をとる一様分布の乱数

- 感度 Sn

$$S_n = \frac{\sharp ofTP}{\sharp ofTP + \sharp ofFN} \tag{5.3}$$

感度＝（正しく検出した制御関係数)/(全制御関係) であり，正しく検出した割合を示す.

- 特異度 Sp

$$S_p = \frac{\sharp ofTN}{\sharp ofTN + \sharp ofFP} \tag{5.4}$$

感度と逆の意味で，実際に無い制御関係を，無いと推定した割合を示す.

- 的中度 P

$$P = \frac{\sharp ofTP}{\sharp ofTP + \sharp ofFP} \tag{5.5}$$

的中度　=(正しい制御関係数)/(検出した制御関係数）であり，制御関係の検出力の的中率を示す.

本研究では推定精度の評価として，総合的な精度を示す正確度 A を第一の評価指標とする．その内わけを示す指標としてその他の 3 指標 Sn, Sp, P を考慮する．GRN 推定ではどの遺伝子間に制御関係があるかを検出することが重要であると考えられるので，A に次いで Sn, P, Sp の順に指標を重視する.

5.2.3　実験結果の観察と考察

実験はそれぞれ 100 回ずつ行い，A, Sn, Sp, P の平均，標準偏差を調べることにより推定手法の性能を調べる．各実験の結果を図 5.2〜5.5 に示す.

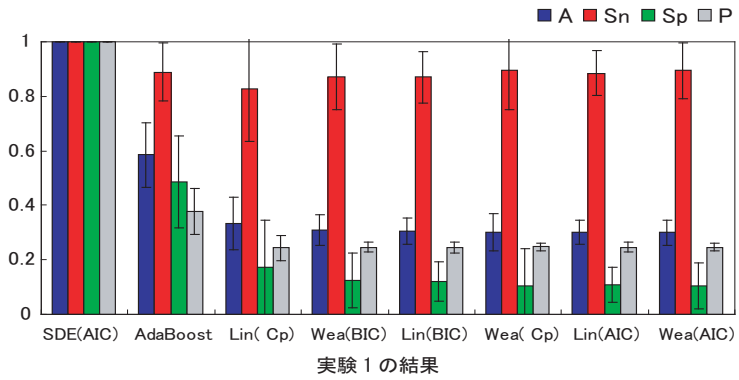


図 5.2: 実験 1 の結果（ノイズ無し）

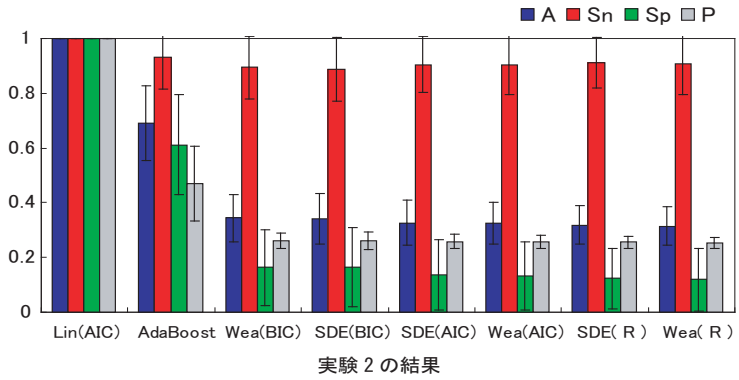


図 5.3: 実験 2 の結果（ノイズ無し）

それぞれのグラフでは 15 個の推定手法（3 モデル× 5 情報量基準) と AdaBoost の計 16 個の推定手法のうち，正確度の平均が高い上位 8 手法を左から順に示している．ただし，実験 1，2，3 では生成するモデルと推定手法が正しい場合，例えば実験 1 の人工データを SDE モデルで作成する場合は推定手法が SDE モデルならば情報量基準の種類に関らず，$A = 1$ となるので，グラフでは SDE(AIC) のみで代表させている．つまり実験 1 では AdaBoost は 6 番目に正確度の平均が高いことを意味している．実験 2，3 も同様である．

実験 1〜3 の結果より AdaBoost の正確度の平均は 6 番目に高いことが分かる．一方，実験 1 で完璧に推定をすることが出来る SDE(AIC や BIC など) 手法は，モデルが異なる実験 2，3 では Sn が非常に高く，かつ P, Sp が低くなっている．このことから，TP, FP が多く TN, FN が少ない状態となっていることが分かる．つまり，遺伝子間の全制御関係のうちの多くを Positive と推定してしまっている．これは，ターゲット遺伝子の時系列データによく回帰するために多くの遺伝子のデータを用いていることを意味している．同様のことが Linear, Weaver モデルでも言える．つまり人工データを作成したモデルと推定に用いるモデルが等しければ完璧に推定することが出

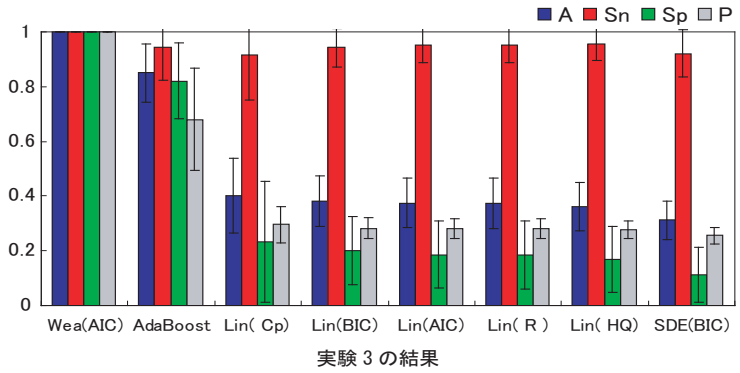


図 5.4: 実験 3 の結果（ノイズ無し）

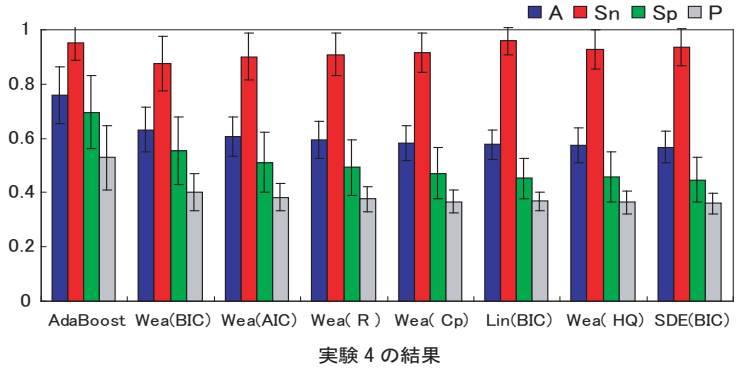


図 5.5: 実験 4 の結果（ノイズ無し）

来るが，異なる場合は著しく Sp, P が下がる．

この点について，AdaBoost は人工データのモデルの種類に関らず安定した精度を得ることが出来る手法であるといえる．つまり，どの実験でも完璧に推定できているわけではないが，全手法の中では常に上位 6 番目に位置しており，常に比較的高い精度を得ることができる．これは，4 でも述べたように，現在同一の GRN に対して様々なモデルが推定に用いられている現状，つまりターゲットとするネットワークを推定するのにどのモデルが適しているか未知である場合に AdaBoost の推定結果を利用する価値があると言える．

また，実験 4 では AdaBoost の正確度の平均は 1 番目に高い．実験 4 のような各遺伝子の時系列データを作成するモデルがそれぞれ異なるとき，つまり実データの場合は推定に適したモデルが各遺伝子によって異なるときも利用価値があると言うことができる．

実験での AdaBoost の学習を調べた結果，実験条件に関らず，AdaBoost アルゴリズムの学習

で式 4.13 の損失関数が最小になる仮説器の選ばれ方について幾つかのパターンに分類できることが分かった．以下の 5 パターンがある．

パターン A：採用された仮説器のモデルが 3 種類

パターン B：採用された仮説器のモデルが 2 種類（正しい仮説器[2]を含む）

パターン C：採用された仮説器のモデルが 2 種類（正しい仮説器を含まない）

パターン D：採用された仮説器のモデルが 1 種類（正しい仮説器を含む）

パターン E：採用された仮説器のモデルが 1 種類（正しい仮説器を含まない）

以下に各パターンについてその学習状況を述べる．

パターン A

仮説器のモデルが 3 種類とは SDE， Linear， Weaver の 3 モデルが全て選ばれるということである．ここで情報量基準の種類は考慮に入れていない．例えば SDE(AIC) と SDE(BIC) は同一と見なしている．特定の正しい仮説器・間違った仮説器がともに何度も選ばれることにより AdaBoost の重み付多数決が上手く作用し，推定精度が高くなる傾向がある．

例として実験 2 の 100 回の実験中 18 回目 ($A = 0.83, Sn = 0.85, Sp = 0.82, P = 0.61$) の遺伝子 1 の制御関係を推定したときの学習状況について，学習回数 ℓ，採用された仮説器番号 $k'(\ell)$，重み付多数決に用いる信頼度 $log(1/\beta_\ell)$ を表に示す．

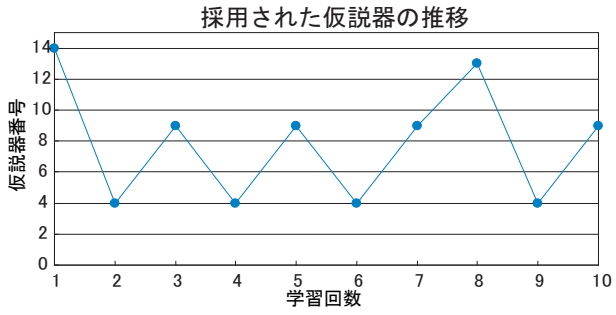


図 5.6: 学習状況，採用仮説器 No

[2]正しい仮説器とは人工データを作成したモデルと仮説器のモデルが等しいという意味で用いている．例えば実験 1 では SDE(AIC, BIC,..Cp) の 5 手法を正しい仮説器と呼ぶ

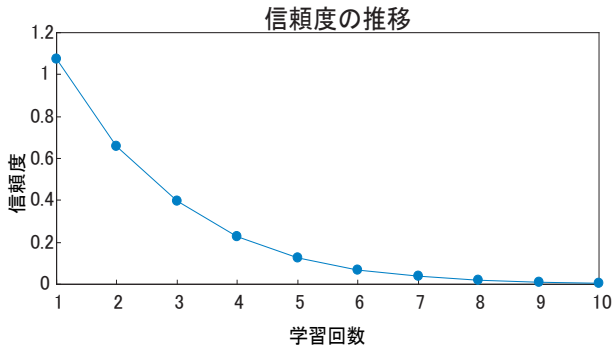


図 5.7: 学習状況, 信頼度

ここで，仮説器番号は以下のとおりである．
仮説器の番号を以下の様に定義する．
SDE　：AIC･･･1，BIC･･･2，HQ･･･3，R･･･4，Cp･･･5
Linear：AIC･･･6，BIC･･･7，HQ･･･8，R･･･9，Cp･･･10
Weaver：AIC･･･11，BIC･･･12，HQ･･･13，R･･･14，Cp･･･15

図 5.6 では仮説器 4，9，13，14 番[3] が採用されている．学習は 100 回行われたが，信頼度が 0 に近づく 10 回目まで表示している．
このときの遺伝子 1 の制御関係について

$$\textbf{正解} = [0,\ 1,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0\ ]$$

1：制御されている，　0：制御されていない [4]
である．また，採用された 5.7 中の仮説器の仮説は以下のようになっている．

$$\mathbf{h_4} = [0, 1, 1, 1, 0, 1, 0, 1, 1]$$
$$\mathbf{h_9} = [0, 1, 1, 0, 0, 0, 0, 0, 0]$$
$$\mathbf{h_{13}} = [1, 1, 1, 0, 1, 0, 1, 0, 0]$$
$$\mathbf{h_{14}} = [0, 1, 1, 0, 1, 0, 0, 1, 0]$$

ターゲットの遺伝子とモデルが異なる $\mathbf{h_4}$ は過検出の状態である．正しい仮説器の仮説である $\mathbf{h_9}$ は過不足無く正しく推定している．また，本研究の主題ではないが，モデルが同じでも情報量基準が異なると，仮説器が異なるという状況が生じることが $\mathbf{h_{13}}, \mathbf{h_{14}}$ より分かる．上記の仮説と信

[3] それぞれ SDE(R)，Linear(R)，Weaver(HQ), Weaver(R)
[4] ここでは見易さのために,「-1」と表示すべき箇所を「0」で表示している

頼度を用いて式 4.17 より，以下のように過不足無く正しく推定を行った．

$$\mathbf{H_{X_1}} = [0, 1, 1, 0, 0, 0, 0, 0, 0]$$

実験 1〜4 で同様の状況により推定の成功が見られた．

パターン B

学習状況はパターン A と類似している．採用される仮説器のモデルが 2 種類という点のみ異なる．この場合も推定精度は高くなる傾向がある．

パターン C

正しくない仮説器は図 5.2〜5.5 から分かるように Sn が高く過検出する傾向にある．よってこのパターンの場合，AdaBoost の推定結果も過検出の状態になり，精度低下の一因となっている．

パターン D

正しい仮説器は過不足無く制御関係を推定することが出来るため，このような学習の場合 AdaBoost の推定結果も完全に正しい推定結果となる．ただしこのパターンが生じる例は希である．

パターン E

パターン C と同様に正しくない仮説器は過検出の傾向になるため，AdaBoost 推定の精度低下の一因となっている．

これらのパターンの割合を表 5.2 に示す．表の数値は各実験で行ったの 900 回の AdaBoost 推定 (= 9 遺伝子 ×100 回の実験) のうちの割合を示している．

表 5.2: 各パターンの割合（ノイズ無し）

	パターン				
	A	B	C	D	E
実験 1	41.3%	26.7%	17.6%	0.8%	13.7%
実験 2	49.2%	20.4%	21.3%	0.0%	9.0%
実験 3	45.4%	41.2%	7.9%	0.0%	5.4%
実験 4	59.8%	21.3%	12.8%	0.0%	6.1%

先に述べたようにパターン A,B,D が AdaBoost 推定の精度が比較的高いパターンである．A，B，D の 3 パターンの割合の和は実験 3，4，2，1 の順に高い．また，図 5.2〜5.5 の AdaBoost の精度も同様の順に高い．つまり正確度 A が実験 3，4，2，1 の順に高い．このことから，パターン A，B，D の学習状況が生じる割合が大きいと正確度が高いという傾向があることが分かる．言い換えると，パターン A，B，D では推定が成功する傾向が高い．

5.3　人工データ実験 2：ノイズ有り

5.3.1　ノイズ有りデータの作成方法

5.2 の人工データ実験 1 で作成したデータに正規分布のノイズを加えて作成を行う．以下に作成方法を示す．

遺伝子 i の時刻 t におけるノイズを含まない発現量を $X_i(t)$，ノイズを含む発現量を $X_i'(t)$ とする．以下の式によりノイズデータを作成する．

$$X(t)' = X(t) \times \Big(1.0 + a \times N(0,1)\Big) \tag{5.6}$$

a はパラメータであり，本実験では a $= 0.05$(5 %ノイズ) とする．

5.3.2　実験結果の観察と考察

人工データ実験 1 と同様に，実験はそれぞれ 100 回ずつ行い，A, Sn, Sp, P の平均，標準偏差を調べることにより推定手法の性能を調べる．各実験の結果を図 5.8〜5.11 に示す．グラフでは 15 個の推定手法（3 モデル× 5 情報量基準) と AdaBoost の計 16 個の推定手法のうち，正確度の平均が高い上位 8 手法を左から順に示している．

表 5.3: 各パターンの割合（ノイズ有り）

	パターン				
	A	B	C	D	E
実験 1	49.6 %	46.8 %	1.4 %	1.4 %	0.8 %
実験 2	32.0 %	28.0 %	31.3 %	2.4 %	6.2 %
実験 3	41.8 %	26.9 %	23.9 %	3.1 %	4.3 %
実験 4	36.7 %	36.4 %	21.6 %	2.0 %	3.3 %

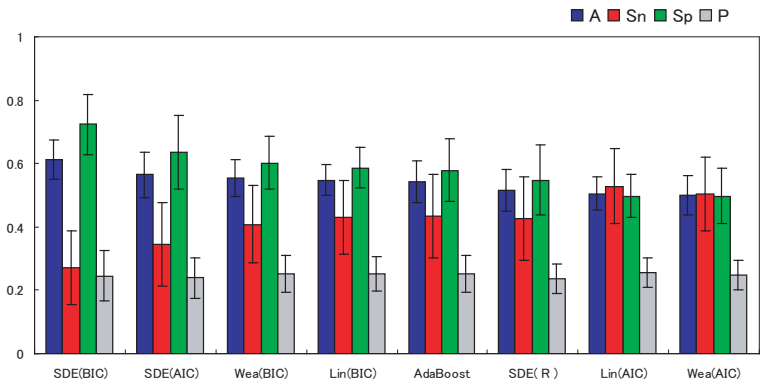


図 5.8: 実験 1 の結果（ノイズ有り）

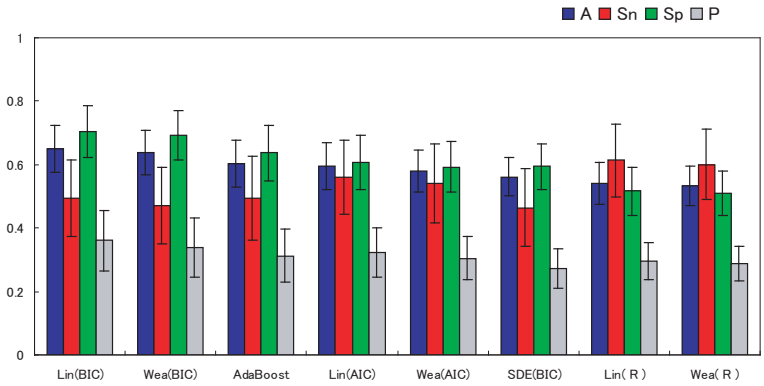


図 5.9: 実験 2 の結果（ノイズ有り）

実験 1 では，標準偏差を考慮に入れると上位 8 手法の正確度には大差がないといえる．しかし，SDE(BIC) のみ他の 7 手法に比べは Sn が低く，Sp が若干高くなっている．また，ノイズ無しの場合に比べて精度が下がっている．これはノイズのために，線形回帰でパラメータが正しく得られなかったことが原因であると考えられる．そのため，これらの仮説器の結果を用いる AdaBoost の推定結果も下がったと考えられる．実験 2，3，4 でも同様のことが言える．

実験 1〜3 では BIC を用いた正しい仮説器の正確度の平均が最も高い．また，BIC を用いた仮説器が正しい・正しくないに関らず，どの実験でも比較的上位を占めていることが分かる．この理由は BIC の式 4.4 から分から推測できる．同式は制御遺伝子の数が多いほどペナルティーが強くなる割合が他の情報量基準より大きい．そのため他の仮説器と比べて，Negative と判定する制御関係が多くなる傾向になる．ターゲットとして用いる人工ネットワークは疎らであるため，Negative と判断する傾向が高い仮説器の正確度が若干高くなる．よって，ノイズのためにオーバーフィッティングし，制御遺伝子の数が多く推定する仮説器に比べて，BIC を用いた各仮説器が比較的上位を占める結果となったと考えられる．ただし，標準偏差を考慮した場合，情報量基準による差

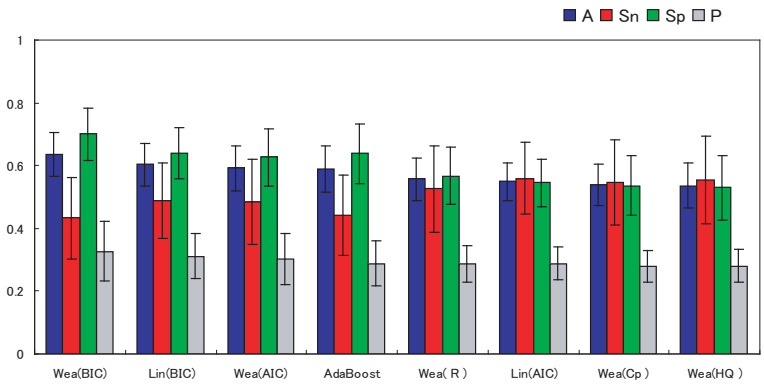


図 5.10: 実験 3 の結果（ノイズ有り）

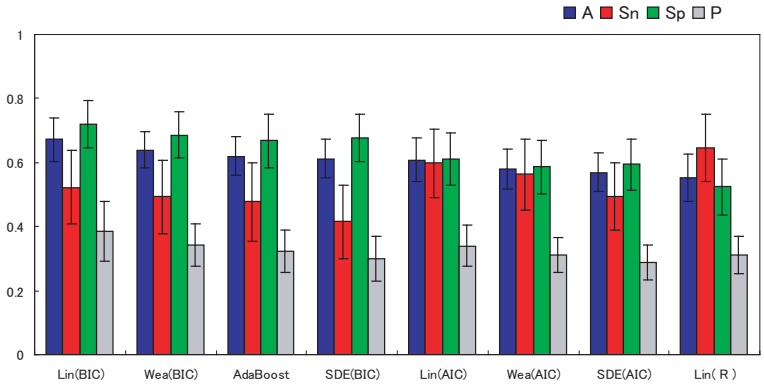


図 5.11: 実験 4 の結果（ノイズ有り）

はあまり無いといえる．

表 5.3 より，実験 1 では，実験 2，3 に比べてパターン A，B の割合が高くなっている．しかし，図 5.8 の AdaBoost の結果は図 5.9 などと同じ水準であり，ノイズ無しの実験のときに見られた関係，つまりパターン A，B の割合が高いと正確度の平均も高い傾向にあるという関係が見られない．これは，仮説器の採用パターンが A あるいは B であっても，正しい仮説器の推定結果の精度が低い傾向にあることが原因と考えられる．

5.4　実データを用いた実験

大腸菌の DNA 修復機構である SOS ネットワーク図 5.12(再掲) をターゲットとして推定実験を行った．SOS ネットワークは生物学実験によって詳細に解明が進んでいるネットワークの 1 つである．測定に用いるデータは Ronen らによって得られたデータであり [18]，SOS ネットワーク内の主要な 8 個の遺伝子の発現時系列データからなる．時系列データは lexA のタンパク質である LexA を放射線で破壊してから 6 分毎に各遺伝子の発現レベルを 300 分間測定した時系列数 50 のデータから成る．放射線の強さによって 4 条件で測定された時系列データからなる (条件 1，2 は $5Jm^{-2}$，条件 3，4 は $20Jm^{-2}$) ．

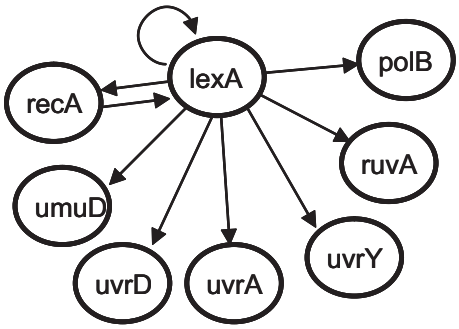


図 5.12: 大腸菌 DNA 修復機構の遺伝子制御ネットワーク (再掲)

5.4.1　データの適用方法

放射線の強さによって時系列データは 4 種類ある．これらの情報を有効に用いるために以下のような方法で仮説器・AdaBoost にデータを適用する．

4.3.1 の仮説器の推定方法と出力のところで，情報量基準 (IC) が最小になるような組合せを推定すると述べた．今回は以下の IC が最小になるような組合せをターゲット遺伝子を制御している遺伝子として推定する．

$$IC = IC_1 + IC_2 + IC_3 + IC_4 \tag{5.7}$$

IC_i：条件 i の時系列データを回帰して得られた情報量基準

4 条件の時系列データを全て同一にみる．つまり学習データの数は 200(=4 条件 $\times$50 サンプル) である．また，[13] では SDE モデルで図 5.12 の SOS ネットワークを推定し，最尤推定により得られた式 3.5 のパラメータ $C_{i,j}$ のうち絶対値の小さい遺伝子は制御していないとみなすというヒューリスティクスを用いることにより主要な制御関係を検出し，同時に間違った制御関係を削除する

ことに成功している．これは時系列データにノイズが含まれると考えられているため用いられている[5]．本実験で用いる SDE の各手法もこのヒューリスティクスを用いる．その詳細は**付録** D に示す．

5.4.2　実験結果の観察と考察

推定結果を正確度 A の高い順にに示す．図 5.13 が上位 8 手法，図 5.14 に下位 8 手法を示している．

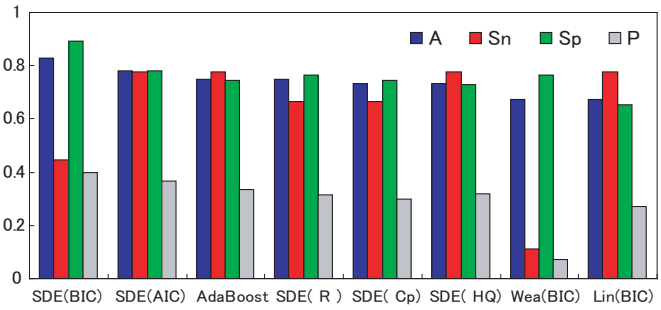


図 5.13: 各手法の精度，(正確度順の上位 8 手法)

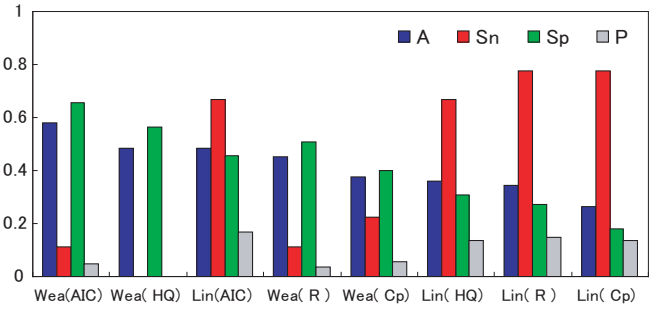


図 5.14: 各手法の精度 (正確度順の下位 8 手法)

次に，上位 5 手法である SDE(BIC)， SDE(AIC)， AdaBoost， SDE(R)， SDE (Cp) によって推定されたネットワークを図 5.15〜5.18 示す．

図 5.13 からわかるように AdaBoost は 3 番目に正確度が高いことが分かる ($A = 0.75, Sn = 0.78, Sp = 0.75, P = 0.33$)．また，SDE(BIC) は $A = 0.83, Sn = 0.44, Sp = 0.89, P = 0.40$ であり，感度 Sn が低いにも拘らず正確度が最も高い．これはネットワークが疎らであり，かつ SDE(BIC) の結果の TrueNegative を大きいためである．ただし的中率 P を見ると両者の差は 0.07

[5]予備実験として Linear,Weaver モデルでも同様のヒューリスティクスを行ったが，精度が低下したため両モデルについては行わない．

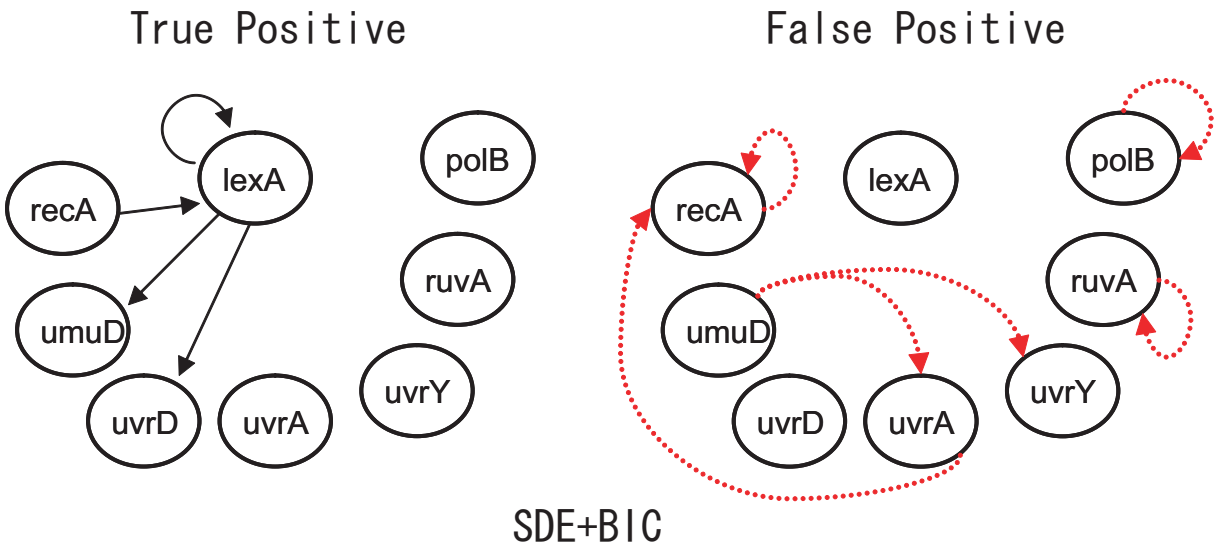


図 5.15: SDE(BIC) によって推定されたネットワーク

であり，検出力の正確さは両者にあまり差は無いと言える．正確度が 2 番目に高い SDE(AIC) は $A = 0.78, Sn = 0.78, Sp = 0.78, P = 0.37$ であり，AdaBoost とほぼ同程度の検出力と言える．

これらの結果より，どの手法の推定結果を用いるべきか分からないとき，AdaBoost の結果を推定結果として利用する価値はあると言える．つまり本実験では，SDE モデルが推定に適しているという情報を得ていない状況で，提案手法を利用する価値があると考えられる．

次に，AdaBoost 推定の学習で採用された仮説器について調べた．各遺伝子での学習状況について，採用された仮説器の番号と信頼度をグラフに示す（図 5.20〜5.27）．

実データでは，人工データ実験のように仮説器が正しいかどうかについて，一概に言うことは出来ない．しかし，15 手法の中で比較的精度の高い SDE モデルによる推定手法を正しい仮説器としたとき，8 個の遺伝子全ての場合で AdaBoost の学習での仮説器の採用パターンは A 又は B であることが図 5.20〜5.27 より分かる．つまり採用された仮説器のモデルは 3 又は 2 種類であり，かつその中に正しい仮説器が含まれていた．これは AdaBoost による推定精度が 15 手法の中で比較的高い理由ではないかと考えられる．

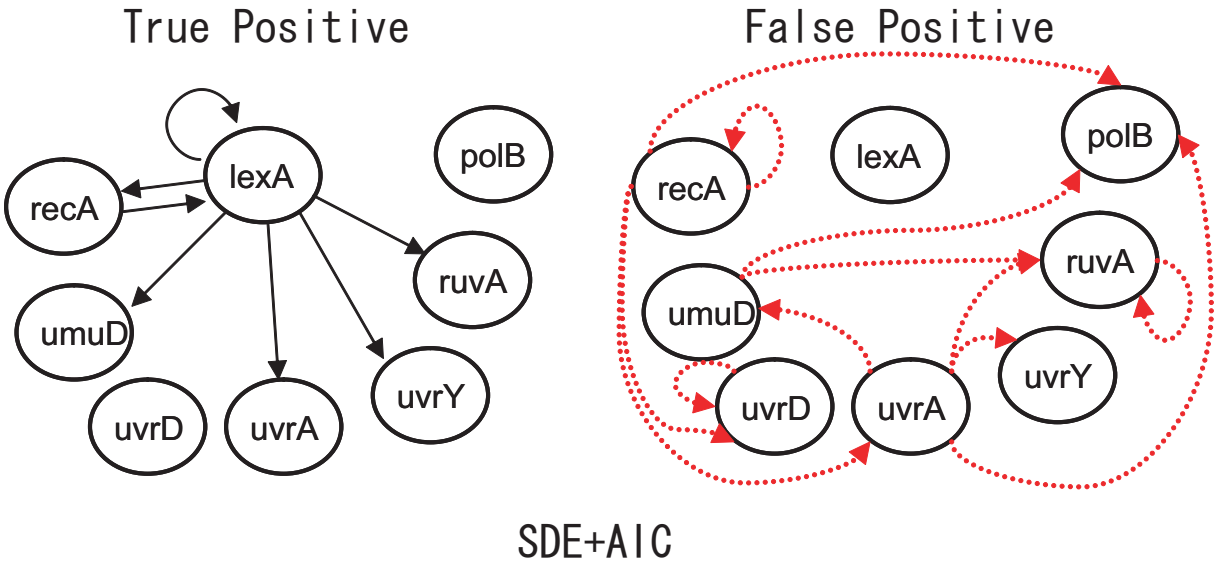


図 5.16: SDE(AIC) によって推定されたネットワーク

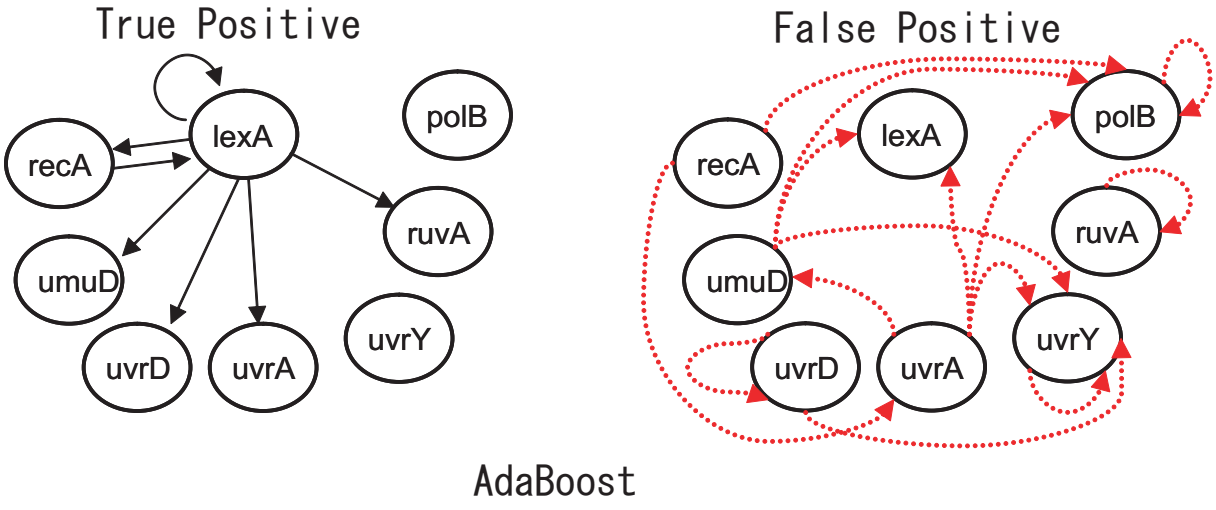


図 5.17: AdaBoost によって推定されたネットワーク

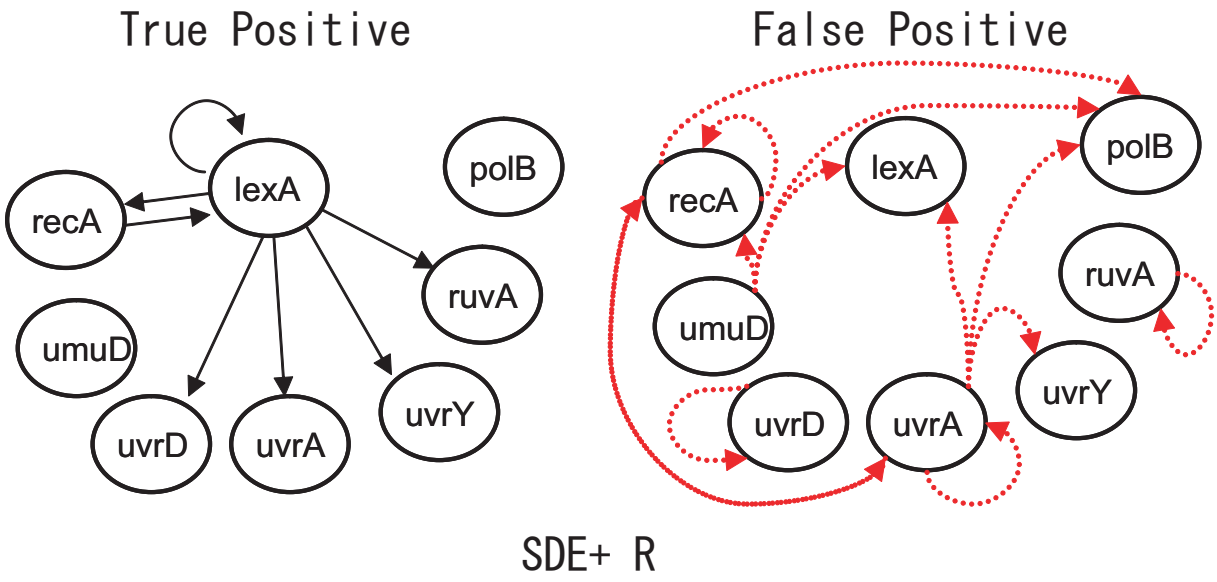


図 5.18: SDE(R) によって推定されたネットワーク

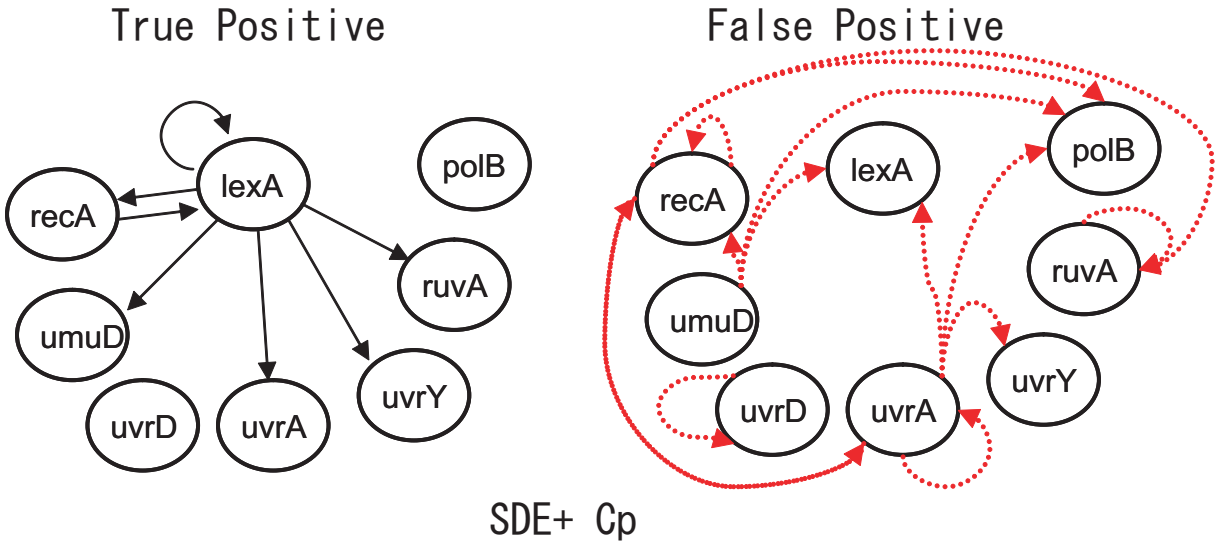


図 5.19: SDE(Cp) によって推定されたネットワーク

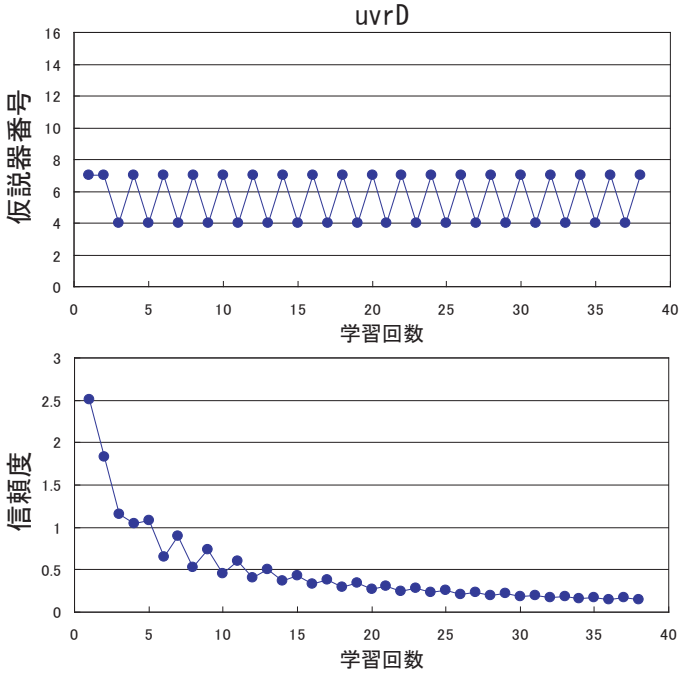


図 5.20: uvrD の制御関係推定時の学習状況

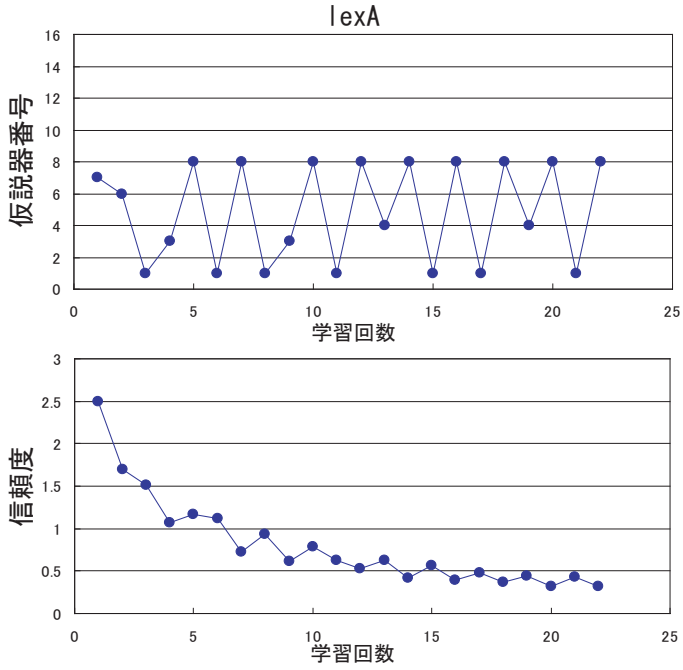


図 5.21: lexA の制御関係推定時の学習状況

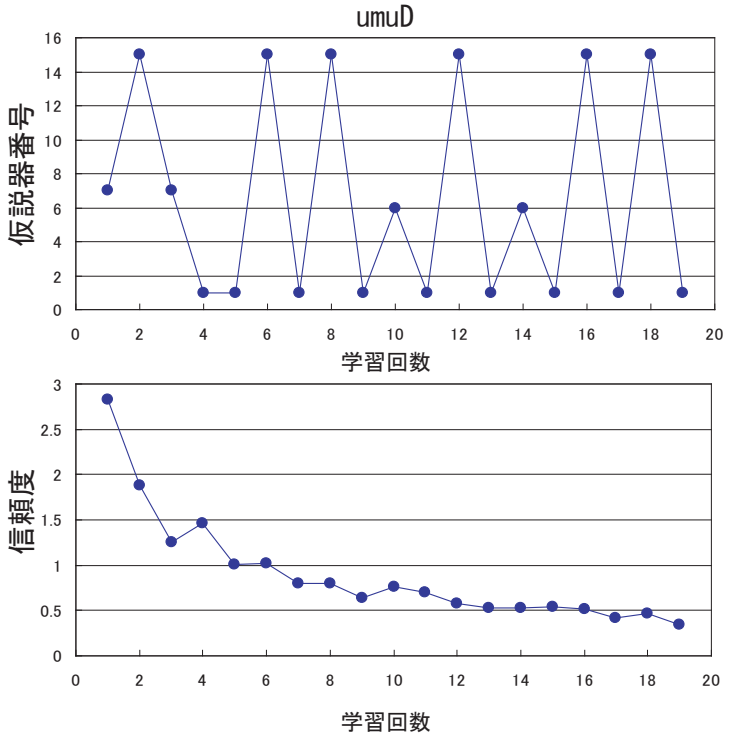


図 5.22: umuD の制御関係推定時の学習状況

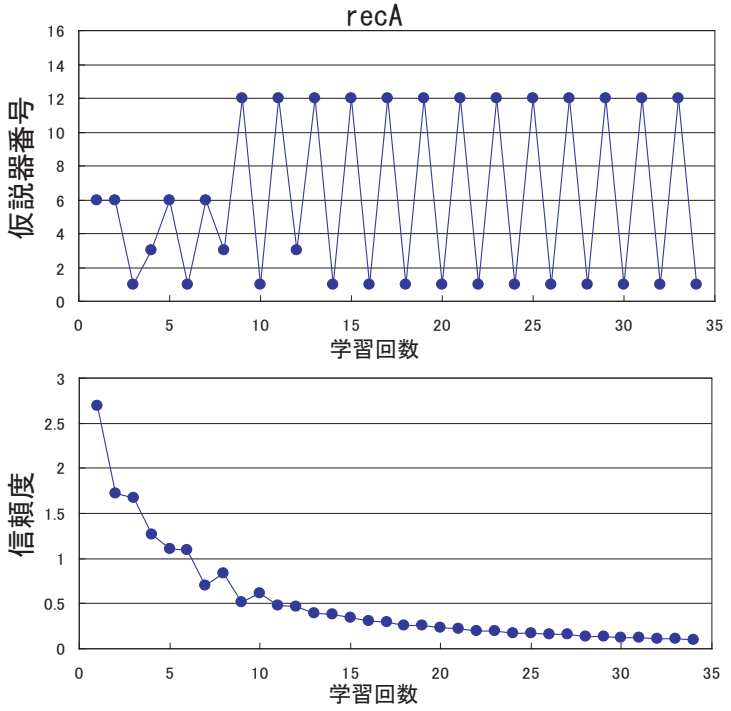


図 5.23: recA の制御関係推定時の学習状況

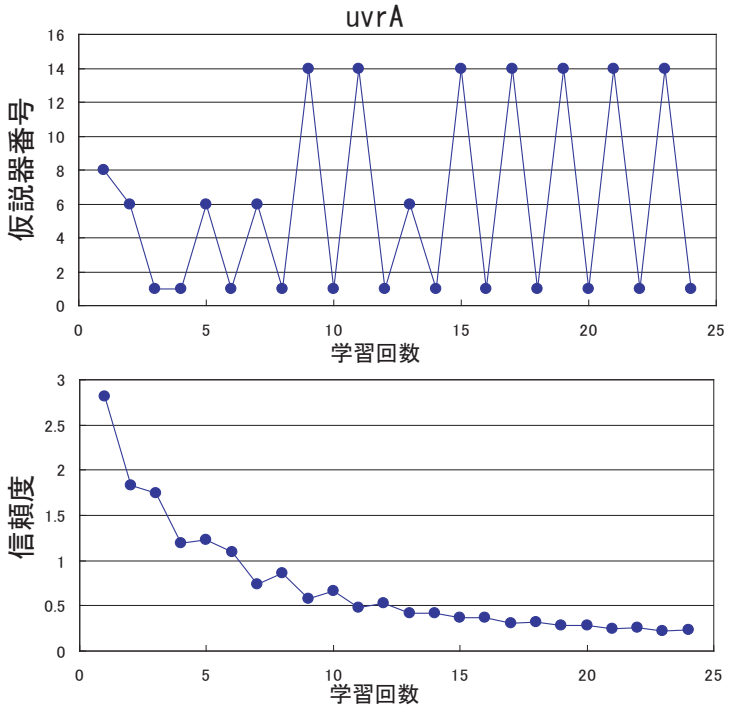


図 5.24: uvrA の制御関係推定時の学習状況

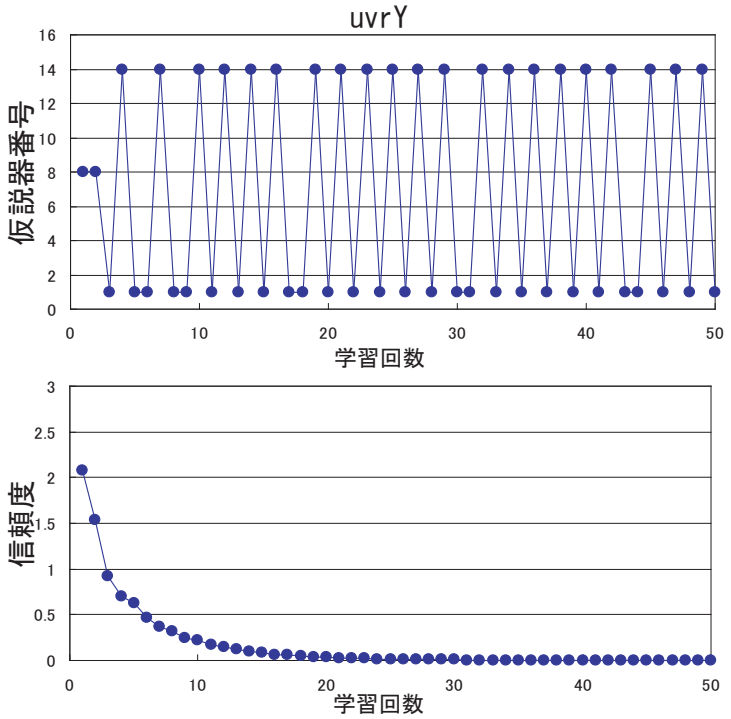


図 5.25: uvrY の制御関係推定時の学習状況

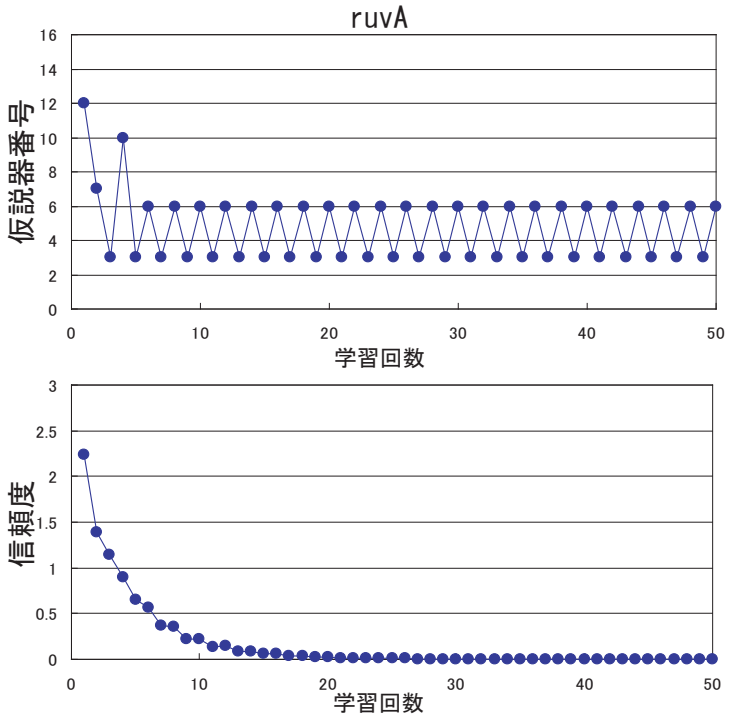


図 5.26: ruvA の制御関係推定時の学習状況

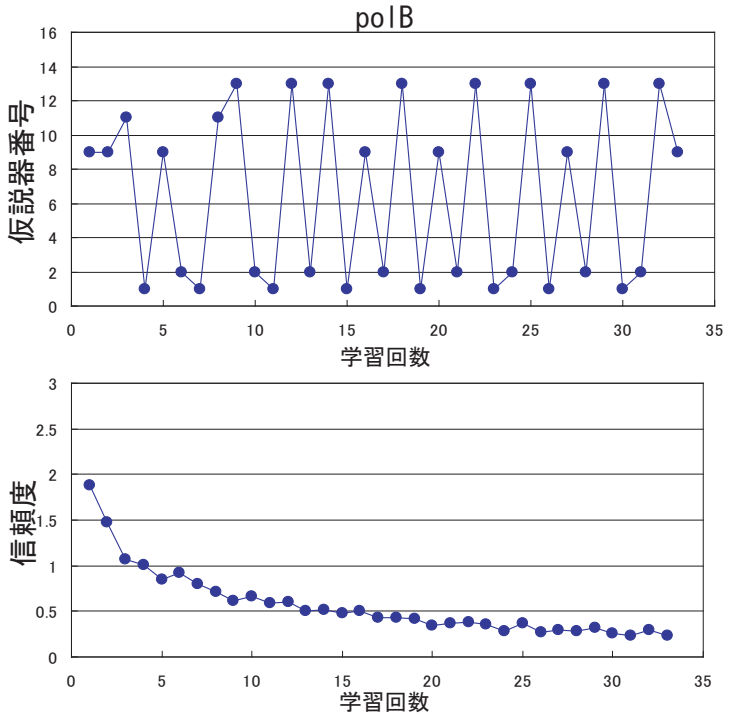


図 5.27: polB の制御関係推定時の学習状況

第6章

考察

本章では提案手法の利点・欠点などについて複数の観点から考察を行う．

6.1 ターゲットネットワークの構造について

人工データ実験として一種類のターゲットネットワークについての結果しか示さなかった．しかしながら，このネットワークの構造は GRN の特徴である制御関係の少なさ，つまり疎なネットワークという特徴を取り入れており，十分に一般性があると考えられる．また，実際の AdaBoost 推定では各遺伝子について，回帰することにより推定している．つまり，ネットワークの構造は推定には関係ない．この点からも人工データ実験で得られた結果は一般性があると考えられる．

6.2 ノイズについて

ノイズを含まない測定データを用いた人工データ実験 1 では，どのモデルが適しているか不明であるときに，AdaBoost を用いた提案手法を利用することに利点があることが述べた．しかし，ノイズを含む人工データ実験 2 では，提案手法の精度は仮説器と比べてあまり差が無かった．ただし，仮説器より低くいわけではなく，ほぼ同水準であり，劣っているわけではない．この原因として，正しい仮説器の精度が低いことが考えられる．言い換えると，パラメータ推定として用いる線形回帰がノイズに弱いため，AdaBoost の精度も他の仮説器と同水準になったと言える．この問題を解決する方法としては，線形回帰ではなく，進化論的計算によるパラメータ推定を用いることが上げられる．同手法では S-system モデルで，ノイズを含むデータからのパラメータに推定にほぼ成功しているため，提案手法の精度向上に繋がると考えられる [14].

6.3 計算量について

今回の提案手法の場合は仮説器の推定方法における計算量に問題がある．4.3.1 仮説器の推定方法の **Step2** で回帰する遺伝子について，遺伝子の全ての組合せについて情報量基準を計算する．つまり全数探索を行う．ターゲットとする GRN の遺伝子の数を n とするとその組合せの数は 2^n 通りある．これは大規模な GRN を推定するときは現実的でないことを意味する．よってこれを解決するために全数探索ではなく組合せ最適化問題に用いる何らかの探索手法を代用する必要があると考えられる．

しかし，別な観点からみると提案手法は計算量に有利な点を持つ．提案手法のアルゴリズムから分かるように，AdaBoost 自体の計算量は比較的小さい．実際，仮説器の推定結果を用いるのみで，学習過程で仮説器に何度も推定させる操作，言い換えると計算量が指数関数的に増加するような操作は無い．つまり一度得られた推定結果は AdaBoost の統合推定に用意に組み込むことが出来る．例えば既に発表されている論文の推定結果を統合できる．これは非常に大きい利点で

あると考えられる．さらに仮説器同士は別々に計算することができる．つまり並列計算が可能である．大規模な遺伝子制御ネットワーク推定にも有望であると考えられる．

6.4　モデルの種類について

評価実験の結果より，推定ターゲットとする GRN の推定に適したモデルの種類に関らず，提案手法は安定した推定精度を獲得できることが利点であることを述べた．しかしこのことを主張できるのはターゲット GRN の推定に適したモデルが SDE，Linear，Weaver モデルであるときに限る．推定に適したモデルを AdaBoost で使用していない場合，つまり推定に適した仮説器が無い場合は精度が著しく下がると考えられる．この問題を解決する方法として，用いるモデルを増やすことが考えられる．つまり S-system など他のモデルを追加することにより，ターゲットの推定に適したモデルを仮説器として使用している可能性を高めることが妥当であると考えられる．今後は他のモデルを追加したときの AdaBoost 推定能力の変化を調べることを予定している．

第7章

結論

本論文は以下の 3 点において GRN 推定研究に貢献していると考えられる．

- 複数の漸化式・微分方程式モデルによる推定方法の AdaBoost による統合
- 生成したモデルの種類に依らない高精度の推定結果の達成
- 実際の生物データでの有効性の検証

本論分では遺伝子制御ネットワーク推定研究の問題点として，対象とする発現量時系列データに対してどのモデルが推定に適しているか不明である点を挙げた．この問題に対して，複数の漸化式・微分方程式モデルによる推定結果を AdaBoost により統合する方法を提案した．評価実験により，提案手法は推定に適したモデルの種類に関らず，安定した推定精度を獲得できることが分かった．さらに実データでその有効性を確認することが出来た．

付録

A Linear モデルによる推定方法

最尤法によりパラメータを求める．

Step1　フィッティングする遺伝子の選択

遺伝子 X_i を選択し，ベクトル Y_i を以下のように定義する．

$$\mathbf{Y_i} = \left[\begin{array}{ccc} X_i(0) & \cdots & X_i(m-1) \end{array} \right]^T \in \Re^{n \times 1} \tag{A.1}$$

Step2　全組合せについて情報量基準の計算

n 個の遺伝子のうち任意の遺伝子を任意の個数だけ選択する．この組合せが遺伝子 X_i を制御していると仮定する．ここでは遺伝子を 3 個任意に選んだとする．これを X_j, X_k, X_l とし，

$$\mathbf{U} = \left[\begin{array}{cccc} 1 & X_j(0) & X_k(0) & X_l(0) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_j(m-1) & X_k(0) & X_l(m-1) \end{array} \right] \in \Re^{m \times 4} \tag{A.2}$$

X_j, X_k, X_l の時系列データを用いて遺伝子 i の時系列データにフィッティングするようにパラメータを求める．

$$\mathbf{Y_i} = \mathbf{UC} + \sigma \mathbf{Z_i} \tag{A.3}$$

ここで $\mathbf{C}, \mathbf{Z_i}$ はそれぞれ

$$\mathbf{C_i} = \left[\begin{array}{cccc} C_{i,0} & C_{i,j} & C_{i,k} & C_{i,l} \end{array} \right]^T \in \Re^{4 \times 1} \tag{A.4}$$

$$\mathbf{Z_i} = \left[\begin{array}{ccc} Z_i(0) & \cdots & Z_(m-1) \end{array} \right]^T \in \Re^{m \times 1} \tag{A.5}$$

である．$\mathbf{C}$ は求めるパラメータ，$\mathbf{Z_i}$ は互いに独立な $N(0,1)$ の正規分布のノイズである．これより対数尤度を求める．

$$logL = -\frac{m}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{UC})(\mathbf{Y} - \mathbf{UC})^T \tag{A.6}$$

尤度が最大になるようにパラメータを求める．つまり，

$$\frac{\partial logL}{\partial \mathbf{C}} = 0, \frac{\partial logL}{\partial \sigma^2} = 0 \tag{A.7}$$

となるような $\mathbf{C}, \sigma^2$ を求める．これを解いて

$$\hat{\mathbf{C}} = (\mathbf{U^T U})^{-1}\mathbf{U Y} \tag{A.8}$$

$$\hat{\sigma} = (\mathbf{Y} - \mathbf{U}\hat{\mathbf{C}})(\mathbf{Y} - \mathbf{U}\hat{\mathbf{C}})^T/2 \tag{A.9}$$

を得る．

$$\hat{logL} = -\frac{m-1}{2}log(2\pi\hat{\sigma^2}) - \frac{m-1}{2} \tag{A.10}$$

これより，組合せ X_j, X_k, X_l の情報量基準を計算する．ここでは AIC を例に取る．

$$AIC = -2\hat{logL} + 2(n' + 1) \tag{A.11}$$

ここで，遺伝子 i を 3 個の遺伝子が制御していると仮定しているので $n' = 3$ である．これを全ての組合せについて行う．

Step3　推定

情報量基準が最小となった組合せに含まれるが遺伝子 i を制御していると推定する．

B　Weaver モデルによる推定方法

Linear モデルと同様に最尤推定によりパラメータを求める．

Step1 の　フィッティングする遺伝子の選択

で，遺伝子 X_i を選択し，$S_i(t)$ を計算する．

$$S_i(t) = log\Big(\frac{M_i}{X_i(t+1)} - 1\Big) \tag{B.12}$$

$$\mathbf{Y_i} = \Big[\ S_i(0) \quad \cdots \quad S_i(m-2)\ \Big]^T \in \Re^{n\times 1} \tag{B.13}$$

と定義する．X_i を制御している遺伝子が X_j, X_k, X_l と仮定して式 A.2 を構成する．これを用いて式 3.4 の関係より，A.3 が成り立つ．以後は Linear モデル時と同様にしてパラメータを計算し，情報量基準が最小になる組合せを求める．4.3.3 で用いる $\hat{X}(t)$ は，回帰して得られた $\hat{S}_i(t-1)$ から式 3.2 を用いて得る．

C　SDE モデルによる推定方法

Linear モデルと同様に最尤推定によりパラメータを求める．

$$\frac{\Delta X}{\sqrt{\Delta t}} = \Big[C_{i,0} + \sum_{j=1}^{n} C_{i,j} f_(X_j(t))\Big] + \sigma\mathbf{Z(t)} \tag{C.14}$$

と置く．

$$Y_i(t) = \frac{X_i(t+1) - X_i(t)}{\sqrt{\Delta t}} \tag{C.15}$$

$$\mathbf{U}_(\mathbf{t}) = \left[\sqrt{\Delta t} \;\; \sqrt{\Delta t} f(X(t)) \;\; \cdots \;\; \sqrt{\Delta t} f(X(t)) \right]^T \tag{C.16}$$

とする．A.3 の関係が成り立つので．以後は Linear モデルと同様にしてパラメータを計算し，情報量基準が最小になる組合せを求める．さらに，回帰して得られた $\hat{Y}_i(t)$ から最小二乗法を用いて $\hat{X}_i(t)$ を得る．

D SDE のヒューリスティクス

最尤推定で得られた $C_{i,j}$ の絶対値が小さいパラメータ，つまりその制御関係は削除する．例として遺伝子 2 について SDE(AIC) でパラメータを求めた結果，遺伝子 1，2，5，7 の組合せが AIC を最小となった場合を挙げる．この時得られるパラメータは $C_{2,0}, C_{2,1}, C_{2,2}, C_{2,5}, C_{2,7},$ である．ここで以下の式を満たさない制御関係は削除する．

$$C_{2,k} < r \times \bar{C}_2, (k = 1, 2, 5, 7) \tag{D.17}$$

ただし， $\bar{C}_2 = \sum_{k=1,2,5,7} |C_{2,k}|$ である．
実験では $r = 0.5$ としている．

謝辞

本研究を進めるにあたり，多くの方からご指導，ご鞭撻を賜った．

伊庭斉志教授には本論文を作成する上で貴重な御指導を賜りった．ヒューマノイドロボット研究を行う予定を翻し，バイオインフォマティクス研究に変更するという我が儘を容認してもらった．全くの新分野での研究は絶望の壁の連続だった（少なくともそう感じた）．特に修士課程一年生時の生活は常に危機感・焦燥感を伴うものであった．しかし，それは常に学び続ける習慣や，どんな厳しい状況でも光を見つけてそれをものにする姿勢を確立する助けとなった．また，株価予測という新しい分野に出会うチャンスを頂いたこと，これは様々な人と出会うきっかけとなった．伊庭研究室という，熱心な放任主義体制 (？) の下で得たこれらの経験は私の貴重な財産である．伊庭斉志教授に謝意を表す．

Masa Planning 代表の佐藤正俊氏には宴会やメール・電話などでしばしば，経験に基く貴重なご助言を頂いた．さらに，就職活動時に相談にも乗っていただいた．私の進路を決めるにあたり大きな助けとなった．

秘書の島津美和氏にはアルバイト代の事務処理を行っていただいた．

博士課程 3 年の Topon Kumar Paul 氏，Nasimul Noman 氏には学会論文の英語チェックで時間を割いていただいた．特に Noman 氏には研究テーマが同じということで，幾度も有益な文献を教えて頂いた．

同じく博士課程 3 年の神尾正太郎氏にはコンピュータ環境について相当迷惑をかけた．論文の添削，さらに研究の相談に乗って頂くなど大変御世話になった．

博士課程 2 年の長谷川禎彦氏にも私の研究に議論や論文に時間を割いてい頂いた．同学年の伊東和紀氏，Deniz Aydemir 氏，柳瀬利彦氏，矢吹崇宏氏のおかげで学生生活を楽しく過ごすことが出来た．これらの方々に謝意を表す．また，ここに書ききれなかった伊庭研究室の学生の方々にも感謝申し上げる．

最後に，私の 6 年間の学生生活を精神的・経済的にも支えていただいた私の家族に感謝する．

参考文献

[1] T. Akutsu, S. Miyano, and S. Kuhara. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. *Proc. of Pacific Symp. Biocomputing'99*, Vol. 1, pp. 80–1000, 1999.

[2] S. Ando and H. Iba. Quantitative modeling of gene regulatory network: Identifying the network by means of genetic algorithms. In *Proc. of Genome Informatics Workshop*, pp. 278–280, 2000.

[3] KC. Chen, TY. Wang, HH Tseng, F. Huang, CY, and CY. Kao. A stochastic differential equation model for quantifying transcriptional regulatory network in saccharomyces cerevisiae. *Bioinformatics*, Vol. 21, pp. 2883–2890, 2005.

[4] H. Drucker. Improving regressors using boosting techniques. In *Proc. of Machine Learning conference.*, pp. 107–115, 1999.

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

[6] Y. Freund and E. Schapire, R. A decision-theoretic generalization of on-line learning and an application to boosting. *Computer and System Sciences*, Vol. 55, pp. 119–139, 1997.

[7] P. Geurts, M. Fillet, D. Seny, d, MA. Meuwis, M. Malaise, MP. Merville, and L. Wehenkel. Proteomic mass spectra classification using decision tree based ensemble methods. *Bioinformatics*, Vol. 21, pp. 3138–3145, 2005.

[8] D'. Haeseleer, P, X. Wen, S. Fuhrman, and R. Somoyogyi. Linear modeling of mrna expression levels during cns development and injury. In *Proc. of Pacific Symp. Biocomputing*, Vol. 145, pp. 41–52, 1999.

[9] H. Iba. Bagging, boosting, and bloating in genetic programming. In *Proc. of the Genetic and Evolutionary Computation Conference, Morgan Kaufmann,*, pp. 1053–1060, 1999.

[10] S. Kim, S. Imoto, and S. Miyano. Inferring gene networks from time series microarray data using dynamic bayesian networks. *Briefings in Bioinformatics*, Vol. 4, pp. 228–235, 2003.

[11] S. Kimura, K. Ide, A. Kashihara, M. Kano, M. Hatakeyama, R. Masui, N. Nakagawa, S. Yokoyama, S. Kuramatsu, and A Konagaya. Inference of s-system models of genetic networks using a cooperative coevolutinary algorithm. *Bioinformatics*, Vol. 21, pp. 1154–1163, 2005.

[12] Y Maki, D. Tominaga, M. Okamoto, S. Watanabe, and Y. Eguchi. Development of a system for the inference of large scale genetic networks. pp. 446–458, 2001.

[13] S. Nabatame and H. Iba. Estimation of gene regulatory network using stochastic differential equation model. In *Proc. of Genome Informatics Workshop*, pp. 150–151, 2005.

[14] N. Noman and H. Iba. Reverse engineering genetic networks using evolutionary computation. In *Proc. of Genome Informatics Workshop*, pp. 205–214, 2005.

[15] H.Iba N.Sugimoto. Inference of gene regulatory networsk by means of dynamic differential bayesian networks and nonparametric regression. In *Proc. of Genome Informatics Workshop*, 2004.

[16] I. Ono, Y. Seike, R. Morishita, N. Ono, M. Natsui, and M. Okamoto. An evolutionary algorithm taking account of mutual interactions among substances for inference of genetic networks. In *Proc. 2004 Congress on Evolutionary Computation*, pp. 2060–2067, 2004.

[17] E. Perrin, B, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. d'Alche Buc. Gene networks inference using dynamic bayesian networks. *Bioinformatics*, Vol. 19, pp. ii138–ii148, 2003.

[18] M. Ronen, R. Rosenberg, I. Shraiman, B, and U. Alon. Assigning numbers to the arrows: Parameterizing a gene regulation network by using accrate expression kinetics. *PNAS*, Vol. 99, pp. 10555–10560, 2002.

[19] A. Savageau, M. *Biochemical systems analysis: a study of function and design in molecular biology.* Addison-Weslye Pub. Co., 1976.

[20] C. Weaver, D, T. Workman, C, and D. Storm, G. Modeling regulatory networks with weighted matrices. In *Proc. on Pasific Symposium on Biocomputing*, pp. 112–123, 1999.

[21] 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.

[22] 小西貞則, 北川源四郎. 予測と発見の科学シリーズ 2，情報量基準. 朝倉書店, 2004.

発表文献

学術雑誌

[J1] 生田目慎也, 伊庭斉志. AdaBoost を用いた遺伝子制御ネットワークの統合的推定. 人工知能学会誌（査読中)

国際会議発表論文

[C1] Shinya Nabatame and Hitoshi Iba. Inference of gene regulatory network using stochastic differencial equation model. In *The 15th International Conference on Genome Informatics(GIW 2005)*, Vol. 16, pp. 150-1 – 150-2, 2005.

[C2] Shinya Nabatame and Hitoshi Iba. Integrative estimation of gene regulatory network by means of AdaBoost. In *The 15th International Conference on Genome Informatics(GIW 2006)*,Vol. 17, pp. 119-1 – 119-2, 2006.

[C3] Shinya Nabatame and Hitoshi Iba. Estimation of gene regulatory network by means of AdaBoost. In *2006 Annual Meeting of Chem-Bio Informatics Socitey*, pp. 74, 2006.

研究会 · シンポジウム等発表論文

[M1] 生田目慎也, 伊庭斉志. 確率微分方程式を用いた遺伝子制御ネットワークの推定. “システムバイオロジー研究のための数理科学的手法”’ バイオインフォマティクス夏の学校 2005, 2005.