細胞分化における時系列発現プロファイルを用いた遺伝子制御ネットワーク推定手法に関する研究

提出先　　大阪大学大学院情報科学研究科

提出年月　2014年1月

中 山　智 義

関連研究論文

1. 査読のある学術論文

   1-1 Tomoyoshi Nakayama, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda. Inference of S-system models of gene regulatory networks using immune algorithm, *Journal of Bioinformatics and Computational Biology*, Vol. 9, No.Suppl. 1, pp.75-86, 2011.

2. 査読のある国際会議

   2-1 Tomoyoshi Nakayama, Hiromi Daiyasu, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda. Reconstruction of dynamic gene regulatory networks for cell differentiation by separation of time-course data, *2013 International Conference on Bioinformatics and Computational Biology (BIOCOMP'13)*, Las Vegas, Nevada, USA, July 25, 2013.

   2-2 Tomoyoshi Nakayama, Yoshiyuki Kido, Hiromi Daiyasu, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda. Estimate Dynamic Gene Regulatory Networks in Adipocyte Differentiation for Detecting Changes of Gene Regulations by Splitting Time Course Data, *23rd International Conference on Genome Informatics (GIW2012)*, Tainan, Taiwan, December 13, 2012

   2-3 Tomoyoshi Nakayama, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda. An Estimation Method of S-system Model of Gene Regulatory Networks Using Immune Algorithm, *18th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB2010)*, Boston, USA, July 12, 2010.

内容梗概

細胞分化はヒトなどの多細胞生物において，細胞が特殊化して分化前の細胞には見られない特有の形質を持つようになる過程のことをいう．多細胞生物は 1 個の受精卵から細胞分裂と細胞分化を繰り返し，細胞数を増やして様々な組織，器官，臓器を形成する．通常，細胞分化は不可逆であり，受精卵の状態では全ての組織へ分化できる全能性を持つが，分化が進むと分化できる細胞の種類が決まってしまう．この細胞分化や分化可能な細胞を決定づけるのは主に遺伝子の発現の変化だと分かってきている．近年，iPS 細胞に始まる再生医療の発展により，細胞分化のメカニズムを解明する要求が高まっており，遺伝子の制御関係を記述した遺伝子制御ネットワークによって細胞分化中の遺伝子の発現の様子を解析することが求められている．

本研究では，細胞分化における動的な遺伝子制御ネットワークに着目し，遺伝子の発現と制御の理解を目指す．細胞分化過程を支配する働きを持つ遺伝子やタンパク質などは同定されてきており，数個の遺伝子が転写制御を司っていることが分かっている．それらの遺伝子に刺激を与えて細胞分化を誘導することが可能になり，細胞分化中の遺伝子の発現量を計測した時系列発現プロファイルが計測できるようになってきたため，遺伝子制御ネットワークを用いての解析が可能になってきている．また，遺伝子制御ネットワークを推定する手法は様々考案されており，目的に応じて使い分けることができるようになっている．しかし，細胞分化中に細胞内の全遺伝子の働きや制御関係がどのように変化するか，といった網羅的な視点での動的な変化への解析は十分とはいえない．また，分化中において細胞分化を司る遺伝子の上流部分の発現量がどのくらい変化すると下流の遺伝子の発現はどのような動態を示すのか，といった定量的な動的な性質は不明な点が多く残されている．本研究では上記 2 点の異なる動性に焦点を絞り，大規模な定性的な遺伝子制御ネットワーク解析と小規模の定量的な遺伝子制御ネットワーク解析といった 2 つの手法によって細胞分化過程の動性を理解できると考え，それぞれの手法について研究を行い，組み合わせることによって，網羅的に動的な性質を理解することを目指した．

小規模な遺伝子群で定量的な遺伝子制御ネットワークを解析する際には S-system モデルを適用して解析することが行われているが，探索アルゴリズムが局所解に収束してしまう問題点が挙げられる．S-system モデルは微分方程式モデルで，適切なパラメータと初期値を与えることで時系列発現プロファイルの再現をすることができる．一般には対象の系のパラメータが分からないため，対象の遺伝子群で構成された時系列発現プロファイル

を用いてパラメータを推定する逆問題を解くことになる．その際，遺伝的アルゴリズム（GA）を用いることが一般的だが，GA の探索方法では一度解候補が局所解に収束してしまうと脱出することができない．本研究では，前述の問題点を解決するため GA を拡張した免疫アルゴリズム（IA）を適用させることで局所解に収束した際に解候補を初期化する手法を提案した．脂肪細胞分化の時系列発現プロファイルを用いて本手法の評価実験を行い，よりパラメータ推定精度の高い手法であることを確認した．

大規模な遺伝子群の定性的な遺伝子制御の変化を解析する手法として，ノードセットセパレーション法があるが，時系列発現プロファイルを全時点使用しており，細胞分化では時系列発現プロファイル中に制御関係の変化が起こっていると考えられ，時点数の多い時系列発現プロファイルを用いると制御関係の変化を推定できない問題点がある．ノードセットセパレーションでは各時点で発現量の高い遺伝子群で遺伝子制御ネットワークを推定するが，全時点を用いて推定をすると制御関係の変化による 2 つの制御関係があることを推定することができない場合がある．提案手法ではノードセットセパレーション法の遺伝子方向で分割するところをスライディングウィンドウ法によって時間方向に分割し，複数の部分時系列発現プロファイルからそれぞれのダイナミックベイジアンネットワークモデルを推定することによって動的な遺伝子制御ネットワークを推定する手法を提案した．本手法を細胞分化の時系列発現プロファイルに適用し，小規模な遺伝子セットで既知関係の推定精度，大規模な遺伝子セットでネットワークの推定精度を確認することにより有効性を示した．そして，得られた結果に小規模な遺伝子制御ネットワークの振る舞いの解析結果を当てはめることで，網羅的な推定手法と定量的な解析手法の整合性を確認した．これによって，遺伝子制御ネットワークの中で最も発現制御に寄与している遺伝子を抽出することができるようになり，網羅的な遺伝子制御ネットワーク推定結果から細胞分化に寄与する未知の遺伝子の発見が期待できる．

目次

図目次

表目次

第 1 章　序論

1.1　本研究の背景

生命の働きを理解するため，分子生物学は生命を構成する遺伝子やタンパク質の働きを解き明かし，その知識を応用して様々な生体の複雑なシステムをも解き明かしてきている．中でも，細胞分化は多細胞生物が多種多様な器官や臓器を形成するためのシステムであり，このシステムを生命の基礎である遺伝情報から説明することで，メカニズムを理解し制御することができると期待されている．しかし，細胞分化のシステムは想像以上に複雑であると分かってきており，情報科学の発展によって可能になった大量の情報処理技術を用いて，これまででは困難であった細胞分化全体のメカニズムを推測し，解き明かすことが求められている．

細胞分化はヒトなどの多細胞生物において，細胞が特殊化して分化前の細胞には見られない特有の形質を持つようになる過程のことをいう．多細胞生物は 1 個の受精卵から細胞分裂と細胞分化を繰り返し，細胞数を増やして様々な組織，器官，臓器を形成する．通常，細胞分化は不可逆であり，受精卵の状態では全ての組織へ分化できる全能性を持つが，分化が進むと分化できる細胞の種類が決まってしまう．この細胞分化や分化可能な細胞を決定づけるのは主に遺伝子の発現の変化だと分かってきている．近年，iPS 細胞に始まる再生医療の発展により，細胞分化のメカニズムを解明する要求が高まっており，遺伝子制御ネットワークによって細胞分化中の遺伝子の発現の様子を解析することが求められている．

遺伝子は親から子孫が持つべき性質を規定した情報を伝達する因子であり，A(アデニン)，T（チミン），G（グアニン），C（シトシン）の 4 種類からなる DNA（デオキシリボ核酸）の塩基配列によって構成されている．遺伝子は生体内で数ステップの反応を経てタンパク質を生成することで役割を果たす．その反応は次のようになる．まず遺伝子領域の塩基配列が mRNA(メッセンジャー RNA) に転写される．次に mRNA が翻訳されてアミノ酸の重合体であるアミノ酸ポリマーが生成され，折りたたまれて特有の立体構造を成す．この生成された立体構造がタンパク質である．遺伝子からタンパク質が生成されるまでの，この一連の過程を遺伝子の発現と呼ぶ．遺伝子の数は生き物によって多岐にわたり，ヒトではおよそ 2 万 5 千と知られ [1]，現在知られている最小のものでは真性細菌である *Candidatus Carsonella ruddii* が 182 であると報告されている [2].

生体内の遺伝子は単体で機能するだけではなく，他の遺伝子に影響を与えて発現を制御する転写因子が存在しており，数百から数万という遺伝子が他の遺伝子を制御しあって生体内で調和のとれたタンパク質生成システムを構成している．その制御は遺伝子から mRNA を生成する転写量を促進，あるいは抑制することによって行われている．この制御関係を転写制御と呼び，遺伝子領域の上流 (あるいは下流) にある特定の DNA 塩基配列と転写因子となるタンパク質が反応することによって実現している．この塩基配列領域に転写因子が結合して重合体をなすことで mRNA の転写量に影響を及ぼすことができる．この働きによって発現するタンパク質の量を調整し，その時々や細胞組織ごとに必要な量を生成する．それによって生命の複雑な機構を調整，制御し維持している．

細胞分化においても転写制御は重要な役目を担っており，遺伝子の発現を調整することにより分化先を決定しているということが分かってきている．分化の刺激を受けた細胞では，特定の遺伝子の発現を促進，抑制されることによって細胞分化が進み，遺伝子の転写制御関係が動的に変化して分化後の細胞に特有の形質を持つようになる．例えば，間葉系幹細胞から脂肪細胞への分化では，刺激を受けて発現した Cebpβ，Cebpδ といった遺伝子が，Cebpα，Pparγ というマーカー遺伝子の発現を促進させ，これらマーカー遺伝子が他の細胞への分化を抑制したり脂肪細胞の形質を表現する遺伝子の発現を促進するといったことが行われる [3]．このように，少数の特定の遺伝子がカスケード状に制御していくことによって細胞分化は進んでいくとされている．分化の進行に応じて制御の変化は他の遺伝子へ波及していき，発現する遺伝子が大きく変化する．その際，クロマチンリモデリングやクロストーク遺伝子といった要因で今まで発現できていた遺伝子が発現しなくなることや，その逆の発現していなかった遺伝子が発現することなどが起こる [4][5][6][7]．また，一度分化した細胞が他の細胞にならない不可逆性を持つ大きな要因としては，分化後の細胞に必要な遺伝子以外の遺伝子の塩基配列がメチル化され，発現ができなくなり遺伝情報を失うといったことが分かっている [8]．

複雑な遺伝子制御関係やその変化を全て生物学的な実験により確かめることは非常にコストが掛かってしまうため，計算機を用いて制御関係を推定することが期待されている．遺伝子の制御関係を推定するために，遺伝子の発現量を測定した遺伝子発現プロファイルを用いて，興味の対象となる遺伝子群の制御関係をネットワークに見立ててネットワークモデルに当てはめることが行われている．その際，遺伝子の制御関係はノードを遺伝子，エッジを制御関係とした遺伝子制御ネットワークとして書かれる．遺伝子制御ネットワークのモデル化は研究の進んでいる分野で，状況や目的に合わせて様々なモデルが提唱され

ている．ネットワークモデルの例として，グラフィカルガウシアンモデル [9]，ブーリアンネットワークモデル [10]，ベイジアンネットワークモデル [11]，ダイナミックベイジアンネットワーク [12]，微分方程式モデル [13][14][15][16] が挙げられる．

ネットワークモデルを用いて遺伝子制御ネットワークを推定するには，実験値である遺伝子の発現量を記録した時系列発現プロファイルからネットワークモデルのパラメータを求める逆問題を解くことになる．このとき，遺伝子数の増加に伴い可能なネットワークの構造が爆発的に増加すること，時系列発現プロファイルは実験値であるためノイズなどの誤差が多く含まれている可能性があること，パラメータ探索の結果得られる解の中には最適解の他に最適解と異なる構造を示す準最適解が存在する可能性があること，転写因子と転写産物が非線形の入出力である可能性があること，発現した遺伝子は多くの転写因子に複雑に影響を与え，その転写因子も他の多くのタンパク質から修飾を受ける複数対複数の複雑な制御関係があること，といった課題が発生する．

遺伝子制御ネットワークモデル推定手法にはそれぞれ特徴があり，前述の問題によって適用可能なデータが異なり，推定の結果得られたネットワークが表す意味が異なる．遺伝子制御ネットワークの解析手法は今も様々な研究が行われているが，全ての問題点を解決するような統一的な手法がなく，それぞれの手法に利点と限界が存在するため，各手法に適した解析や適用するシステムの規模があり目的や興味の対象に応じて用いるデータやモデルを選ぶ必要がある．モデルの適用可能性は，扱う入力が静的か動的か，離散値か連続値か，制御関係を決定的に表すか確率的に表すか，表されたネットワークは定性的か定量的か，近似的か正確かといった尺度から比較して考えることができる [17]．入力が静的ならそのネットワークは共発現関係や相関関係にある遺伝子のネットワークを表し，動的であれば時間変化に依存して制御を行う転写制御関係や因果関係を表していると考えられる．また，遺伝子発現プロファイルには遺伝子が発現したか発現していないかの 2 値で扱う離散値として表現したものと発現量を計測したマイクロアレイデータなどでは連続値として発現の強さを表現したものがある．ネットワークモデルの制御関係には遺伝子の発現状態が同じであれば常に一定の制御が掛かると考える決定的な制御関係とある確率分布に従って確率的に制御が掛かると考える確率的な制御関係があり，決定的な制御関係では同じ遺伝子の状態には同じ制御が掛かるためシミュレーションによってネットワークの状態を考えることが容易だが，発現の状態が変化すると制御関係も大きく変わる可能性があるためノイズに弱い．確率的な制御関係では逆のことが言える．定性的なネットワークは遺伝子間の制御関係があるかないかだけを表す．定量的なネットワークは制御の強さや重

みまで表すため，強い制御関係の発見や発現量のシミュレーションなどに利用することができる．近似的な手法は遺伝子の制御をオン・オフの二値によって表したりすることで遺伝子制御反応を近似し，正確な手法は遺伝子の制御をできる限り生物学的な反応を反映させてモデルを構築している．目的に応じた遺伝子制御ネットワーク推定手法を決定するには，これらの特徴の持つ理転と欠点を考慮することが必要である．

細胞分化の遺伝子制御ネットワークを理解するためには，遺伝子全体の制御関係を網羅的に推定できる近似的な手法と遺伝子の発現量の振る舞いを解析できる詳細な手法のどちらも欠かすことはできないが，統一的な手法は提案されておらず，それぞれに適した手法を適用して独立に解析するに留まっている．また，独立に遺伝子制御ネットワーク推定手法を用いる際に，細胞分化特有の問題を考慮して解析を行わなくてはならない．細胞分化においては，制御関係が一定でないため分化前と分化後で遺伝子制御ネットワークが異なってしまう問題や，網羅的な遺伝子間の制御による生物学的な反応が分かっていないため，制御における力学的なパラメータや分子の密度などの定量的な情報はほとんど利用できないが起こる．細胞分化過程の遺伝子制御ネットワークを解析するためには，上記問題点を考慮しなければならないが，現在では分化前のネットワークと分化後のネットワークを独立に推定したり，変化を無視して分化過程全体の時系列発現プロファイルを用いて推定が行われている．それぞれの手法を細胞分化に適した手法として開発し，統合することができると示すことができれば細胞分化の遺伝子制御ネットワークの構造全体からを把握することができるようになると考えられる．

1.2 本研究の目的

本研究では，細胞分化における動的な遺伝子制御ネットワークに着目し，網羅的かつ詳細な細胞分化の遺伝子制御ネットワークを推定する統一的な手法を開発することによって細胞分化過程での大規模な遺伝子制御ネットワークの構造と遺伝子発現の振る舞いの理解を目指す．細胞分化過程を支配する働きを持つ遺伝子やタンパク質などは同定されてきており，刺激を与え細胞分化を誘導することが可能になってきている．しかし，細胞分化中に細胞内の全遺伝子の働きや制御関係がどのように変化するか，といった網羅的な視点での動的な変化への解析は十分とはいえない．また，分化中においてカスケード上流の遺伝子の発現量や制御関係がどのくらい変化すると下流の遺伝子の発現はどのような動態を示すのか，といった定量的な動的な性質は不明な点が多く残されている．本研究では上記 2 点の異なる特性に焦点を絞り，大規模で定性的な遺伝子制御ネットワーク動態解析と小規

模で定量的な遺伝子制御ネットワーク推定といった 2 つのアプローチによって細胞分化過程の動性を理解できる手法の提案を目的とした．これによって，遺伝子制御ネットワークが未知の細胞分化系列において，まず大規模な遺伝子制御ネットワークを推定した後に重要性の高いと考えられる小規模な部分を詳細に解析するといった，統一的な遺伝子制御ネットワーク推定手法となる．

まず，小規模の定量的な遺伝子制御ネットワーク解析に関して取り組む．遺伝子の発現量は常に一定ではなく，その時々で必要な量を発現することによって生体の複雑な機構を制御している．細胞分化では，少数の特定の遺伝子によって制御の振る舞いを決定していることが報告されている．そのような遺伝子発現の動的な振る舞いを解析する際には微分方程式モデルを用いての解析が行われてきた．微分方程式モデルは遺伝子の制御関係だけではなく時系列での遺伝子の発現量の再現が可能なモデルで，適切なパラメータを与えることで系内の遺伝子発現量の初期値を与えることでの動態のシミュレーションができる．このパラメータは一般的に明らかではないため，遺伝子の発現量を時間に沿って計測した時系列発現プロファイルを入力として，入力データを再現するパラメータを推定する逆問題を解くことになる．微分方程式モデルの中でも S-system モデルは化学反応を基にして遺伝子制御関係を説明したモデルで，探索によって得られたパラメータから制御関係の理解が簡単なため少数の遺伝子における遺伝子制御ネットワークの動態解析に用いられている．しかし，パラメータを推定する際に，時系列発現プロファイルを入力した場合，局所解が多くなり，探索が局所解に収束してしまう問題がある．本研究では，上記の要因によってパラメータの推定精度が低下してしまうのを軽減するため，探索が局所解に収束してしまった際に局所解を解候補として記憶しておき再探索を開始することで収束を避け，S-system モデルの推定精度の向上を目指す．

次に，大規模の定性的な遺伝子制御ネットワーク解析に関して取り組む．細胞分化中における遺伝子の発現制御は一定ではなく，分化の進行に伴って様々な遺伝子の制御関係が変化していく．そのような変化を追って遺伝子の制御関係を解析していく手法にノードセットセパレーション法 [18] がある．ノードセットセパレーション法は薬剤を投与した後の遺伝子制御の変化を解析する手法であり，入力された時系列発現プロファイルの連続した 2 時点毎に閾値を超えた発現を示した遺伝子だけを抽出して全時点を用いてネットワーク推定を行い，時間で順序付けられた複数のネットワークを得ることで動的なネットワークを表現している．しかし，この手法は連続した 2 時点で発現量が上昇した遺伝子間の制御関係のみを推定するため，長い期間をかけて変化する制御関係は連続した 2 時点で

閾値を超えない可能性があり，他の遺伝子を抑制する働きなどにより発現量を低下させる制御関係は制御を受ける側の遺伝子の発現量が上昇しないため閾値を超えない可能性があるため，推定が困難であると考えられる．また，時間方向の解像度が低いデータを対象としており入力された時系列発現プロファイル全時点を用いているため，データ内で制御関係が変化してごく一部だけ制御関係が見られるような場合に全時点を用いて推定すると，残りの部分では制御関係が無い挙動をする影響によって，その一部分の制御関係が推定できない可能性が考えられる．細胞分化においては上記問題点を引き起こす制御関係の変化が多くの遺伝子間で起こっていると考えられる．そこで本研究では，上記の要因によって推定精度が低下してしまうのを軽減するため，時系列発現プロファイルを時間方向に分割し，それぞれのデータからネットワークを推定することによって，長い期間をかけて変化する制御関係やパラメータの変化に対応する手法を提案する．これによって，大規模で動的な細胞分化の遺伝子制御ネットワークの推定を目指す．

この 2 つの手法に取り組むことによって，細胞分化に適した大規模な遺伝子制御ネットワークの推定と小規模な遺伝子制御ネットワークの動的な振る舞いの解析を達成する．それにより，大規模な遺伝子制御ネットワークを推定した後に小規模な部分をさらに細かく発現量のシミュレーションや動態の解析を行うような応用に利用できることを述べる．

1.3 本論文の構成

本論文は 4 章構成である．第 1 章では，本研究の背景である細胞分化における遺伝子制御ネットワークを推定する意義，および本研究の方針を述べる．第 2 章では，細胞分化における遺伝子発現の動態を解析するための遺伝子制御ネットワークモデルの推定精度を向上させる手法の提案を行う．第 3 章では，遺伝子制御の遷移を解析できるように動的な大規模遺伝子制御ネットワーク推定を行う手法を提案する．

第 2 章では，S-system モデルによる遺伝子制御ネットワーク推定手法に存在した問題点である，推定が局所解に陥り精度が低下してしまう点に関して，局所解に陥った際に最初から探索を行う手法を適用することで解決する手法を提案する．まず遺伝子の働きと細胞分化，時系列発現プロファイル，遺伝子制御ネットワークと遺伝子制御ネットワークモデルについて延べ，遺伝子制御ネットワーク推定手法における従来手法について説明する．その後，遺伝子制御ネットワークモデルの一つである S-system モデルに着目し，従来の推定手法である遺伝的アルゴリズムを改良した免疫アルゴリズムを適用することにより精度を向上させることを説明する．S-system モデルは遺伝子間の相互作用を微分方程

式モデルで表したものであり，パラメータを適切に設定することにより遺伝子発現の動態を再現しシミュレーションすることができる．ある遺伝子群での相互作用をシミュレーションしたい場合，その遺伝子群に対してのモデルのパラメータは一般に分からないため，調査したい系の時系列発現プロファイルを再現するパラメータを探索するような逆問題を解くことになる．しかし，細胞分化中に見られる遺伝子の発現量が大きく変動するような反応では，局所解が多く存在し時系列発現プロファイルの再現精度が高くならない問題点がある．そこで，免疫アルゴリズムを適用することにより，局所解に陥った際に局所解を保存しながら探索を初期状態にすることによってより最適解を探索する可能性を高める手法を提案する．シミュレーションデータと脂肪細胞分化データを用いた評価実験により，免疫アルゴリズムの適用によって推定されるパラメータの適合性が上昇することを検証する．

第 3 章では，遺伝子の制御関係の変化を追跡できる動的な遺伝子制御ネットワークを推定する手法を，従来手法のノードセットセパレーション法の問題点である全時点を使用してしまう点に関して，時間方向に時系列発現プロファイルを分割する手法を提案する．まず，細胞分化における遺伝子の制御関係変化とそれに対応した動的なネットワークを解析する手法について説明する．その後，遺伝子制御関係の変化を追跡してネットワークを推定する従来手法であるノードセットセパレーション法についての問題点を挙げる．ノードセットセパレーション法は入力の時系列発現プロファイルを全時点使うため，データ中に制御関係が変化していた場合にはその変化を捉えることができない．提案手法ではスライディングウィンドウ法を取り入れ時系列発現プロファイルを分割し，分割されたデータそれぞれに対してダイナミックベイジアンネットワーク推定手法を適用することにより，変化の時期毎にネットワークの推定を行いデータと推定されるネットワークとの整合性を高めることで従来より時期を捉えたネットワークを推定する手法を提案する．分割手法について，他の分野で用いられている手法を紹介した後に，時系列発現プロファイルに適用できる手法を示す．評価実験として，小規模なネットワーク推定と大規模なネットワーク推定によって既知の制御関係をより良く推定し，ネットワークの推定精度が向上することを検証する．最後に，2 章で得られた小規模なネットワーク推定結果と 3 章の遺伝子制御ネットワーク推定手法との整合性を確認し，提案手法が網羅的な遺伝子制御ネットワークの推定結果を定量的な解析手法に適用可能であることを検証する．

第 4 章では本研究の成果を纏めるとともに，本研究の適用範囲と今後の課題について述べる．

第 2 章　免疫アルゴリズムによる遺伝子制御ネットワーク動態解析のための S-system モデル推定手法

2.1　緒言

生物の設計図となるゲノムの全塩基配列はマウスや線虫などの主なモデル生物やヒトについては既に解読が終了しており，現在は様々な生物種での塩基配列が解読されデータベースに情報が蓄積されている．このゲノム上には生物を構成するために必要な遺伝情報が書き込まれており，ゲノム上の遺伝子は生物にとって必要不可欠であるタンパク質を生成する．遺伝子間には相互作用が存在し，その表現方法として遺伝子制御ネットワークが存在する．遺伝子制御ネットワークは遺伝子間の制御関係をネットワーク状にして可視化したものである．制御関係の一つに転写制御関係があり，ある遺伝子が生成したタンパク質が他の遺伝子のタンパク質の生成を促進，抑制する関係のことを指す．転写制御は生体内で起こる複雑な生命現象の一つであり，遺伝子の働きを調整する重要な役目を担っているため，その全容の解明に対しての要求は非常に強い．

細胞分化はそのような複雑な転写制御によって実現している機構であり，多細胞生物にとって無くてはならない過程である．多細胞生物は 1 つの細胞である受精卵から，細胞分裂と細胞分化を繰り返すことによって成体となる．このとき，細胞分裂によって細胞数を増やし，細胞分化によって細胞の機能を血液や皮膚や骨など特殊化させる．細胞分化により多様な形態をもつことができるため，近年の急速に発展している iPS 細胞に始まる再生医療の分野では，そのメカニズムを究明しコントロールすることが求められている．細胞分化のメカニズムは転写制御によって成り立っているということが分かっており，特定の遺伝子が機能して他の遺伝子へ転写制御をすることで対応した細胞種へと分化が進む．そのような転写制御関係は従来では生物学的実験手法を用いて解析していたが，生物の遺伝子数は数万ほど存在すると言われており，その全てに対して実験を行うには莫大なコストがかかる．そのためバイオインフォマティクスの分野では計算機を用いて情報科学的，統計学的解析手法に基づき実験コストを削減して遺伝子制御ネットワークを解明する研究が求められている．

遺伝子が生成した mRNA の量を知る方法の一つにマイクロアレイ技術が存在する．マイクロアレイ技術を用いることで一度に数千から数万の遺伝子について観測することができる．さらに近年では次世代シーケンサーと呼ばれる高速かつ大量に塩基情報を読み取る

技術も発達しつつあり，これら大量のデータを処理する必要がある．しかし，実際の生物情報を扱うには生命現象の複雑さや情報量の多さ，要求される演算能力の高さ等により困難を極めるため，高速かつ高精度に情報を解析する方法論やアルゴリズムの開発が急務となっている．観測されたデータの中でも時間変化に応じた遺伝子の転写量を観測した時系列発現プロファイルは動的な遺伝子制御ネットワークを解明するために非常に重要である一方，そのデータは非線形でありかつ時間差による遺伝子の挙動の変化が解析を困難にしている．さらに，どのような過程を経て遺伝子の転写量を制御しているのかは遺伝子毎に異なり，未知である．これらの問題点を解決するようなアルゴリズムができれば遺伝子制御ネットワークの推定精度は向上し広範な応用が期待できる．

生命現象を数値化し，システムとして理解することを目的としたシステムバイオロジーの分野では，遺伝子の挙動を一つのシステムとして捉えてシミュレーションによって再現しようとしている．推定した遺伝子制御ネットワークは遺伝子発現のシミュレーションに用いられ，細胞分化のような複雑な挙動を示す遺伝子発現のダイナミクスを理解する手助けとなる．今までの生物学で行われていた解析的，直感的な理解とは異なり，実際にシミュレーションモデルを構築して遺伝子発現の挙動を再現することで，その挙動を再現するための合成的な理解ができるようになる．しかし，今までの遺伝子制御ネットワーク推定手法のようにノード間の辺の有無の推定だけでは遺伝子の挙動を再現できない．そのため，ネットワークの存在のみならず，その強度まで推定する必要がある．遺伝子制御ネットワークの推定方法にはベイジアンネットワーク [11] やブーリアンネットワーク [10] などが存在しているが，それらはネットワークから遺伝子の発現量の変化を推定することはできない．しかし，S-system モデルは未知の反応系での相互作用を推定する数理モデルで，各遺伝子間の制御関係を再現する数理モデルを構築することによって遺伝子制御ネットワークを推定して遺伝子発現量の挙動を再現することができるため，非常に有用である．このモデルを用いれば現実の遺伝子制御反応に近いシミュレーションを行うことができると期待されている．しかし，S-system モデルを時系列発現プロファイルから推定する際，モデルを表すネットワーク構造のパラメータが局所解に収束してしまう問題点がある．

上記問題点の解決のため，推定中に局所解へ収束してしまった際に探索を初期化して収束を避ける免疫アルゴリズムを適用する手法を提案する．免疫アルゴリズムは探索が局所解に収束した際，その局所解を記憶し，探索を初期状態から始めることによって，局所解の収束を避けて複数の記憶した解から最適な解をアルゴリズムである．時系列発現プロ

ファイルを再現する S-system モデルでは，局所解の数が非常に多く，局所解を回避することで，推定精度の向上が期待できる．

本章では，脂肪細胞の時系列発現プロファイルを用いて，S-system モデル推定に免疫アルゴリズムを適用することによって推定し，その結果と評価を行う．

2.2 遺伝子の発現と遺伝子制御ネットワーク推定

2.2.1 遺伝子発現と転写制御

遺伝子は A(アデニン)，T（チミン），G（グアニン），C（シトシン）の 4 種類からなる DNA 塩基の配列で構成される，生物の重要な構成要素であるタンパク質を生産する設計図である．この遺伝子は生体が持つ全 DNA 塩基配列であるゲノム上に点在している．遺伝子は生体内で数ステップの反応を経てタンパク質を生成することで役割を果たす．その反応は次のようになる．まず遺伝子領域の塩基配列が mRNA(メッセンジャー RNA) に転写される．次に mRNA が翻訳されてアミノ酸の重合体であるアミノ酸ポリマーが生成され，折りたたまれて特有の立体構造を成す．この生成された立体構造がタンパク質である．遺伝子からタンパク質が生成されるまでの，この一連の過程を遺伝子の発現と呼ぶ．

遺伝子から発現するタンパク質の量は遺伝子毎に一定ではなく，その時々や細胞組織ごとに必要な量を生成する．それによって生命の複雑な機構を調整，制御し維持している．その制御は遺伝子から mRNA を生成する転写量を促進，あるいは抑制することによって行われており，何千，何万という遺伝子が他の遺伝子を制御しあって生体内で調和のとれたタンパク質生成システムを構成している．この制御関係を転写制御と呼び，遺伝子領域の上流 (あるいは下流) にある特定の DNA 塩基配列と転写因子となるタンパク質によって実現している．この塩基配列領域に転写因子が結合して重合体をなすことで mRNA の転写量に影響を及ぼすことができる．

2.2.2 細胞分化

細胞分化は多細胞生物の細胞が特殊な機能を持つ細胞に変化する過程のことで，遺伝子の発現制御がコントロールしているとされている．分化前の細胞がどの細胞に分化するかはどの遺伝子がどの時点で機能するかが重要で，細胞ごとの遺伝子の発現量の差異などによって 1 つの受精卵から複雑な生体システムを形成することができる．例えばヒトは 60 兆個の細胞で構成されると知られているが，細胞増殖だけではなく細胞分化によって脂肪や筋肉などの様々な器官や臓器に 1 つの受精卵から分化していく．分化能を持つ細胞は細

胞種毎にどの細胞へ分化できるかが決まっており，ヒトなどの動物細胞では基本的に分化した後に元の細胞に戻ることや他の分化経路の細胞へ分化することはない．

2.2.3 遺伝子発現プロファイル

遺伝子は基本的にタンパク質として発現することで生体内で機能を果たす．細胞分化においても同様で，転写因子のタンパク質が発現して転写制御を行うことで分化が進んだり異なる分化経路を辿らないようにしている．その遺伝子の発現量を測定する手法は多種多様に存在しており，現在も性能は日進月歩で向上している．その中で現在一般的に用いられているのがマイクロアレイを用いたものである．

マイクロアレイは数万から数十万に区切られたスライドガラス上にプローブと呼ばれるDNA の部分配列を固定した物である．対象の生体の細胞から抽出して蛍光色素を付けたmRNA とプローブをハイブリダイゼーションと呼ばれる方法に掛けることでマイクロアレイの区切り毎に対象の遺伝子発現量に応じた蛍光が出る．この光量によって mRNA の転写量を測定することができ，この転写量から遺伝子の発現量を推定する．1 回のハイブリダイゼーションで数万個の遺伝子発現情報を測定できるので，網羅的な発現解析に広範に利用されている．

遺伝子発現プロファイルはマイクロアレイによって様々な条件下で取得した発現量のデータである．この条件を時間軸に設定したものは時系列発現プロファイルと呼ばれ，時点毎に一枚のマイクロアレイを使用することによって得られる．時系列のデータでは遺伝子間の転写制御のダイナミクスを観測することができ，遺伝子制御ネットワーク解析にとって非常に有用なデータである．しかし，マイクロアレイによって得られるデータは非常に量が多く，計算機による解析が必須となっている．

2.2.4 遺伝子制御ネットワーク

時系列発現プロファイルを見ると，ある遺伝子の発現量が変化した後に他の遺伝子の発現量が有意に変化する事が確認できる．これにより，転写制御関係の有無を推測できる．単純化すると促進，抑制の関係を持つ 4 つの変化が考えられる．

まず，遺伝子 A の発現量が増加した後に遺伝子 B の発現量が増加する関係があった場合，遺伝子 A は遺伝子 B の発現を促したと考えられる（図 2.1.I）．逆に，遺伝子 A の発現量が低下したときに遺伝子 C の発現量が低下するような関係が見られた場合，遺伝子 A が遺伝子 C の発現を促していたが，遺伝子 A の発現量が減少したため遺伝子 C の発

現量も低下してしまう，と考えられる．その為，この関係では遺伝子 A は遺伝子 C を促進していると考えられる（図 2.1.II）．次に，遺伝子 B の発現量が増加した後に遺伝子 C の発現量が減少する様な関係があった場合，遺伝子 B は遺伝子 C を抑制しているとなる（図 2.1.III）．逆に，遺伝子 C の発現量が減少したときに遺伝子 D の発現量が増加した場合，遺伝子 D は今まで遺伝子 C の抑制下にあったが，遺伝子 C の発現量の減少に従って遺伝子 D が強く発現する様になったと考えられる（図 2.1.IV）．

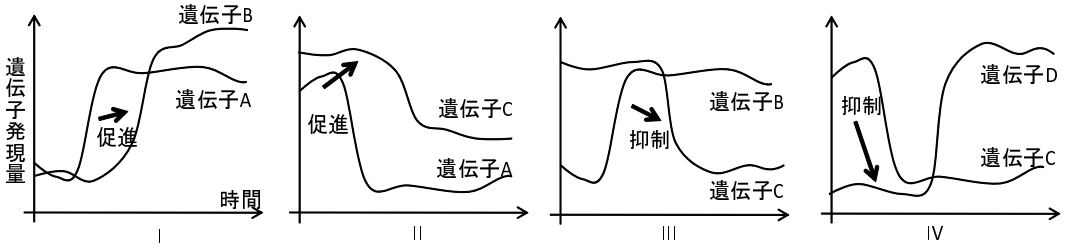

図 2.1　遺伝子発現量と転写制御

遺伝子制御ネットワークはこの促進抑制をネットワーク状に記述した物で，ノードを遺伝子，エッジを転写制御と見なす．エッジの書き方には無向辺，有効辺，重みつき有効辺の様な種類があり，図 2.2 では図 2.1 の転写制御関係を有向グラフで表現した．

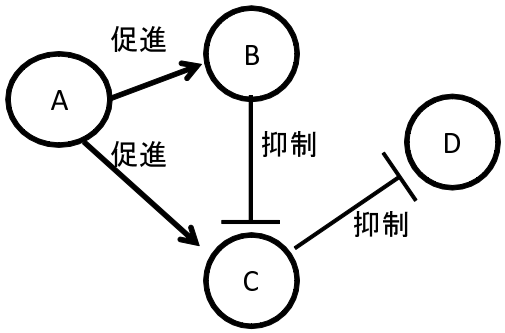

図 2.2　遺伝子発現量ネットワークの例

このように生体内での制御関係をネットワークとして捉えることによって，遺伝子間の間接的な影響も把握できるようになり，対象の遺伝子に影響のある遺伝子群をより広く理解することが容易になる．

2.2.5 遺伝子制御ネットワークモデル

動的な遺伝子発現の挙動を再現するような遺伝子制御ネットワークは前述したように時系列遺伝子発現プロファイルから得られる．しかし，時系列遺伝子発現プロファイルから実際に遺伝子制御ネットワークを求めるには，マイクロアレイを観測することによって得られるデータは非常に膨大である事，転写制御は時間遅延を伴う間接的な影響も存在する事，制御の強度を測定して定量化することは困難な事などが問題となり，人の手で解析することは難しい．その為，現在では計算機を用いて時系列遺伝子発現プロファイルから遺伝子制御ネットワークを推定することが広く行われている．

推定は対象の遺伝子群からなるネットワークをモデル化する事によって行われる．遺伝子制御ネットワークのモデル化は研究の進んでいる分野で，遺伝子制御ネットワークを表すための決まった 1 つのモデルというものは存在しない．その代わり，様々なモデルが提案されており，そのそれぞれに特徴がある．ネットワークモデルの例として，グラフィカルガウシアンモデル [9]，ブーリアンネットワークモデル [10]，ベイジアンネットワークモデル [11]，微分方程式モデルが挙げられる．この中で，微分方程式モデルにはネットワークの構造だけでなく，時系列遺伝子発現プロファイルをシミュレーションして再現する事ができる特徴があり，より詳細な遺伝子制御ネットワークの挙動を知ることができるようになる．

微分方程式モデルの中にも ANN モデル [13]，GRLOT モデル [14]，SSM モデル [15]，S-system モデル [16] といったように様々な種類がある．この中でも S-system モデルは化学反応の反応速度を求める際に使用される一般質量作用則 [19] を近似したもので，対象にしている化学反応系での反応形式や化学量論式が明らかでない，または相互作用の存在が明確でない場合に用いられるモデルである．遺伝子の発現過程は促進と抑制から成り立っており，一種の化学反応系であるとみなせるため，S-system モデルを用いて遺伝子制御ネットワークを推定できるとされる．S-system モデルは生成項と分解項の 2 つの項から成り立っており，その内，分解項は他の遺伝子からの制御を受けることと発現量の大きさに制約がないことが他の微分方程式モデルに見られない特徴的な記述である．その為，より複雑な挙動を再現できる [20]．また，各パラメータの値が遺伝子制御ネットワークの構造やエッジの強度，遺伝子の発現速度を指し，物理現象に即しているため，理解しやすい形になっている．さらに，遺伝子操作を加えていない野生型の時系列発現プロファイルからでもネットワークを解析できるモデルであるため，精度良く推定できれば実験コ

ストの軽減が期待できる．

2.3 S-system モデルのパラメータ推定における従来手法

ここでは，S-system モデルと S-system モデルのパラメータを解説し，従来までのパラメータを推定する手法とその問題点を述べる．

2.3.1 S-system モデルとパラメータ

S-system モデルは化学反応方程式に用いられる一般質量作用則の近似式で，非線形微分方程式によって記述される．遺伝子制御ネットワークのモデルとして用いる際は，化学反応速度の代わりに各遺伝子の発現量の変化速度を表す．発現量の変化速度は促進過程，分解過程の 2 項で式 (2.1) の様に記述される．S-system モデルでは，遺伝子制御ネットワークは全ての遺伝子間に関係を持つと仮定したときに，各エッジの重みはどの様に変化するかという全結線モデルで表現される．

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{N_g} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{N_g} X_j^{h_{ij}} \qquad (i = 1, 2, \ldots, N_g) \tag{2.1}$$

ここで，N_g は対象の遺伝子制御ネットワークのノード数，すなわち遺伝子数である．式 (2.1) の左辺は i 番目の遺伝子発現量 X_i の変化速度 dX_i/dt を表わしている．右辺第 1 項は時間変化 dt に対する X_i の増加量，第 2 項は減少量を表している．パラメータ α_i，β_i，g_{ij}，h_{ij} はネットワークの構造および反応の強さを表していて，α_i，β_i は増加，減少の反応速度を表す正の定数，g_{ij}，h_{ij} は反応次数になる．g_{ij}，h_{ij} は遺伝子 j が遺伝子 i へと与える影響の強さである．g_{ij} の値が正であれば，遺伝子 j は遺伝子 i の発現を促進している事を指し，負であれば発現を阻害する働きをしている事を指す．同様に，h_{ij} の値が正なら，遺伝子 j は遺伝子 i の発現を抑制している事を指し，負であれば発現を促す働きをしている事を指す．また，遺伝子 i の転写過程に X_j が関与していない場合 g_{ij}，もしくは h_{ij} はそれぞれ 0 となる．

2.3.2 従来のパラメータ推定手法

S-system モデルのパラメータが分かれば，式 (2.1) によって，対象の遺伝子制御ネットワークの初期発現量を与える事で時系列遺伝子発現プロファイルをシミュレーションで

き，遺伝子制御ネットワークの動態解析ができる．しかし，この非線形微分方程式のパラメータ数は $2N_g + 2N_g^2$ となり遺伝子数の二乗に依存しており，遺伝子数が増えると共役勾配法などを用いて解析的に高精度で求める事は現実的ではなくなる．また，S-system モデルを時系列発現プロファイルに適用した場合，非常に多くの局所解があり，異なるパラメータの組み合わせであっても似たような動的挙動を示すことが多くなる事，パラメータは相互に影響を与えているため逐次的に決定することができない事が問題となっている．その為，このような方程式を解くには，入力されたデータを再現するようなパラメータを見つけ出す探索アルゴリズムを用いたリバースエンジニアリング手法がよく用いられる．

S-system モデルの優れた表現能力を利用するため，$2N_g + 2N_g^2$ 個のパラメータ全ての値を実数で表現される．また，時系列発現プロファイルを再現するには，初期発現量とパラメータを与え，次時点での発現量の増加量を計算して加算していく事を繰り返すことで達成できる．そのため，前時点での誤差が非常に大きな影響を持つ．誤差はノイズがあると増大するためノイズに非常に弱いという欠点がある．また，遺伝子数が増加するほど誤差の影響は大きくなるため，せいぜい数個の遺伝子セットにしか適用されていない．

ここでは先行研究として従来までに研究された S-system モデルのパラメータ推定手法を挙げる．

2.3.2.1　実数値遺伝的アルゴリズム

遺伝的アルゴリズム（Genetic Algorithm，GA）[21] は生物が環境に適応して遺伝子を変化させて進化する過程を模倣した手法である．GA では目的関数における各パラメータを遺伝子，解候補を個体と呼ぶ．問題に対して複数の個体を生成して，解がより良くなるように個体へ遺伝的操作を加えて遺伝子を進化させ，問題への適応度が低い個体を淘汰していくことで最適解を求める．

GA には対象とするデータの構造に応じてビットストリング GA と実数値 GA がある．ビットストリング GA はパラメータ群をビットの列で表現する為，パラメータが単純で二値に対応付けできるデータに対して適している．実数値 GA はパラメータを実数値で表現するため複雑になるが，解が実数値である関数最適化問題のような数値最適化に対してはビットストリング GA よりも高い性能を示す [22][23]．S-system モデルで対象としてる問題は実数値で表現されるパラメータ最適化問題にあたるので，実数値 GA を用いる．

実数値 GA の動作は主に 6 つのステップに分かれる．それぞれ初期世代生成，評価，終

了判定，選択，遺伝的操作，淘汰である（図 2.3）．

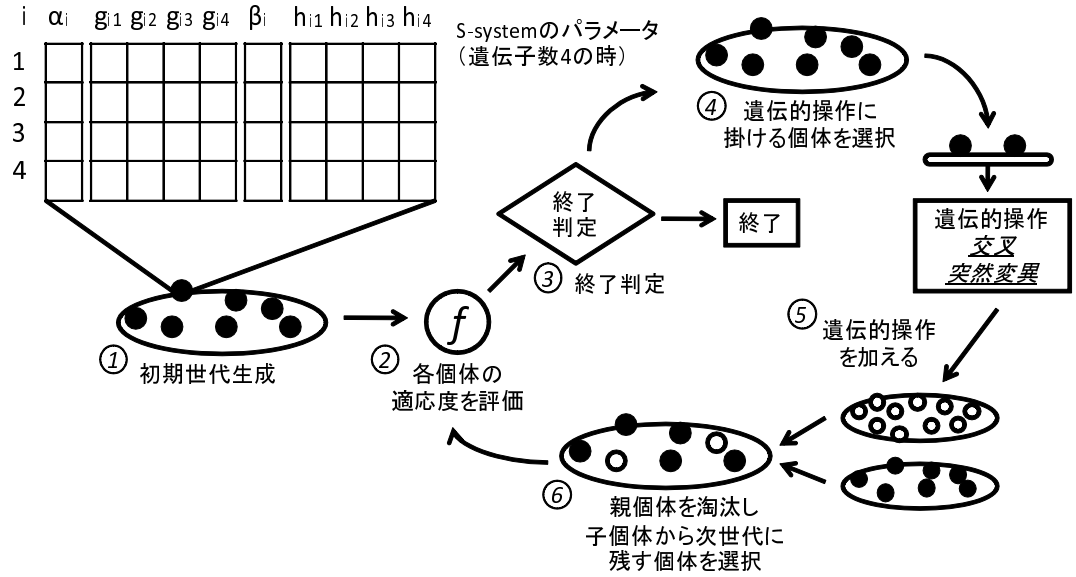

図 2.3　遺伝的アルゴリズムの動作の概念図

1．初期世代生成　まず，初期世代として S-system モデルのパラメータ $2N_g(N_g+1)$ 個の値をランダムに設定した個体を複数生成する．

2．評価　その後，各個体の問題への適応度を評価する．評価には，その個体が持つパラメータから計算される時系列データと入力されたデータとの動的挙動の誤差がどの位あるか，といった式 (2.2) による手法が一般的に用いられる．

$$Fitness = \sum_{i=1}^{N_g} \sum_{t=1}^{T} \left(\frac{X_{i,t}^{cal} - X_{i,t}^{exp}}{X_{i,t}^{exp}} \right)^2 \tag{2.2}$$

ここで $X_{i,t}^{exp}$ はマイクロアレイ装置によって得られる時刻 t における遺伝子 i の遺伝子発現量，$X_{i,t}^{cal}$ は式 (2.1) を解くことによって得られる時刻 t における遺伝子 i の遺伝子発現量，T は観測した時系列データの時点数である．$Fitness$ は 0 に近いほど測定値との誤差が小さく，そのパラメータは良い親和度を持つといえる．その為，GA ではこの評価値 $Fitness$ を最小にする個体を探索する事になる．

また，式 (2.3) の様にして評価関数の値を $[0,1]$ の範囲に正規化することもある．この操作によって評価の順序が変わることはなく，高い値ほど良い親和度を持ち，1 が入力

データと一致する事となる．推定問題ではこの評価関数を最大化するパラメータを探索する．これは評価関数の値が高いほど良い評価を持つ事となり，議論が分かりやすくなる．以降，評価値とはこの Φ の値を指す事とする．

$$\Phi = \frac{1}{1 + Fitness} \tag{2.3}$$

評価関数は個体の持つ遺伝子の良し悪しを決めるため，目的の方向へ推定を誘導する為にペナルティ項を入れる事もよく行われる．例えば，遺伝子制御ネットワークが疎な構造である事を導入した式 (2.4) を用いることも試されている [24].

$$Fitness = \sum_{i=1}^{N_g} \sum_{t=1}^{T} \left(\frac{X_{i,t}^{cal} - X_{i,t}^{exp}}{X_{i,t}^{exp}} \right)^2 + c \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (\mid g_{ij} \mid + \mid h_{ij} \mid) \tag{2.4}$$

ここで，c は 2 つの評価項のバランスをとる重み係数である．この式は式 (2.2) に第 2 項を加えた評価式である．第 2 項はパラメータ g_{ij}, h_{ij} の絶対値の和が小さいほど良い事を表現しており，遺伝子間の関係を減らす方向に推定を進める力を持つ．遺伝子制御ネットワークは疎であることが知られているので，この手法によって，より遺伝子ネットワークの構造を表すパラメータを得やすくなる．

3. 終了判定　終了条件として設定した条件を満たせば一番良い評価値を持つ個体を出力し，アルゴリズムは終了となる．条件には一般的に，規定の世代数を経過する，あるいはある程度の評価値を持つ個体が生成される，もしくは一定の世代数間で，評価値が基準より変化しなかった場合等が設定される．

4. 選択　個体を評価した後は，より問題への親和性が高い個体を生成するための遺伝子操作が行われるが，その前に遺伝子操作の対象となる親を選択する．この選択方法は交叉の方法で変化する．

5. 遺伝的操作　ビットストリング GA では主に突然変異によって解の多様性を増していたが，実数値 GA では交叉法が解の多様性を得る主な手段となっている．これは，ビットストリングでは遺伝子型と表現型が一対一対応しないことがある一方で，実数値では各パラメータの値そのものに意味があり，パラメータをランダムに設定しては前世代での特徴を受け継ぎつつ新しい個体を作る事ができなくなるからである．実数値 GA の交叉方法

には様々なものがあり，[22]，BLX-α [25]，単峰性正規分布交叉（UNDX）[26]，シンプレックス交叉（SPX）[27] 等が存在する．

6．淘汰　交叉によって産生した子個体群の一部は親個体群の一部を淘汰する．その時の子，親個体の選択方法は世代の進化方針を決定付けるため，重要な役割を持っておりこの方法も様々考案されている．その中で，世代間最小ギャップモデル（MGG）[28] は多様性維持に優れ，計算量も少ないため，現在では広く使われている．距離依存世代交代モデル（DDA）[29] は個体間の距離情報を利用して多様性が失われないように個体を交換するので，MGG よりも多様性を維持し，局所解へ陥ってしまうことを避けられるとされている．

また，今までの世代で一番評価値が高い個体をエリート個体として保存しておき，次の世代へ無条件で残すエリート保存戦略 [30] が存在する．これによって遺伝的操作で評価値の高い個体が生まれても次の世代で消えてしまう，といったことがなくなる．

以上を 1 世代とし，終了条件を満たすまで 2 の選択から操作を繰り返すことで目的の最適解を得る．

2.3.2.2　遺伝的プログラミング

遺伝的プログラミング（Genetic Programming，GP）[31] は GA を拡張したアルゴリズムで，データの構造を配列ではなく木構造によって表現するという特徴がある．S-system モデルに適用するには，各パラメータの実数値を実数と四則演算式からなる 2 分木で表現し，各木を遺伝操作によって進化させるようにする [32]．動作の流れは GA と同じように，初期世代生成，評価，終了判定，選択，交叉，淘汰の 6 ステップで行われる．図 2.4 はその概念図である．以降，松村ら [32] が考案した手法を説明する．

1．初期世代生成　個体は S-system モデルのパラメータセットを持ち，各パラメータを式の木で表現し，遺伝子と呼ぶことにする．初期世代生成では個体の構成要素である式の木を遺伝子数だけ作成する．各木はノードを四則演算（$+$，$-$，$\times$，$\div$），終端子を複数個の実数値として表現される．まず，遺伝子 1 つに対して乱数を発生させる.，また，非終端子 1 つと終端子 2 つからなる高さ 2 の木を 50 個生成する．次に，作成した木の値を計算し，その値が遺伝子の乱数値に近ければ木の候補に加える．その後，木の候補が 20 個を超えるまで作成した木に交叉操作を加えて，木を新たに作成し，値を計算して乱数値に近ければ木の候補に加える事を繰り返す．ただし，作成した木の総数が 1000 を超えればう

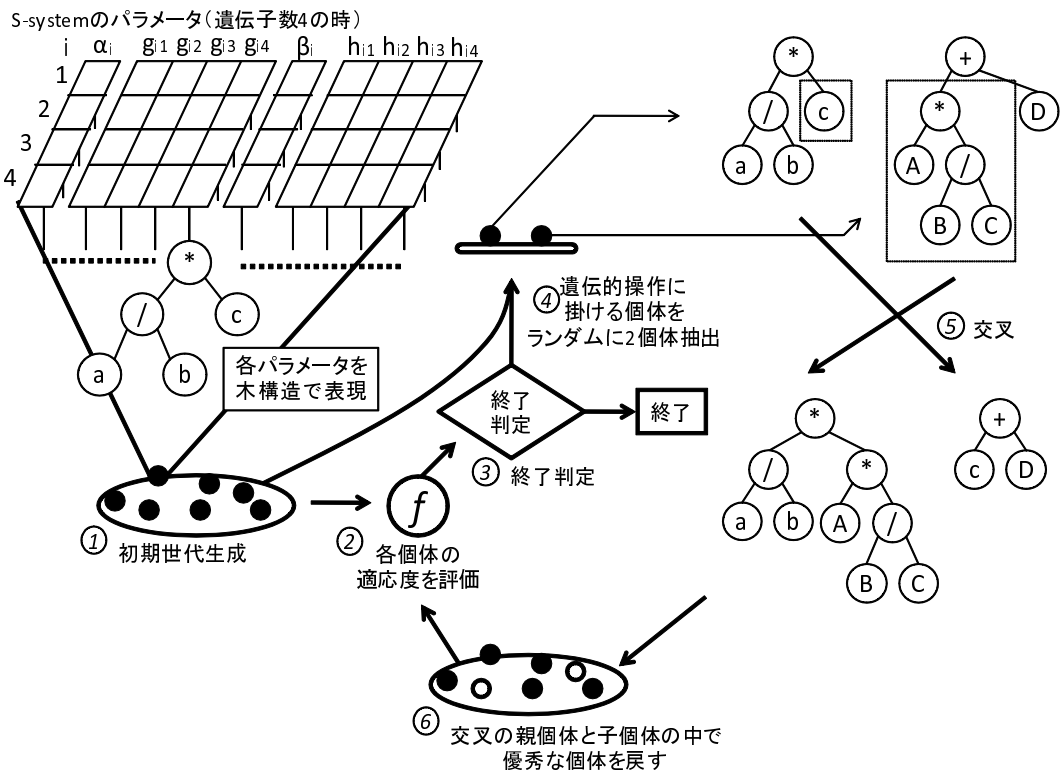

図 2.4　遺伝的プログラミングの動作の概念図

ち切る．作成した木の候補からランダムに 1 つ選び，それを対象の遺伝子を表現する木とする．ただし，木の候補が 1 つもなければ高さ 2 の木の作成からやり直す．以上の工程を全遺伝子において実行し，1 つの初期個体を生成する．これを個体群の数だけ繰り返して初期個体群を生成する．

2. 評価　評価は GA と同じように，パラメータから計算される時系列データと入力データの誤差の小ささを見る．

3. 終了判定　GA と同様に終了条件を満たせば，一番良い評価値を持つ個体のパラメータを出力してアルゴリズムは終了となる．

4. 選択　交叉の対象となる親を個体群の中から 2 個体ランダムに選出する．

5. 交叉　交叉は図 2.4 のように木の一部を入れ替えることによって行われる．選択された個体の中で，交叉を行う遺伝子を 1 つランダムに選択する．そして，木の中で交叉を行

うノードをランダムに決定し，2 個体間でそのノード以下を入れ替える．この操作によって子個体は 2 個体生成される．

6．淘汰　交叉によって生成された個体と親となった個体を入れ替える．親個体を P1，P2，子個体を C1，C2 とすると，次世代に残す個体は P1，C1 の内で評価値の高い個体と P2，C2 の個体で評価値の高い方の 2 個体である．

以上の操作 1 回を 1 世代とし，終了条件を満たすまで操作を繰り返す．

2.3.2.3　微分進化法

微分進化法 (Differential Evolution，DE)[33] は GA などと同じような進化的計算アルゴリズムの 1 つである．基本的な流れは GA と同様であるが，選択，遺伝的操作，淘汰の過程が異なる [34].

まず，個体群の数を N，入力データの遺伝子数を N_g，世代数を G とする．各個体は S-system モデルのパラメータセット $2N_g(N_g+1)$ 個を持ち，各パラメータを遺伝子として表現する．親個体群 x^i_G，$i=1,\ldots,N$ から子個体群 y^i_{G+1} を N 個，次式 (2.5) によって産生する．

$$y^i_{G+1} = x^j_G + F\left(x^k_G - x^l_G\right) \tag{2.5}$$

ここで，j, k, $l \in 1,\ldots,N$ は $i \neq j \neq k \neq l$ を満たす様にランダムで決定する．この操作は進化アルゴリズムでの突然変異にあたり，F は倍率や増幅定数などと呼ばれる定数で x^i_G の変化割合を指す．次に，次世代の個体 x^i_{G+1} を進化アルゴリズムでの交叉にあたる方法によって作成する．x^i_G と y^i_{G+1} の各遺伝子について，範囲 [0,1] の一様乱数を発生させ，その値が交叉割合定数 CF を下回ればその遺伝子は x^i_G の値，そうでなければ y^i_{G+1} の値が x^i_{G+1} の遺伝子となる．その操作を繰り返して x^i_{G+1} の全遺伝子が求まれば，x^i_G と x^i_{G+1} の評価値が高い方が実際の次世代 x^i_{G+1} として置き換わる．以上の操作を繰り返すことでより評価の高い個体を得る．

また，式 (2.5) で計算される個体を，より収束性を高く頑強にした Trigonometric Mutation DE（TDE）[35] を用いた研究も行われている [34]．式 (2.5) では元になる 1 個体を他の 2 個体で決定するベクトル方向に変化を加える方法であったが，個体の評価値を考慮に入れておらず，局所探索性が乏しく収束が遅い．そこで，3 個体を選んだ際に評価値の勾配を計算し，次世代の個体の評価がより良くなるように変更することで，上記の

問題を解決している．その式は次のようになる．

$$
\begin{aligned}
y^i_{G+1} &= \left(x^j_G + x^k_G + x^l_G\right)/3 + (p_k - p_j)(x^j_G - x^k_G) \\
&\quad + (p_l - p_k)(x^k_G - x^l_G) + (p_j - p_l)(x^l_G - x^j_G) \\
p_o &= \mid f(x^o_G) \mid /\acute{p}, \qquad \acute{p} = \mid f(x^j_G) \mid + \mid f(x^k_G) \mid + \mid f(x^l_G) \mid
\end{aligned} \tag{2.6}
$$

ただし，$f(x)$ は式（2.2）の値で，個体 x のパラメータセットから計算されるデータと入力データとの誤差値である．

2.3.3　従来手法の問題点

上記のアルゴリズムは全て進化アルゴリズムにのっとって探索を行う．つまり，親となる個体の特徴が子個体へと受け継がれていくことになる．これらの手法では一度個体群が局所解へ落ち込んでしまうとその局所解から抜け出すことは困難になる．局所解とは本来の解の他に存在する部分的に最適な解である（図 2.5）．GP，DE では突然変異の様な強制的に値を変えてしまう機構が備わっていないため，進化の過程で個体群がほぼ同じような遺伝子を持ってしまうと局所解に収束するしかない．GA では突然変異や，一世代での変化が小さくなるように個体を淘汰させる事で局所解への収束を避けているが，突然変異確率を上げ過ぎると特徴の遺伝という機能が失われてしまい，一世代での変化が小さすぎるために収束の遅さが問題になる．

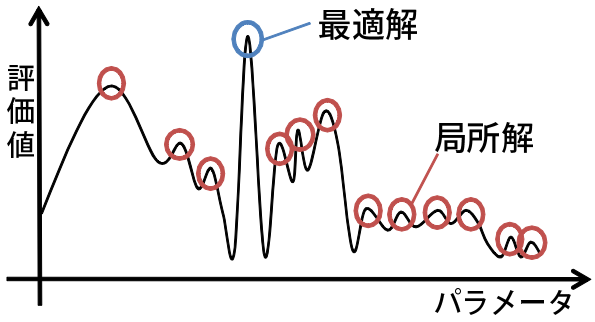

図 2.5　局所解と最適解

また，これらのアルゴリズムでは初期解に依存して探索が行われるので最初に生成される乱数が悪いと局所解へと収束する問題がある．S-system モデルは多峰性関数である事が知られており，局所解が多く解空間が膨大である．その上，乗数であるパラメータ g，h の値を少し変化させただけでも動的挙動が大きく変わり，発散を防ぐ機構がないため，解

空間の大半は動的挙動が発散するような値になっている．その為，一度発散しないような解候補を持つと，その値に個体群が集中してしまうと考えられる．

以上より，従来手法は局所解へ収束した際の復帰方法が無い点が問題であると言えるため，復帰能力のある，早い収束性を持った探索アルゴリズムを適用することで精度の向上が見込まれる．

2.4 提案手法

ここでは免疫アルゴリズム（Immune Algorithm，IA）[36] の S-system モデルパラメータ推定への適用方法を述べる．

2.4.1 目的と概要

従来手法の問題点を解決することにより遺伝子制御ネットワークの推定精度を向上させるのが本研究の目的である．前節で述べたように，GA 等の進化アルゴリズムには大域的に探索できる利点があったり，局所解に陥りにくい工夫がされていたりするものの，初期解に依存して探索を行うため初期収束が起こってしまったり，一度局所解へ陥ってしまえば抜け出すことが難しくなるといった問題点があった．提案手法である IA は脊椎動物の獲得免疫機構をモデル化した探索アルゴリズムであり，解の多様性を維持し，世代交代回数に対する探索範囲が広いとされる．IA を S-system モデルに適用することによって GA の利点である大域的な探索を行いつつ，複数の準最適解を一度に探索し類似した解が過剰に発生するのを抑制できる [37] ため，問題点を解決できると考えられる．

2.4.2 免疫アルゴリズム

IA は獲得免疫機構を元にした近似解を求めるアルゴリズムである．獲得免疫機構とは，ヒトの様な脊椎動物に見られる防御機構で，病原体（抗原）の体内への侵入を防ぐ働きをする．未知の抗原であっても，免疫機構が抗体の遺伝子を組み替えることによって対象の抗原へ効果を発揮できるように適応する仕組みを持っており，その過程は非常に複雑な反応によって成り立っている．ここでモデルとした獲得免疫機構は図 2.6[36] の様に簡略化したものを用いる．

まず抗原に対して B 細胞がヘルパー T 細胞によって増殖する．次に，B 細胞が自分自身の抗体を産生する遺伝子を組み替えて，抗原への親和性の高い抗体を作り出して抗原を攻撃する．抗原を攻撃した後，B 細胞が増殖しすぎないようにサプレッサー T 細胞に

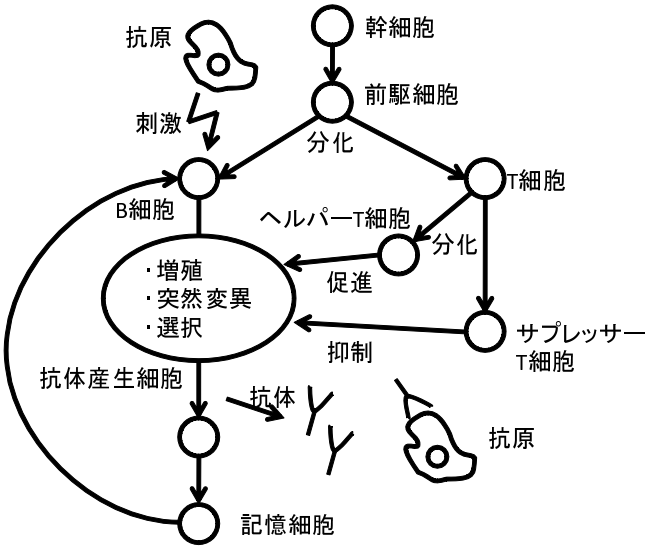

図 2.6　獲得免疫機構の概念図

よって死滅する．ただし，一部は記憶細胞として残り，次に同じ抗原がやってきたときに素早く攻撃できるようにする．以上の動作を繰り返すことによって，抗原により効果のある抗体を産生する．

IA では対象の問題を抗原，解候補を B 細胞および抗体と見なして抗原に親和性の高い抗体を探索し，記憶抗体とすることで問題を解く．初期世代となる抗体群を生成した後，評価，終了判定，濃度計算，記憶細胞分化，サプレッサー T 細胞分化，抗体選択，抗体産生といった操作を繰り返して目的の抗体を探索する．GA に無い特徴的な機能としては，濃度を計算し，記憶細胞やサプレッサー T 細胞への分化がある（図 2.7）．

濃度は，抗体群中での抗体がどれだけ集中しているかを指している．濃度が一定以上の抗体は記憶細胞候補へ分化させて，その中で最も良い評価値を持つ個体を記憶細胞として保存しておく．その為，記憶細胞は局所解として集中してしまった個体を消してしまわずに解候補として保持する役割を持つ．また，記憶細胞候補として分化した抗体をサプレッサー T 細胞へと分化させ，サプレッサー T 細胞と近似した抗体は削除する．その後，削除した抗体と同じ数だけ新規抗体を補充する．この動作で局所解から脱出する．これらの機構によって，解の多様性を保持し，局所解への収束を避ける事ができる．

2.4.3　免疫アルゴリズムの S-system モデル構築への適用

ここでは S-system モデルの推定に用いる際の IA のアルゴリズムを記す．概要図としては図 2.8 のようになる．

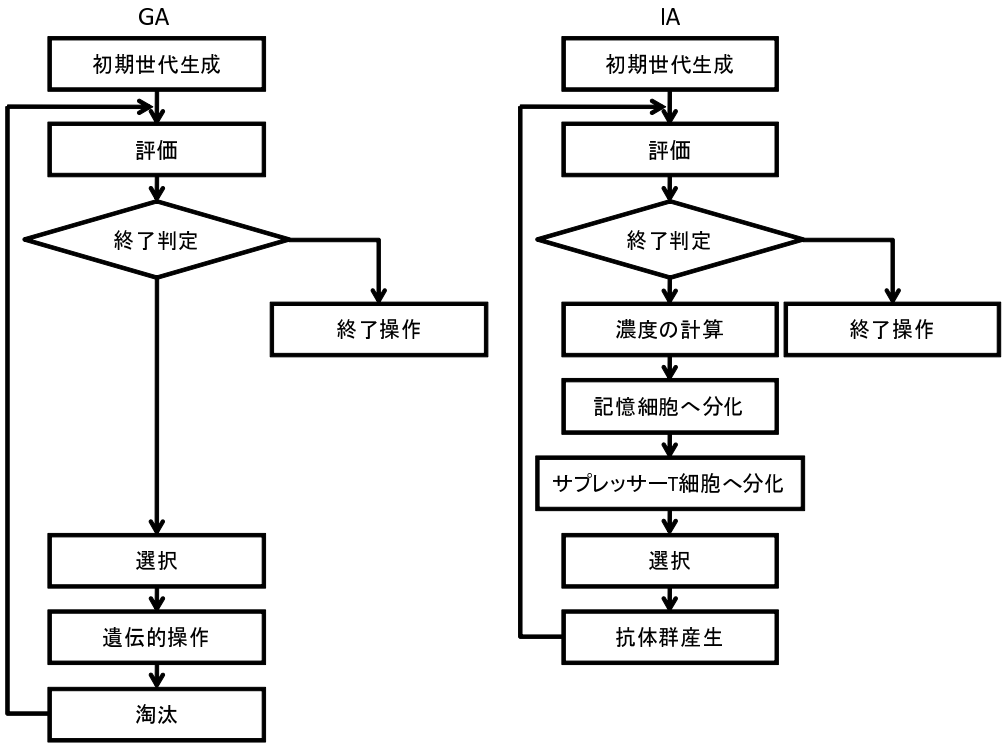

図 2.7　GA と IA の違い

1. 初期世代生成　抗体は遺伝子として S-system モデルのパラメータセット α, β, g, h, 計 $2N_g(N_g+1)$ 個を持つ．ここでは初期世代として，各パラメータそれぞれに実数値をランダムに設定した N_Υ 個の抗体を産生する．

2. 評価　抗体と抗原の親和性を評価する．抗体 υ は入力された時系列データと同じ時点数を持つ時系列データをシミュレーションし，入力データとどれだけ当てはまるかで評価される．その評価値 Φ_υ は GA 同様に式 (2.1)，(2.3) によって計算される．式 (2.1) の微分方程式に入力データの初期発現量を与えて，各時点を古典的 4 次のルンゲクッタ法によって計算して近似解を求める．

3. 終了判定　世代数 G が上限回数 N_R と等しくなれば，規定回数動作を繰り返したとして探索を終了し，後述する終了処理をする．

4. 濃度計算　抗体群中での抗体の濃度を計算する．抗体 υ の濃度 Θ_υ は，υ の持つパラメータに近い値を持つ抗体が抗体群中にどの位存在するか，で計算する．具体的には υ と

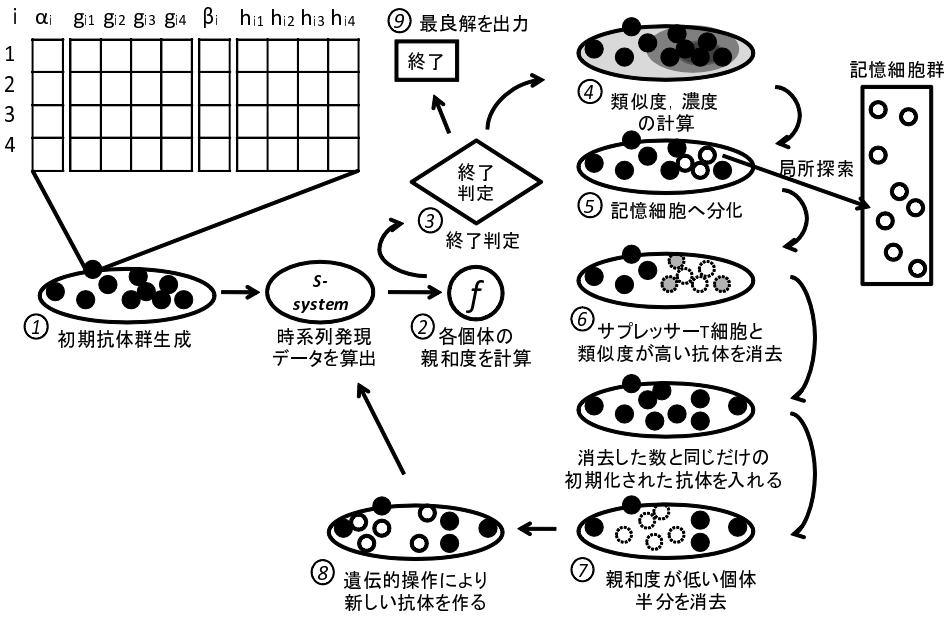

図 2.8　免疫アルゴリズムの動作の概念図

の類似度が濃度閾値 T_θ を上回る抗体の割合を濃度とする．抗体 υ と抗体 ω との類似度 $\Psi_{\upsilon,\omega}$ は次の式 (2.7) で求められる．

$$\Psi_{\upsilon,\omega} = \frac{1}{1 + D(\upsilon,\omega)} \tag{2.7}$$

$D(\upsilon,\omega)$ は 2 抗体間のユークリッド距離を表している．類似度の値域は $[1/N_\Upsilon, 1]$ となり，値が大きいほど類似していることを表す．この類似度より，濃度 Θ_υ を次式によって計算する．

$$\Theta_\upsilon = \frac{1}{N_\Upsilon} \sum_{\omega=1}^{N_\Upsilon} u_{T_\theta}(\Psi_{\upsilon,\omega}) \tag{2.8}$$

ここで，$u_l(t)$ は次式 (2.9) で表されるステップ関数である．

$$u_l(t) = \begin{cases} 1 & t \geq l \\ 0 & t < l \end{cases} \tag{2.9}$$

5．記憶細胞分化　抗体を解候補となる記憶細胞へ分化させる．Θ_υ が記憶細胞候補分化閾値 T_μ を超える抗体群中の最高評価値を持つ抗体を記憶細胞候補 μ とする．その後山登り法によって最良解候補 $\acute{\mu}$ を生成する．

ここで，山登り法は次のようにして構成した．現在の最良解候補 $\acute{\mu}$ の各遺伝子それぞれに 50% の確率で突然変異を起こす．突然変異の対象になった遺伝子には，-0.1 から 0.1 の範囲で生成される一様乱数を足す．この操作を N_{hn} 回繰り返して解候補群 $\mu_k\,(k = 1, 2, \ldots, N_{hn})$ を得る．その後，候補解群の中で，最高評価値を持つ解を新たな最良解候補 $\acute{\mu}$ とする．以上の操作を N_{hc} 回繰り返すことで山登り法を実現した．

記憶細胞が上限数 M に達していない場合は $\acute{\mu}$ を新たな記憶細胞 m として格納する．記憶細胞数が M に達している場合は，$\acute{\mu}$ との類似度 $\Psi_{m,\acute{\mu}}$ が最も高い記憶細胞 m を $\acute{\mu}$ と入れ替える．ただし，m の親和度が $\acute{\mu}$ より高い場合は入れ替えない．

6．サプレッサー T 細胞産生　類似した抗体が多く出ることを防ぐために抗体をサプレッサー T 細胞へと分化させる．記憶細胞候補 μ の全てをサプレッサー T 細胞 s_k (k = 1，2，．．．，N_h) へ分化させ，各 s_k との類似度 Ψ_{υ,s_k} がサプレッサー T 細胞類似閾値 T_s を超える抗体を全て消滅させる．その後，消滅した抗体と同じ数の，パラメータセットをランダムに決定した抗体を集団に加える．サプレッサー T 細胞は抗体を消滅させた後に全て消去される．

7．選択　新しい抗体を産生する前に評価値の低い抗体を更新する．抗体群を評価値順にソートし，評価値の低い $N_\Upsilon/2$ 個の抗体を淘汰対象とする．残った $N_\Upsilon/2$ 個の抗体は次の抗体産生で用いる．

8．抗体産生　ここでは抗体の生成を行う．交叉の親は淘汰対象にならなかった $N_\Upsilon/2$ 個から選ぶ．親は評価値が高い抗体が選ばれるように各抗体 μ_i の選出確率 P_i を求める．

$$P_i = \frac{\varepsilon_{\upsilon_i}}{\sum_j^{N_\Upsilon/2} \varepsilon_{\upsilon_i}} \tag{2.10}$$

$$\varepsilon_{\upsilon_i} = \frac{\Phi_{\upsilon_i}}{\sum_j^{N_\Upsilon/2} \Phi_{\upsilon_j}} \frac{1}{\Theta_{\upsilon_i}} \prod_j^{N_h} \left(1 - \Psi_{\upsilon_i,s_j} u_{T_e}(\Psi_{\upsilon_i,s_j})\right) \tag{2.11}$$

ここで ε_{υ_i} は抗体 υ_i の性能の期待値である．評価値 Φ_{υ_i} が高く，濃度 Θ_{υ_i} が低い抗体は高い期待値を持つ．またサプレッサー細胞が存在する場合，サプレッサー細胞 s に期待値閾値 T_e 以上近似した抗体はその類似度 $\Psi_{\upsilon_i,s}$ に従って期待値が低くなる．

確率によって選ばれた親が交差して生まれた子抗体群 N_c 個を評価値が高いものから順に $N_\Upsilon/2$ 個選び，淘汰対象の抗体群と入れ替える．抗体の生成に使用した交叉アルゴリズ

ムは SPX を用いた．ただし，生まれた子抗体群には交叉の他に突然変異率 p_m で突然変異を起こしている．α，β に起こった変異ついては値をランダムに，g，h に起こった変異には値を 0 になるように変異させた．これは遺伝子制御ネットワークが疎な構造となっていることから，g，h に値を持たないように強制させ，遺伝子間の関係を減らそうとする働きを持つ．

9. 終了処理　解候補である記憶細胞と現在推定中の抗体群の内，最も良い評価値を持つ個体を出力して終了する．

2.4.4　提案手法のアルゴリズム

これまでに説明した S-system モデル，免疫アルゴリズムを用いた提案手法のアルゴリズムをまとめると以下のようになる．

入力 時系列発現プロファイル E_E

終了世代数 N_R， 集団内抗体数 N_Υ

局所探索個体数 N_{hn}，局所探索回数 N_{hc}，　生成子個体群数 N_c

記憶細胞上限数 M，突然変異確率 p_m

濃度閾値 T_θ， 記憶細胞候補分化閾値 T_μ

サプレッサー T 細胞類似閾値 T_s， 期待値減少類似度閾値 T_e

出力 S-system モデルのパラメータセット

1. 初期世代生成

1.1. 世代数 $G = 1$ とする

1.2. 初期抗体群 υ_i, $(i = 1, \ldots, N_\Upsilon)$ を生成する

2. 評価

2.1. 各遺伝子の初期時点での発現量 X と υ_i の持つパラメータセットから時系列発現プロファイル E_{υ_i} を各抗体について計算する

2.2. E_E と E_{υ_i} より抗体それぞれの評価値 Φ_i を計算する

3. 終了判定 世代数 G が N_R 以上であれば 14 へ

4. 濃度計算

4.1. 全抗体 υ に対して類似度 $\Psi_{\upsilon,.}$ を求める

4.2. 全抗体に対して類似度 $\Psi_{\upsilon,\upsilon_i}$ が T_θ を上回った抗体 υ_i の数の割合により濃度 Θ_{υ_i} を計算する

5. **記憶細胞分化**

5.1. 濃度が T_μ 以上の抗体を記憶細胞候補群 μ へ分化させる

5.2. μ の中で最も評価値の高い個体 $\acute{\mu}$ を得る

5.3. $\acute{\mu}$ の全遺伝子に 50% の確率で範囲 $[-0.1, 0.1]$ の一様乱数を加えて N_{hn} 個の記憶細胞候補群 μ を作成する

5.4. このステップへの到達回数が N_{hc} 回になるまで v.ii. へ戻る

5.5. μ から最も評価値の高い個体を記憶細胞 m へと分化させる

6. **サプレッサー T 細胞分化**

6.1. 記憶細胞候補群 μ を全てサプレッサー T 細胞 s へ分化させる

6.2. s との類似度 $\Psi_{s,\upsilon}$ が閾値 T_s より高い抗体 υ を初期化する

7. **選択** υ_i を評価値順にソートし，下位半分を淘汰対象，上位半分を交叉親候補とする

8. **抗体産生**

8.1. 交叉親候補に対して Φ，Θ，$T_{\mu s}$ より期待値 ε を求めて交叉する

8.2. 生成された子抗体の内，評価が高い個体 N_Υ 個で淘汰対象の抗体を取り除く

8.3. G に 1 を加算する

8.4. 2. 評価へ戻る

9. **終了処理** 抗体群の中と記憶細胞群の中で Φ が最も高いパラメータセットを出力する

2.5 実験と考察

提案方法が従来手法に比べ，精度の向上が見込まれたかどうかを調査するため実験を行った．ここではその方法と結果，および考察を述べる．

2.5.1 実験条件

ここでは実験で使用するデータと実行環境について述べる．

入力として与える時系列データとして，シミュレーションデータと実データの二つを用いた．シミュレーションデータは，S-system モデルを用いた遺伝子制御ネットワーク推

定の際によく用いられる [20, 32, 34]，表 2.1 に示す S-system モデルのパラメータセットからシミュレーションした 5 遺伝子の時系列データ（図 2.9）を用いる．ただし，初期遺伝子発現量を表 2.2 として 30 時点までの時系列遺伝子発現プロファイルを計算した．

i	α_i	g_{i1}	g_{i2}	g_{i3}	g_{i4}	g_{i5}	β_i	h_{i1}	h_{i2}	h_{i3}	h_{i4}	h_{i5}
1	5.0	0.0	0.0	1.0	0.0	-1.0	10.0	2.0	0.0	0.0	0.0	0.0
2	10.0	2.0	0.0	0.0	0.0	0.0	10.0	0.0	2.0	0.0	0.0	0.0
3	10.0	0.0	-1.0	0.0	0.0	0.0	10.0	0.0	-1.0	2.0	0.0	0.0
4	8.0	0.0	0.0	2.0	0.0	-1.0	10.0	0.0	0.0	0.0	2.0	0.0
5	10.0	0.0	0.0	0.0	2.0	0.0	10.0	0.0	0.0	0.0	0.0	2.0

表 2.1　S-system モデルのパラメータ

遺伝子 1	遺伝子 2	遺伝子 3	遺伝子 4	遺伝子 5
0.7	0.12	0.14	0.16	0.18

表 2.2　初期発現量

実データとして，マウス未分化間葉系細胞株 ST2（理研、細胞番号 RCB0224）を脂肪細胞へ分化誘導した際の RNA をマイクロアレイによって計測した時系列発現プロファイルを用いた．この ST2 細胞株は RPMI1640 培地から，10 % FBS（ウシ胎仔血清），0.5 m M の DMEM（3-イソブチル-1-メチルキサンチン），0.25μM の DEX（デキサメタゾン），5μg/ml のインスリンを含むインスリン-トランスフェリン-セレニウム添加剤，1μM のロシグリタゾンを添加した DMEM 培地（ダルベッコ・フォークト変法イーグル最小必須培地）へ交換して誘導し，誘導から 48 時間後に 10% の FBS を添加した DMEM 培地へ交換することによって脂肪細胞分化を誘導している．その際の発現量を Affymetrix GeneChip Mouse Genome 430 2.0 Array のマイクロアレイから取った，全 30 時点からなる時系列発現プロファイルを用いた．このプロファイルは分化刺激後から 1 時間間隔で脂肪細胞分化刺激後 0 時間から 30 時間までの 31 時点存在する（図 2.10）．解析に用いた遺伝子は PPAR γ と C/EBP α，C/EBP β，C/EBP δ，C/EBP γ の C/EBP ファミリーの計 5 遺伝子で，これらの遺伝子は制御関係にあることが知られている [38]．

今回使用したプログラムは c++ で開発し，24CPU，Intel Xeon 2.67GHz のクラスタマシンで実行した．

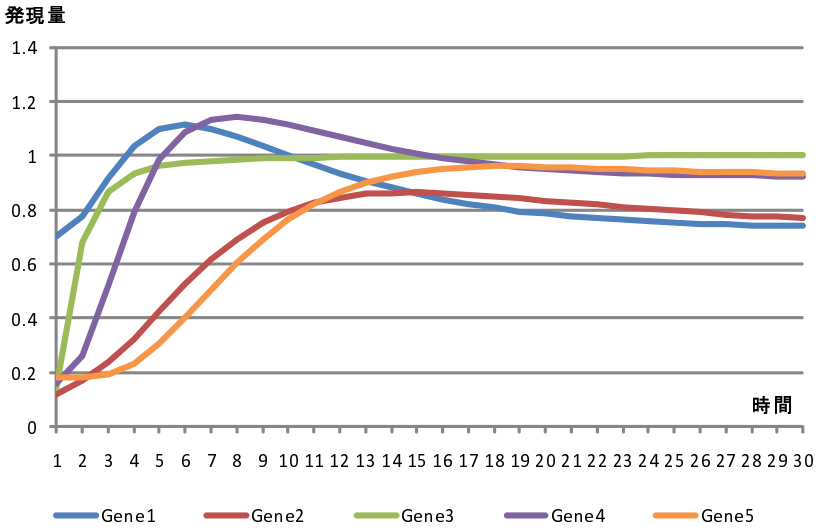

図 2.9　表 2.1 のパラメータに表 2.2 を与えて得られた時系列プロファイル

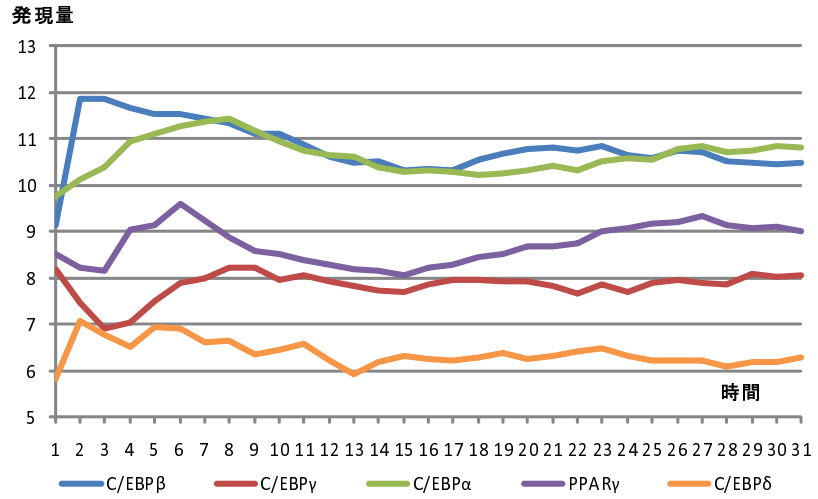

図 2.10　Mouse の間葉系幹細胞の脂肪細胞分化刺激後の時系列プロファイル

2.5.2　実験内容

入力として，前述のシミュレーションデータと実データを用いる．各データを表現する S-system モデルパラメータを GA と IA の二つの手法で推定し，提案手法の有効性を示す．シミュレーションデータではそれぞれの手法で 10 試行づつ行い，実データではそれぞれの手法で 30 試行づつ行う．評価の指標としては，各データでの各試行の評価値による解の安定性，および 10 試行中の最良解でのシミュレーション精度とする．また，シ

ミュレーションデータでの実験については解となるパラメータが分かっているので，得られた最良解と解との差も用いて評価する．

ここで，解の安定性とは，10 試行で得られた評価値にどの位ばらつきが見られるかを示す尺度で，標準分散 SD の大小で判定する．精度 A は 10 試行での最高評価値である．また，パラメータ間の誤差 E_p は平均二乗平方根誤差とする．SD，E_p は小さいほど良い推定能力であると言え，A は高い値ほど入力データを再現できていることを指す．具体的に E_p は，解となるパラメータを $\vec{x_0}$，得られたパラメータを $\vec{x}$ とし，次のように求める．

$$E_p = \frac{\mid \vec{x_0} - \vec{x} \mid}{\sqrt{2N_g(N_g+1)}} \tag{2.12}$$

実験に際して，入力として与えた閾値などのパラメータを表 2.3 に示す．

データ	シミュレーションデータ		実データ	
手法	GA	IA	GA	IA
世代数 N_R	100000	100000	100000	100000
個体数 N_Υ	200	200	600	600
突然変異率 p_m	0.005	0.005	0.005	0.005
交叉法	SPX	SPX	SPX	SPX
世代交代法	DDA	—	DDA	—
α，β の値域	0.00～15.00	0.00～15.00	0.00～30.00	0.00～30.00
g，h の値域	-3.00～3.00	-3.00～3.00	-3.00～3.00	-3.00～3.00
記憶細胞上限数 M	—	50	—	50
局所探索個体数 N_{hn}	—	20	—	20
局所探索回数 N_{hc}	—	10	—	10
生成子個体群数 N_c	400	400	1200	1200
濃度閾値 T_Θ	—	0.35	—	0.35
記憶細胞候補分化閾値 T_μ	—	0.70	—	0.70
サプレッサー T 細胞類似閾値 T_s	—	0.3	—	0.3
期待値現象類似度閾値 T_e	—	0.3	—	0.3

表 2.3　実験で設定したパラメータや条件

2.5.3　実験結果

シミュレーションデータ　各試行の評価値のヒストグラムをあらわしたものが図 2.11 で，結果をまとめたものが表 2.4 である．ただし，実行時間は 10 試行の平均値である．図の x 軸は評価値を表し，高いほど良い推定である．y 軸が頻度を指している．

手法	SD	A	E_p	実行時間 (s)
GA	0.165	0.950	88.0	47592.19
IA	0.0363	0.971	61.1	46736.15

表 2.4　シミュレーションデータでの結果

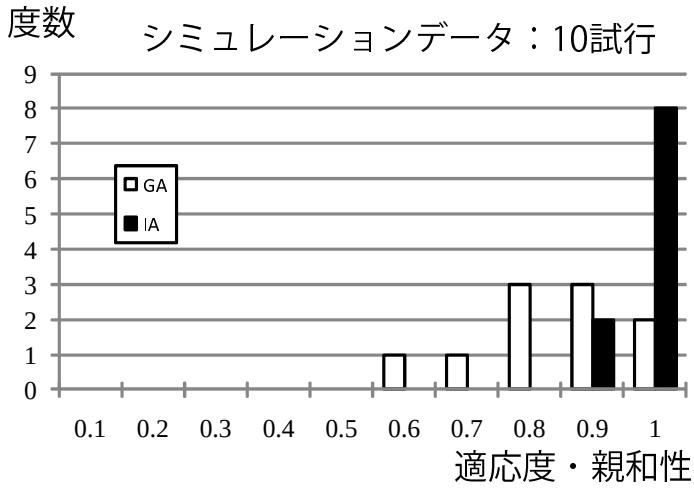

図 2.11　シミュレーションデータでの各 10 試行の結果

また，GA と IA の推定結果に有意差が見られるかどうかをノンパラメトリック検定である Wilcoxon の順位和検定で検定した．帰無仮説として 10 試行の平均が等しい，対立仮説として 10 試行の平均は等しくない，となる．また，有意水準は 5% とする．結果，p-value<0.000325 となり，帰無仮説が棄却されるため有意差があると示された．

推定の結果，最良評価値として得られたパラメータは，GA が表 2.5，IA が表 2.6 となっている．このパラメータに入力データの初期発現量を与えてシミュレーションした結果が図 2.12 と図 2.13 である．濃い色が各手法で得られたパラメータからシミュレーションしたデータ，薄い色が入力データである．

i	α_i	g_{i1}	g_{i2}	g_{i3}	g_{i4}	g_{i5}	β_i	h_{i1}	h_{i2}	h_{i3}	h_{i4}	h_{i5}
1	2.54	-2.87	1.50	0.00	0.29	-2.15	11.86	2.56	0.00	3.00	1.11	2.05
2	6.51	0.93	-2.31	0.00	1.65	1.13	11.66	0.00	2.40	0.00	0.00	0.00
3	8.6	-2.26	1.19	-2.34	1.91	-1.73	13.93	0.45	0.00	3.00	0.50	0.00
4	3.6	-2.59	1.21	0.08	0.76	-2.37	9.89	0.00	1.29	2.51	1.66	0.56
5	5.34	0.00	3.00	0.00	2.21	-2.77	3.60	-1.62	2.86	0.00	0.00	0.87

表 2.5　GA によって得られたパラメータ

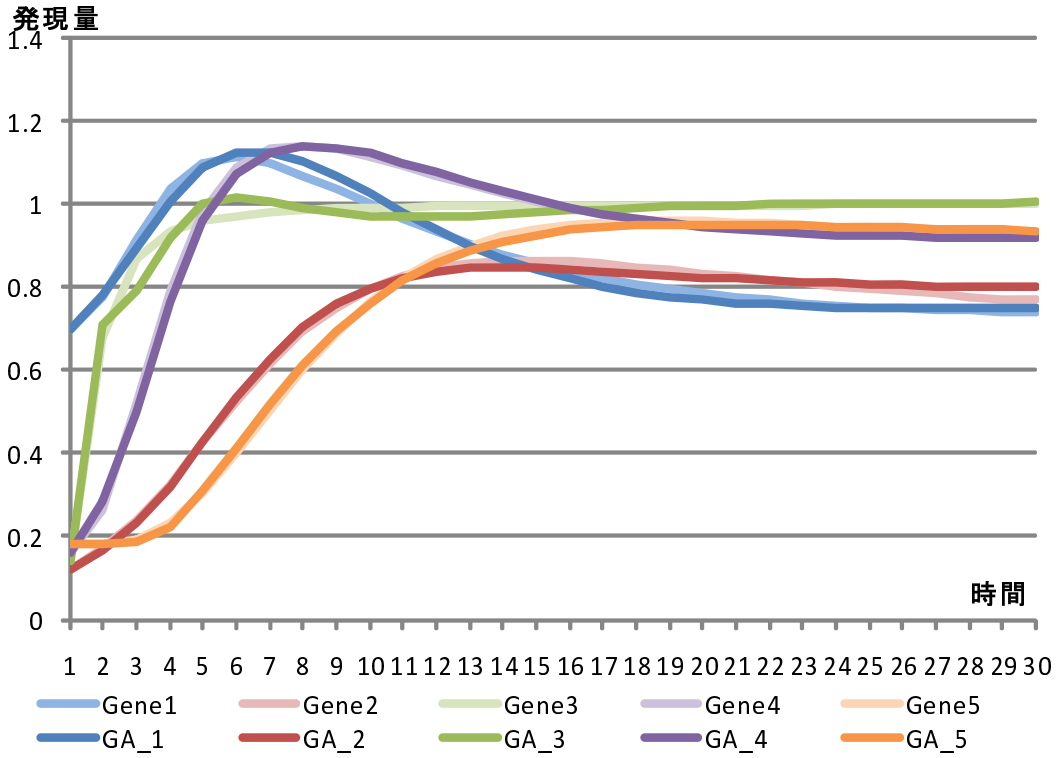

図 2.12　GA によって得られたパラメータからシミュレーションした時系列プロファイル

i	α_i	g_{i1}	g_{i2}	g_{i3}	g_{i4}	g_{i5}	β_i	h_{i1}	h_{i2}	h_{i3}	h_{i4}	h_{i5}
1	3.28	-0.03	0.21	1.41	-0.55	-0.93	11.2	2.18	2.47	0.43	-0.58	-0.75
2	5.30	1.63	-1.35	-0.48	1.36	0.36	7.34	0.04	1.75	-0.06	1.12	0.64
3	13.52	-0.22	0.05	0.00	-1.34	0.45	13.03	-0.55	-0.05	2.75	-0.14	0.62
4	7.64	1.93	-1.57	2.68	-0.45	0.52	13.45	1.11	0.87	1.09	2.77	1.03
5	9.19	-1.42	2.20	-0.92	2.08	-1.66	12.12	0.51	0.67	0.41	0.23	2.67

表 2.6　IA によって得られたパラメータ

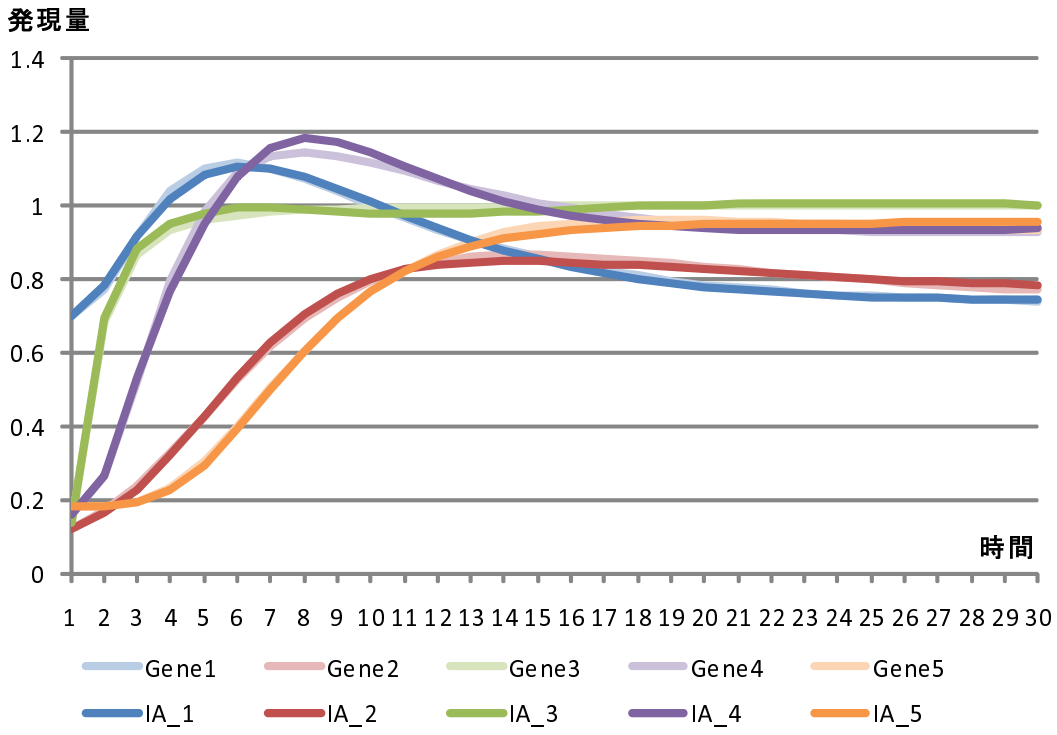

図 2.13　IA によって得られたパラメータからシミュレーションした時系列プロファイル

また，推定がどの様に進んだかを可視化するため，最良評価値を出した試行の全個体の評価値の推移をグラフにプロットする．また，個体の持つパラメータを色で表現する．赤色が強ければ反応速度が早く，緑色が強ければ正の促進，分解関係が強く，青色が強ければパラメータをネットワーク状に表したときに遺伝子制御ネットワークの様に疎である度合いが強い，とする．加えて，シミュレーションデータでは彩度を解との距離を表す事にする．色が鮮やかになればなるほどその個体の持つパラメータは解と近い．色とパラメータの関係をまとめると図 2.14 のようになる．

各手法の推定の推移を 1 世代ごとにプロットした（図 2.15，図 2.16）．x 軸が世代数で右に行けば推定が進んでいることになる．奥行きである y 軸は世代内の評価値順にソート

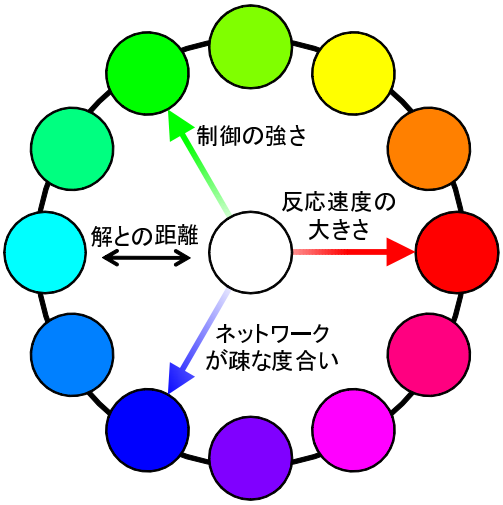

図 2.14　パラメータから色への変換

した個体を表す．z 軸は評価値で，上にある個体ほど良い評価を持っている．

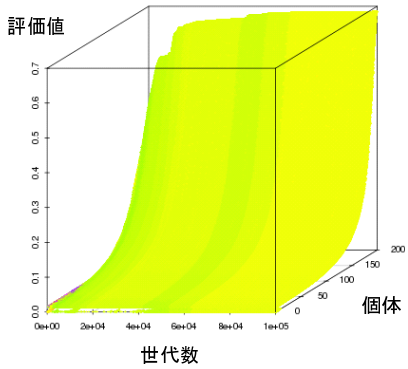

図 2.15　GA での推定の推移

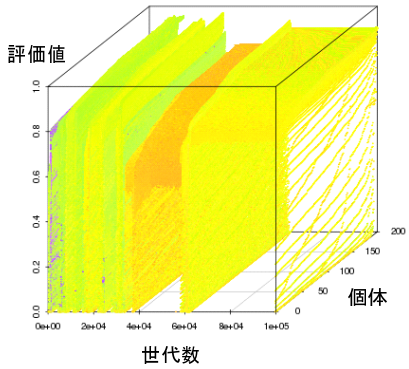

図 2.16　IA での推定の推移

実データ　シミュレーションデータ同様，各試行の評価値をヒストグラムにしたものが図 2.17 で，結果をまとめたものが表 2.7 である．ただし，実行時間は 30 試行の平均値である．

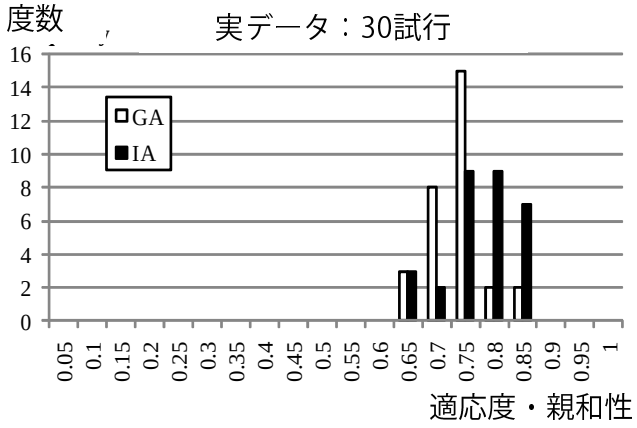

図 2.17　実データでの各 30 試行の結果

手法	SD	A	実行時間 (s)
GA	0.0411	0.808	122,807
IA	0.0573	0.839	143,843

表 2.7　実データでの結果

また，GA と IA の推定結果に有意差が見られるかどうかをノンパラメトリック検定である Wilcoxon の順位和検定で検定した．帰無仮説として 30 試行の平均が等しい，対立仮説として 30 試行の平均は等しくない，となる．また，有意水準は 5% とする．結果，p-value<0.00013 となり，帰無仮説が棄却されるため有意差があると示された．

推定の結果，最良評価値として得られたパラメータは，GA が表 2.8，IA が表 2.9，このパラメータに入力データの初期発現量を与えてシミュレーションした結果が図 2.18 と図 2.19 である．

i	α_i	g_{i1}	g_{i2}	g_{i3}	g_{i4}	g_{i5}	β_i	h_{i1}	h_{i2}	h_{i3}	h_{i4}	h_{i5}
1	9.65	0.18	0.28	0.15	0.79	1.56	24.74	1.12	0.19	0.2	0.14	0.65
2	24.31	0.67	0.05	0.14	0.12	0.3	20.48	0.11	0.89	0.11	0.17	0.15
3	4.63	-2.51	0	1.73	0.58	-0.18	30	0.03	0.03	-2.55	0	0
4	0.02	0.5	0.01	0.02	1.5	-0.21	0	0.01	0.02	0.59	0	0.76
5	25.59	0.05	0.22	0.9	0.05	0.09	0.8	0.16	0	0.12	0.12	2.96

表 2.8　GA によって得られたパラメータ

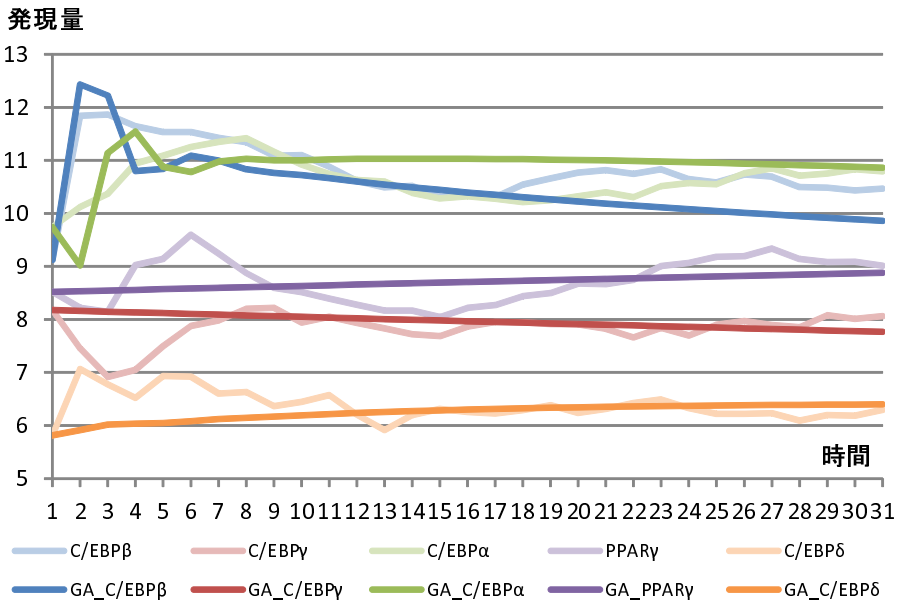

図 2.18　GA によって得られたパラメータからシミュレーションした時系列プロファイル

i	α_i	g_{i1}	g_{i2}	g_{i3}	g_{i4}	g_{i5}	β_i	h_{i1}	h_{i2}	h_{i3}	h_{i4}	h_{i5}
1	29.92	-0.81	0	0	2.84	0	29.99	0.02	0.2	1.9	1.72	-2.39
2	28.16	2.97	-1.58	-0.05	-1.01	0	11.2	0.66	3	-1.93	-1.03	0.79
3	1.35	0.99	0.12	-0.2	1.32	0.74	29.65	0.29	0.01	3	-0.43	-1.94
4	8.12	-1.92	-0.04	3	-1.98	0.01	0	-2.42	0	2.13	1.68	-2.99
5	29.87	0.83	2.41	-0.8	-0.93	0.02	1.54	0.11	0.78	3	-1.52	0.2

表 2.9　IA によって得られたパラメータ

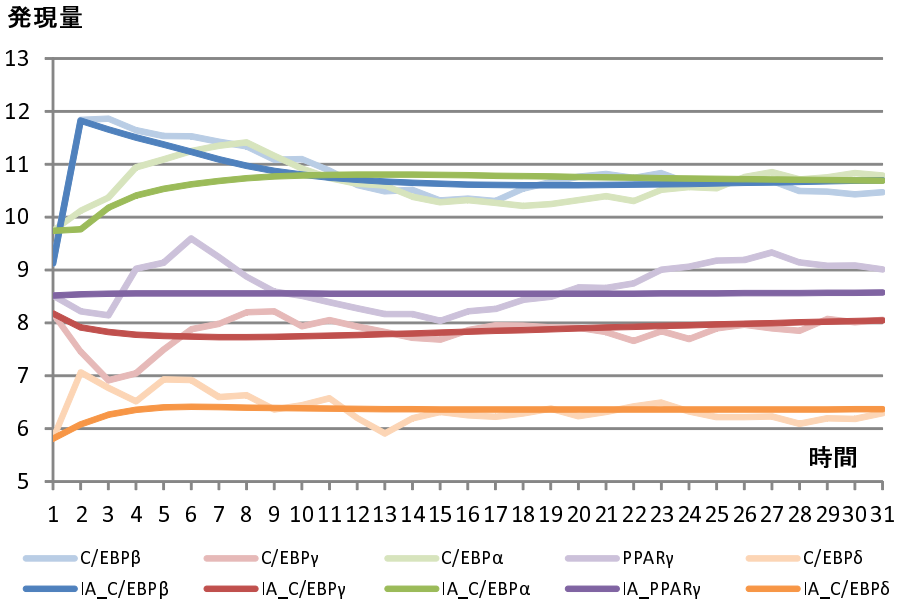

図 2.19　IA によって得られたパラメータからシミュレーションした時系列プロファイル

また，各手法の推定の推移を 100 世代ごとにプロットしたものが図 2.20 と図 2.21 である．

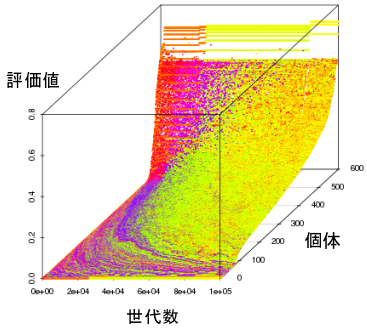

図 2.20　GA での推定の推移

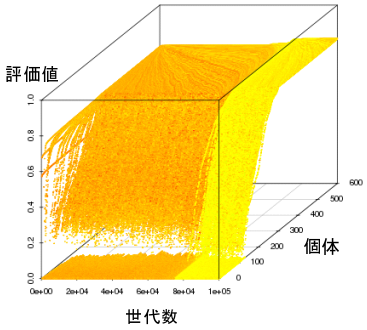

図 2.21　IA での推定の推移

2.5.4 提案手法の評価と考察

ここでは，まずデータごとに結果の評価をした後に，提案手法の能力について考察する．

シミュレーションデータでの実験についての評価 図 2.11，表 2.7 を見ると，シミュレーションデータにおいては IA は GA と比較して良い推定をしたと言える．図 2.11 から，IA は GA に比べて高い評価値を持つように分布しており，表 2.7 の A から IA の GA の最大の評価値より良い評価値を持つことが分かる．得られたパラメータの誤差 E_p も IA の方が小さく，パラメータの同定を改善していると言える．シミュレーション結果を見ても遺伝子 3 の挙動が GA よりもよく再現できており，E_s が小さくシミュレーションによる発現量の再現が上手くいっている．実行時間についても IA は GA より若干早い．これは計算量の多い交叉に用いる個体が GA の半分になっているためであると考えられる．

図 2.15 と図 2.16 の推定の推移を見ると GA がほぼ同じような色で推移しているのに比べて，IA は変化に富んだ推定をしているのが分かる．所々で評価値が落ちているのは解の記憶と初期化が起こり，局所解へ収束した個体を削除したためである．その結果，パラメータの探索範囲が広がり，より良い解を持ちやすくなると考えられる．

実データでの実験についての評価 実データにおいても IA は GA より良い推定結果を出した．図 2.17 より，IA は GA に比べて評価値が向上している．また，シミュレーションデータの時よりも若干のばらつきが見られるものの，標準誤差はシミュレーションデータと実データとでほぼ同じで安定した性能を持っていることが分かる．

シミュレーション結果を見ると GA と IA の両方共 C/EBPβ，C/EBPα の二遺伝子の挙動は再現されていて，IA は GA に比べて良くフィットしている．しかし，その他の遺伝子はどちらの手法でも初期に見られる特徴ある挙動を再現せず定常状態に陥っている．これは，推定の世代数が十分でなかった，発現量の変動が小さいために挙動を捉えられなかった，設定したパラメータの範囲内に解が無かった，という可能性がある．

実行時間が GA より IA の方が大きくなっているのは個体数を増やしたためである．IA では抗体の濃度を評価するために全抗体間で類似度を計算する必要がある．この操作は GA に無く，n を抗体数とすると計算量が $O(n^2)$ となる．この計算が交叉に用いる個体数が GA の半分であることの計算削減量を上回り，実行時間の増加を引き起こしたと考えられる．

図 2.20 と図 2.21 の推定の推移を見ると GA は色の変化に富んだ推移をしているが，IA

はほぼ同じ色で推移している．また，シミュレーションデータでの推定では見られた記憶細胞分化による急激な評価値の低下が見られない．これは記憶細胞分化が十分起こらなかったことを指し，IA の特徴をシミュレーションデータでの実験より生かしきれていない事を指している．また，図 2.20 を見ると，シミュレーションデータよりも評価値を持った個体が多く存在しているのが分かる．これはデータがもつ摂動によって，シミュレーション結果の挙動に幅を持たせることができるため，局所解がより多く存在してしまっている事を示していると考えられる．IA で記憶細胞分化が十分行われなかった原因としては，局所探索が十分行われない状況が考えられ，局所解として取り得るパラメータの範囲が近似解と見なされないほど広い，もしくは近似解と見なされない範囲にある複数の局所解に収束してしまい，抗体が分散してしまい濃度が高くならず解の消去が起こらなかった可能性が考えられる．

提案手法の能力　シミュレーションデータと実データの結果から，IA による S-system モデルのパラメータ推定は従来手法である GA を用いたものより推定精度が安定し，向上した．また，得られたパラメータからのシミュレーション結果も GA より IA の方が良くフィッティングしており，IA は S-system モデルのパラメータ推定問題に効果のあるアルゴリズムだと言える．

しかし，IA の特徴を生かして探索を行うには局所解の形状による問題点が見つかっている．データが持つ局所解の形状によってパラメータを適切に変えるなどの対応によりこの問題点を解決することで，更なる精度向上が見込まれる．

2.6　結言

本研究では遺伝子制御ネットワークモデルの一つである S-system モデルを細胞分化の時系列遺伝子発現プロファイルへ適用した際の推定精度の向上を目的とし，従来の探索手法である GA を S-system モデル推定へ適用した際の問題点である局所解への収束を避けて，探索範囲が広く解の多様性を持つ IA を適用する手法を提案した．これによって，入力された時系列発現プロファイルをよく再現する高精度のパラメータ推定が可能になり，従来より細胞分化における遺伝子発現のシミュレーションや動態の解析に利用することが可能となった．

提案手法の性能を評価するため従来手法である GA と比較実験を行った結果，提案手法は従来手法よりも良い評価値を得て，推定精度の向上が認められた．しかし，IA のパラ

メータによる性能変化が大きく，また，S-system モデルのパラメータの値の範囲の設定やデータに依存した問題点も存在する．実験条件を変更した場合の推定性能の変化について調査し，よりパラメータ設定を簡略化できるような効率的なアルゴリズムを開発することができれば更に良い推定結果が得られると期待できる．

第 3 章　時系列プロファイル分割による細胞分化の動的な遺伝子制御ネットワーク構築手法

3.1　緒言

近年の生命科学の発展により多細胞生物において器官や臓器などで異なる細胞が生まれる過程である細胞分化は，遺伝子の発現によって制御されていると分かってきた．遺伝子は親から子へ受け継がれる生体の形質などの情報を格納しており，細胞内の遺伝子の働きによって器官や臓器などの一般に遺伝子は発現という過程を経て，タンパク質となることによって機能を果たす．遺伝子の発現量を計測したデータを遺伝子発現プロファイルと呼び，様々な実験環境下での遺伝子発現プロファイルが計測されてきている．中でも，遺伝子の発現量の時間的変化を計測した時系列発現プロファイルは細胞内のどの時期にどのような遺伝子が働いているのかを計測できるため，薬剤応答や刺激への反応などの条件下での遺伝子の振る舞いを観測する重要なデータとなっている．これらのデータを用いて細胞分化のメカニズムを解析することは，再生医療の分野のみならず，創薬においては分化のコントロールによる様々な疾病の改善や分化の影響範囲の予測による副作用軽減，農学においては有用物質の生産や増産などの応用に貢献することが期待される．

遺伝子の発現量を計測した遺伝子発現プロファイルは一般に公開されているデータベースにも登録されるようになり，細胞分化においても多数の実験による結果を解析することが容易になってきている．遺伝子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バイオインフォマティクスの分野において，遺伝子間の制御関係を理解するための基本的な解析でありながら課題の多い分野である．これまで数々の手法が提案されてきており，第 2 章では定量的な解析手法を提案した．しかし，この手法は実験データに含まれるノイズの影響を強く受けるため，ごく小規模な遺伝子制御ネットワークにしか用いることができない限界がある．大規模な遺伝子制御ネットワーク解析にはノイズに強い手法を用い，得られた制御関係をもとに定量的な解析手法を適用することで，網羅的に制御関係の強さを求めることができるようになると考えられる．

時系列発現プロファイルを用いた大規模な遺伝子制御ネットワーク解析としてダイナミックベイジアンネットワークモデル [12] による推定が広く使われてきた．しかし，一般的なダイナミックベイジアンネットワークモデルの推定手法は細胞分化過程のように遺伝子制御ネットワークが変化することを考慮していない．そのため，細胞分化などの遺伝子

の制御が変化することに対応した手法が求められている．

遺伝子制御ネットワークが変化することへ対応した手法としてはノードセットセパレーション法 [18] が存在する．ノードセットセパレーション法は 1 つの時系列発現プロファイルから複数の時間的な順序関係を持つ遺伝子制御ネットワークを推定することによって，遺伝子制御関係の変化を捉えた動的な遺伝子制御ネットワークを推定する手法である．この手法は薬剤応答のような短時間に遺伝子の働きが大きく変わる現象を対象としており，時系列発現プロファイルの時点数が少ない場合でも，どの遺伝子がどの時点で働いているのかということが分かる．しかし，遺伝子の制御関係が時系列発現プロファイルの中で変化しないという仮定を含んでおり，分化のように数日掛かるような長い時間計測したような時点数の大きなデータになると，制御の変化を無視することができない．

上記問題点を解決するため，時系列発現プロファイルの分割方向を遺伝子方向ではなく時間方向にすることで，データ中の期間に制御関係の変化が起こっていたと場合に対応した手法を提案する．時間方向に分割した場合，制御の変化が起こった前後で分割することができれば，変化前と変化後の遺伝子制御関係の推定が可能になると考えられる．そこで，等間隔に分割する方法であるスライディングウィンドウ方法 [39] の考え方を適用することによって動的な遺伝子制御ネットワークの推定精度の向上が期待できる．

本章では，脂肪細胞の時系列発現プロファイルを用いて，時系列分割手法であるスライディングウィンドウ法を適用することで動的な遺伝子制御ネットワークを推定し，その結果と評価を行う．また，2 章で得られた定量的な解析手法による結果との整合性を確認し，制御の強さを当てはめることで，ネットワークにおける重要な遺伝子を抽出できることを示す．

3.2 細胞分化における遺伝子制御関係の変化

3.2.1 細胞分化と遺伝子制御ネットワーク

細胞分化はヒトなどの多細胞生物において，細胞が特殊化して分化前の細胞には見られない特有の形質を持つようになる過程のことをいう．多細胞生物は 1 個の受精卵から細胞数を増やして様々な組織，器官，臓器を形成する際に細胞分裂と細胞分化を繰り返して複雑な生体機能を構成する．通常，ヒトなどの動物細胞では細胞分化は不可逆であり，受精卵の状態では全ての組織へ分化できる全能性を持つが，分化が進むと分化できる細胞の種類が決まってしまう．この細胞分化や分化可能な細胞を決定づけるのは要因の 1 つに遺伝子の発現の変化が関わっていると分かってきている．

遺伝子は親から子孫が持つべき性質を規定した情報を伝達する因子であり，A(アデニン)，T（チミン)，G（グアニン)，C（シトシン）の 4 種類からなる DNA（デオキシリボ核酸）の塩基配列によって構成されている．遺伝子は生体内で数ステップの反応を経てタンパク質を生成することで役割を果たす．その反応は次のようになる．まず遺伝子領域の塩基配列が mRNA(メッセンジャー RNA) に転写される．次に mRNA が翻訳されてアミノ酸の重合体であるアミノ酸ポリマーが生成され，折りたたまれて特有の立体構造を成す．この生成された立体構造がタンパク質である．遺伝子からタンパク質が生成されるまでの，この一連の過程を遺伝子の発現と呼ぶ.

遺伝子は他の遺伝子の発現を制御し，発現の促進や抑制といった働きかけをするものがある．そのような遺伝子は他の遺伝子の転写を制御する因子であることから転写因子と呼ばれ，細胞分化などの生体機能の調節に重要な役割を持つ．複数の遺伝子間の制御関係をより理解するため，遺伝子をノード，制御関係をエッジとして見なしたネットワーク状に書いた遺伝子制御ネットワークが用いられている．遺伝子制御ネットワークによって，複雑な遺伝子制御関係の構造をグラフ理論や制御理論によって解析して理解することが可能になる.

遺伝子が発現した量をそのまま計測することはタンパク質の複雑さから困難であり，中間生成物である mRNA の量を発現量として計測することが行われてきた．遺伝子が生成した mRNA の量を知る方法の一つにマイクロアレイ技術があり，これを用いることで一度に数千から数万の遺伝子についての発現量を測定した遺伝子発現プロファイルを得ることができる．様々な実験に合わせて遺伝子発現プロファイルを取ることがされており，時系列に沿った遺伝子発現量の推移を見る際には時系列発現プロファイルを用いる．遺伝子の制御関係はこの時系列発現プロファイルから推定することができ，ネットワークモデルに当てはめることで遺伝子制御ネットワークを推定することができる.

3.2.2 動的な遺伝子制御ネットワーク

細胞分化において，遺伝子制御ネットワークの構造が変化することが知られている．細胞の形質を決定付ける大きな要因として，細胞種特有の遺伝子が機能していることが挙げられる．それを実現するために分化中に遺伝子制御ネットワークが変化していく．制御関係の変化には，制御関係があった遺伝子間に制御が見られなくなる，制御関係が見られなかった遺伝子間に制御が起こる，という 2 種類が考えられる．現在はクロストーク遺伝子 [6][7] やクロマチンリモデリング [4][5] などの要因が制御関係の変化に関わっていると考

えられている．

クロストーク遺伝子は複数の分化経路を持つ細胞がある細胞へ分化する際に，他の分化経路を誘導する遺伝子を抑制し，1 つの分化経路を進める働きをする遺伝子である．分化先が決定した後は他の分化経路を促進する刺激を加えたとしても他の分化経路を進める遺伝子が転写，発現できなくなる．また，クロマチンリモデリングは遺伝子が制御を行うための DNA 領域に結合できるかどうかを変化させ，遺伝子制御ネットワークの構造を大きく変化させる．そのため，遺伝子が同じように発現していたとしてもクロマチンの状態によって転写制御ができるかできないかが変わってくる．このように，ネットワークの構造が時間とともに変化していくことを動的な遺伝子制御ネットワークと呼ぶ．

3.3 動的な遺伝子制御ネットワークの推定手法

動的な遺伝子制御ネットワークとは，時間の経過とともに遺伝子の制御関係が変化することをネットワーク状に表したものである．一般的な遺伝子制御ネットワークモデルの推定手法は制御関係の変化が無いものと仮定しており，そのまま動的な遺伝子制御ネットワークの推定へ適用することができない．そのため，制御関係の変化を推定するための手法が必要となる．この節では，動的な遺伝子制御ネットワークを推定するため考案された手法であるノードセットセパレーション法とその問題点について述べる．

3.3.1 ノードセットセパレーション法

ノードセットセパレーション法 [18] は，時系列発現プロファイルの時点ごとに働いている遺伝子の組を作り，その遺伝子群でダイナミックベイジアンネットワークモデル [12] に基づいてネットワーク推定を行っていくことで動的な遺伝子制御ネットワークを推定する手法である（図 3.1）．ダイナミックベイジアンネットワークは時系列発現プロファイルを 1 次のマルコフ過程とみなしており，前時点で働いている遺伝子の影響を受けて次の時点で働いている遺伝子が決まるとしている．以降，ダイナミックベイジアンネットワークとその推定手法，ノードセットセパレーション法による動的な遺伝子制御ネットワーク推定手法の説明を行う．

ダイナミックベイジアンネットワーク（DBN）[12] はベイジアンネットワーク（BN）[11] を拡張したネットワークモデルである．まず，BN を説明した後，DBN について説明する．

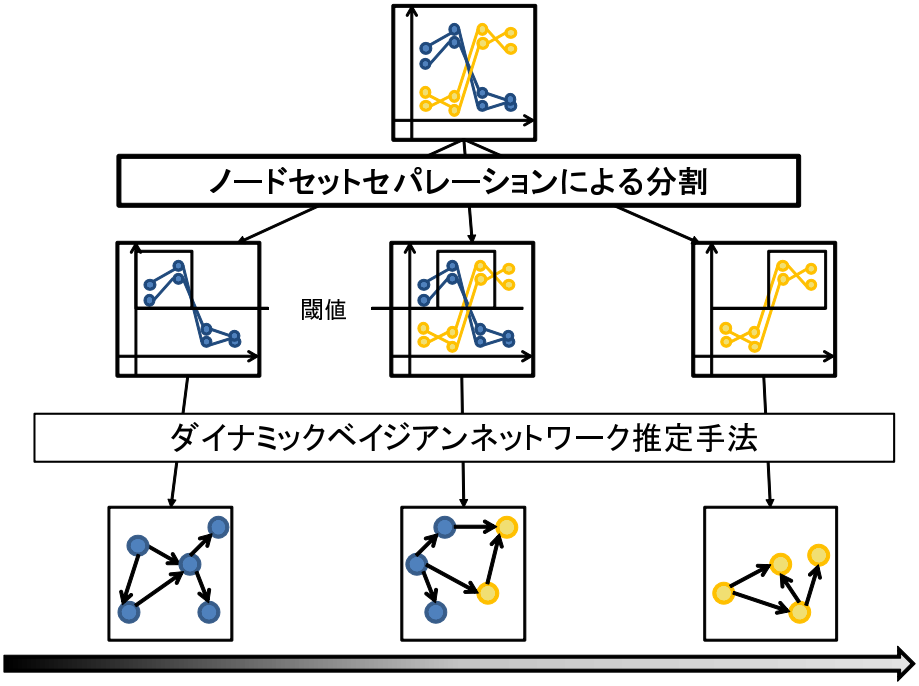

図 3.1　ノードセットセパレーション法概要図

ベイジアンネットワーク　BN はノードとなる確率変数間の条件付き確率をエッジとして閉路なし有向グラフによって表現する．ノード X_i からノード X_j にエッジが伸びているとき，X_j の取る値が X_i に依存していることを表している．依存関係にある確率変数同士がどのような確率関数に従って依存するかは，各変数に対する条件付き確率分布で表す．このとき，各変数は親となる変数以外の変数とは独立であると仮定する．

図 3.2 は確率変数が 0 か 1 の値を持つ離散値の場合の例で，変数 W，X，Y，Z の間の依存関係を表している．X は W に，Z は X と Y に依存している事を有向辺によって表現している．各変数は条件付き確率分布の表を持ち，この例では変数 W が 0 のとき X は 30% で 0，70% で 1 になり，W が 1 のとき X は 80% で 0，20% で 1 になることを表している．

BN は，データセット D を最も再現するようなネットワーク G を求める D が与えられた条件下でのネットワーク G のもっともらしさを事後確率 $p(G|D)$ として表す．この事後確率を直接的にデータセットからネットワークを決定するのは困難であるため，BN の推定ではベイズの定理から式 3.1 のようにして分解する．

$$p(G|D) = \frac{p(G) \cdot p(D|G)}{p(D)} \tag{3.1}$$

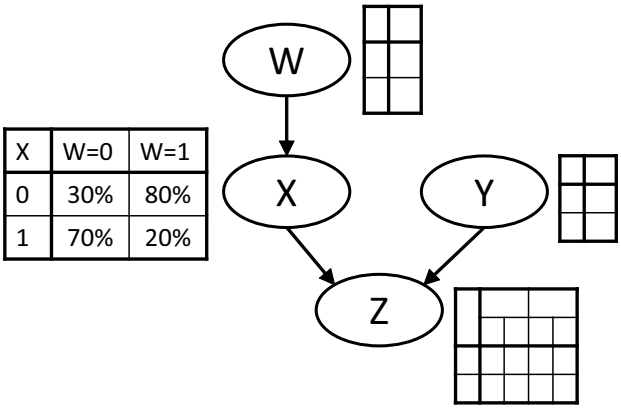

図 3.2　離散値でのベイジアンネットワーク例

$p(G)$ はネットワーク G の事前確率を表しており，入力のデータセットにおいて辺の有無が既知の場合，ネットワーク G にその辺があるかどうかの重みを付けることができる．$p(D|G)$ はネットワーク G が与えられた条件下でのデータセット D のもっともらしさである事後確率で，前述の確率分布から計算することができる．$p(D)$ はデータセット D の事前確率を表しているが，入力データである D は固定であることから $p(B|D)$ を求める際には省略される．従って，$p(G|D)$ は $p(G)$ と $p(D|G)$ に依存することになる．

BN の推定では，ネットワークの推定精度を $p(G|D)$ で定量的に評価する．このとき，ベイズの定理によって直接計算することが困難であった $p(G|D)$ をネットワーク G に依存する $p(G)$ と $p(D|G)$ で求めるように分解することで，評価関数 $p(G|D)$ を最大化するネットワーク G を求める問題と考えることができる．ネットワーク G は閉路なし有向辺で構成されるため，全件探索による計算量はノード数を N とすると $O(3^{N^2})$ という非常に膨大な数になる．そのため，一般的に BN の推定はネットワークスコアとして $p(G|D)$ を最大化するようなネットワーク G を探索的手法を用いて求めることが行われている．

ダイナミックベイジアンネットワーク　DBN[12] は，ベイジアンネットワークでは時間とともに推移するような時系列のデータを扱うことができない問題点を解決した手法である．

各確率変数 X_i に対して，時点 t での状態 $X_{i,t}$ を考える．時点 t と次の時点 $t+1$ での確率変数を別のノードとして考え，BN を推定することを繰り返し，得られたネットワークを結合することで DBN の推定が実現する．最終的に結合して得られるネットワークには閉路があっても問題ない．例を図 3.3 に示す．

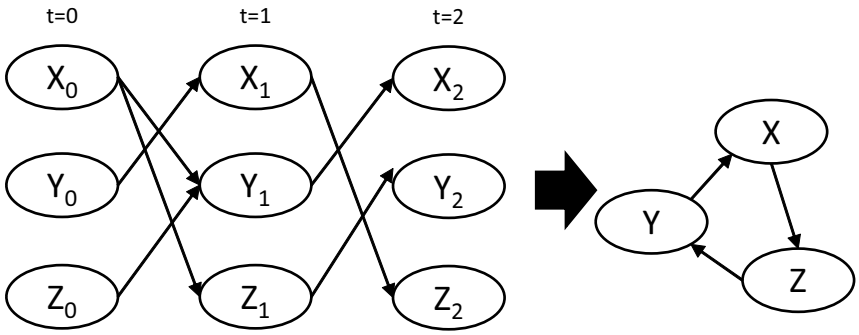

図 3.3　ダイナミックベイジアンネットワークへの拡張概要図

BN と比較して，DBN の利点としては時系列データを扱えること，閉路の推定ができることが挙げられる．特に，遺伝子制御ネットワークにおいては閉路の存在が確認されているため，時系列発現プロファイルを用いた遺伝子制御ネットワーク推定において広範に利用されている．

動的な遺伝子制御ネットワーク推定　ノードセットセパレーション法では各時点毎にコントロール群と比較して有意に働いているアクティブ遺伝子セットを決める．アクティブ遺伝子セットを $A_t = \{g_i : pv(gi, t) \leq \theta_t\}$ として表す．ここで，t は 1 から T までの時点，T は入力した時系列発現プロファイルの時点数，g_i は i 番目の遺伝子，$pv(g_t, t)$ は遺伝子 g_i の t 時点での p 値，θ_t は FDR から決定される t 時点での閾値である．その後，連続した 2 時点毎にノードセット $N_t = A_{t-1} \cup A_t$ を作る．ただし A_0 は空集合とする．この N_t を用いて全時点の時系列発現プロファイルからダイナミックベイジアンネットワークを推定し，動的な遺伝子制御ネットワーク G_t を得る．これによって，t 時点のネットワーク G_t ではどの遺伝子制御関係が働いているのかを知ることができる．時点 t の前後での遺伝子制御関係の変化は G_t とその前後のネットワーク G_{t-1}, G_{t+1} とを比較することで知ることができる．

3.3.2　従来手法の問題点

ノードセットセパレーション法は，時点数が少なく時点間で発現量が大きく変わるような時系列発現プロファイルに対して適用可能で，時系列発現プロファイルの時点数が多くなると制御関係の変化を正しく推定できないと考えられる．この手法で時点数が多い時系列発現プロファイルの全時点を用いてダイナミックベイジアンネットワーク推定手法を用

いると，2 つの遺伝子間で制御関係の変化が起こっていたとしても制御関係ありかなしかのどちらかしか推定結果として出力されず，偏った結果しか推定されない可能性がある．図 3.4 では，黒い両矢印で示した期間に制御関係が見られ，白い両矢印で示した期間では制御関係が見られないような遺伝子の時系列発現プロファイルの例を挙げている．この場合，2 つの遺伝子が閾値を超えてダイナミックベイジアンネットワーク推定手法にかけられたとしても，閾値を超えないような発現量が多く見られた場合や制御関係がなく無関係な挙動を多く見せた場合，どの時期で遺伝子を分割しても結果として推定されるのは制御関係がないネットワークになる．従って，この手法を時点数の多い時系列発現プロファイルに適用することは困難であると考えられる．

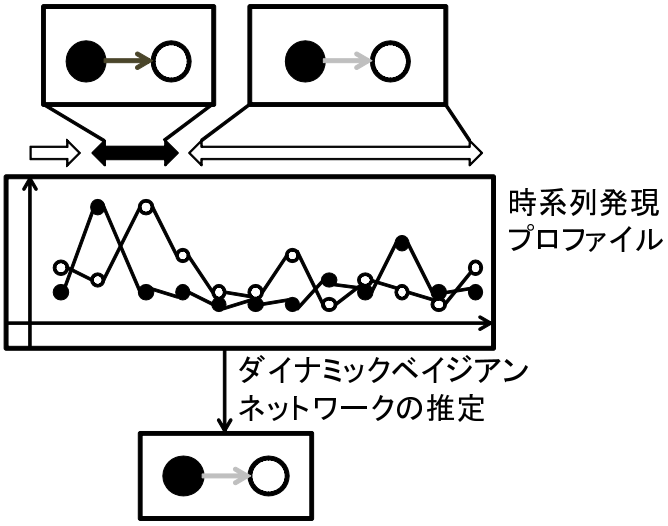

図 3.4　ノードセットセパレーション法の問題点の例

また，ノードセットセパレーション法では 2 つの遺伝子の発現量が上昇した場合のみを考えている点に問題点がある．他の遺伝子を抑制する働きなどにより発現量を低下させる制御関係は制御を受ける側の遺伝子の発現量が上昇しないため閾値を超えないと考えられ，推定が困難である．細胞分化においては，現在進行している分化経路以外の分化経路で発現する遺伝子を抑制して現在の分化を進める働きを持つクロストーク遺伝子 [6][7] が存在していると報告されており，抑制の制御関係を無視することはできない．

3.4　時系列発現プロファイルの時間方向分割による手法の提案

動的な遺伝子制御ネットワークを推定する手法であるノードセットセパレーション法に存在する制御の変化を推定できない問題点を解決するため，本章ではノードセットセパ

レーションの分割手法を変更し，時系列発現プロファイルを時間方向に分割することによって動的な遺伝子制御ネットワークを推定する手法を提案する．本節では，時間方向分割の有効性の説明と既存手法であるスライディングウィンドウ法の適用方法について説明する．

3.4.1 目的と概要

本研究では，細胞分化における網羅的で動的な遺伝子制御ネットワークを推定することが目的である．細胞分化ではクロストーク遺伝子やクロマチンリモデリングによって制御関係が動的に変化する．ネットワークの構造が変化する過程を 1 つの遺伝子制御ネットワークで記述することは困難である．そのため，制御関係の変化毎に複数のネットワークを記述することで，制御関係の構造の変化を捉えた動的な遺伝子制御ネットワークとして記述することを考える．例えば，遺伝子の制御関係が 3 段階に変化するなら 3 つのネットワークで制御関係の構造変化を表すことで，初期，中期，後期の 3 段階それぞれの遺伝子制御関係を把握することができる．このようにして，動的な遺伝子制御ネットワークを複数の遺伝子制御ネットワークで表すことができる．

細胞分化における動的な遺伝子制御ネットワークを推定するには，細胞分化の時系列発現プロファイルから制御関係が変化するごとに時系列発現プロファイルを時間方向に分割し，分割された部分時系列発現プロファイルから遺伝子制御ネットワークを推定することで達成できる．得られる複数の遺伝子制御ネットワークを時期ごとに見ることで，制御関係の変化がどのように起こったのかを把握することができるようになる．また，制御関係の変化毎に分割された部分時系列発現プロファイルには制御関係の変化が含まれていないと考えられるため，異なる制御関係が含まれる事による従来手法の問題は起こらない．

しかし，制御関係の変化する時期は一般にはっきりと分かっていないため，分割する位置を適切に決めることができない．そのため，時系列発現プロファイルを分割する位置を決定するために他の手法を用いる必要がある．本研究では，スライディングウィンドウ法を用いて重なりを持たせて等間隔に分割し，分割された部分時系列発現プロファイルそれぞれに対してダイナミックベイジアンネットワーク推定手法を行うことで，動的な遺伝子制御ネットワークを推定する．等間隔に分割することによって，分割された時系列発現プロファイルそれぞれが持つ時点数は同じになり，時点数による推定結果のバイアスを考えずに推定結果の遺伝子制御ネットワークを比較することができる．得られた遺伝子制御ネットワークを比較，解析することによって，細胞分化における網羅的な遺伝子制御ネッ

トワークの構造から，制御関係の変化がどの時点でどの遺伝子に起こっているか，どの遺伝子が細胞分化を司るような制御を行っているかを検討できるようになる．

3.4.2　時間方向分割

動的な遺伝子制御ネットワークは複数の時間的な順序を持つ遺伝子制御ネットワークを推定することによって推定できる．そのため，従来手法であるノードセットセパレーション法は遺伝子方向に時系列発現プロファイルを分割し複数の部分時系列発現プロファイルを抽出し，部分時系列それぞれに対して遺伝子制御ネットワークモデルである DBN を推定していた．このとき，時系列発現プロファイル中に遺伝子の制御変化が起こるような期間が含まれている際に制御関係が変化したことを推定できない問題点が挙げられる．

提案手法では，ノードセットセパレーション法の考え方を踏襲し，時系列発現プロファイルを分割し，分割された時系列発現プロファイルそれぞれから複数の遺伝子制御ネットワークを推定することで動的な遺伝子制御ネットワークとする．ただし，時系列発現プロファイルを分割する際に遺伝子方向ではなく時間方向に分割することで前述の問題点を解決することができると考えられる．手法の比較を図 3.5 で示すように，ノードセットセパレーション法と提案手法は時系列発現プロファイルを遺伝子方向で分割するか，時間方向で分割するかの違いがある．

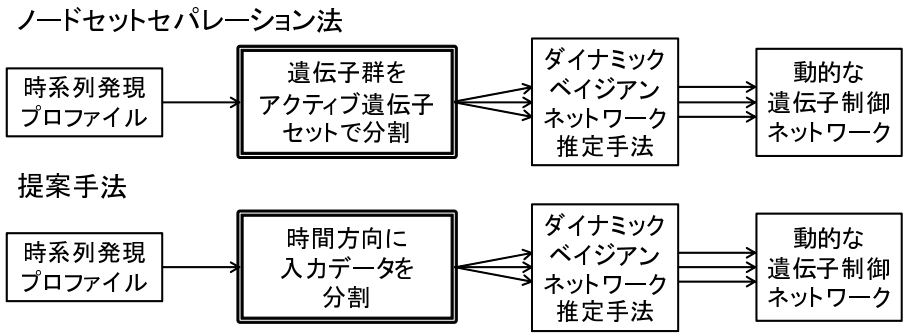

図 3.5　ノードセットセパレーション法と提案手法の比較図

時系列発現プロファイルを時間方向に分割することで，制御関係の変化があったとしても，制御関係のある時期とない時期に分けて遺伝子制御ネットワークの推定が可能になる．また，遺伝子方向には分割しないため，抑制の制御関係を持つ遺伝子があり発現量が低下していたとしても同じ部分時系列発現プロファイルにその遺伝子は含まれるため抑制の制御関係の推定が可能になる．

以降，時系列データを分割する際に用いる手法であるスライディングウィンドウ法について述べた後に時系列発現プロファイルへ適用することを説明する．

3.4.2.1 スライディングウィンドウ法

スライディングウィンドウ法 [39] は TCP（トランスミッション コントロール プロトコル）通信処理において通信の高速化のために考案された手法で，応用として，時系列データから部分系列を抽出する際に用いられる．各部分系列は時系列データから等間隔にウインドウサイズと呼ばれる時点数を持つように抽出される．1 時点目からウインドウサイズ時点までの時系列データを抽出し，その後スライド幅分の時点数をずらして抽出するということを繰り返すことによって複数の部分系列を得ることができる．T 時点の時系列データ $X_t(t = 1,\ldots,T)$ をウィンドウサイズ $w < T$，スライド幅 $s < T - w$ で分割する場合，$\frac{T-w}{s} + 1$ 以下の最大の整数 N 個の部分系列 $Y_i = X_{1+s(i-1)},\ldots,X_{w+s(i-1)}\,(i = 1,\ldots,N)$ とすることで分割された時系列データ Y_i が抽出される．

この手法は入力された時系列データを同じ時点数の部分系列を等間隔に抽出するため，入力データによらず時点数が十程度の小規模のデータセットから用いることができる．

3.4.3 時系列分割手法の時系列発現プロファイルへの適用

一般的に得られる時系列発現プロファイルは時点数が少なく，多くとも数十点である．そのため，時系列発現プロファイルを分割する手法として時点数が少ないデータセットへ応用できるスライディングウィンドウ法を用いる．スライディングウィンドウ法を時系列発現プロファイルへ適用し，提案手法の全体は図 3.6 のようになる．

3.5 評価実験

脂肪細胞分化系列の時系列発現プロファイルに対して提案手法を適用し，その評価結果を従来手法と比較した結果を示す．また，2 章で得られた定量的な制御関係を利用するため，2 章で用いたデータセットに提案手法を適用し，定量的な解析の結果を当てはめ，その結果を示す．以下ではまず評価実験に用いた遺伝子数が小規模，大規模の 2 つのデータセットについて説明した後，大規模ネットワークの解析結果を利用して 2 章で用いたデータセットを用いて行う実験について説明する．次に，小規模と大規模なデータセットに対する実験条件それぞれに対して，提案手法と従来手法による動的な遺伝子制御ネット

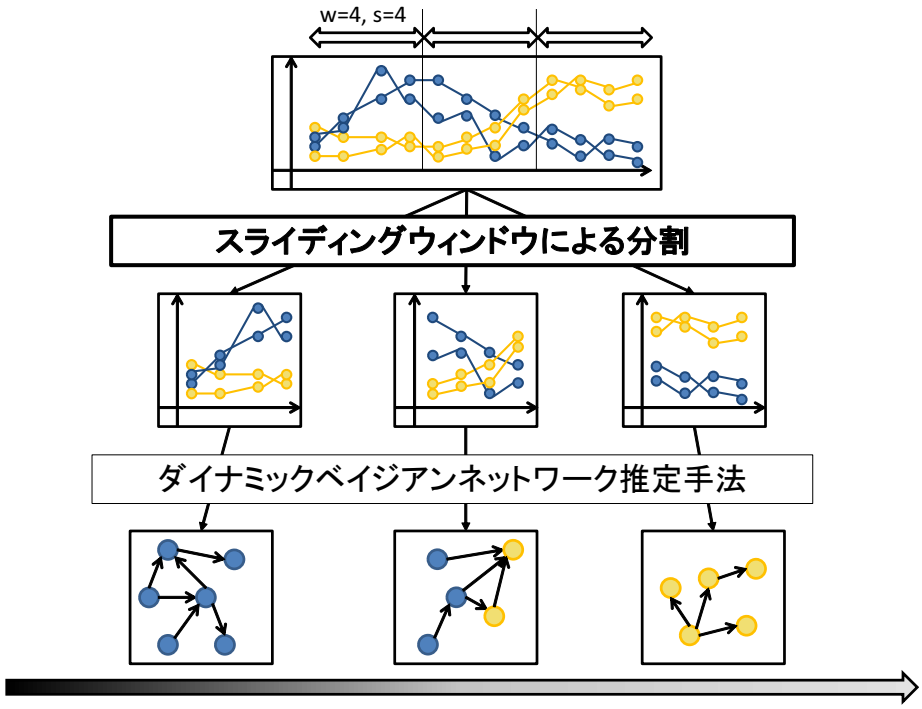

図 3.6　スライディングウィンドウ法の適用による提案手法概要図

ワーク推定を行い，その評価結果を両手法の間で比較した後，2 章との結果の整合性を確認する．

3.5.1　実験に用いたデータセット

実験ではマウス未分化間葉系細胞株 ST2（理研、細胞番号 RCB0224）を脂肪細胞へ分化誘導した際の RNA をマイクロアレイによって計測した時系列発現プロファイルを用いた．この ST2 細胞株は RPMI1640 培地から，10 % FBS（ウシ胎仔血清），0.5 m M の DMEM（3-イソブチル-1-メチルキサンチン），0.25μM の DEX（デキサメタゾン），5μg/ml のインスリンを含むインスリン-トランスフェリン-セレニウム添加剤，1μM のロシグリタゾンを添加した DMEM 培地（ダルベッコ・フォークト変法イーグル最小必須培地）へ交換して誘導し，誘導から 48 時間後に 10% の FBS を添加した DMEM 培地へ交換することによって脂肪細胞分化を誘導している．その際の発現量を Affymetrix GeneChip Mouse Genome 430 2.0 Array のマイクロアレイから取った，全 60 時点からなる時系列発現プロファイルを用いた．このプロファイルは分化刺激後から 5，15，30，45 分の 4 点，1 時間から 30 時点まで 1 時間刻みの 30 点，36 時点から 186 時点まで 6 時

間刻みの 26 点からなる計 60 時点の時系列発現プロファイルである．得られたマイクロアレイデータを Bioconductor[40] の affy パッケージによる RMA 法 [41] によって正規化し，その後各遺伝子を遺伝子方向に Zscore 化した．この正規化によって時系列発現プロファイルは遺伝子方向に平均 0，分散が 1 になり，遺伝子発現の量ではなく，遺伝子の発現変動の値を用いることができる．そのため，コントロールデータが無い場合でも発現の変化量として時系列発現プロファイルを用いることができ，ノードセットセパレーション法の閾値を設定することができる．

ダイナミックベイジアンネットワークの推定には SiGN[42] と呼ばれるソフトウェアを用いた．SiGN はスーパーコンピュータ上で実装され，並列計算によって高速にダイナミックベイジアンネットワークを推定，推定を複数回行った結果を統合するブートストラップによる高い精度の推定が可能という特徴を持つ．この際用いるネットワークスコアには BNRC[43][44] を用いた．BNRC においては値が小さいほど入力データへの当てはまりが良いネットワークを推定したことになる．また，ネットワーク事前確率 $p(G)$ は操作せず全ての制御関係に対して重みを乗せずに推定をした．

スライディングウィンドウ法ではデータに最適なパラメータを用いるため，入力時系列発現プロファイルに対して小実験を行う．ウィンドウサイズ w とスライド幅 s を決めることにより分割された部分時系列発現プロファイルは一意に決まるため，各パラメータに対する BNRC を取り得る全てのパラメータで求めることで，最良のパラメータを決定することができる．このとき取り得る全てのパラメータとは，入力データの全時点を覆うように分割することと重なりを持つことを満たすパラメータの組である．

小規模なデータセットとして，図 3.7 で示す脂肪細胞分化において制御関係が 23 個既知である遺伝子 14 個 [45] を用いて，既知ネットワークを正しく推定できるかを確認する．また，ネットワークスコアの値が従来手法と比較して良い推定結果かを確認する．また，大規模なデータセットとして脂肪細胞分化において有意に発現が上昇したとされる遺伝子 279 個 [46] を用いて，ダイナミックベイジアンネットワークが持つネットワークスコアの値が従来手法と比較して良い推定結果かを確認する．

提案手法の効果を確認した後，定量的な解析である 2 章の結果との整合性を評価する．大規模なデータセットの結果から得られたネットワーク構造から重要な遺伝子を抽出できるかどうかを 2 章で取り上げた遺伝子のノードが持つ出次数によって評価する．また，2 章で用いた 5 遺伝子 31 時点の実データから提案手法を適用し，ダイナミックベイジアンネットワークモデルによって得られた制御関係に S-system モデルの推定結果から得られ

た制御の強さを当てはめることができると示すことで整合性の確認を行う．

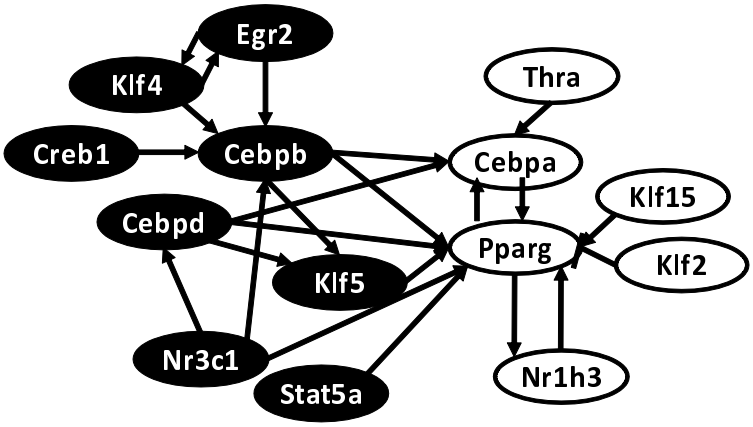

図 3.7　脂肪細胞分化中の既知の制御関係

3.5.2　小規模なデータセットでの評価

小規模なデータセットでは，制御関係が既知である遺伝子群を用いて提案手法が既知の関係を正しく推定できるかどうかを確認した．また，ネットワークスコアを従来手法と比較し改善しているかを確認した．

今回の実験では，従来手法であるノードセットセパレーション法のパラメータである閾値を 0 として遺伝子を分割した 60 の時系列発現プロファイル，提案手法であるスライディングウィンドウ法のパラメータをウィンドウサイズ w=15，スライド幅 s=5 に設定して 10 の部分時系列発現プロファイルを用いた．これらの設定は小実験で一番よい結果を出力したものを用いた．SiGN のパラメータは以下のようにした；ブートストラップ数=10,000，レプリカ数=3，ブートストラップ閾値=0.05 B スプライン補間ハイパーパラメータ (hn, hb, hi)=（2,1.0,2.0）, その他はデフォルト値．計算環境にはヒトゲノム解析センター（Human Genome Center; HGC）および京コンピュータ (Advanced Institute for Computational Science, RIKEN) を用いた．

図 3.8 は得られたネットワーク結果を纏めたものである．推定によって得られた複数のネットワークに 1 つでも既知の関係があれば矢印を引いている．既知関係のネットワーク（図 3.7）のうち，従来手法でのみ推定できたエッジを黒い点線，提案手法でのみ推定できたエッジを赤，どちらの手法でも推定できたエッジを黒の矢印で表している．従来手法でのみ推定できたエッジは 1 本，提案手法でのみ推定できたエッジは 5 本，どちらの手法でも推定できたエッジは 4 本である．

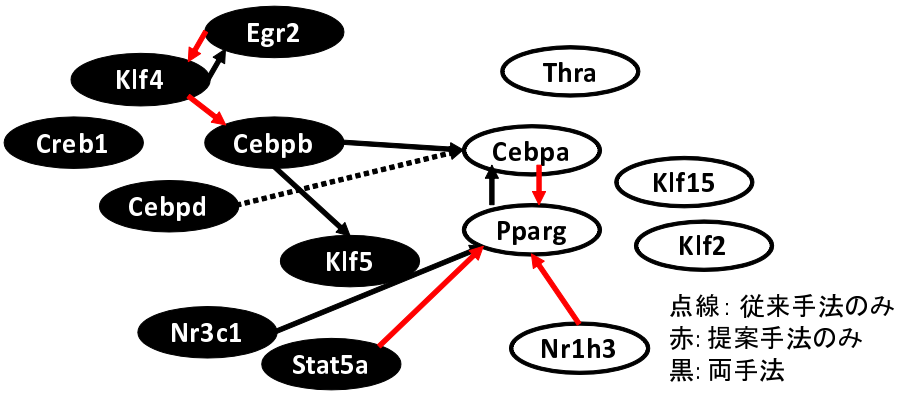

図 3.8　小規模なデータセットでの実験結果

また，得られた複数のネットワークどのネットワークがどれだけの正解率を持っていたかを表す尺度として F-measure がある．F-measure はエッジがあると推定した結果が既知関係であるという適合度 (Precision) と既知関係の内どれだけ推定したかという再現率（recall）から計算される値で，次のようにして求められる．

$$
\begin{aligned}
F-measure &= 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \\
Precision &= \frac{true\ positive}{true\ positive + false\ positive} \\
Recall &= \frac{true\ positive}{true\ positive + false\ negative}
\end{aligned}
\tag{3.2}
$$

F-measure は高いほど正解のエッジだけを多く推定したと評価する尺度であり，0 から 1 の値を取る．この F-measure を各手法から得られたネットワークに対して計算し，プロットしたものが図 3.9 になる．

この時のネットワークスコア BNRC は箱ひげ図 3.10 で示しており N_1 から N_{10} は提案手法による推定の結果得られた 10 個のネットワークの分布である．N_1 は 1 時点目から 15 時点目，N_2 は 6 時点目から 21 時点目というように対応している．N は従来手法による推定結果を纏めた際の分布である．N_4 のプロットが無いのは負の無限大や非常に小さく表示すると他の箱が見えなくなるために除外したためである．

3.5.3　大規模なデータセットでの評価

大規模なデータセットでは，遺伝子セットを小規模から拡張した場合の提案手法の性能を評価する．小規模なデータセットとは異なり，正解の制御関係が無いデータセットなので，推定手法によって得られるネットワークスコアをもとに大規模なデータへの適用可能

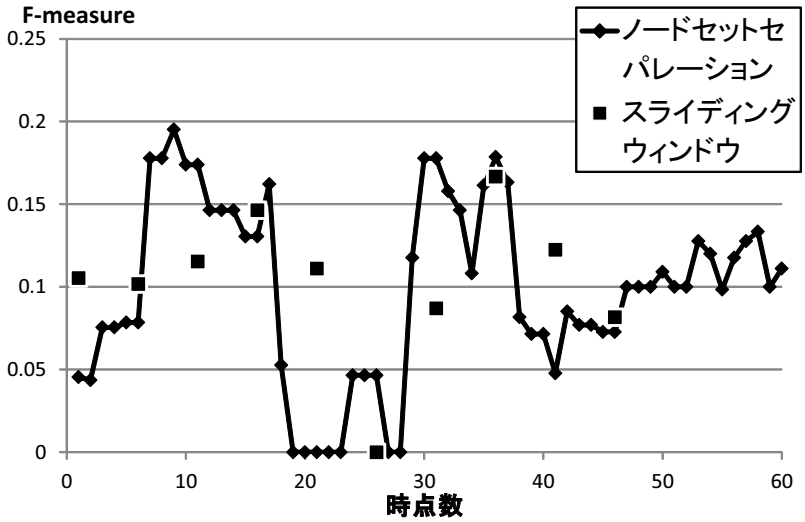

図 3.9　小規模なデータセットでの F-measure 値の推移

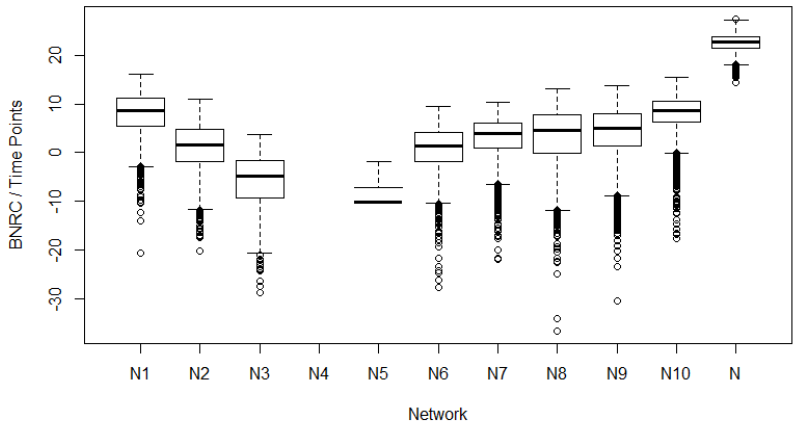

図 3.10　小規模なデータセットでの BNRC

性を確かめる．

今回の実験では，従来手法であるノードセットセパレーション法のパラメータである閾値を 0 として遺伝子を分割した 60 の時系列発現プロファイル，提案手法であるスライディングウィンドウ法のパラメータをウィンドウサイズ w=39，スライド幅 s=1 に設定して 22 の部分時系列発現プロファイルを用いた．これらの設定は小実験で一番よい結果を出力したものを用いた．SiGN のパラメータは以下のようにした；ブートストラップ数=4,000，レプリカ数=3，ブートストラップ閾値=0.1 B スプライン補間ハイパーパラメー

タ (hn, hb, hi)=（2,1.0,2.0）, その他はデフォルト値．計算環境にはヒトゲノム解析センター（Human Genome Center; HGC）および京コンピュータ (Advanced Institute for Computational Science, RIKEN) を用いた．

提案手法の結果は箱ひげ図 3.11 り N_1 から N_{22} は提案手法による推定の結果得られた 22 個のネットワークの分布である．N_1 は 1 時点目から 39 時点目，N_2 は 2 時点目から 40 時点目というように対応している．N_{23} は従来手法による推定結果を纏めた際の分布である．

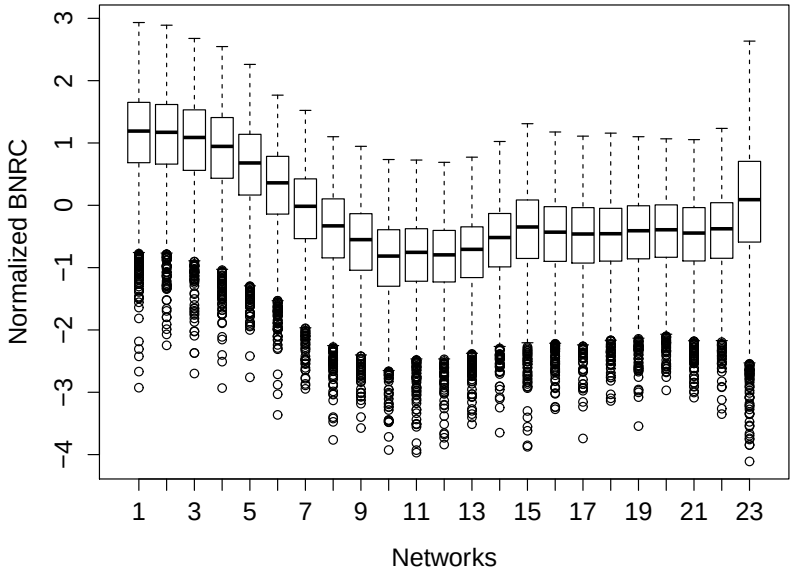

図 3.11　大規模なデータセットでの BNRC

また，従来手法と提案手法の結果得られた全ネットワークの BNRC の中央値は表 3.1 のようになった．等分散性や正規分布性を仮定しないウィルコクソンの順位和検定を有意水準 5% にて中央値検定を行った所，p 値が 2.2e-16 以下であるとされたため，提案手法は従来手法より有意に BNRC が小さいと言える．

手法	ノードセットセパレーション法	スライディングウィンドウ法
BNRC の中央値	0.0880	-0.0747

表 3.1　大規模なデータセットでの BNRC 中央値

3.5.4 定量的な解析の当てはめ

大規模ネットワーク推定の結果，22 の遺伝子制御ネットワークが得られた．この中から重要と考えられる遺伝子を抽出することによって詳細で定量的な解析を細胞分化において制御を司る働きを持つ遺伝子間に対して行うことができる．重要と考えられるかどうかは遺伝子のノードが持つ出次数によって決めることができる．出次数とは他の遺伝子へ制御を行っている数である．出次数が多ければ，他の多くの遺伝子に制御を行う遺伝子であるため，そのネットワーク構造を構成するために重要な遺伝子であると考えられる．今回の大規模ネットワーク推定実験には，発現量が有意に上昇した遺伝子 [46] として報告された中に 2 章で取り上げた Cebpβ，Cebpγ，Cebpα，Pparγ の 4 つの遺伝子が含まれている．これらの遺伝子の持つ出次数の順位を確認すると，時期によって変化するものの上位 1/3 以内に含まれていた．特に Cebpβ は前半の N_5 と N_7 において最も出次数の高い遺伝子として抽出された．

また，2 章のデータセットに提案手法を適用し，定量的な解析手法の結果の当てはめを行い整合性を確認する．ダイナミックベイジアンネットワークによる推定結果に S-system モデルの解析結果である制御の強さを当てはめることで，どの遺伝子がそのネットワークに寄与しているかを解析できるようになる．この実験では，提案手法であるスライディングウィンドウ法のパラメータをウィンドウサイズ w=18，スライド幅 s=12 に設定して 2 つの部分時系列発現プロファイルを用いた．SiGN のパラメータは以下のようにした；ブートストラップ数=4,000，レプリカ数=3，ブートストラップ閾値=0.1 B スプライン補間ハイパーパラメータ (hn, hb, hi)= （2,1.0,2.0）, その他はデフォルト値．計算環境にはヒトゲノム解析センター（Human Genome Center; HGC）および京コンピュータ (Advanced Institute for Computational Science, RIKEN) を用いた．

図 3.12 は得られたネットワーク結果である．左のネットワークが 1 時点目から 18 時点目の前半，右のネットワークが 13 時点目から 30 時点目までの後半の部分時系列発現プロファイルから推定した結果である．この 2 つのネットワークを統合したものが図 3.13 になる．Cebpα から Cebpδ と Cebpβ から Cebpγ への制御は前半と後半のネットワークが指す促進抑制の制御が不一致だったため，不明として点線で表している．

S-system モデルのパラメータ g_{ij} は遺伝子 j が遺伝子 i の発現を促進する反応次数，h_{ij} は遺伝子 j から遺伝子 i の発現を抑制する反応次数である．遺伝子 j が遺伝子 i へ促進か抑制の働きをするかどうかは，$g_{ij} - h_{ij}$ の値で決めることができる．$g_{ij} - h_{ij}$ の値

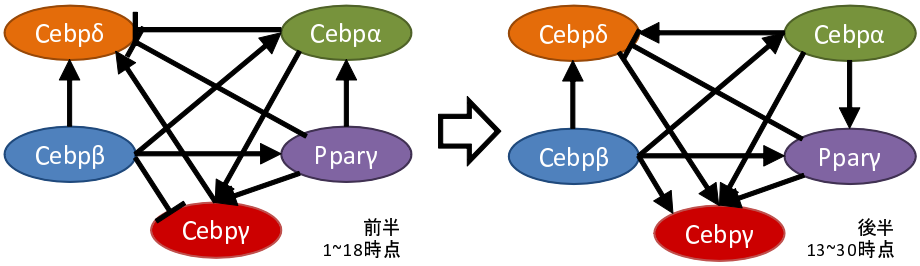

図 3.12　2 章の実データへの提案手法の適用結果

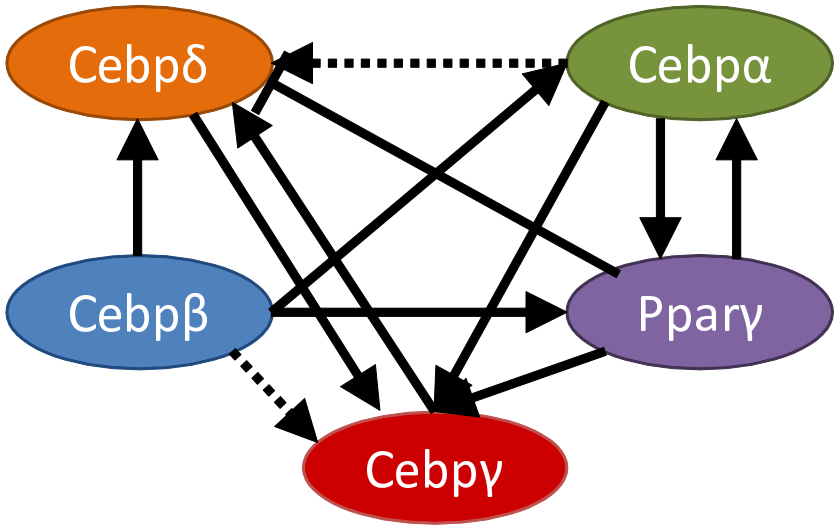

図 3.13　2 章の実データへの提案手法を適用した結果を統合したネットワーク

が正であれば，g_{ij} の値が大きいか，h_{ij} の値が負であるかとなるため，促進の働きをすると考えることができる．逆に，$g_{ij} - h_{ij}$ の値が負であれば，g_{ij} の値が負であるか，h_{ij} の値が大きいため，抑制の働きをすると考えることができる．$g_{ij} - h_{ij}$ の値の絶対値は制御の強さを表しており，絶対値が大きいほど遺伝子 j が遺伝子 i に与える影響が大きい．

表 3.2 は 2 章の IA を適用した結果得られた実データでの S-system パラメータである $g_{ij} - h_{ij}$ の値の表である．このパラメータを，提案手法で推定した結果得られた遺伝子制御ネットワークに適用したのが図 3.14 になる．図 3.14 では，制御の強さに応じて制御辺の太さを変えている．また，Cebpα から Cebpδ への制御は $g_{5,3} - h_{5,3}$ が負であることから，抑制の制御があるとし，Cebpβ から Cebpγ への制御は $g_{2,1} - h_{2,1}$ が正であることから，促進の制御があるとした．

i	$g_{i1}-h_{i1}$	$g_{i2}-h_{i2}$	$g_{i3}-h_{i3}$	$g_{i4}-h_{i4}$	$g_{i5}-h_{i5}$
1(Cebpβ)	-0.83	-0.20	-1.9	1.1	2.4
2(Cebpγ)	2.3	-4.6	1.9	0.02	-0.79
3(Cebpα)	0.70	0.11	-3.2	1.8	2.7
4(Pparγ)	0.50	-0.04	0.87	-3.7	3.0
5(Cebpδ)	0.72	1.63	-3.8	0.59	-0.18

表 3.2　適用する S-system モデルのパラメータ

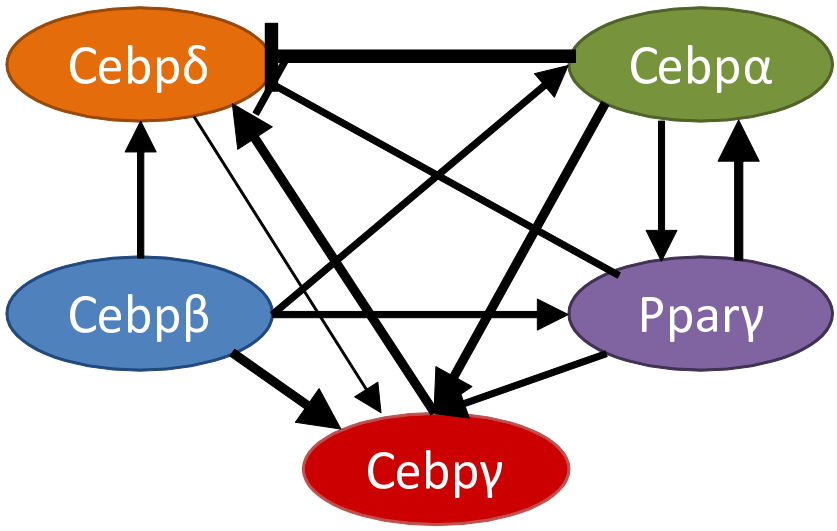

図 3.14　2 章の定量的な解析結果を提案手法の結果へ適用

3.5.5　考察

まず，小規模なデータセットと大規模なデータセットを用いて行った提案手法の有効性を確かめる実験についての考察の後，提案手法で推定されたネットワークに 2 章の定量的な解析手法による制御の強さを当てはめる実験について考察する．

提案手法の有効性　小規模なデータセットにおいて，提案手法は従来手法よりも実際の生物の細胞分化に即した制御関係を推定することができ，推定されたネットワークは入力データに対する当てはまりがよく推定精度が向上したことを確認できた．各ネットワークの推定精度である F-measure の結果（図 3.9）からは提案手法と従来手法はそれぞれのネットワークに注目した場合ほぼ同じであると言える．これはダイナミックベイジアンネットワーク推定を同じ条件で行ったためと考えられ，遺伝子方向に分割した場合と時間方向に分割した場合でそれぞれのネットワークの推定精度が変わらなかったと考えられる．しかし，図 3.7 から，提案手法によって推定されたネットワーク全てを纏めると従来

手法によって推定されたネットワーク全てを纏めた結果より既知関係を 4 本多い 9 本のエッジを推定している．正解の制御関係数は 23 本であるため，従来手法が 22%，提案手法が 39% の既知関係を推定できたことになる．これから，提案手法は既知関係を従来手法より制御関係の変化をより多く推定できていることが確認できたと言える．また，図 3.10 におけるネットワークスコア BNRC から，提案手法は従来手法より値が小さく分布しており，データに対する当てはまりがよくなっていることが確認できた．今回はダイナミックベイジアンネットワーク推定手法に SiGN を用い，同じ条件設定で推定を行った結果，提案手法が従来手法よりも推定精度が向上していたため，ネットワーク推定手法が同じであっても時系列発現プロファイルを分割手法を時間方向に変更することによって推定精度を向上させることができたと考えられる．

大規模なデータセットにおいては，提案手法は従来手法よりも推定精度が向上したことが確認できた．表 3.1 とウィルコクソンの順位和検定の結果から提案手法は従来手法より有意に小さいネットワークスコアでネットワークを推定したことが分かった．BNRC は低ければネットワークがデータに対してよく当てはまっているとなるため，推定精度が向上したと言える．

これらの結果から，提案手法であるスライディングウィンドウ法による時系列発現プロファイル分割手法は大規模で動的な遺伝子制御ネットワーク推定に対して従来手法であるノードセットセパレーション法による分割より効果のある手法だといえる．今回の実験では時点数が 60 点の時系列発現プロファイルを用いたように時点数の多い時系列発現プロファイルを用いる必要があるが，細胞分化におこる制御関係の変化を捉えるためには制御関係の変化が見られる時期ごとに時間方向で区切られた時系列発現プロファイルを用いた方がよいということが分かった．

定量的な解析手法の当てはめ　大規模なデータセットによって推定された遺伝子制御ネットワークを 2 章の手法に適用するには，推定された遺伝子制御ネットワークの構造から解析対象となる少数の遺伝子を抽出しなければならない．他の遺伝子にどれだけ制御を行っているかを表す出次数の高い遺伝子順に抽出することによって，他の多くの遺伝子に制御を行う重要な遺伝子を抽出できることを確認した．2 章で取り上げた遺伝子における大規模なデータセットから推定された遺伝子制御ネットワークの中での出次数の順位を確認すると，Cebpβ などの脂肪細胞において重要な働きを持つ遺伝子は全て出次数が全体の順位の中で 1/3 以上と高く，ネットワーク内で最も出次数の高い遺伝子も存在した．これか

ら，提案手法の推定結果から出次数の高い転写因子を抽出することによって，細胞分化を制御する重要な遺伝子を抽出し，定量的な解析手法へ掛けるための遺伝子を特定することができると考えられる．

提案手法の推定結果に 2 章の結果を当てはめた図 3.14 と表 3.2 から，ダイナミックベイジアンネットワークによる推定結果と S-system モデルによる制御の強さの解析の結果は大きく変わらないことが確認できる．Cebpδ から Cebpγ への制御が提案手法の推定結果では促進の制御として推定されたが，S-system モデルでは抑制の制御であった以外の 11 本の制御関係については促進と抑制の関係は一致していた．これは，推定に用いたデータセットが同じ為，同じ制御関係があると推定された結果と考えることができる．同じデータを用いることで，提案手法によって推定されたダイナミックベイジアンネットワークの結果は S-system モデルによって推定された制御の強さと整合性が取れた推定ができると考えられる．

また，推定された Cebpα がこのネットワークにおいて強い制御を多く持つことから，Cebpα の働きが全体に最も大きな影響を与えると推測できる．Transcriptional Regulatory Element Database[47][48] によると，Cebpα は *mouse* の Cebp ファミリーの中で最も遺伝子制御を行っていることで知られているため，妥当な結果だと言える．このようにして，大規模な遺伝子制御ネットワークの推定結果に定量的な解析手法を適用することで，制御関係の強さが推測でき，遺伝子発現に寄与する遺伝子を抽出できるようになった．

3.6 結言

本研究では，細胞分化に適した動的な大規模遺伝子制御ネットワークを推定する目的のため，時系列発現プロファイルをスライディングウィンドウ法によって時間方向に分割する手法を提案した．提案手法では入力された時系列発現プロファイルを同じ時点数を持つ部分時系列発現プロファイルに等間隔に分割し，部分時系列発現プロファイルそれぞれに対してダイナミックベイジアンネットワークの推定を行い，得られた複数のネットワークを時間の順序で並べることで動的な遺伝子制御ネットワークを推定する．

脂肪細胞分化系列における時系列発現プロファイルを用いて従来手法であるノードセットセパレーション法と比較実験を行うことによって，時系列発現プロファイルの時間方向分割によって遺伝子制御ネットワークの推定結果の正解率とネットワークスコアが向上することを確認した．このネットワークスコアの向上は大規模な遺伝子制御ネットワークの

推定でも同様に見られたため，細胞分化における制御関係の変化を推定するための手法として提案手法は有効であると考えられる．

本研究での網羅的な遺伝子制御ネットワーク推定手法を 2 章で行った定量的な解析手法と組み合わせることで，どの遺伝子が遺伝子制御ネットワークを司る働きを持つ可能性が高いかを推測できるようになった．提案手法の結果得られたネットワークから出次数の高い遺伝子を抽出することで，定量的な解析手法を適用できることが分かった．また，提案手法による結果に定量的な解析手法を当てはめることで，推定されたネットワークに制御の強さが分かるようになり，遺伝子の発現に最も寄与する遺伝子を抽出することができるようになった．

第 4 章　結論

本論文では，細胞分化における動的な遺伝子制御ネットワークに着目し，遺伝子の発現と制御の理解を目指した．そのための手法として，細胞分化中の遺伝子制御ネットワークの動態解析を行う S-system モデルへ免疫アルゴリズムを適用する手法と，細胞分化中に制御関係が変化する動的な遺伝子制御ネットワークの推定手法を提案した．

本研究の 1 番目の成果として，遺伝子制御ネットワークの動態解析を行うため，S-system モデルの推定に免疫アルゴリズムを適用する手法を提案した．S-system モデルは遺伝子制御ネットワークの動態解析に用いられていたが，従来の探索手法である遺伝的アルゴリズムでは局所解へ収束してしまうため，高い推定精度でのパラメータ推定を行うのが困難であった．時系列発現プロファイルを用いた S-system モデルには局所解が多く存在するため，この問題を解決することで推定精度を向上することができると考え，局所解へ収束した際に解候補の初期化による再探索を行う探索手法である免疫アルゴリズムの適用による S-system モデル推定手法を提案した．解候補の初期化によって局所解へ収束してしまうことによる推定精度の低下を防ぐことができる．シミュレーションデータと脂肪細胞分化における実データに対して提案手法を適用し，どちらのデータに対しても提案手法が高い推定精度をもつ S-system モデルパラメータを出力した．これによって，細胞分化系列の時系列発現プロファイルを用いた S-system モデルパラメータ推定における提案手法の有効性を確認した．

本研究の 2 番めの成果として，大規模で動的な遺伝子制御ネットワークを推定するために，時系列発現プロファイルを時間方向に分割する手法を提案した．細胞分化では制御関係が大きく変わることが分かってきており，大規模な遺伝子間で動的な遺伝子制御ネットワークを形成している．また，クロストーク遺伝子のように他の分化経路で働く遺伝子を抑制する働きを持つようになる遺伝子も報告されてきた．そのような細胞分化の時系列発現プロファイルを用いて動的な遺伝子制御ネットワークを推定する際に，従来のノードセットセパレーション法による時系列発現プロファイルの遺伝子方向分割ではうまく制御関係が推定できない場合が生まれていた．提案手法は，時系列発現プロファイルを時間方向に分割するスライディングウィンドウ法を適用することで，制御関係の変化の時期を捉えてネットワーク推定をすることができるようになった．脂肪細胞分化系列の時系列発現プロファイルを用いて確認実験を行い，小規模で動的な遺伝子制御ネットワークの推定によって脂肪細胞分化において既知の関係を従来手法よりよく推定することを示し，大規模

で動的な遺伝子制御ネットワークの推定によって遺伝子数を増やした場合でも従来手法よりよい推定を行うことを示した．

上記 2 点の成果によって細胞分化で重要な働きを持つ遺伝子における遺伝子制御ネットワークの動態解析と細胞分化中の遺伝子制御ネットワークの大規模な変化を捉えた動的な遺伝子制御ネットワークの推定において，推定精度を向上させることが可能になった．この 2 点を組み合わせることで，網羅的な遺伝子制御ネットワークの推定結果を利用して重要な遺伝子の抽出をすることで定量的な解析手法を適用して詳細な解析を行えるようになり，定量的な制御関係の強さを当てはめることができることでどの遺伝子が全体の制御変化に最も寄与している可能性があるかを推測できるようになったことが，本研究の 3 番目の成果である．

以上の 3 点によって，細胞分化の研究における遺伝子制御解析において網羅的な視点と詳細を解析する視点を繋いで遺伝子制御ネットワークを推定できる可能性を示し，細胞分化の研究に貢献できたと考えている．

謝辞

本研究は，著者が 2009 年から 2010 年まで大阪大学基礎工学部情報科学科在学中，2010 年から 2011 年まで大阪大学大学院情報科学研究科博士前期課程在学中，2011 年から 2014 年まで大阪大学大学院情報科学研究科博士後期課程在学中に行ってきた、生物情報科学に関する研究成果をまとめたものです．

本研究の全課程の遂行ならびに本論文をまとめるにあたり，終始懇切なる御指導，御鞭撻を賜りました大阪大学大学院情報科学研究科バイオ情報工学専攻 松田秀雄 教授には，ここに厚く御礼申し上げます．貴重なるお時間を割いていただき丁寧なる御教示を賜りました大阪大学大学院情報科学研究科バイオ情報工学専攻 清水浩 教授，前田太郎 教授，四方哲也 教授，若宮直紀 教授に厚く御礼申し上げます．

本研究の遂行にあたり，数多くの御指導，御助言を頂きました大阪大学大学院情報科学研究科 竹中要一 准教授に厚く御礼申し上げます．

本論文をまとめるにあたり，全過程において様々な御支援，御指導を頂きました，大阪大学大学院情報科学研究科 瀬尾茂人 助教には，厚く御礼申し上げます．

本研究の遂行にあたり，第 3 章で述べた実験環境について様々な御支援，御指導頂きました，東京大学大学院情報理工学系研究科 玉田嘉紀 助教に深く感謝致します．

大阪大学大学院情報科学研究科 大安裕美 特任研究員には本研究を遂行するにあたり，数多くの御指導，御助言を頂きました．心より御礼申し上げます．

大阪大学サイバーメディアセンター 木戸善之 特任講師には本研究の遂行におきまして多くの助言と議論をしていただきました．心より御礼申し上げます．

大阪大学大学院情報科学研究科バイオ情報工学専攻松田研究室の皆様には，本研究をまとめるに際して，様々な御支援，御指導頂きましたことを感謝いたします．

最後に，本研究の遂行に際し，著者を御激励，御支援下さいました方々へ心より感謝致します．

参考文献

[1] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature, Vol. 431, pp. 931–945, 2004.

[2] A. Nakabachi, A. Yamashita, H. Toh, H. Ishikawa, H. E. Dunbar, N. A. Moran, and M. Hattori. The 160-kilobase genome of the bacterial endosymbiont carsonella. Science, Vol. 314, No. 5797, p. 267, 2006.

[3] R. Siersbaek, R. Nielsen, and S. Mandrup. Pparγ in adipocyte differentiation and metabolism-novel insights from genome-wide studies. FEBS Letters, Vol. 584, No. 15, pp. 3242–3249, 2010.

[4] J. L. Golob, S. L. Paige, V. Muskheli, L. Pabon, and C. E. Murry. Chromatin remodeling during mouse and human embryonic stem cell differentiation. Developmental Dynamics, Vol. 237, pp. 1389–1398, 2008.

[5] R. Siersbaek, R. Nielsen, S. John, M. Sung, S. Beak, A. Loft, G. Hager, and S. Mandrup. Extensive chromatin remodelling and establishment of transcription factor 'hotspots' during early adipogenesis. The EMBO Journal, Vol. 30, pp. 1459–1472, 2011.

[6] S. Muruganandan, A. A. Roman, and C. J. Sinal. Adipocyte differentiation of bone marrow-derived mesenchymal stem cells: Cross talk with the osteoblastogenic program. Cellular and Molecular Life Sciences, Vol. 66, No. 2, pp. 236–253, 2009.

[7] 吉澤陽志, 瀬尾茂人, 竹中要一, 松田秀雄. 細胞分化クロストークのモデル化と細胞分化クロストーク遺伝子の推定手法. 情報処理学会研究報告. MPS, 数理モデル化と問題解決研究報告, Vol. 2011, No. 10, pp. 1–8, 2011.

[8] K. Arita, M. Ariyoshi, H. Tochio, Y. Nakamura, and M. Shirakawa. Recognition of hemi-methylated dna by the sra protein uhrf1 by a base-flipping mechanism. Nature, Vol. 455, pp. 818–821, 2008.

[9] H. Toh and K. Horimoto. Inference of a genetic network bv a combined approach of cluster analysis and graphical gaussian modeling. Bioinformatics, Vol. 18, pp. 287–297, 2002.

[10] S. A. Kauffman. The Origins of Order: Self-Organization and Selection in

Evolution. Oxford University Press, 1993.

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

[12] N. Friedman, K. Murphy, and S. Russell. Learning the structure of dynamic probabilistic networks. in Proc. UAI' 98, pp. 139–147, 1998.

[13] J. Vohradsky. Neural network model of gene expression. FASEB J, Vol. 15, pp. 846–854, 2001.

[14] P. Mendes, W. Sha, and K. Ye. Artificial gene networks for objective comparison of analysis algorithms. Bioinformatics, Vol. 19(90002), pp. 122–129, 2003.

[15] C. Koh, F.-X. Wu, G. Selvaraj, and A. J. Kusalik. Using a state-space model and location analysis to infer time-delayed regulatory networks. EURASIP Journal on Bioinformatics and Systems Biology, 2009.

[16] M. A. Savageau. Biochemical System Analysis : a study of function and design in molecular biology. Addison-Wesley, 1976.

[17] H. D. Jong. Modeling and simulation of genetic regulatory systems: A literature review. Journal of Computational Biology, Vol. 9, No. 1, pp. 67–103, 2002.

[18] Y. Tamada, H. Araki, S. Imoto, M. Nagasaki, A. Doi, Y. Nakanishi, Y. Tomiyasu, K. Yasuda, B. Dunmore, D. Sanders, S. Humphreys, C. Print, D. S. Charnock-Jones, K. Tashiro, S. Kuhara, and S. Miyano. Unraveling dynamic activities of autocrine pathways that control drug-response transcriptome networks. Pacific Symposium on Biocomputing, Vol. 14, pp. 251–263, 2009.

[19] E. O. Voit. Canonical nonlinear modeling: S-system approach to understanding complexity. Van Nostrand Reinhold, 1991.

[20] M. T. Swain, J. J. Mandel, and W. Dubitzky. Comparative study of three commonly used continuous deterministic methods for modeling gene regulation networks. BMC Bioinformatics, Vol. 11, p. 459, 2010.

[21] D. E. Goldberg. Genetic Algorithm in search, optimization and machine learning. Addison-Wesley, 1989.

[22] 小野功, 山村雅幸, 喜多一. 実数値 GA とその応用. 電子情報通信学会, Vol. 100(88), pp. 61–68, 2000.

[23] 菊池進一, 富永大介, 有田正規, 冨田勝. S-system を用いた GA による遺伝子ネット

ワーク予測. IPSJ SIG Notes, Vol. 110, pp. 13–18, 2001.

[24] D. Tominaga and K. Takahashi. Inference biological network structure by genetic algorithm with structure restriction using time-series data. IPSJ SIG Notes, Vol. 2003, No. 122, pp. 25–28, 2003.

[25] L. J. Esheleman and J. D. Schaffer. Real-coded genetic algorithms and interval-schemata. Foundations of Genetic Algorithms 2, pp. 187–202, 1993.

[26] I. Ono, H. Satoh, and S. Kobayashi. A real-coded genetic algorithm for function optimization using the unimodal normal distribution crossover. Journal of Japanese Society for Artificial Intelligence, Vol. 14, No. 6, pp. 1146–1155, 1999.

[27] 樋口隆英, 筒井茂義, 山村雅幸. 実数値 GA におけるシンプレクス交叉の提案. 人工知能学会誌, Vol. 16, No. 1, pp. 147–155, 2001.

[28] 佐藤浩, 小野功, 小林重信. 遺伝的アルゴリズムにおける世代交代モデルの提案と評価. 人工知能学会誌, Vol. 12(5), No. 5, pp. 734–744, 1997.

[29] O. Takahashi, H. Kita, and S. Kobayashi. A real-coded genetic algorithm using distance dependent alternation model for complex function optimization. GECCO'00, pp. 219–226, 2000.

[30] D. Jong and K. Alan. An Analysis of the Behavior of a Class of Genetic Adaptive Systems. PhD thesis, University of Michigan, 1975.

[31] J. Koza. Genetic Programming. MIT Press, 1992.

[32] K. Matsumura, H. Oida, and S. Kimura. Inference of s-system models of genetic networks by function optimization using genetic programming. Transactions of Information Processing Society of Japan, Vol. 46, No. 11, pp. 2814–2830, 2005.

[33] R. Storn and K. Price. Differential evolution a simple and efficient heuristic for global optimization over continuous spaces. Journal of Global Optimization, Vol. 11(4), pp. 341–359, 1997.

[34] N. Noman and H. Iba. Inference of gene regulatory networks using s-system and differential evolution. In GECCO '05, pp. 439–446, New York, NY, USA, 2005. ACM.

[35] H.-Y. Fan and J. Lampinen. A trigonometric mutation operation to differential evolution. Journal of Global Optimization, Vol. 27, No. 1, pp. 105–129, 2003.

[36] K. Mori, M. Tsukiyama, and T. Fukuda. Application of an immune algorithm to

multi-optimization problems. Transactions of IEE of Japan, Vol. 117-C, No. 5, pp. 593–598, 1997.

[37] 本間俊雄, 加治広之, 登坂宣好. 免疫アルゴリズムによる構造システムの最適化と解の多様性. 日本建築学会構造系論文集, Vol. 588, No. 588, pp. 103–110, 2005.

[38] 春日雅人, 阪上浩, 森要之. 脂肪細胞分化の転写制御, 第 124 巻. 肥満の科学, 2003.

[39] S. Savage, N. Cardwell, D. Wetherall, and T. Anderson. Tcp congestion control with a misbehaving receiver. SIGCOMM Comput. Commun. Rev., Vol. 29, No. 5, pp. 71–78, 1999.

[40] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. Genome Biology, Vol. 5, No. R80, 2004.

[41] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, Vol. 4, No. 2, pp. 249–264, 2003.

[42] Y. Tamada, T. Shimamura, R. Yamaguchi, S. Imoto, M. Nagasaki, and S. Miyano. Sign: large-scale gene network estimation environment for high performance computing. Genome Inform. '11, Vol. 25, No. 1, pp. 40–52, 2011.

[43] S. Kim, S. Imoto, and S. Miyano. Inferring gene networks from time series microarray data using dynamic bayesian networks. Brief Bioinform, Vol. 4, No. 3, pp. 228–235, 2003.

[44] S. Kim, S. Imoto, and S. Miyano. Dbn and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data. BioSystems, Vol. 75, pp. 57–65, 2004.

[45] R. Siersbaek, R. Nielsen, and S. Mandrup. Transcriptional networks and chromatin remodeling controlling adipogenesis. Trends in Endocrinology and Metabolism, Vol. 23, No. 2, pp. 56–64, 2012.

[46] Y. Tokuzawa, K. Yagi, Y. Yamashita, Y. Nakachi, I. Nikaido, H. Bono, Y. Ni-

nomiya, Y. Kanesaki-Yatsuka, M. Akita, H. Motegi, S. Wakana, T. Noda, F. Sablitzky, S. Arai, R. Kurokawa, T. Fukuda, T. Katagiri, C. Schonbash, T. Suda, Y. Mizuno, , and Y. Okazaki. Id4, a new candidate gene for senile osteoporosis, acts as a molecular switch promoting osteoblast differentiation. PLoS Genetics, Vol. 6, No. 7, 2010.

[47] C. Jiang, Z. Xuan, F. Zhao, and M. Q. Zhang. Tred: a transcriptional regulatory element database, new entries and other development. Nucleic Acids Research, Vol. 35, pp. D137–D140, 2007.

[48] C. S. H. Laboratory. Transcriptional Regulatory Element Database. http://rulai.cshl.edu/TRED/GRN/CEBP.htm, (参照 2013-12-22).