修 士 論 文

グラフ構造に基づいた遺伝子相互作用の推定に関する研究

北陸先端科学技術大学院大学
情報科学研究科情報システム学専攻

大谷 俊朗

2007 年 3 月

修 士 論 文

グラフ構造に基づいた遺伝子相互作用の推定に関する研究

指導教官　平石邦彦 教授

審査委員主査　平石邦彦 教授
審査委員　金子峰雄 教授
審査委員　上原隆平 助教授

北陸先端科学技術大学院大学
情報科学研究科情報システム学専攻

510020 大谷 俊朗

提出年月: 2007 年 2 月

目 次

第 1 章　はじめに **1**
1.1　遺伝子間の相互作用について 1
1.2　従来研究 3
1.3　本研究の目的 4
1.4　本研究の内容 5

第 2 章　使用する菌類と各種データの紹介 **6**
2.1　枯草菌について 6
2.2　DNA 塩基配列データについて 6
2.3　マイクロアレイデータについて 7

第 3 章　既存手法の紹介 **10**
3.1　既存手法の特徴 10
3.2　使用するデータと各種定義 10
3.2.1　マイクロアレイによる発現強度の相関係数 10
3.2.2　ウインドウ類似度 11
3.2.3　ピーク位置 13
3.3　推定手順 14
3.4　既存手法における問題点 15

第 4 章　既存手法の改良 **17**
4.1　ピーク位置決定方法の改良 17
4.1.1　既存手法のピーク位置決定方法について 17
4.1.2　現在のピーク位置決定方法における問題点 18
4.1.3　ピーク位置決定方法の改良 19
4.1.4　改良手法によるピーク位置の計算例 20
4.2　グラフ構造による同一転写因子の被制御遺伝子候補の絞込 24
4.2.1　既存手法における被制御遺伝子とバインディングサイトの絞込みについて 24
4.2.2　改良方法 25
4.2.3　クリークを用いたピーク位置の絞込みの効果 29
4.3　バインディングサイトの特定方法の提案 31

4.3.1 バインディングサイトの可能性の高いピーク位置 31
4.3.2 ランダム配列との比較によるウインドウ類似度の特異性による判別 31
4.3.3 実験方法 33
4.3.4 実験結果 33

第 5 章 各種パラメータの設定 36
5.1 ウインドウ類似度における部分文字列の移動幅の最適化 36
5.1.1 移動幅について 36
5.1.2 移動幅とウインドウ類似度の関係 36
5.1.3 移動幅の変化に対するピーク位置候補の増加数 38
5.1.4 各移動幅おけるウインドウ類似度の増加幅 39
5.1.5 推定に最適な移動幅 39
5.2 発現強度の相関による同一転写因子の被制御遺伝子の絞込みについて . . . 43
5.2.1 相関係数絶対値と相関係数順位 43
5.2.2 発現強度分布 43
5.2.3 被制御遺伝子の相関係数順位と発現強度順位の関係 45
5.3 ウインドウ長さに関する調査 46
5.3.1 ウインドウ長さとウインドウ類似度の関係についての調査 46
5.3.2 ウインドウ長さの推定への影響に関する調査 47

第 6 章 まとめ 50
6.1 ピーク位置について 50
6.2 類似関係のグラフ構造を利用した絞込みについて 50
6.3 バインディングサイトの特定方法について 51
6.4 各種パラメータについて 51
6.4.1 ウインドウ類似度における移動幅について 51
6.4.2 マイクロアレイデータによる同一転写因子の絞込み 51
6.4.3 ウインドウ長さについて 52
6.5 今後の課題 52

謝辞 54

図 目 次

1.1 遺伝子の発現過程 1
1.2 転写制御因子の動作 2
1.3 遺伝子間の相互作用 3
1.4 既存手法と本研究による推定手法 5

2.1 相補配列生成 7
2.2 マイクロアレイ 8

3.1 ウインドウ 13
3.2 ウインドウ類似度 13
3.3 ピーク位置 14

4.1 既存手法によるピーク位置（比較元：mcpC 比較対象：lytD） 18
4.2 ピーク位置とバインディングサイトの不一致（比較元：mcpC 比較対象：lytD） 19
4.3 改良手法によるピーク位置（比較元：mcpC 比較対象：lytD） 22
4.4 改良前と改良後のピーク位置の比較（比較元：mcpC 比較対象：lytD） . . 23
4.5 比較元遺伝子側からの類似関係 25
4.6 対象遺伝子側からの類似関係 25
4.7 双方向の類似関係 25
4.8 対象遺伝子の類似関係 26
4.9 グラフ構造による判別 26
4.10 ピーク位置の相互関係 27
4.11 クリークを構成するピーク位置の関係 28
4.12 クリークによる判別の効果 (比較元：katA) 29
4.13 クリークによる判別の効果 (比較元：lytD) 30
4.14 特異的に高いウインドウ類似度を持つピーク位置（比較元：katA） 32
4.15 ランダム配列との比較（比較元：katA） 34
4.16 ランダム配列との比較（比較元：lytD） 34
4.17 ランダム配列との比較（比較元：kyuN） 35

5.1 部分文字列の前後への移動による一致 37
5.2 移動幅 Δ の場合の比較対象文字列 38

5.3 移動幅 Δ における閾値 8000 以上のウインドウ増加数 39
5.4 バインディングサイトにおける移動幅の効果（比較元：katA 比較対象：mrgA） 42
5.5 相関係数絶対値分布 44
5.6 相関係数順位分布 44
5.7 発現強度分布 45
5.8 発現強度順位分布 46
5.9 相関係数順位と発現強度順位の分布 47
5.10 ウインドウ長さに対する平均ウインドウ類似度 48
5.11 ウィンドウ長さを変えた場合の推定結果比較 (比較元：yneA) 49

表 目 次

3.1 相関係数の絶対値が大きい上位遺伝子 11

4.1 比較元の各位置 i に対し最大ウインドウ類似度をとる対象遺伝子の位置 j . . 21
4.2 各塩基の出現確率 32

5.1 ウインドウ類似度変化（ykvW） 40
5.2 ウインドウ類似度変化（cheV） 40
5.3 ウインドウ類似度変化（lytD） 40
5.4 ウインドウ類似度変化（mcpC） 40
5.5 ウインドウ類似度変化（mrgA） 41
5.6 ウインドウ類似度変化（katA） 41
5.7 移動幅によるウインドウ類似度変化の平均値 41
5.8 遺伝子 yneA のバインディングサイトデータ 48

第1章　はじめに

1.1　遺伝子間の相互作用について

我々の生存を支えている生命現象はきわめて複雑であり, それには多種類の物質が関わっている．中でも重要なのは，タンパク質と核酸である．タンパク質は，各種の生命現象の直接の担い手であり，核酸，特にDNAは遺伝子の本体である．そこにはタンパク質を作るための設計図が書き込まれている．1990年代後半以降，さまざまな生物種の全ゲノム解析が行われてきた（100種以上の微生物ゲノム，出芽酵母，線虫，ヒトゲノム等）．塩基配列の決定は, 遺伝子の機能を予測したり, その活性や発現を操作することを目的として進められている [1]．DNAは，4つの塩基（A：アデニン，T：チミン，G：グアニン，C：シトシン）からなり，それらが互いに相補的な二重らせん構造を形成している．遺伝子とはDNA中の機能的な役割を持つ領域を示し，発現によって生体機能にかかわるタンパク質を生成するための設定図としての役割を持っている．タンパク質の生合成に対する遺伝子の発現（manifestation）の機構は, 転写（transcription）と翻訳（translation）の二つの段階に分けられる．転写は，DNAを鋳型としてmRNAを合成する反応であり，遺伝子の塩基配列から，RNAの相補的塩基配列へ, 遺伝情報が単に移し変えられる．この転写反応はRNAポリメラーゼと呼ばれる酵素の関与のもとに進行し，最終的にタンパク質の合成に必要のない部分は切り離され，翻訳段階でこの配列を元にタンパク質が生成される (図 1.1).

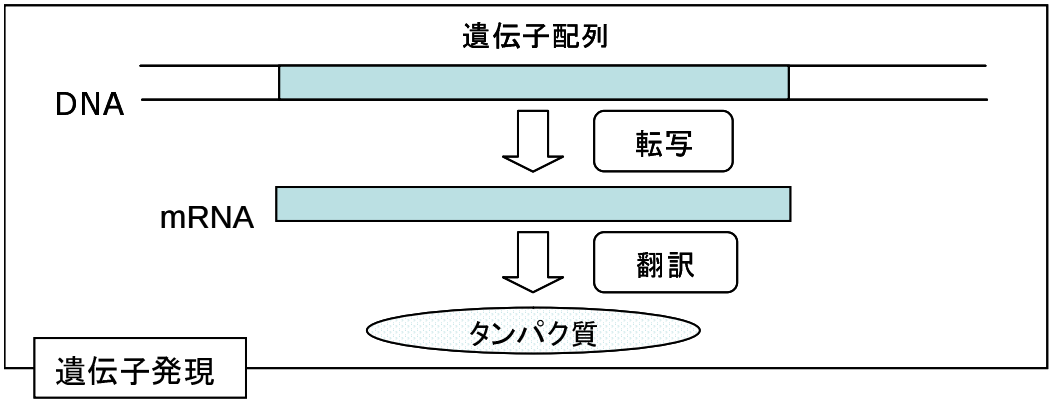

図 1.1: 遺伝子の発現過程

生成されたタンパク質の大部分は酵素を構成し，生命活動を維持する重要な役割を担っているが，ほかの機能として自身をコードしていた遺伝子や他の遺伝子の転写に影響を与える場合がある．このとき，影響を与えるタンパク質を転写制御因子と呼ぶ．転写制御因子が発現を促進する場合には図 1.2 のように RNA ポリメラーゼにより転写が開始され遺伝子の発現が起きる．転写制御因子は，遺伝子上流に存在する転写制御領域と呼ばれる転

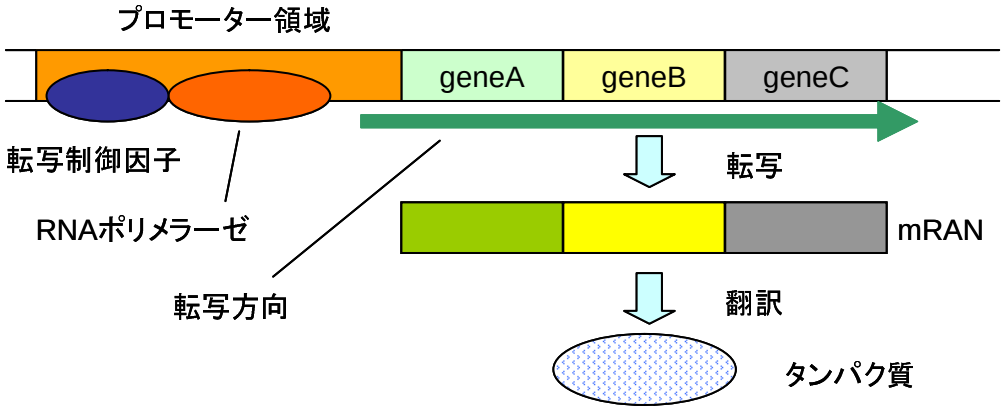

図 1.2: 転写制御因子の動作

写の制御に関与する領域で特異的な塩基配列に結合し，転写因子等に影響を与えることで，転写を促進または抑制する [2]．転写制御領域は，一般に転写開始位置から上流数百塩基対の長さの領域を指し，例外として遺伝子の中や離れた上流に存在する場合もある．また，転写制御因子は複数のタンパク質からなる複合体を形成し，大きく分けて DNA 塩基配列を認識し，特異的に結合する結合ドメインと転写因子の働きを制御する制御ドメインの二つのドメインからなる．結合ドメインは異なる塩基配列を共通の配列として認識し結合する能力を持ち，結合する塩基配列の長さは 6 から十数塩基であることが知られている．また，一つの転写制御因子に結合ドメインが複数ある場合があり，それらは連続または離散的な領域を認識し結合する [3].

ある遺伝子 A から生成されたタンパク質がある遺伝子 B の転写制御領域に結合し，遺伝子 B の転写因子に影響を与えることで，遺伝子 B の発現を制御する．細胞内では，数多くの遺伝子間にこのような関係があり，各遺伝子から生成されたタンパク質により互いの発現を制御し合う関係が存在する．このような関係を遺伝子の相互作用と呼ぶ.

通常，DNA 中には複数の遺伝子が存在し，遺伝子間の依存関係は多対多の関係であり，図 1.3 のように遺伝子 A から生成された転写制御因子が，生成元の遺伝子の発現を抑制したり, あるいは他の複数の遺伝子の発現に対し，抑制や促進を行う場合がある．また，遺伝子 A が直接の制御遺伝子でなくとも，遺伝子 C や遺伝子 D のように遺伝子 A の被制御遺伝子がさらに次の遺伝子の制御遺伝子になる場合も多く，遺伝子間に間接的な作用が存在する場合もある．このような遺伝子間の相互作用をネットワークとしてとらえたものを遺伝子調節ネットワークと呼ぶ．遺伝子発現の遺伝学的な解析にかわって，遺伝子発現

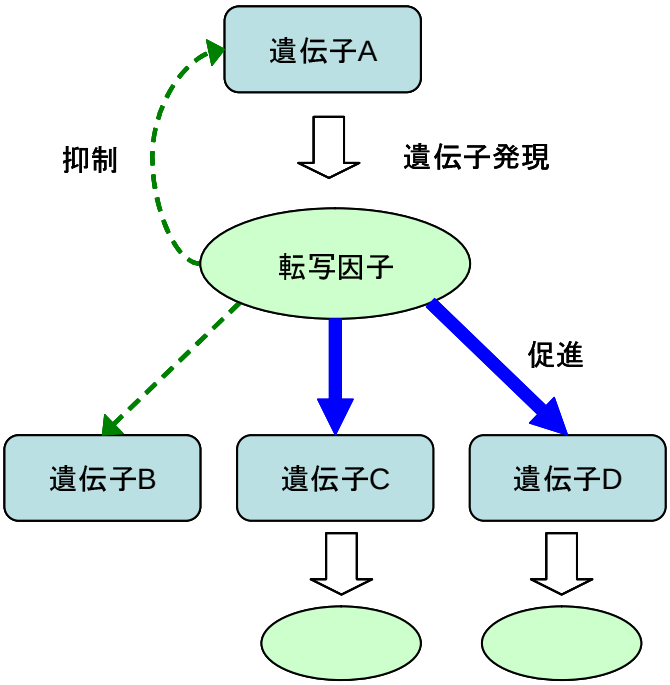

図 1.3: 遺伝子間の相互作用

データを用いて，計算機科学的アプローチにより調節ネットワークを推定する研究が近年盛んに行われている [4]．遺伝子の相互作用を解明することで，細胞内の様々な反応を一つのシステムとしてモデル化することが可能になり，新薬の開発などに繋がると考えられ，遺伝子相互作用の推定精度のさらなる向上が望まれている．

1.2 従来研究

本節では遺伝子間の依存関係を推定する従来研究について説明する．遺伝子間の依存関係推定に関する従来研究では，大きく分けて 2 種類の方法により推定が行われている．一つは従来から行われていた生物学的な性質に基づいた推定である．そして，もう一つは近年盛んに行われるようになった，ある種の数理的モデルを応用した解析である．

まず一つ目の生物学的性質を用いた推定は，プロモーター領域に関する研究が中心である．転写制御因子は特異的な塩基配列を認識，結合することで遺伝子の発現に影響を与えるという生物学的知見に基づき，依存関係が既知の遺伝子の転写制御領域の間に類似した塩基配列を発見する．そして依存関係が未知の遺伝子群の中から転写制御領域にそれらと類似した塩基配列を持つ遺伝子を探索，それらの遺伝子が実際に被制御遺伝子であるかどうかを実験により確認するという方法である．多重配列アラインメントやモチーフを利用した方法があり，これらの方法により得られた依存関係やバインディングサイトの配列，

DNA における位置などの情報は，データベースとして蓄積され公開されており，現在の研究の基礎となっている [1].

これに対し，近年開発されたマイクロアレイ技術により得られたデータを元に，遺伝子間の相互作用や調節機構を解釈するための遺伝子発現パターンのモデルを構築しようとする試みが，二つ目の計算機モデルを応用した方法である．マイクロアレイ技術は，生物種によっては数千から数万もの遺伝子の種類があるため，調節機構や遺伝子配列，タンパク質の変化の解析を数千の遺伝子について一度に行うために開発された．計算機モデルを利用した推定は，マイクロアレイ実験によって得られる遺伝子発現変化のパターンの情報により行われる．代表的なものに，ブーリアンネットワーク [5] やベイジアンネットワーク [6][7] を用いた推定がある．ブーリアンネットワークは，遺伝子の発現を 2 値化して遺伝子間の依存関係を論理関数として表現する手法である．しかし，推定には測定が困難な発現状況の時系列データが必要であるうえ，推定する遺伝子数に対し多量の DNA マクロアレイデータを必要とする．ベイジアンネットワークは，ブーリアンネットワークとは異なり，遺伝子の発現を静的で確率的な事象としてモデル化する．確率変数間の依存関係を非循環な有効グラフで表現する手法であり，現在，この分野において広く用いられている手法であるが，同じデータを説明可能な複数のモデルが存在するなどの問題がある．これらの手法はマクロアレイデータの精度に強く依存しているが，現在の実験技術ではマイクロアレイデータには 30〜40 %程度の誤差が存在しており [8]，マイクロアレイデータのみから正確なモデルを構築することは難しいと考えられる．

1.3 本研究の目的

本研究の目的は，平石・土居らの提案による DNA マイクロアレイと DNA 配列比較の組合せによる遺伝子間相互作用の推定方法を以下の方法により改良することである．既存手法では，選択した一つの遺伝子 a と比較して，破壊株による DNA マイクロアレイデータを用いた遺伝子発現強度の相関が高く，かつ，DNA 配列の遺伝子制御領域内に類似性の高い部分的な文字列を含む遺伝子の集合 B を，a と同じ制御因子の制御を受ける遺伝子と推定している．これはマイクロアレイデータによる発現傾向の相関という統計的性質と，同一転写制御因子は特異的文字列を認識，結合することで遺伝子の発現に影響を与えるという生物学的性質の二つを組合わせて推定していることになる．

これに対し，本研究では，既存手法で得られた遺伝子間の類似性をグラフに表し，同じ制御因子の制御を受ける遺伝子の集合が持つと考えられるグラフ構造を被制御遺伝子の判別に利用することで候補遺伝子を絞り込む方法を提案する．これにより既存手法で得られた被制御遺伝子の数をさらに絞込み推定精度の向上を図る．つまり，図 1.4 に示すように，遺伝子間の類似関係が持つグラフ構造を新たな性質として判別に用いることで，同一転写制御因子の被制御遺伝子をより正確に推定可能にすることを目的とする．

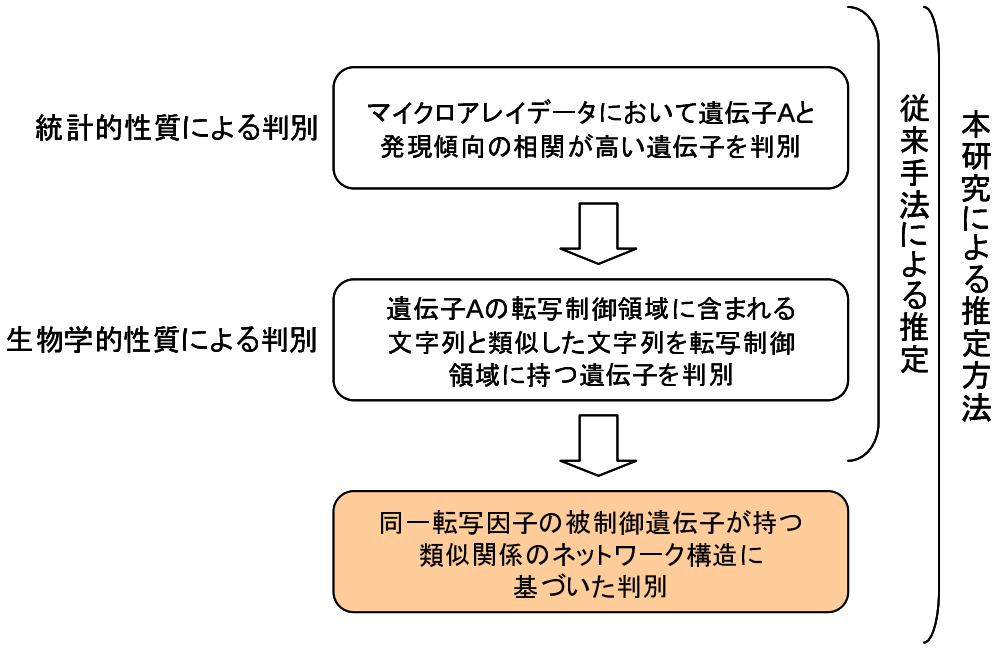

図 1.4: 既存手法と本研究による推定手法

1.4 本研究の内容

本研究では以下のようなアプローチで推定精度向上を図る.

- 既存手法の改良
  - 各遺伝子間の類似関係によるグラフ構造を用いた被制御遺伝子の判別方法の提案.
  - 結合サイト候補位置の決定方法の改良.
  - 結合サイト特定方法の提案.
- 各種パラメータに関する調査と最適化
  - 部分文字列の移動幅とウインドウ類似度の関係に関する調査と移動幅の最適化.
  - 発現強度の相関を用いた被制御遺伝子判別における閾値の最適化.
  - ウインドウ長さとウインドウ類似度の関係についての調査.

第2章　使用する菌類と各種データの紹介

本章では本研究で推定対象とする菌類とそのDNA配列データ，DNAマイクロアレイデータについて説明する．対象とする菌類は枯草菌とし，全塩基配列および各遺伝子の位置などの情報を使用する．また，DNAマイクロアレイデータは，九州大学大学院生物資源科学研究府遺伝子資源工学専攻遺伝子制御講座から提供された106種類の破壊株データを使用する．

2.1　枯草菌について

本研究の実験に用いる枯草菌(Bacillus subtilis)は，細菌の一種であり原核生物である．枯草菌のDNAは日本とヨーロッパの国際共同研究により，すでに全塩基が決定されており，全ての遺伝子の位置についても判明している[9]．また，制御関係が既知である被制御遺伝子や転写制御因子の結合する結合サイトの塩基配列，DNA上での位置などのデータはDBTBS(http://dbtbs.hgc.jp)[10]などのデータベースで公開されており，推定結果についての検証が比較的容易に可能であるため，本研究では推定対象として枯草菌を使用する．

2.2　DNA塩基配列データについて

遺伝子の本体であるDNAは，正式にはデオキシリボ核酸と呼ばれる．DNAは，ヌクレオチドと呼ばれる単位が次々に連結した鎖状の高分子物質である．通常，その鎖が2本より合わさって二重らせん構造となる．DNAは，塩基，デオキシリボースと呼ばれる糖，リン酸の三者から構成される．ここでは本研究に関係する塩基についてのみ説明する．

まずDNAの一本鎖について説明する．塩基には，アデニン（A），グアニン（G），シトシン（C），チミン（T）の4種類があり，その並び方がDNAの塩基配列である．例えばAGTCTACCGT……のように読み取ることができる．この塩基配列が遺伝情報を担っている．そして，遺伝情報が発現される過程では，DNA上の遺伝子部分の塩基配列はタンパク質のアミノ酸配列をコード（指定）していることになる．DNAの一本鎖に説明してきたが，実際にはDNAは，二本鎖がより合わさった二重らせん構造を持つ．鎖には方向があり，二本鎖は互いに逆方向に配置し，二本鎖の骨格から内側に突き出た互いの塩基の枝の部分で二本鎖が結合する．

４種類の塩基のうち，A に対してはＴ（逆に T に対してはA），G に対してはC（逆に C に対してはG）のように結合する相手は決まっている．これらの関係を相補性と呼び，対になった塩基同士は，相補的塩基対と呼ぶ (これら以外の組合せでは結合しない)．これにより片方の鎖があれば，図 2.1 のように相補鎖の塩基配列を機械的に求めることができる．

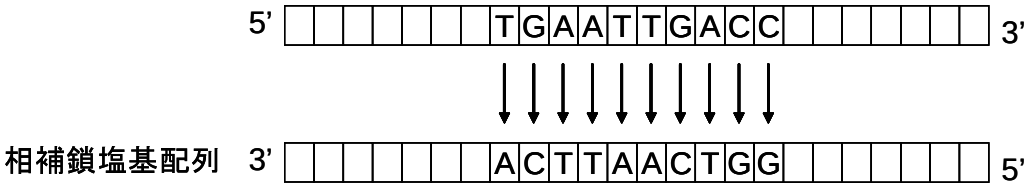

図 2.1: 相補配列生成

2.3　マイクロアレイデータについて

マイクロアレイデータとは，マイクロアレイ技術により遺伝子の発現量を測定したデータである．マイクロアレイ技術は調節機構や遺伝子配列，タンパク質の変化の解析を数千の遺伝子について一度に行うために開発された技術である．本研究で使用する遺伝子発現量を解析するマイクロアレイは，主として cDNA アレイ（スポットアレイ）と高密度オリゴヌクレオチドアレイの 2 種類に大別される．ここでは本研究で使用する cDNA アレイについて説明する．

スポットアレイの作り方と使用方法を図 2.2 に示した．原則として，ある生物種がもつ多くの遺伝子について，エキソン部分の代表となる DNA 断片をポリメラーゼ連鎖反応を用いて増幅し，スライドガラス上に高密度のグリッドパターンでスポットする．一方で，比較したい 2 つの生物試料から mRNA を抽出し，その mRNA に対応する DNA のコピー（相補的DNA，すなわち cDNA）を作成する．cDNA を作成するときに Cy3 や Cy5 といった蛍光色素で標識しておく．この標識された cDNA をスライドとハイブリダイズし，それぞれの DNA スポット上の蛍光色素のシグナル値を測定する．こうして得られた 2 種類の蛍光色素のシグナル値の比はもともとの 2 種類の試料に含まれる mRNA の量の相対比を反映している [1].

今回使用するデータは，特定の遺伝子の破壊株と野生株を２つの生物試料としたマイクロアレイデータである．破壊株とは特定の遺伝子を実験的手法により破壊した株のことをいう．また野生株とは破壊操作を施していない通常株のことをいう．株とは，遺伝的形質が同じ生物の別の個体をいう．破壊株において破壊した遺伝子が調節遺伝子であった場合，その遺伝子から直接的または間接的影響を受ける遺伝子の発現量は，野生株における発現量と比較して増加あるいは減少するため，マイクロアレイにより観測できる．した

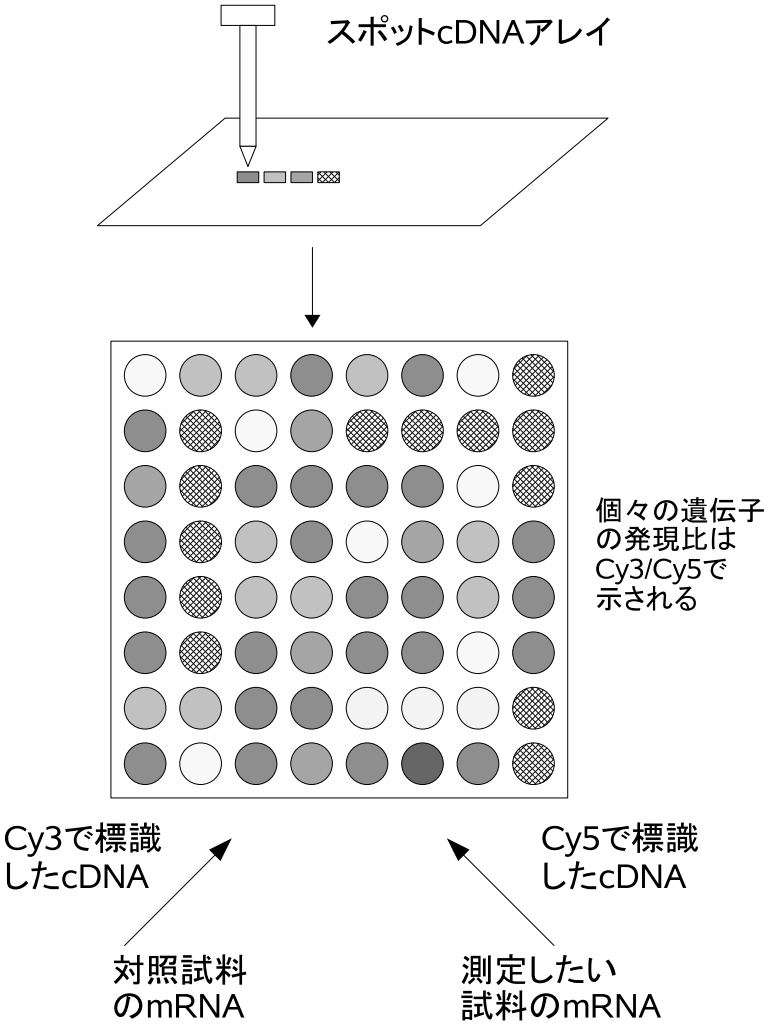
スポットcDNAアレイ
個々の遺伝子
の発現比は
Cy3/Cy5で
示される
Cy3で標識
したcDNA
Cy5で標識
したcDNA
対照試料
のmRNA
測定したい
試料のmRNA

図 2.2: マイクロアレイ

がって，式 2.3 は破壊株 X における遺伝子 i の発現量の比を表す．

$$ratio_i^X = \frac{\text{野生株での遺伝子 } i \text{ の発現量}}{\text{破壊株 } X \text{ での遺伝子 } i \text{ の発現量}}$$

理論上，破壊した遺伝子と遺伝子 i の間に依存関係がない場合 1 となり，遺伝子 i が正の制御を受ける場合は 1 よりも小さく，また負の制御を受ける場合は 1 よりも大きくなる．この発現量の比をすべての破壊株に対して一つの遺伝子の値の分布をとると対数正規分布に近い分布になる [3]．そのため，一般的には対数により正規化した値を用いるため，本研究でもそのような値 $log(ratio_i^X)$ を用いる．

マイクロアレイは遺伝子発現解析に大変有効であるが，マイクロアレイ実験を行ううえでは誤差を生じる可能性のある原因がいくつも存在する．具体的には，蛍光色素の取り込み効率が異なったり，スライドに標識された DNA をハイブリダイズする効率の差，スライドの蛍光シグナルを検出する際の正確性やばらつきなどである．また試料間におけるあるレベルでの生物学的ノイズにより，生命現象がしばしば実際の個々のマイクロアレイ実験の結果に反映されないこともある．また遺伝子発現マイクロアレイは，mRNA という本来は不安定な分子を定常状態にして測定していることも理由の一つである．マイクロアレイ実験で測定される個々の mRNA の細胞内における定常状態レベルは，転写と分解の割合に依存している．また，mRNA をマイクロアレイの定量のために標識する際には，個々の mRNA がすべて同等の効率で cDNA コピーとして標識されるという仮定に基づいているが，実際に個々の mRNA がコピーされて標識される量にはばらつきが生じている可能性もある [1]．これらの理由から本研究ではマイクロアレイデータの個々の値については用いず相対的な順位や発現の相関関係の調査に用いている．また破壊株と野生株の発現量の比をマイクロアレイにより観測した各遺伝子への影響は，直接的な影響だけでなく途中に別の遺伝子をいくつか介して影響を与えた場合も含まれていることに注意する必要がある．

第3章　既存手法の紹介

本研究で用いる破壊株 DNA マイクロアレイデータと塩基配列比較による遺伝子相互作用の推定方法について説明する．同時に，既存手法における問題点の指摘を行う．

3.1　既存手法の特徴

既存手法の大まかな推定手順は以下の通りである．

1. DNA マイクロアレイデータは発現傾向の相関から発現パターンが近い同一の転写制御因子の被制御遺伝子である可能性の高い遺伝子群を大まかに選別することのみに使用する．そして，得られた遺伝子群の中で特定の遺伝子と比較して転写制御領域における長さ数十程度の文字列の比較を類似性を比較し，残った遺伝子の中で，制御領域内に閾値以上の類似パターンをもつ遺伝子グループを抽出する．

2. 特定のパターンを仮定せずに，指定した遺伝子と類似した部分領域をもつ遺伝子を網羅的に検索する．また，類似度の計算には，部分文字列の出現頻度に関する統計的特異性を用いて，結合サイトに含まれる可能性の低い部分を除外し，類似度が高い領域は共通の因子が結合するサイトであると推定する．

この方法では，DNA マイクロアレイによる発現強度の相関と，転写制御領域における塩基配列の類似性という二つの要素を組合わせることで，DNA マイクロアレイや塩基配列の類似性という誤差の大きい一つのデータに依存しすぎることなく，推定することが可能である．また，転写制御領域における類似文字列の発見により遺伝子間の制御関係とバインディングサイトを同時に推定可能であることも特徴である．

3.2　使用するデータと各種定義

3.2.1　マイクロアレイによる発現強度の相関係数

マイクロアレイにより得られた野生株と破壊株の発現比から，遺伝子 a とその他の遺伝子間の発現比の相関係数を計算し，得られた相関係数の対数を取り正規化する．正規化した数値の絶対値が高い上位 300 遺伝子 B を同一の転写因子の被制御遺伝子候補として選択する．例として遺伝子 lytD との相関係数の絶対値が大きい上位遺伝子を表 3.1 に示す．

表 3.1: 相関係数の絶対値が大きい上位遺伝子

順位	遺伝子名	相関係数
1	lytD	1
2	mcpB	0.780334908875107
3	yhdD	0.753614008050558
4	cheV	0.730230324268497
5	yfmS	0.714937397596026
6	yvzB	0.681842514780249
7	hag	0.679194998519859
8	motA	0.667723244125916
9	ywtD	0.667458696448107
10	yfmT	0.653573847214765
11	mcpA	0.652530094416397
12	tlpB	0.647735473483387
13	motB	0.643756930483576
14	yviA	0.638212350922099
15	epr	0.631171607378781
		...

3.2.2 ウインドウ類似度

転写制御因子は各因子ごとに特異的な塩基配列を認識し，結合することが生物学上知られており，それらは数十塩基以内の長さの場合が多い．同一の転写制御因子のバインディングサイトは類似しており，転写制御領域に類似した文字列を含む遺伝子は同一の転写制御因子の被制御遺伝子である可能性が高いと考えられる．このため転写制御領域における文字列の類似性の比較する評価基準を定める．

同一の転写制御因子のバインディングサイトであっても，結合する遺伝子により文字の挿入，削除，置換などがあり，完全に一致していない場合も多々あり，このような違いにも対応して文字列の類似性を評価する必要がある．文字の挿入や削除に対応するため，文字列を短い部分文字列の集合と考え，各位置での部分文字列の類似性を総合して文字列全体の類似性を評価する．

部分文字列の類似性を以下のように各位置における一致文字数の数により，評価する．

- 部分文字列の類似度
  長さ 6 の文字列 s が与えられたとき，i 番目の文字を $s[i]$ により表す．長さ 6 の二つの文字列 s_1，s_2 が与えられたとき，各位置 i において一致する文字の合計を

$$MPOS(s_1, s_2) := \{0 \leq i \leq 5 | s_1[i] = s_2[i]\}$$

とする．ここで $k = MPOS(s_1, s_2)$ としたとき，文字列 s_1，s_2 間の類似度 $ssim(s_1, s_2)$ を以下のように定義する．

$$ssim(s_1, s_2) := ((1/4)^k (3/4)^{6-k})^{-1}$$

同一の転写制御因子のバインディングサイトであっても，被制御遺伝子により，文字の挿入，削除などにより連続した一致をしないことがある．これに対応するため，比較対象の文字列が移動幅 Δ の範囲で移動すること許容する．ここで，遺伝子 g の位置 k からの連続した 6 文字を $g[k]$ とするとき，位置 j からの比較対象文字列を以下のように集合 $g^{\Delta}[j]$ とする．

$$g^{\Delta}[j] := \{g[j-\Delta], ..., g[j-1], g[j], g[j+1], ..., g[j+\Delta]\}$$

比較元の文字列と最も類似度の高い位置の文字列との類似度を位置 i に対する類似度とする．また，遺伝子制御領域において期待値よりも出現頻度の高い文字列は，バインディングサイトになる可能性が低いと考えられる．このため統計的特異性の低い文字列は類似していても評価しない．長さ 6 の文字列の統計的特異性を表すため以下のように特異度 $\theta(s)$ を定義する．

- 特異度
  遺伝子制御領域における各塩基の出現確率を p_A，p_T，p_G，p_C としたとき，文字列の出現がランダムならば， A, T. G, C がそれぞれ a, t, g, c 回出現する文字列 s の出現確率は，$p_s = p_{A^a} p_{T^t} p_{G^g} p_{C^c}$ である．遺伝子制御領域に出現する長さ 6 の文字列の総数を N とすると，文字列 s が制御領域含まれる個数の期待値は $E_s = N_{p_s}$ あり，長さ 6 の文字列 s の特異度を次のように定義する．

$$\theta(s) := (O_s - E_s)/E_s$$

以上のことを踏まえ，部分文字列の多少の前後を考慮した長さ 6 の文字列の類似度を以下のように定義する．

$$sim_{a,b}(i, j) := \mathrm{Max}\left[\{ssim(a[i], s) \mid \theta(a[i]) \leq v, s \in b^{\Delta}[j], \theta(s) \leq v\} \cup \{0\}\right]$$

ここで，v は文字列の特異度閾値である．次に，バインディングサイト全体の類似性を評価するために長さ 20〜30 程度の文字列全体の類似性を評価するためにウインドウ類似度を以下のように定義する．

- ウインドウ類似度
  遺伝子 g の転写制御領域における位置 i からはじまる長さ l の文字列をウインドウと呼び，$W_g[i]$ により表す．図 3.1 のように二つのウインドウ $W_a[i]$，$W_b[j]$ が与えられ，a を比較元遺伝子，b を比較対象遺伝子としたとき，ウインドウ類似度 $wsim_{a,b}(i, j)$ を以下のように定義する．

$$wsim_{a,b}(i, j) := \sum_{k=0}^{l-6} sim_{a,b}(i+k, j+k)$$

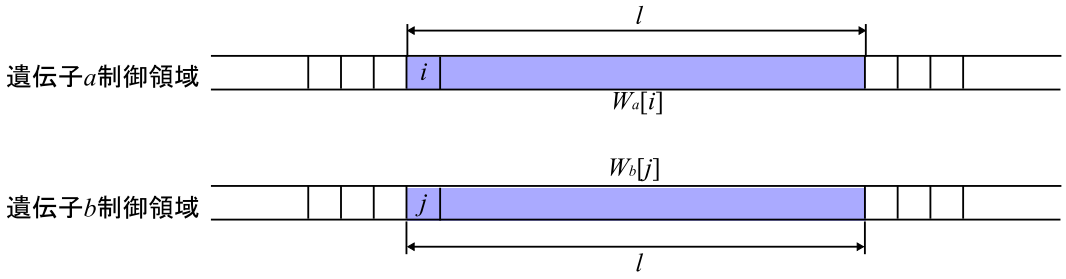

図 3.1: ウインドウ

ウインドウ類似度は図 3.1 のように，ウインドウ内の部分文字列の類似度の総和を表す.

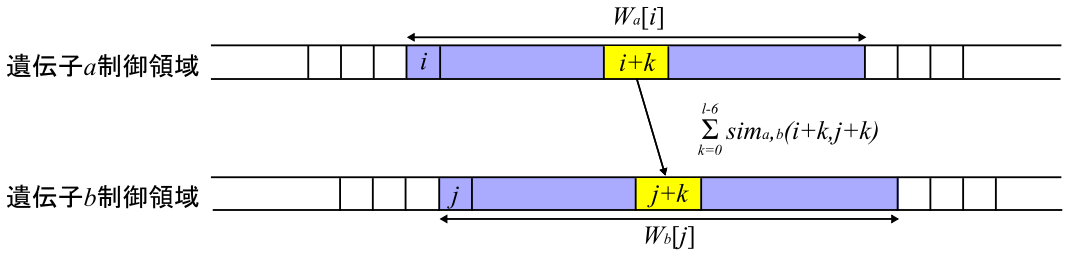

図 3.2: ウインドウ類似度

3.2.3 ピーク位置

遺伝子 a と遺伝子 b が同一の転写制御因子の被制御遺伝子であるならば，比較元の遺伝子を a，対象遺伝子を b と考えた場合，以下のように考えることができる.

- 遺伝子 a の制御領域の位置 i がバインディングサイトであるならば，遺伝子 b の全てウインド位置と比較した場合，最もウインドウ類似度の値が高い位置 j が遺伝子 b のバインディングサイトである可能性が高い.
- 遺伝子 a の制御領域内では位置 i がバインディングサイトであるならば，前後と比べて，より高いウインドウ類似度を位置からのウインドウ内にバインディングサイトは含まれている可能性が高い.

よって，以下の手順で選択された位置 i バインディングサイトの可能性が高いピーク位置と呼ぶことにする.

1. 遺伝子 a のウインドウ $W_a[i]$ を固定し，それに対する遺伝子 b の最大類似度 $Max_{ab}[i]$ を計算.

2. i の変化に対する $Max_{ab}[i]$ が，極大値をとり，かつウインドウ類似度が 15000 以上の位置 i のみをピーク位置候補とする．

3. 位置 i から距離 $l/2$ 内にピーク位置候補が複数存在する場合には，より大きいウインドウ類似度を持つ位置のみをピーク位置候補として残し，最終的に残った位置 i をピーク位置とする．

ウインドウ類似度の値に閾値を設けたのは，あまりウインドウ間に類似性がない場合には，遺伝子 a と b の共通の転写因子が結合する可能性は低いと考えられるからである．

例として図 3.3 のような最大類似度 $Max_{ab}[i]$ をとる遺伝子についてついて考えると，矢印で示した部分がピーク位置となる．

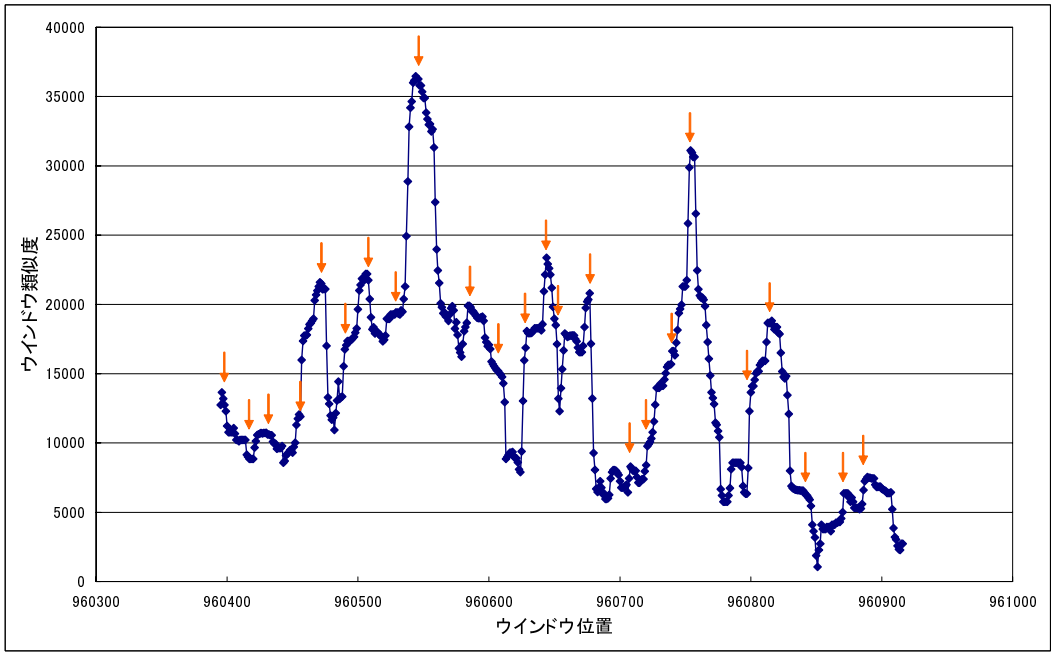

図 3.3: ピーク位置

3.3 推定手順

本節では同一の転写制御因子の被制御遺伝子の推定方法を例として調査対象の遺伝子を a とした場合について説明する．

ステップ 1. 入力データとして破壊株によるマイクロアレイデータ，DNA 塩基配列，遺伝子位置のデータを与える．

ステップ 2. 遺伝子 a と相関係数の絶対値が大きい上位 300 遺伝子群 B を選択．

ステップ 3. 遺伝子 a と遺伝子群 B の各遺伝子のペアについて，制御領域の全ての位置についてウインドウ類似度を計算する．このとき，遺伝子 a の制御領域における各位置 i に対する遺伝子群 B の類似度の最大値 $\mathrm{Max}_{aB}[i]$，平均値 $\mathrm{Avg}_{aB}[i]$ を計算する．

ステップ 4. 遺伝子 a の制御領域において，$\mathrm{Max}_{aB}[i] - \mathrm{Avg}_{aB}[i]$ が大きく，かつ，ピーク位置を含むような領域 R を探す．

ステップ 5. 領域 R にピーク位置があり，かつ，類似度の高い遺伝子群 B_R を求める．破壊株データにおいて遺伝子 a に強く影響を与えた因子郡を求める．各因子 f について，その影響を強く受ける遺伝子グループを抽出し，それを B_{Rf} とする．

ステップ 6. 遺伝子 a および B_{Rf} の各遺伝子について，マッチした位置からのウインドウを取り出し，多重配列アラインメントを行う．共通パターンの存在が確認できたらバインディングサイトである可能性が高い．

ステップ 7. 遺伝子 a を制御する因子 f, 制御方向 $(+/-)$，因子 f により制御される a 以外の遺伝子群 B_{Rf}，遺伝子 a および B_{Rf} の各遺伝子のバインディングサイトを推定結果として出力．

ステップ 1〜7 までの手順を全ての遺伝子について繰り返す．

3.4 既存手法における問題点

既存手法における問題点は大きく分けて二つ考えられる．

問題点 1 バインディングサイト候補の数が，バインディングサイトを特定するほど十分に絞り込まれていない．

原因 既存手法で用いている，マイクロアレイデータと転写制御領域における部分文字列の類似性評価による絞込みだけでは十分に数を絞り込めていない．

問題点 2 既存手法では正しく推定できない遺伝子が存在する．

原因 1 バインディングサイトを含むウインドウ位置のウインドウ類似度がピーク位置の閾値として設定している 15000 に達しない場合がある．sigD のようなバインディングサイト長さが 6〜8 程度と短い場合，ウインドウ類似度の値は 10000 前後と閾値に達しない場合が多い．

原因 2 ウインドウ類似度はピーク位置の閾値に達しているが，ピーク位置として選択されない．これはピーク位置選択方法の問題である．

原因 3 同一の転写因子の被制御遺伝子であるにも関わらず，マイクロアレイデータによる発現傾向の相関が高い上位遺伝子 300 に含まれない．

第4章　既存手法の改良

4.1　ピーク位置決定方法の改良

ピーク位置は遺伝子間の転写制御領域のウインドウ類似度を元に，バインディングサイトである可能性が高い位置を絞り込むために使用される．しかしながら，既存手法におけるピーク位置の決定方法では，実際にはバインディングサイトを含むウインドウであるにもかかわらずピーク位置として選択されないケースがいくつか認められた．ピーク位置として選択されないウインドウ位置は，以降のバインディングサイトの推定からもれてしまうため，正しく推定が行えない．既存手法では単一の方法で被制御遺伝子の可能性の高い遺伝子に絞るのではなく，異なる情報を複数組合わせて用いることにより，被制御遺伝子をもれなく推定することに重点をおいている．このため，たとえバインディングサイトの候補の数が増加しても，現在のバインディングサイトの決定方法を改良し，より正確にバインディングサイトを含むウインドウ位置が，ピーク位置として選択できるよう改良する必要がある．本章では，現在の選択方法でピーク位置に選択されなかったバインディングサイトの数値データを調査し，ピーク位置決定方法を改良する．

4.1.1　既存手法のピーク位置決定方法について

比較元遺伝子を遺伝子a，対象遺伝子を遺伝子bとしたとき，遺伝子aのウインドウ位置iを固定し，このウインドウと最大のウインドウ類似度をとる遺伝子bのウインドウ位置j，およびこのときのウインドウ類似度を得る．以降ここで得られた最大ウインドウ類似度を位置iに対するウインドウ類似度と呼ぶことにする．このとき現在のピーク位置の決定方法は，単純に遺伝子aの各位置iに対しする最大ウインドウ類似度$Max_{ab}[i]$について，周囲より高いウインドウ類似度を持つ部分にバインディングサイトは含まれる可能性が高いという仮定から，位置iを移動させたときに$Max_{ab}[i]$が極大値を取るような位置iをピーク位置候補として考えている．また，二つのピーク位置i_1, i_2があるとき，ピーク位置の距離$|i_1 - i_2|$が近すぎる場合には，ウインドウ同士が重なり同じ文字列が含まれている．ピーク位置の総数はバインディングサイトを除いてしまわない限り,少ないほうが最終的にバインディングサイトの特定につながる可能性が高まる．このため，

$$|i_1 - i_2| \leq l/2$$

の場合には，より大きいウインドウ類似度をとるピーク位置のみを残すことで，重複したウインドウ部分のピーク位置を減らし, ピーク位置の総数の減少を図っている．この方法により求めたピーク位置は図 4.1 のように前後のピーク位置とある程度の間隔を保ち，その周辺領域で，ウインドウ類似度が極大となるような部分が選択される．さらに既存手法では, ウインドウ類似度で 15000 程度を閾値とし，閾値以上のウインドウ類似度を持つピーク位置のみを考慮している. これはピーク位置の数を絞り込むことで, バインディングサイト候補を絞り込んでいる.

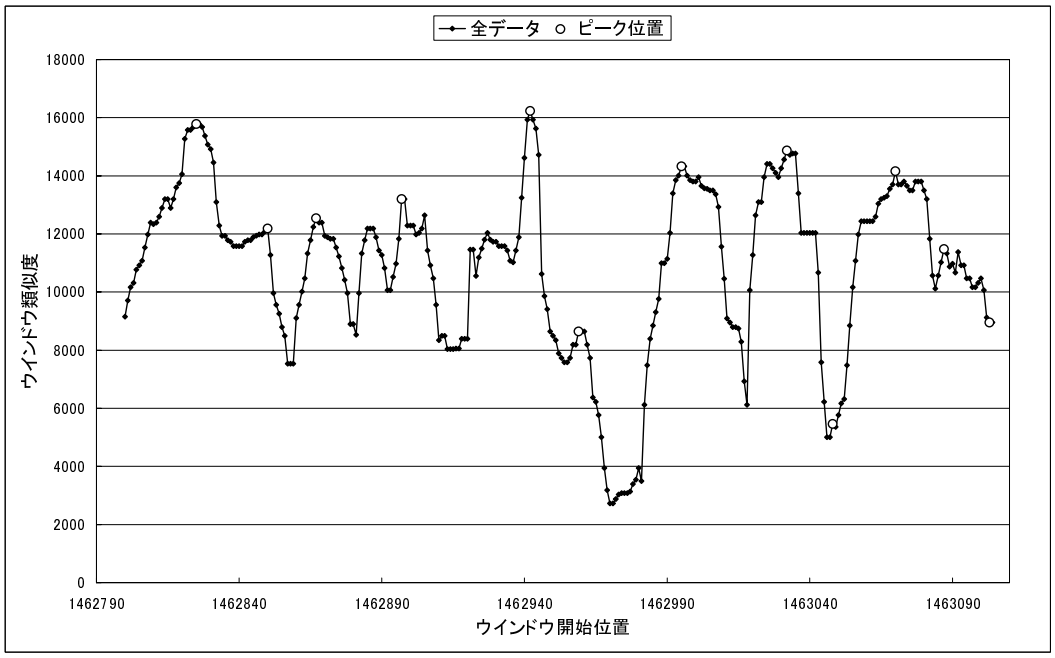

図 4.1: 既存手法によるピーク位置（比較元：mcpC 比較対象：lytD）

4.1.2 現在のピーク位置決定方法における問題点

既存手法の推定結果の中で, 既知のバインディングサイトについて正しく推定できなかったケースについて調査した結果，最大ウインドウ類似度には含まれていることが分かった．つまり，既存手法のピーク位置の決定方法では，バインディングサイトを含むウインドウ部分は前後と比較して高い類似度，つまり極大値を持つだろうという仮定に基づいている．しかしながら，この仮定に当てはまらない場合があるということである．この場合の例として, 比較元の遺伝子を mcpC，比較対象の遺伝子が lytD の場合のピーク位置について図 4.2 に示す.

図 4.2 は図 4.1 を部分的に拡大した図である．極大値の位置にあるピーク位置が前後のバインディングサイトを含むウインドウ部分に含まれていないことがわかる．バインディ

ングサイトを含む前後のウインドウ位置とピーク位置のウインドウ類似度の差はごく小さいものであるが，現在のピーク位置の決定方法では極大値をとる部分をピーク位置と選択するため，ピーク位置とバインディングサイトが一致していない．このため，正しく推定を行うためには，ピーク位置の決定方法を見直し, バインディングサイトを含むウインドウがピーク位置に含まれるよう修正する必要がある.

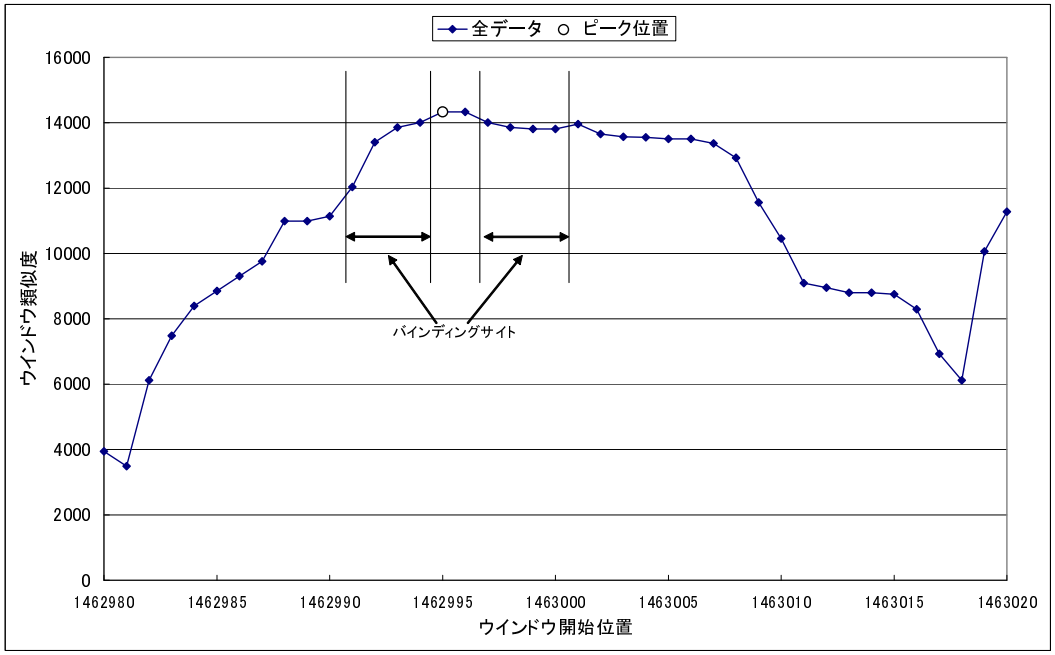

図 4.2: ピーク位置とバインディングサイトの不一致（比較元：mcpC 比較対象：lytD）

4.1.3 ピーク位置決定方法の改良

ピーク位置の決定方法を改良するため，図 4.2 の数値データ（表 4.1）を詳しく見ていくことにする．これまで，ピーク位置の決定には比較元のウインドウ位置 i に対するウインドウ類似度の変化についてのみ着目し，バインディングサイトはウインドウ類似度が極大値をとるような位置にくることが多いという性質を利用していた．ここではバインディングサイトを含むウインドウ位置にそれ以外の共通性がないかどうかについて調査する.

表 4.1 を詳しく見ていくと，比較元遺伝子のウインドウ位置 i を移動させたとき，対象遺伝子のウインドウ位置がある程度近い領域が連続して現れていることがわかる．これは一箇所で高いウインドウ類似度をとる，つまり一致文字数が多い箇所があったならば，当然両方のウインドウを同じだけ移動させても高い類似度ととるということと，さらに，ウインドウ類似度の定義において，比較するウインドウ間に長さ 6 文字の類似した部分文字列が含まれていた場合，文字の挿入や削除によりウインドウ間で部分文字列の位置に多少

のずれがある場合にも，ウインドウ類似度が高くなるように定められていることが原因である．この移動幅により，類似した部分文字列のウインドウ内における位置が完全に一致する場合，つまり比較元と比較対象のウインドウを同じだけ移動させた場合だけでなく，対象遺伝子側のウインドウ位置が前後に移動幅分変化した場合にも高いウインドウ類似度をとる．よって比較元の遺伝子のウインドウ位置 i が連続的に変化した場合に，同様に連続して変化する場合と，多少前後した近い領域にあるウインドウ位置がくる場合とがあると考えられる．表 4.1 では，各位置 i に対応する最大ウインドウ類似度をとる対象遺伝子の位置 j は区切り線で仕切られた領域ごとに，各ウインドウが重なる程度の距離の部分が連続していることがわかる．

表 4.1 においてバインディングサイトの位置は，連続した領域のなかでの最大値をとるウインドウ位置が最もバインディングサイトを多く含んでいた．他の遺伝子についても，既知のバインディングサイトを含む領域について同様に調査したところ，バインディングサイトは，このような連続した領域における最大値，あるいは連続した領域がウインドウ長さを超えるような場合には極大値となっていることが分かった．このため比較元のウインドウ位置 i の変化に対し，対象遺伝子のウインドウ位置 j がある程度近距離の部分が集まっている部分を一つの領域と考え，各領域中でウインドウ類似度が極大値を取るような位置 i をピーク位置とすることで，バインディングサイトを含むウインドウがピーク位置からもれてしまうのをこれまでの方法よりも防ぐことが可能になると考えられる．

具体的には，これまでは遺伝子制御領域全体に適用していたピーク位置の決定方法をこの連続した領域ごとに適用することにする．ここで比較元のウインドウ位置 i を 1 ずらした際に，連続していると判定する基準は，比較元のウインドウ位置 $i, i+1$ の各々に対する最大ウインドウ類似度を取る対象遺伝子位置を各々 j_i，j_{i+1} としたとき

$$|j_i - j_{i+1}| \leq 6$$

とする．但し, 連続した領域が極端に短い場合には，ピーク位置になっているケースは見当たらなかった. このため，領域長さが 2 以下の場合にはその領域は無視するものとする．次に各領域に対し，これまで制御領域全体に適用していたピーク位置の決定手順を適用する．以上の手順により選択されたウインドウ位置を新しいピーク位置と定める．

4.1.4 改良手法によるピーク位置の計算例

改良手法の効果の確認のため図 4.3 に比較元遺伝子 mcpC，比較対象遺伝子 lytD の場合の改良手法よるピーク位置の計算結果を示す．バインディングサイトを含むウインドウをピーク位置として正しく認識できていることがわかる. その他の遺伝子についても，元々旧手法で正しく認識されていたバインディングサイトについても変わらず認識できた. これは今回の改良手法は，比較元遺伝子のウインドウ位置 i の移動に対する対象遺伝子の位置 j を考慮した定義に変更されているが，i に対するウインドウ類似度の変化という点から考えた場合, 既存のピーク位置（遺伝子制御領域全体からみた極大値）に加えて，遺伝

表 4.1: 比較元の各位置 i に対し最大ウインドウ類似度をとる対象遺伝子の位置 j

比較元遺伝子			対象遺伝子			
遺伝子名	向き	位置 i	遺伝子名	向き	位置 j	ウインドウ類似度
mcpC	+	1462980	lytD	-	3686703	3944.3
mcpC	+	1462981	lytD	-	3686702	3489.2
mcpC	+	1462982	lytD	-	3686576	6118.7
mcpC	+	1462983	lytD	-	3686575	7484.0
mcpC	+	1462984	lytD	-	3686574	8394.3
mcpC	+	1462985	lytD	-	3686573	8849.4
mcpC	+	1462986	lytD	-	3686572	9304.5
mcpC	+	1462987	lytD	-	3686571	9759.6
mcpC	+	1462988	lytD	-	3686549	10990.1
mcpC	+	1462989	lytD	-	3686548	10990.1
mcpC	+	1462990	lytD	-	3686547	11141.8
mcpC	+	1462991	lytD	-	3686567	12035.2
mcpC	+	1462992	lytD	-	3686566	13400.5
mcpC	+	1462993	lytD	-	3686565	13855.6
mcpC	+	1462994	lytD	-	3686564	14007.3
mcpC	+	1462995	lytD	-	3686542	14327.6
mcpC	+	1462996	lytD	-	3686541	14327.6
mcpC	+	1462997	lytD	-	3686561	14007.3
mcpC	+	1462998	lytD	-	3686560	13855.6
mcpC	+	1462999	lytD	-	3686560	13805.0
mcpC	+	1463000	lytD	-	3686559	13805.0
mcpC	+	1463001	lytD	-	3686558	13956.7
mcpC	+	1463002	lytD	-	3686557	13653.3
mcpC	+	1463003	lytD	-	3686534	13569.1
mcpC	+	1463004	lytD	-	3686533	13552.2
mcpC	+	1463005	lytD	-	3686532	13501.6
mcpC	+	1463006	lytD	-	3686531	13501.6
mcpC	+	1463007	lytD	-	3686530	13366.8
mcpC	+	1463008	lytD	-	3686529	12928.5
mcpC	+	1463009	lytD	-	3686528	11563.2
mcpC	+	1463010	lytD	-	3686523	10456.3
mcpC	+	1463011	lytD	-	3686522	9091.0

子制御領域を細かく分割した領域中の最大値を新たにピーク位置に加えたことになるため当然の結果といえる．極端に短い連続した領域を除いたが，そのような部分にはバインディングサイトは含まれていなかったため，元々のピーク位置決定方法でピーク位置であったバインディングサイトへの影響は与えずにバインディングサイト候補の数を減少させることができた．

ピーク位置の数，つまりバインディングサイトの候補数という観点から考えてみると，新手法ではピーク位置を追加した形になるため，図 4.4 の制御領域全体における計算結果をみてみるとピーク位置の数が元の方法の場合よりもかなり増加している．これは元の方法では最大でもウインドウ幅 l 内で一つのピーク位置を定めていたが，改良手法では連続した領域長さがウインドウ長さに比べ短いものが多いため，結果としてピーク位置の総数が増加してしまった．

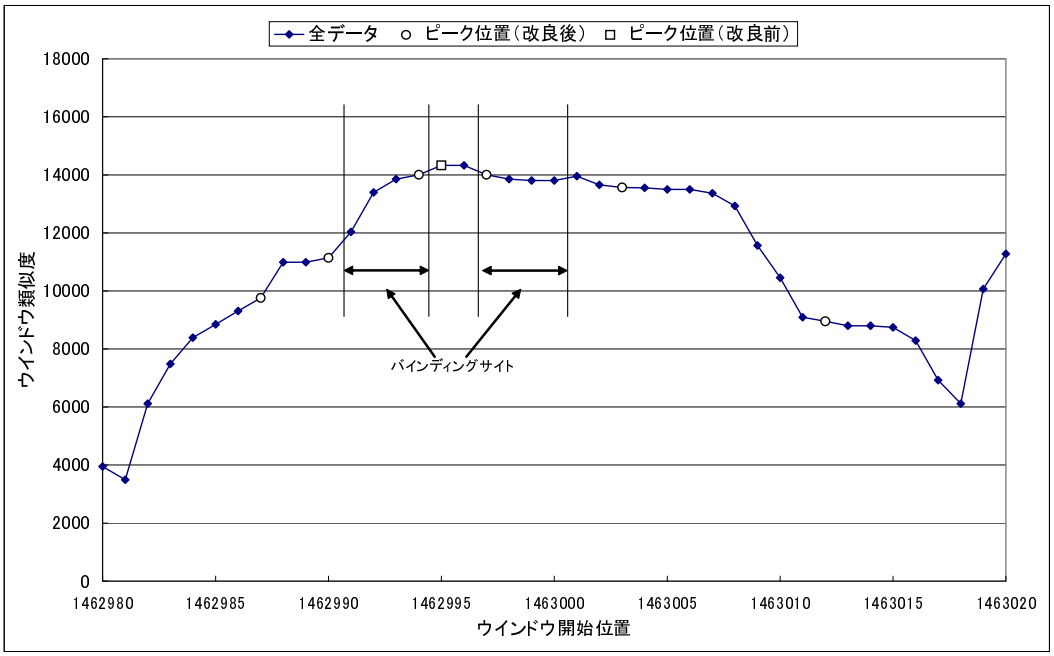

図 4.3: 改良手法によるピーク位置（比較元：mcpC 比較対象：lytD）

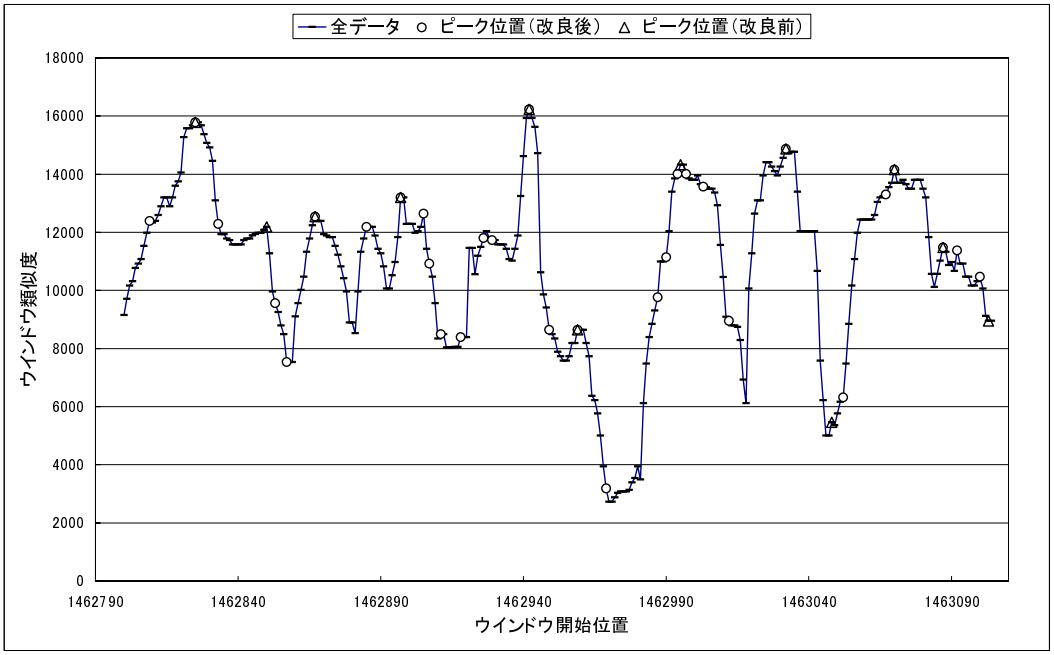

図 4.4: 改良前と改良後のピーク位置の比較（比較元：mcpC 比較対象：lytD）

4.2 グラフ構造による同一転写因子の被制御遺伝子候補の絞込

既存手法により得られたバインディングサイトの推定結果には，実際には関係のないウインドウ位置が多数含まれており，このままではバインディングサイトの特定が困難である．このため本章では既存手法により得られた遺伝子間の類似関係を用いて，候補遺伝子とそのバインディングサイト候補を絞り込む方法を提案する．

4.2.1 既存手法における被制御遺伝子とバインディングサイトの絞込みについて

既存手法では，大きく分けて二つの段階で同一の転写因子の被制御遺伝子群を推定している．ある遺伝子 a について考えると，第一段階としてマイクロアレイデータから得られた発現比のデータにおいて，遺伝子 a との相関が高い上位遺伝子群を抽出することで，同一の転写因子の被制御遺伝子の可能性が高い遺伝子をある程度絞り込む．第二段階として，第一段階で得られた候補遺伝子群の中から同一のバインディングサイトが結合する可能性が高いと考えられる類似した部分文字列を持つ遺伝子群へと，ウインドウ類似度を用いて部分文字列の比較により絞り込んでいる．これは転写制御因子は特異的文字列を認識して結合し，遺伝子の発現に影響を与えるという生物学的特長を利用して，同一の転写制御因子の被制御遺伝子を推定している．遺伝子の総数が 4000 個以上あるため，確率的に類似した文字列を含む転写制御領域を持った遺伝子は多数存在するはずであるが，マイクロアレイによる統計的性質により比較遺伝子数を減少させ偶然一致する数を減らしているのである．つまり，現在の手法では統計的性質と生物学的性質の異なる二つの性質を利用して同一転写因子の被制御遺伝子の推定を行っていることになる．

これら二つの絞込みについて考えてみると，どちらも比較元遺伝子から考えた類似性により，同一転写因子の被制御遺伝子とそのバインディングサイトを推定している．つまり第一段階ではある比較元遺伝子に対する発現傾向の相関が高い上位遺伝子を同一の転写制御因子の被制御遺伝子である可能性が高いとしているが，対象遺伝子側から考えた相関については考えていない．第二段階についても，比較元遺伝子の遺伝子制御領域の部分文字列と類似した文字列を遺伝子制御領域に含む対象遺伝子群を同一転写制御因子による被制御遺伝子としている．ウインドウ類似度の定義は，比較対象のウインドウ内の長さ 6 の文字列比較において，比較対象側で移動幅を導入しているため，同じウインドウ間のウインドウ類似度計算であってもどちらを比較元とするかにより値は変化することになる．つまり，既存手法では，どちらの段階の推定においても一方向から考えた類似性のみを利用しており，逆方向のデータは利用していない．比較元遺伝子から比較対象遺伝子への高い類似性をグラフとして表せば，比較元遺伝子を a とした場合，その関係を図 4.5 のようなグラフとして表すことができる．

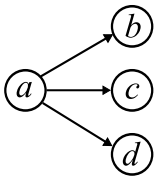

図 4.5: 比較元遺伝子側からの類似関係

4.2.2 改良方法

これまで利用していなかった双方向の類似性を考慮することで，推定結果の精度を高めることができると考えられる．先に述べたように，既存手法では一方向の情報しか利用していなかったが，同一転写制御因子の被制御遺伝子同士であるならば，どちらの側から考えても高い類似性をもつと考えられ，片方向のみの類似関係しか持たない遺伝子同士は，その可能性が低いと考えられる．例えば，図 4.5 で示した類似関係に対し，遺伝子 a から考えた対象遺伝子の側から考えた類似関係が図 4.6 のように得られたとするならば，双方

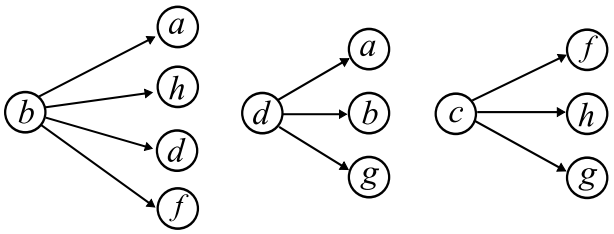

図 4.6: 対象遺伝子側からの類似関係

向に矢印がある場合を無効辺で表し，片方向の矢印を除去すれば，遺伝子 a については図 4.7 のようにその関係を表せる．つまり，遺伝子 a と同一の転写制御因子の被制御遺伝子である可能性の高い遺伝子は b，d であり，遺伝子 c は可能性が低いとして除外できる．

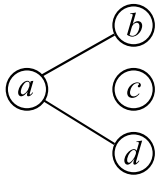

図 4.7: 双方向の類似関係

この相互に高い類似性を持つという関係を全ての遺伝子についてグラフとして表すことを考えた場合，同一の転写制御因子の影響を受ける遺伝子の推定に，新たな性質を利用できると考えられる．ある遺伝子と同一の転写制御因子の全ての被制御遺伝子間には辺

が存在すると考えられるので，全ての類似関係をグラフ上に表せば，同一の転写制御因子の影響を受ける遺伝子集合はクリーク（完全部分グラフ）を構成すると考えられる．例として図 4.7 の場合について考えると，遺伝子 a と同一の転写因子の被制御遺伝子である可能性がある遺伝子は b，d であり，相互の類似関係が各々図 4.8 のように得られたとする．これをまとめるとその関係は図 4.9 のようになり，最大クリークを構成する遺伝子集合が同一の転写因子の影響を受ける可能性が高いと考えるならば，$\{a, b, d\}$ が同一の転写制御因子の被制御遺伝子である可能性が高いと推定することができる．これまでの統計的性質と生物学的性質に加えて，同一の転写制御因子の被制御遺伝子が持つ類似関係のグラフ構造を判別に利用することで，推定精度の向上が図れると考えられる．

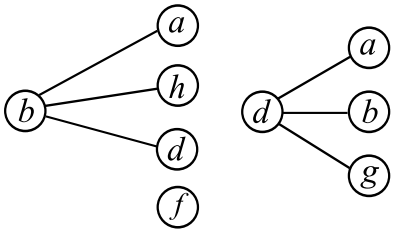

図 4.8: 対象遺伝子の類似関係

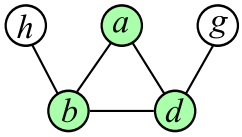

図 4.9: グラフ構造による判別

このように，従来手法により得られた遺伝子間の類似性を，グラフとして表しクリークを構成する遺伝子群を同一の転写因子の被制御遺伝子であると推定することで，候補遺伝子の絞込を行い推定精度の向上を図る．

ところで，従来手法における高い類似性は，最終的にピーク位置として表されている．ピーク位置は絞込みの第一段階で判別された遺伝子群の中で，さらに第二段階の転写制御領域における文字列の高い類似性を持つという条件を満たしたウインドウ位置である．このため各遺伝子の推定結果におけるピーク位置をグラフの頂点と考え，ピーク位置の総数を絞り込むことで最終的な推定結果の向上につなげる．つまり，各遺伝子間で相互にピーク位置であるという推定結果が得られた部分に限り，グラフ上での辺が存在すると定義することで，第一，第二段階両方の相互類似性を評価することが可能になる．但し，各ピーク位置の相互類似性を評価するにあたり，完全なピーク位置の一致を用いることはできない．ウインドウ内に特異的に類似した文字列が存在する場合，ウインドウが多少前後してもそのウインドウ類似度は大きく変化しない，あるいは変化しないため，図 4.10 の

ようにどちらを比較元にするかによって多少異なる位置をピーク位置として選択する可能性があるからである．

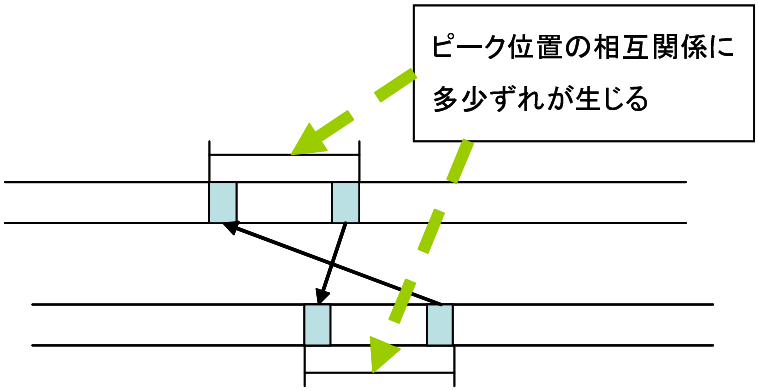

図 4.10: ピーク位置の相互関係

クリークを用いてバインディングサイトの可能性の高いピーク位置かどうかを判別するために，何らかの基準を定めなければならないが，考慮しなければならないことがある．一つ目はバインディングサイトとは関係がない文字列であっても，偶然高い部分文字列を含む転写制御領域を持った遺伝子が複数あれば，クリークを構成してしまうということ，つまり，最大クリークを構成する部分が必ずしも被制御遺伝子の可能性が高いとはいえないということである．二つ目は，一つの転写因子の被制御遺伝子の数は一つだけのものから数十種の遺伝子の発現に影響を与えるものまでさまざまであるということである．既存手法では一つの比較元遺伝子と類似性の高い遺伝子を同一転写因子の被制御遺伝子と推定していたため，最低でも，同一転写因子の被制御遺伝子数が 2 以上あれば理論上は推定可能である．しかしながら，クリークを構成する頂点数，つまり遺伝子数に制限を設けてピーク位置の絞込みを行うならば，その頂点数以下の被制御遺伝子しかもたない転写制御因子の場合は，推定不可能ということになってしまう．

以上のことを踏まえたうえで，本研究ではクリークを構成する頂点数の最低値を定め，最低頂点数以上の頂点からなるクリークを構成する遺伝子を，バインディングサイトである可能性の高いピーク位置として判別することにする．既知の転写因子の被制御遺伝子に関していえば，被制御因子数が数十種となっているのは，Sigma factor と呼ばれる特殊な因子が殆どで，Sigma factor に関しては，バインディングサイトの位置が遺伝子配列との相対位置でほぼ決定していることが知られており，位置を元に推定することが可能である．それら以外の転写因子の被制御遺伝子は大部分は 2〜3 個程度のものが多い．また，クリークを利用することで判別した場合，最低頂点数以下の被制御遺伝子しか持たない転写因子の場合ピーク位置を除外してしまい推定できなくなっていしまう．このため，本研究ではクリークを構成する最低頂点数を 3 と固定する．これは最低頂点数を 2 とした場合，クリークによる判別の効果があまり得られないと考えられることと，できるかぎりバインディングサイトの含むピーク位置を除外しないようにするためには最低頂点数は低い

ほうが望ましいと考え決定した．

次に，クリークを構成するとみなす条件を定義する．理想的にはバインディングサイトを含むウインドウ位置は一致する，つまり，ピーク位置が一点に集中することであるが，実際には各遺伝子に対するピーク位置にはずれが生じる．このためある程度のずれを許容した定義を与えなければならない．ここで遺伝子 x を比較元，遺伝子 y を比較対象としたとき，遺伝子 x のピーク位置を m，ピーク位置に対応する遺伝子 y のウインドウ位置を n としたとき，遺伝子 y のピーク位置を m_y，対応する遺伝子 x のウインドウ位置を n'_x と書くことにする．三つの遺伝子 a，b，d のピーク位置がと各々に対応する対象遺伝子側の位置が図 4.11 に示した関係にあるとする，このとき以下の 2 つの条件を満足するとき，かつそのときに限り 3 頂点のクリークを構成していると定義する．

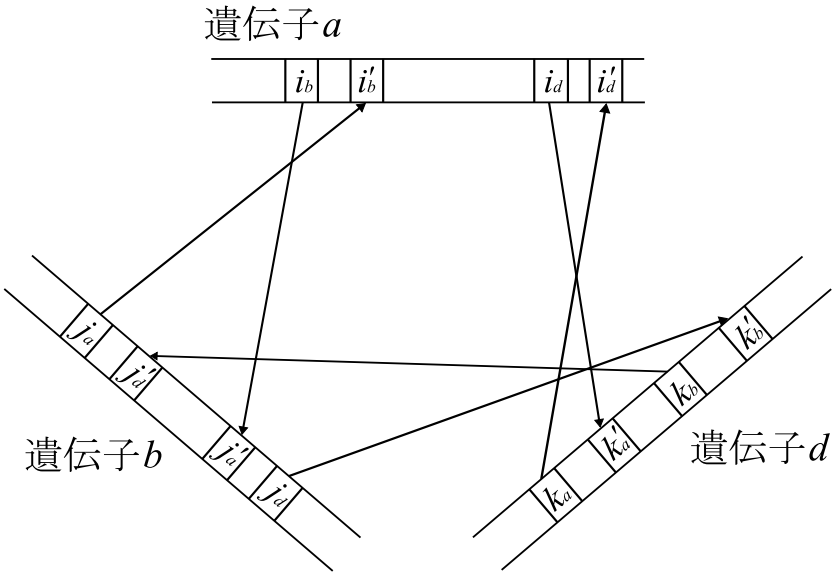

図 4.11: クリークを構成するピーク位置の関係

- 相互にピーク位置である条件

$$| i_b - i'_b |, | i_d - i'_d |, | j_a - j'_a |, | j_d - j'_d |, | k_a - k'_a |, | k_b - k'_b | \leq l - 6$$

  但し，l はウィンドウ長さ．

- 3 頂点間でクリークを構成する条件

$$| i_b - i_d |, | j_a - j_d |, | k_a - k_b | \leq l - 6$$

各ピーク位置の最大のずれ幅は，最低でも長さ 6 の文字列が一致することを条件として設定した．これ以上短い部分が重なっていても同じ位置の文字列によって相互に高い類似性を持っているとはいえないと考えられる．

4.2.3　クリークを用いたピーク位置の絞込みの効果

既存手法で得られた推定結果と，3 頂点以上からなるクリークを構成するピーク位置という条件を判別に用いた場合について，比較を行い効果を確認するため実験を行う．実験対象として，転写因子及びバインディングサイトが既知であり，バインディングサイトの長さの異なる遺伝子 lytD，katA を使用する．

各遺伝子の推定結果におけるピーク位置の頻度分布を図 4.12，図 4.13 に示す．図 4.12 は転写制御因子 perR，比較元遺伝子 katA の頻度分布であるが，領域によりばらつきはあるものの全領域で減少しており，全体としてピーク位置の総数を 1/3 程度にまで絞り込むことができた．図 4.13 は転写制御因子 sigD，比較元遺伝子 lytD の場合のピーク位置頻度分布であるが同様の結果が得られている．どちらの場合も，バインディングサイトを含むピーク位置を除外することなく，実際にはバインディングサイトでないピーク位置を減少させることができた．つまり，従来手法と比較してバインディングサイトを約 1/3 にまで絞り込めたことになり，推定結果の精度を既存手法よりも高めることができた．

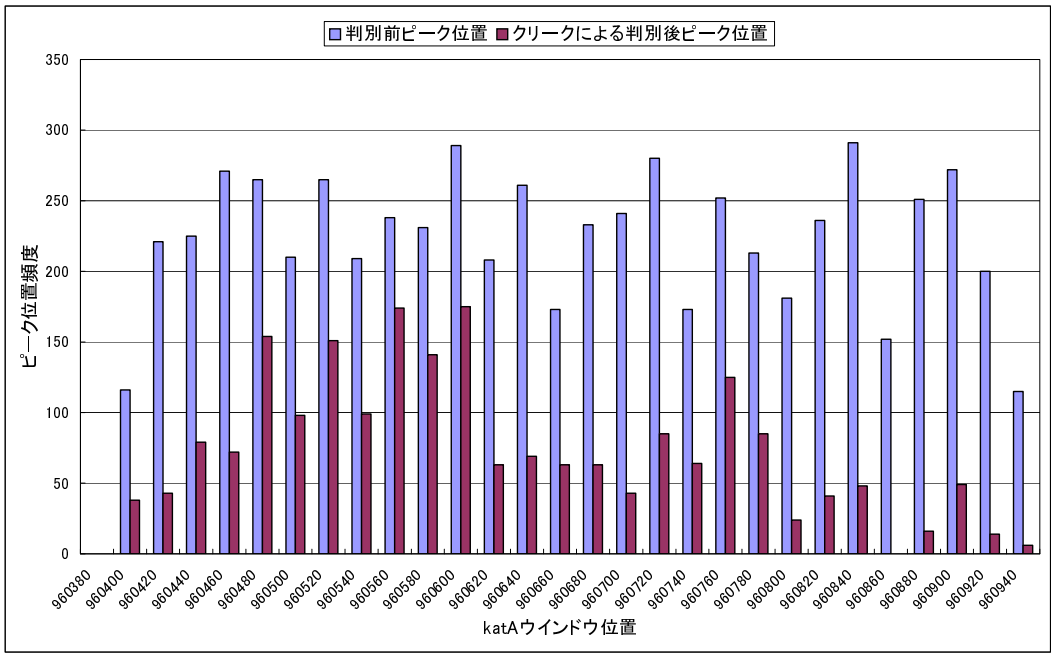

図 4.12: クリークによる判別の効果 (比較元：katA)

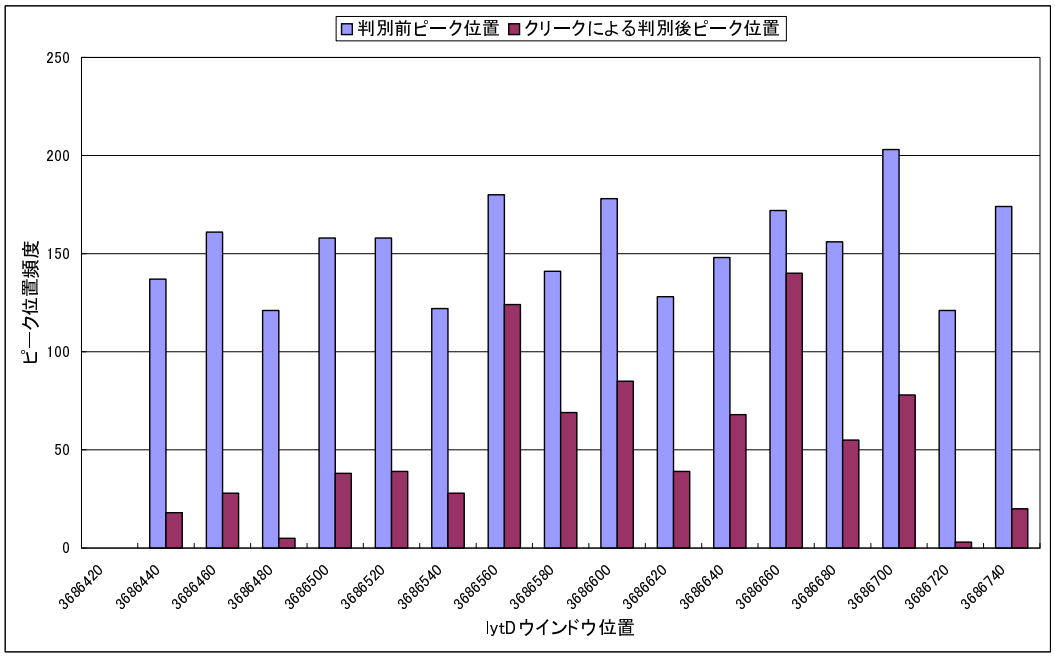

図 4.13: クリークによる判別の効果 (比較元：lytD)

4.3 バインディングサイトの特定方法の提案

バインディングサイトが短い被制御遺伝子についても推定可能にするために，ピーク位置として採用する基準となるウインドウ類似度閾値を，15000 から 8000 にまで引き下げたことにより, ピーク位置の総数が増加したため，既存手法で用いていたピーク位置の連続した領域をバインディングサイトの可能性が高い領域する方法は利用できなくなった. このため，本章では現在の推定方法で得られたピーク位置の中から，バインディングサイトの可能性が高い位置を絞り込む方法について考える.

4.3.1 バインディングサイトの可能性の高いピーク位置

現在の手法は，バインディングサイトの可能性の高いウインドウのみを推定するのではなく，可能性のあるウインドウ位置を極力排除しないように，判別のための一つ一つの条件は厳しくせず複数の異なる性質を用いることで推定結果の絞込みを行っている．現段得られている推定結果では，ピーク位置の総数が非常に多く，そのままでは推定結果のグラフから特定の位置がバインディングサイトであるとは判断できない場合が多い．このため，バインディングサイトの可能性が高いピーク位置を判別する方法が必要である．生物学的な実験によりバインディングサイトを特定するにしても，推定結果として得られたバインディングサイト候補の中で，可能性が高い順序付けが行われている必要がある.

順位付けに利用できる情報としてウインドウ類似度があるが，単純にウインドウ類似度が高い順にバインディングサイトの可能性が高いのかというと，推定結果を調査した結果では必ずしもそうではないことが分かった．例えば，転写因子が sigD の被制御遺伝子のように，バインディングサイトが 6〜7 文字程度と短い場合には，ウインドウ類似度の値が 8000〜11000 程度と，推定結果として得られたピーク位置全体のウインドウ類似度の分布から考えるとそれほど高くない．つまり，バインディングサイトの長さによってウインドウ類似度の値はほぼ決定されるが, これらの長さは各転写因子ごとに固有の長さであり，ウインドウ類似度の値も必ずしも全体の中で，高い値をとるとは限らないのである. このように，単純にウインドウ類似度の値のみでバインディングサイトの可能性を測ることは困難である．このため別の方法を考える必要がある.

4.3.2 ランダム配列との比較によるウインドウ類似度の特異性による判別

推定結果について精査していくと他の部分に比べ，明らかに高いウインドウ類似度を持つ位置が見られた．例として比較元遺伝子を katA とした場合の推定結果を図 4.14 に示す．図 4.14 において四角で囲まれた部分に含まれるピーク位置は明らかに周囲よりも高いウインドウ類似度である.

転写制御領域に含まれる文字列の出現頻度に関する調査により，長さ 6 の文字列は全体の 9 割に存在しているが，長さ 8，10 と長くなるに従い，その出現頻度は非常に少なくな

る [3]．このため著しく高いウインドウ類似度を持つピーク位置はバインディングサイトが含まれている可能性が高いと考えられる．

本研究では，ウインドウ類似度が特異的であるかどうかを判別するために，各塩基の出現する期待値を元にランダムに生成した遺伝子制御領域の塩基配列を用いた推定結果と比較を行う．各塩基の出現確率を基にしたランダム配列とオリジナルの推定結果を比較することで，出現する可能性の低い文字列が一致する場合に限り，特異性が高いという結果を得られると考えられる．また，表 4.2 をみるとわかるように，全塩基配列と制御領域では各塩基の出現頻度が異なるため，転写制御領域における各塩基の出現確率に基づき生成したランダム配列を用いて比較を行う．

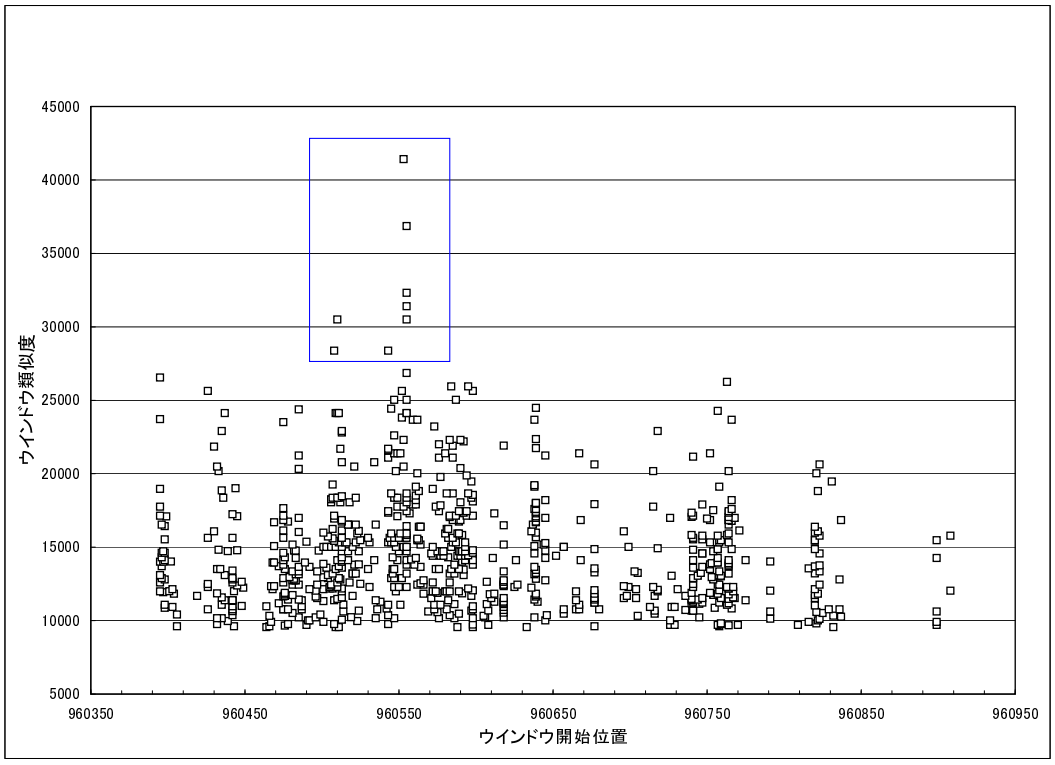

図 4.14: 特異的に高いウインドウ類似度を持つピーク位置（比較元：katA）

表 4.2: 各塩基の出現確率

塩基名	全塩基配列	転写制御領域
A	0.281805318099446	0.319810040907179
T	0.28301770849	0.286518164541068
G	0.217096887312228	0.213776789150459
C	0.218080086096326	0.179895005401293

4.3.3 実験方法

本実験の目的は，推定結果の中から各塩基の出現確率基づきランダムに生成した配列を用いた推定結果と，オリジナルの推定結果の比較によりウインドウ類似度の特異性を発見することで，バインディングサイトの可能性の高い部分を発見することが可能であるかどうかを確認することである．

オリジナルの推定結果との比較に用いるランダム配列を用いた推定方法には以下のように行う．

1. 遺伝子 a と発現強度の相関が高い上位 300 遺伝子 B を求める．
2. B の各遺伝子の転写制御領域と同じ長さのランダム配列を表 4.2 の出現確率に基づいて生成．
3. 遺伝子 a の制御配列は実際の配列，その他の遺伝子の制御配列はランダム配列を用い通常の推定を行う．
4. ランダム配列を用いた推定結果と通常の推定結果を比較．

ランダム配列を用いた推定結果とオリジナル配列の推定結果を比較し，ランダム配列による推定結果と比較して，明らかに高いウインドウ類似度を持つピーク位置を判別する．この明らかに高いウインドウ類似度をとる部分にピーク位置が含まれていれば，ランダム配列の推定結果と比較して高いウインドウ類似度を持つピーク位置はバインディングサイトとなる可能性が高いと考えることができる．なお実験対象として，転写因子の異なる三つの遺伝子 katA，lytD，kyuN について実験を行う．

4.3.4 実験結果

各遺伝子についての実験結果を図 4.15，図 4.16，図 4.17 に示す．なお実験におけるウインドウ長さは 20 とした． 比較元の遺伝子が katA の場合，図 4.15 の実験結果をみるとランダム配列による推定結果と比較して高いウインドウ類似度をとるピーク位置がバインディングサイトとなっていることがわかる．これにより，期待した通り出現可能性の低い長い文字列が一致する場合には，バインディングサイトである可能性が高いと考えられる．しかしながら，バインディングサイトが短い転写因子 sigD の被制御遺伝子 lytD の場合について比較結果の図 4.16 を見てみると，バインディングサイトの部分はランダム配列による推定結果と比較して特異性を持っているとは考えにくい．これはバインディングサイトの長さが 6〜7 文字程度と短い場合には出現頻度が非常に高く，ランダムに生成した転写制御配列において類似した文字列が多数存在するためだと考えられる．次に図 4.17 の比較元遺伝子が ykuN，転写因子が fur の場合であるが，バインディングサイト長さは 15〜20 程度とかなり長いが，被制御遺伝子ごとに若干の差異があるため 10 文字以上の完全な一致という部分はなく，6 文字程度の類似文字列がウインドウ内にいくつか存在する

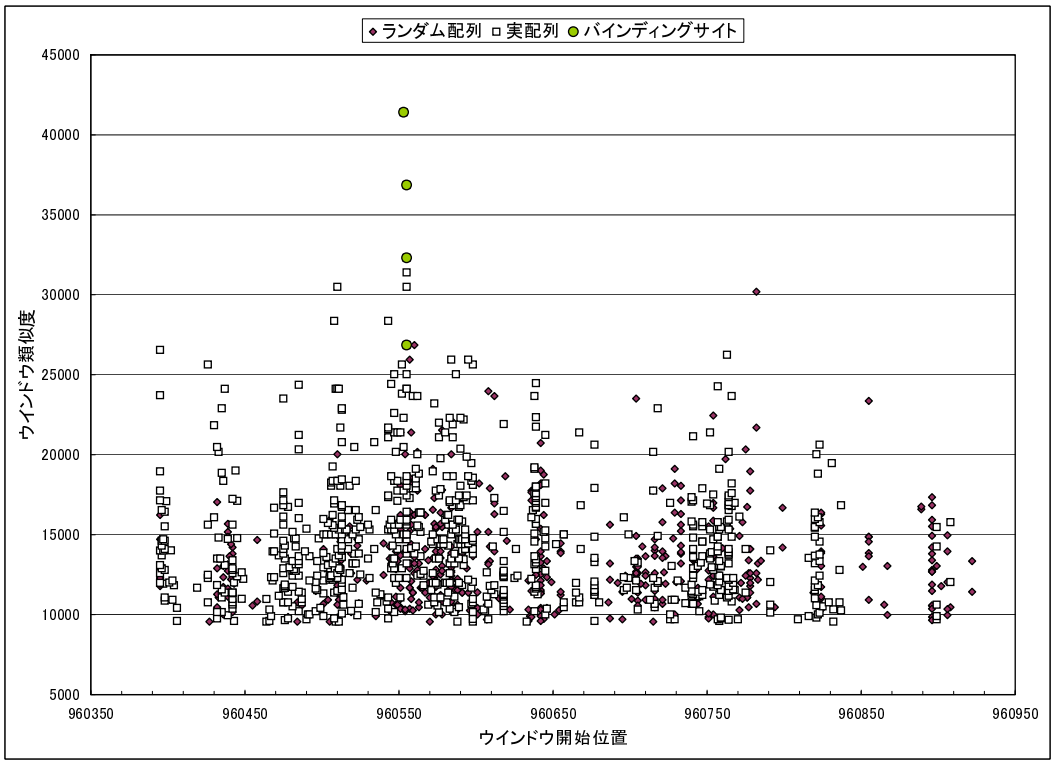

図 4.15: ランダム配列との比較（比較元：katA）

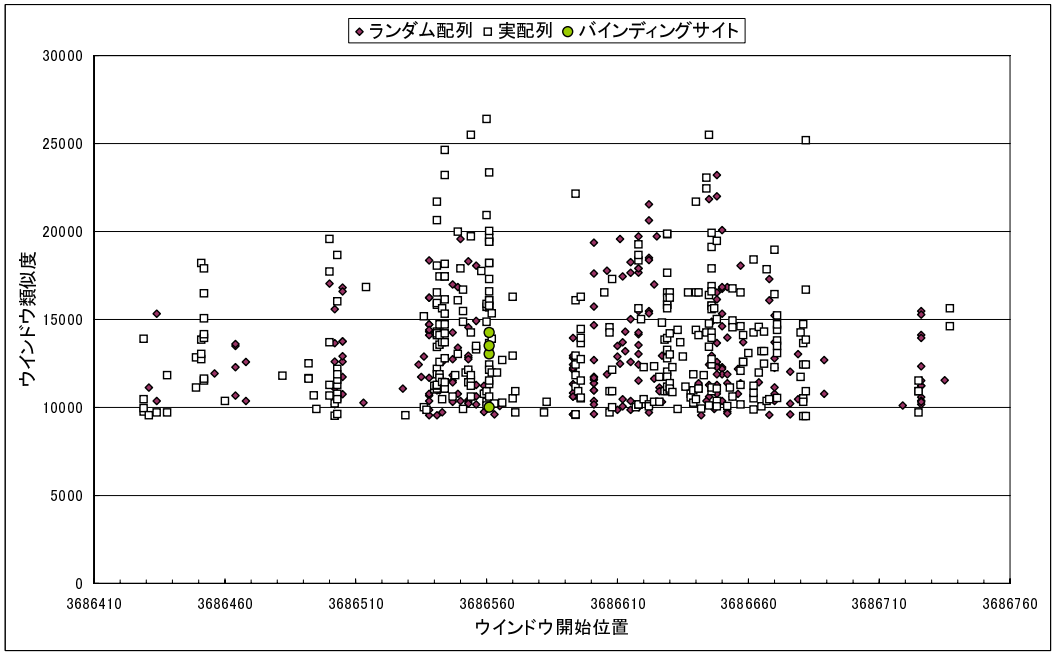

図 4.16: ランダム配列との比較（比較元：lytD）

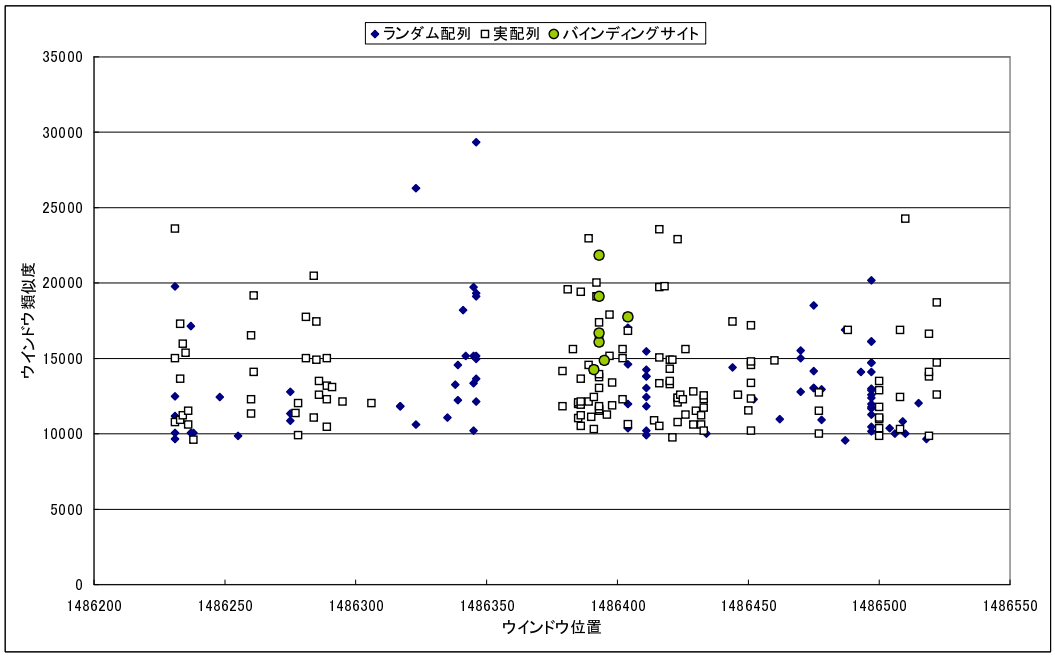

図 4.17: ランダム配列との比較（比較元：kyuN）

という状態になっており，lytD の場合ほどバインディングサイトにおけるウインドウ類似度の値は高くなく，ランダム配列による推定結果と比較しても特異性があるとは考えられない．

以上の結果からランダム配列による推定結果との比較によりバインディングサイトが長い被制御遺伝子については，ピーク位置の中から可能性が高い位置をある程度絞り込むことが可能であることが明らかになった．しかしながら，バインディングサイトの長さが短いものや，バインディングサイトが連続した文字列ではなく少しはなれて二箇所にある場合などは，この方法では，特定することはできない．

第5章　各種パラメータの設定

本研究の推定方法では，ウインドウ長さなど設定しなければならないパラメータが複数存在する．しかしながら，これまで十分な調査を行っておらず，これらのパラメータが推定結果にどのように影響するのか，また，全体の推定を行う上で，現在の設定値が最適であるのか不明である．このため本章では，各種パラメータの影響と推定に最適と思われる値の調査を行う．

5.1　ウインドウ類似度における部分文字列の移動幅の最適化

5.1.1　移動幅について

ウインドウ類似度は，ウインドウ内における各位置からの長さ 6 の文字列同士の比較を行い，各部分文字列における類似度の合計をウインドウ類似度としている．. ただし，定義のように比較対象側の文字列の位置は前後にある程度の移動を許し，その中で一致する文字数の最も多い位置の文字列と比較することでウインドウ類似度を求めている．同一の転写因子の結合領域であったとしても，多少の文字のずれ，置換，削除，挿入などがある場合がある．また，バインディングサイトが離れて二箇所に存在する場合に，被制御遺伝子によってバインディングサイトの間隔が多少異なる場合がある．これらのケースにではウインドウの各位置における文字を比較した場合，一部分の文字列にしか一致することはない．図 5.1 の例のように部分文字列のずれを許容した場合とそうでない場合では，ウインドウ間の類似度に大きな差が出ることになる．この部分文字列の比較における移動幅を設定することで，遺伝子毎に部分的に異なるバインディングサイト部分に関してもウインドウ間の類似性を評価可能にしている．この部分文字列のずれを許容する最大値を移動幅 Δ と呼ぶ.

5.1.2　移動幅とウインドウ類似度の関係

部分文字列の比較において多少の移動を許容した理由である文字の挿入や削除は実際には 1，2 文字程度のことが多い．これまでは，全ての文字の挿入や削除に対応することを考え，許容する移動幅を $\Delta = 3$ と余裕をとって大きく設定し推定を行っていた．しかしながら，移動幅 Δ に対し比較される位置の数は，図 5.2 に示すように $2\Delta + 1$ 箇所となり，$\Delta = 3$ の場合には全部で 7 箇所と比較されることになる．Δ を大きくとるということ

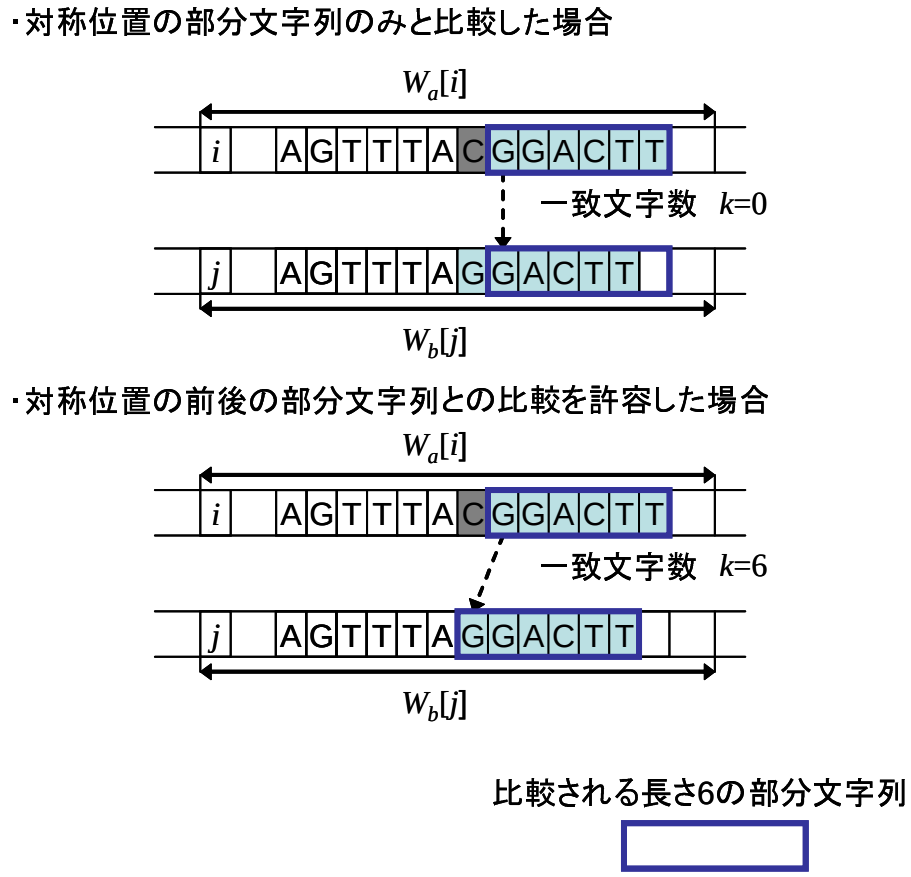

図 5.1: 部分文字列の前後への移動による一致

は，本来一致すべきでない位置の文字列同士で高い類似度をとる可能性が高くなると考えらる．バインディングサイトとは関係のないウインドウ間のウインドウ類似度が高くなり，閾値 8000 を超える部分が不必要に増えるようであれば推定を困難にするため望ましくない．

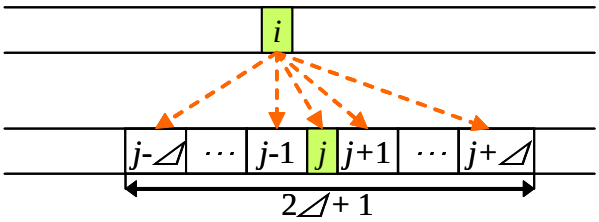

図 5.2: 移動幅 Δ の場合の比較対象文字列

これらの理由により，移動幅の変化に対し高いウインドウ類似度をとる位置が極端に増加するようであれば，移動幅を実際の 1，2 程度に設定するほうが，より現実に則した比較となり，バインディングサイトにおける文字列の類似性も正確に評価が可能になると考えられる．このため，移動幅によるウインドウ類似度の分布の変化と，バインディングサイトにおけるウインドウ類似度を高くする効果のバランスを考えて最適な移動幅について決定する必要がある．

5.1.3 移動幅の変化に対するピーク位置候補の増加数

移動幅の変更によりウインドウ類似度が高い位置が増加する．このウインドウ類似度の高さが推定結果に影響を与える場合として，まず考えられるのは，元々はピーク位置の閾値にウインドウ類似度が達していなかった部分が，移動幅によるウインドウ類似度の増加により，ピーク位置候補となる場合である．ここでは，配列における文字の挿入や削除は 1 または 2 文字程度であることを考慮して，$\Delta = 1, 2, 3$ とした場合と，$\Delta = 0$ ，つまり，部分文字列を許容しなかった場合について比較を行い．移動幅の変化によりどの程度ピーク位置候補が増加しているのか調査する．遺伝子 ykvW，cheV，lytD，mcpC，mrgA，katA について，各遺伝子 x と相関の高い上位遺伝子 300 を B としたとき，遺伝子 $y \in B$ に対する最大ウインドウ類似度 $Max_{xy}[i]$ が閾値 8000 を越える位置について，$\Delta = 0$ と比較した場合の増加数の合計を図 5.3 に示す．

各遺伝子の制御領域長さが異なるため増加数自体は差があるものの，増加割合は各遺伝子でほぼ一致している．$\Delta = 3$ の場合の増加数を見みると全ての遺伝子で，増加数が 10000 を超えている．グラフは 300 遺伝子に対する調査の合計を取っているので，平均を取ると，一つの遺伝子に対し最低でも 30 箇所が移動幅の影響により新たなピーク位置の候補となっていることになる．しかしながら，これらの大部分は実際にはバインディング

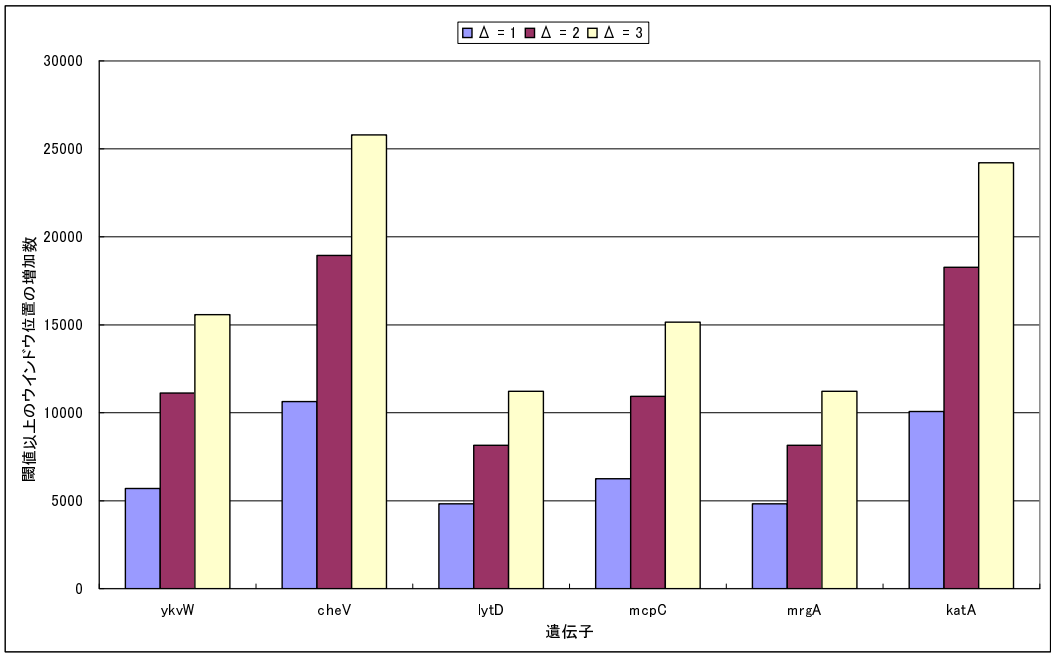

図 5.3: 移動幅 Δ における閾値 8000 以上のウインドウ増加数

サイトとはまったく関係のないウインドウであると考えられ，ピーク位置の決定への影響も少なくないと考えられる．

5.1.4 各移動幅おけるウインドウ類似度の増加幅

次に各ウインドウにおける移動幅の変化によるウインドウ類似度の上昇幅について調査する．遺伝子 ykvW，cheV，lytD，mcpC，mrgA，katA について，$\Delta = 1, 2, 3$ とした場合と，$\Delta = 0$ の場合の各位置におけるウインドウ類似度の差の分布を表 5.1 〜表 5.6 に示す．移動幅を増すごとに，表 5.7 に示した各移動幅における平均値よりも高い増加幅をとる位置が非常に多くなっている．ピーク位置の閾値が 8000 であることを考えると，増加幅が非常に大きい場合は問題である．特に $\Delta = 3$ の場合については，増加幅が 3000 以上の数が全ての遺伝子において 5000 箇所を超えており，推定を困難にする一因になっているものと考えられる．

5.1.5 推定に最適な移動幅

以上の結果から部分文字列の移動幅はできる限り短いほうが良いと考えられる．ただし，文字の挿入や削除が 2 文字程度起きる場合も多いので，

$$\Delta = 2$$

表 5.1: ウインドウ類似度変化（ykvW）

分布範囲	$\Delta = 1$	$\Delta = 2$	$\Delta = 3$
0 - 499	87056	57889	42942
500 - 999	20531	27145	26135
1000 - 1500	8078	15183	17475
1500 - 1999	3917	9022	12025
2000 - 2499	2066	5561	7810
2500 - 2999	1135	3396	5542
3000 - 3499	644	2196	3698
3500 - 3999	313	1351	2551
4000 - 4499	246	887	1743
4500 以上	398	1754	4463

表 5.2: ウインドウ類似度変化（cheV）

分布範囲	$\Delta = 1$	$\Delta = 2$	$\Delta = 3$
0 - 499	108795	66745	45695
500 - 999	33865	42333	37806
1000 - 1500	14157	25160	29486
1500 - 1999	7409	15173	20706
2000 - 2499	3999	9290	13978
2500 - 2999	2255	5584	9052
3000 - 3499	1245	3601	6092
3500 - 3999	659	2013	3860
4000 - 4499	435	1258	2186
4500 以上	900	2562	4858

表 5.3: ウインドウ類似度変化（lytD）

分布範囲	$\Delta = 1$	$\Delta = 2$	$\Delta = 3$
0 - 499	61705	42664	30251
500 - 999	16354	21039	20049
1000 - 1500	6738	11910	13701
1500 - 1999	3183	6553	9467
2000 - 2499	1769	3767	6326
2500 - 2999	1032	2379	4241
3000 - 3499	618	1419	2851
3500 - 3999	374	906	1844
4000 - 4499	219	569	1162
4500 以上	399	1185	2499

表 5.4: ウインドウ類似度変化（mcpC）

分布範囲	$\Delta = 1$	$\Delta = 2$	$\Delta = 3$
0 - 499	58014	33434	21385
500 - 999	19156	23915	20092
1000 - 1500	8040	15133	17237
1500 - 1999	3983	8982	12780
2000 - 2499	1964	5031	8111
2500 - 2999	1175	3003	5283
3000 - 3499	714	1687	3314
3500 - 3999	428	1120	2067
4000 - 4499	235	616	1310
4500 以上	476	1264	2606

表 5.5: ウインドウ類似度変化（mrgA）

分布範囲	$\Delta = 1$	$\Delta = 2$	$\Delta = 3$
0 - 499	51158	32680	21936
500 - 999	15099	18363	17249
1000 - 1500	6370	10705	11901
1500 - 1999	3220	6345	8899
2000 - 2499	1739	4020	5961
2500 - 2999	1134	2731	3978
3000 - 3499	664	1835	2764
3500 - 3999	358	1173	2037
4000 - 4499	257	813	1497
4500 以上	731	2065	4508

表 5.6: ウインドウ類似度変化（katA）

分布範囲	$\Delta = 1$	$\Delta = 2$	$\Delta = 3$
0 - 499	95278	55414	38856
500 - 999	32403	38277	33854
1000 - 1500	14133	24340	26431
1500 - 1999	7539	15061	18983
2000 - 2499	3967	9560	13119
2500 - 2999	2184	5783	8690
3000 - 3499	1321	3842	6115
3500 - 3999	774	2229	4196
4000 - 4499	467	1476	2772
4500 以上	1002	3086	6052

表 5.7: 移動幅によるウインドウ類似度変化の平均値

遺伝子名	データ数	$\Delta = 1$	$\Delta = 2$	$\Delta = 3$
lytD	92391	524.68	882.87	1249.77
mrgA	80730	602.56	1077.67	1540.75
katA	159068	631.78	1120.42	1480.29
mcpC	94185	591.15	1037.83	1429.50
ykvW	124384	473.33	904.32	1273.11
cheV	173719	585.55	1022.91	1371.92

とするのが，推定全体のバランスを考えるとよいと考えられる．

実際のバインディングサイト付近のウインドウ類似度と移動幅の関係について図 5.4 を見みると，移動幅の効果により，バインディングサイト付近のウインドウ類似度が非常に高くなっていることがわかる．この場合は移動幅により一致する文字数が増加したと考えられる．移動幅 2 の場合と移動幅 1 の間で大きな差がでているのは文字のずれが 2 であることを示している．バインディングサイトの前後における移動幅によるウインドウ類似度の変化はバインディングイトにおけるウインドウ類似度の上昇分に比べ十分小さいことから，結果としてバインディングサイトのウインドウ類似度を特異的にしていることが確認できる．

移動幅を導入したことにより，ウインドウ類似度の分布が全体的に上昇してしまうが，移動幅 2 ならば表 5.7 をみると平均の上昇幅は 1000 前後であり，通常のバインディングサイトの推定に対する影響も最小限に抑えられると考えられる．また，$\Delta = 3$ の場合には 7 箇所と比較していたが，$\Delta = 2$ ならば比較文字列は 5 箇所と減少するため，計算時間の短縮にもつながる．

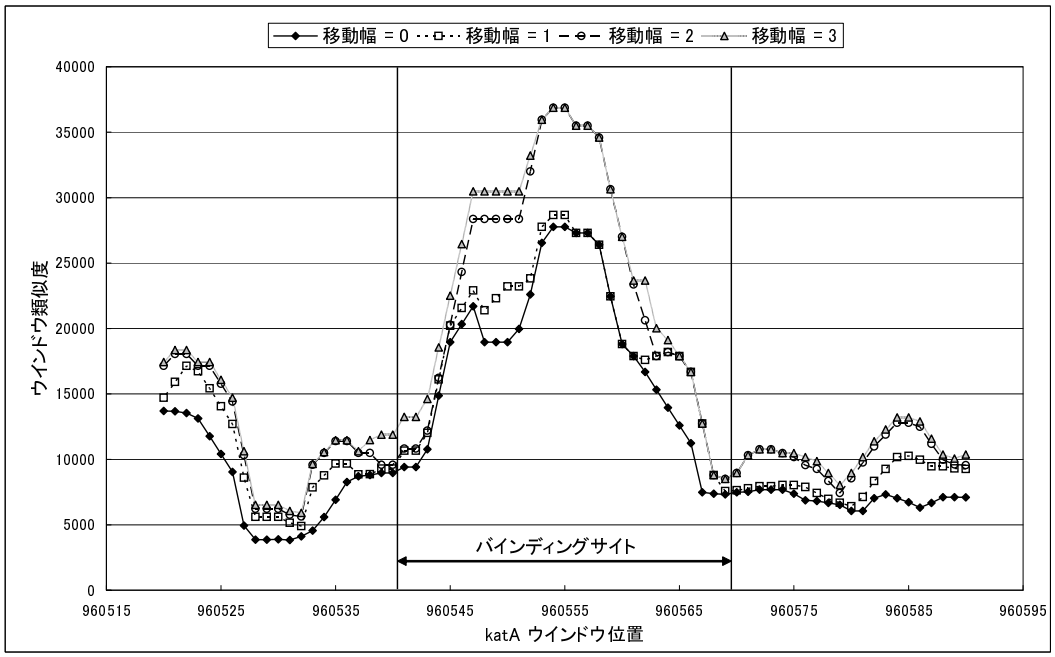

図 5.4: バインディングサイトにおける移動幅の効果（比較元：katA 比較対象：mrgA）

5.2 発現強度の相関による同一転写因子の被制御遺伝子の絞込みについて

既存手法では，全遺伝子数約 4000 個の中で，比較元の遺伝子 a と発現強度の相関の絶対値が高い上位 300 遺伝子を遺伝子 a の制御因子 α の被制御遺伝子候補としている．しかしながら，推定結果の中で上位 300 個の中に含まれてない遺伝子が確認されており，この絞込みの方法の効果の検証と最適な候補の数を調査により明らかにする必要がある．また現在は利用していないが，DNA マイクロアレイデータによる遺伝子発現強度のデータも同一転写因子が各遺伝子に与える影響の大きさを表すため，同一転写因子の被制御遺伝子候補の絞込に利用可能であると考えられる．よって，ここで合わせて調査する．

5.2.1 相関係数絶対値と相関係数順位

現在の方法では，全遺伝子の中からマイクロアレイデータにおいて相関係数の絶対値の高い順に上位 300 を初期の候補として選択している．この方法による効果と妥当性を調査するため，DNA マイクロアレイデータにおいて破壊株として使用した 106 個の遺伝子の中で，既知の被制御遺伝子が存在する 47 個について制御因子として，各制御因子に対し既知の被制御遺伝子を一つ選択し相関係数絶対値および相関係数絶対値順位について調査を行った．

同一転写因子における被制御遺伝子の間の相関係数の頻度分布は図 5.5 のようになった．このグラフから同一の転写因子の被制御遺伝子であっても必ずしも高い相関係数をとるわけではないとわかる．むしろ相関係数の絶対値が低いほうに多く分布しており，閾値を設定して判別に利用することは困難である．次に図 5.6 に相関係数順位分布の調査結果を示す．グラフから相関係数の絶対値順位が 400 程度に集中的に分布していることが見て取れる．しかしながら，400 番目までに入っている被制御遺伝子は 40 %程度にすぎず，それ以下の順位においても平均的に分布してしまっている．これはマイクロアレイが誤差が大きくでてしまう実験方法であるということと，全遺伝子数が 4000 個以上もあるのに対し，マイクロアレイデータの数が 106 と非常に少ないため，データの信頼性があまり高くないことも影響していると考えられる．現在は相関係数上位 300 程度としているが，調査結果から上位 400 程度の遺伝子に絞り込むのが望ましいと考えられる．相関係数順位の分布については，マイクロアレイ実験装置の進歩により，精度が向上していることや実験データの数を増やすことによって変動することが考えられるため，データにあった閾値を設定することが不可欠である．

5.2.2 発現強度分布

相関係数は発現傾向の類似性を評価しているが，その元となっている発現比についても同一の転写因子の被制御遺伝子については大きくその値が変化するため，利用可能と考え

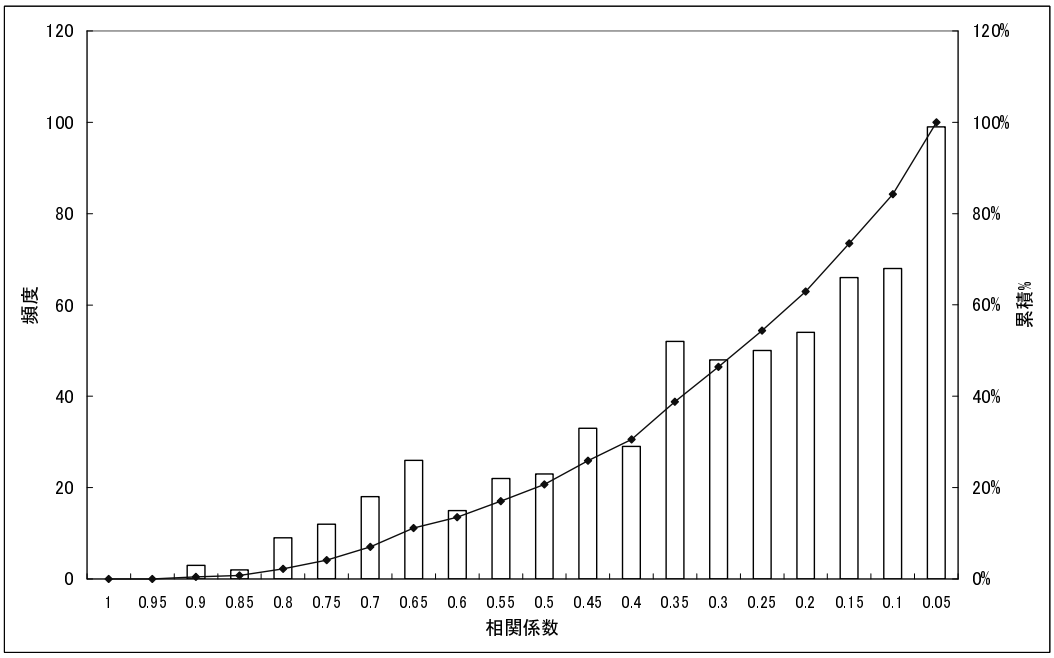

図 5.5: 相関係数絶対値分布

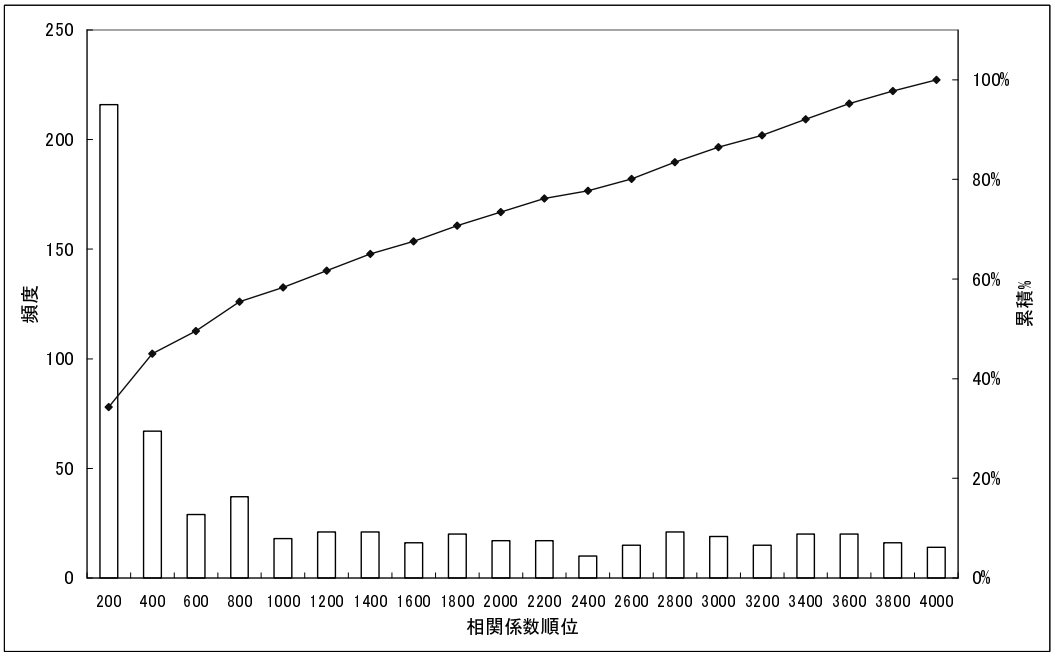

図 5.6: 相関係数順位分布

られる．ただし，相関係数が全データを通しての発現傾向の類似性を表すのに対し，特定の遺伝子を破壊した場合の発現量の変化を表すので，破壊株が転写因子でなかったとしても間接的な影響により発現量が変化しているのかどうか判別できないことは注意すべきである．しかしながら，特定の遺伝子の影響の強さを測定しているという点では，相関係数とは異なる情報源として利用可能であると考えられる．

各制御因子を破壊したマイクロアレイにおける被制御遺伝子の発現比の絶対値分布（図 5.7）をみると，高い値をとるとは限らず広く分布している．また値が低いほうに多く分布しており，値により被制御遺伝子の可能性が高いかどうか判断することは難しい．これに対し，発現比の順位分布（図 5.8）をみると相関係数の場合と同様に上位 400 個程度の部分に集中していることがわかる．

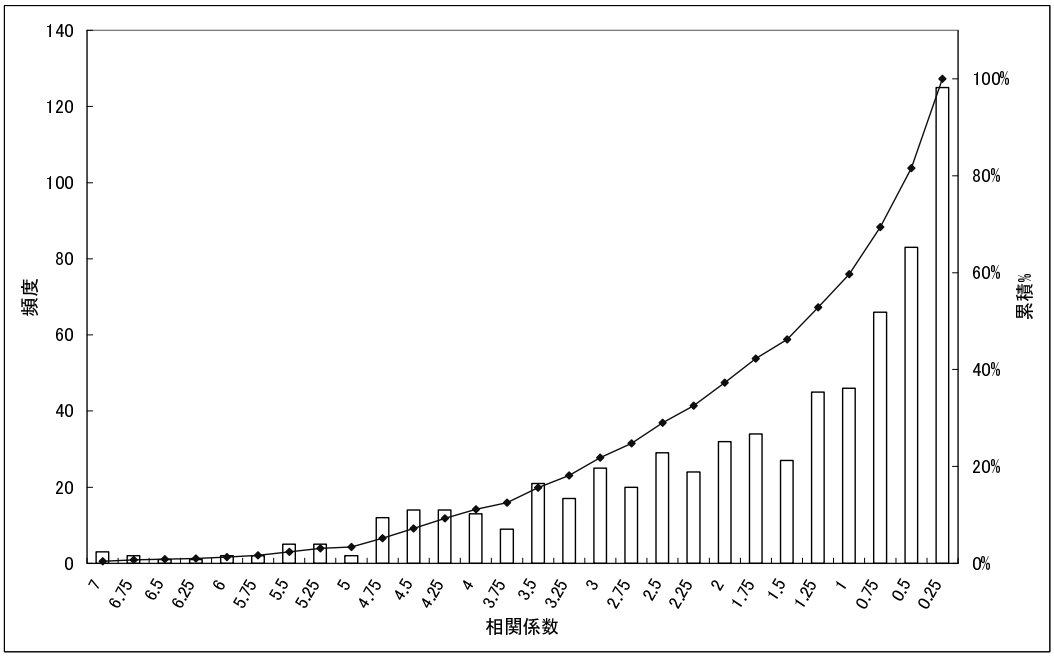

図 5.7: 発現強度分布

5.2.3 被制御遺伝子の相関係数順位と発現強度順位の関係

これまでの調査結果から，同一転写因子の被制御遺伝子の絞込みに利用できるのは，相関係数，発現強度ともに相対的な順位であることが明らかになった．次に，この二つのデータの関係について調査する．図 5.9 に相関係数順位と発現強度順位の関係について示す．分布傾向としてどちらかで上位 400 番目程度の中に含まれているものは，もう片方の順位においても 400 番目程度までに含まれている場合が多いことがわかる．逆にどちらかで 400 番目以降に含まれているものは互いの順位に関係があるとは考えられない．このた

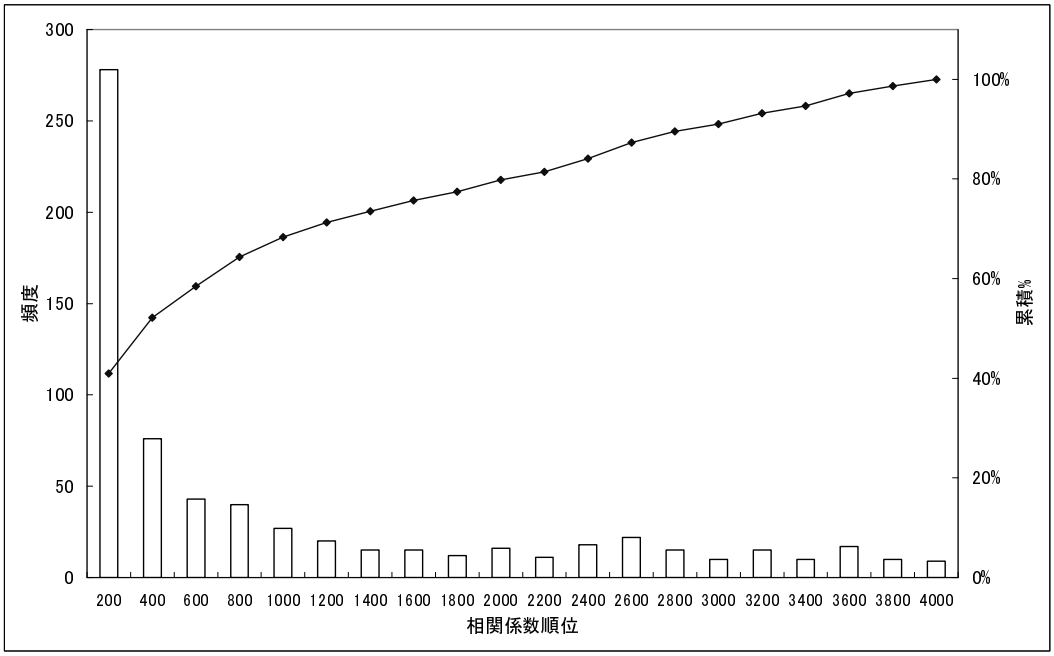

図 5.8: 発現強度順位分布

め，両方の順位において上位 400 番目以内に含まれている場合には，同一の転写因子の被制御遺伝子の可能性が高いと考えられる．

5.3 ウインドウ長さに関する調査

本節では，ウインドウ長さとウインドウ類似度の関係について調査を行う．本研究で用いているウインドウ類似度の定義は，単純に一致する文字列の数をカウントするのではなく，ウインドウ内における長さ 6 の各文字列の特異度に依存して類似度が変化する．また移動幅についてもウインドウ長さの変化に関係なく固定されており，ウインドウ長さを変化させた場合にウインドウ類似度にどのように影響を与えるのか不明である．よってウインドウ長さとウインドウ類似度の関係，そして推定に用いる長さが推定結果に及ぼす影響について調査する．

5.3.1 ウインドウ長さとウインドウ類似度の関係についての調査

ウインドウ長さについてはこれまで，既知のバインディングサイトの多くが 30 以下であることから，バインディングサイト全体を含むことができる長さとして，20〜30 程度を用いて実験を行ってきた．また，またウインドウ類似度の定義上，長さ 6 の文字列を基準とした比較であること，そして移動幅の効果を十分に得られることを考慮して最低ウイ

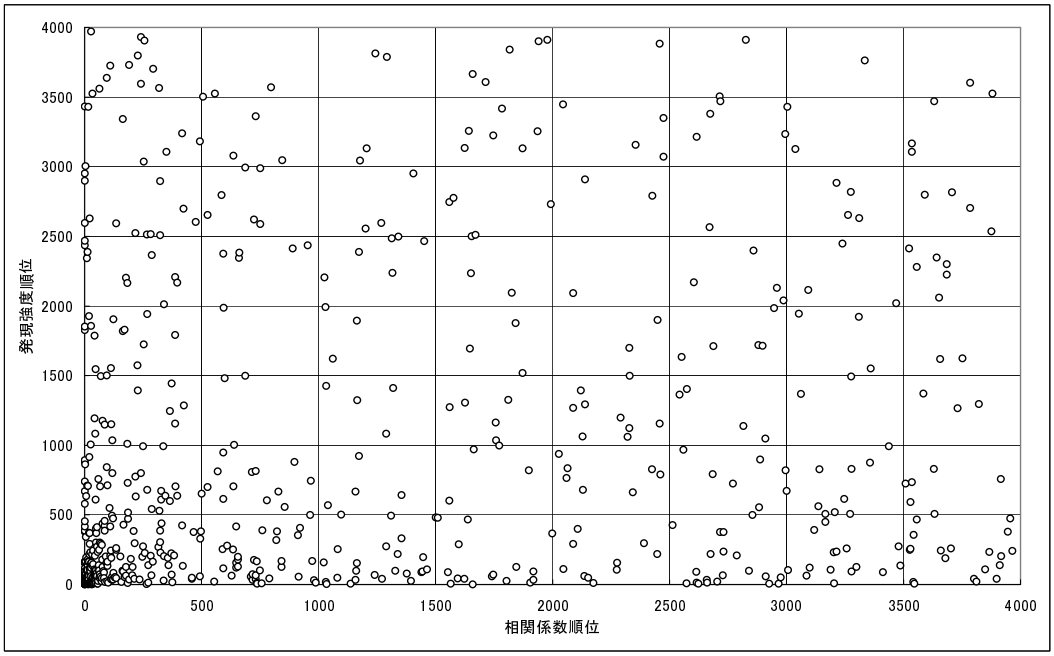

図 5.9: 相関係数順位と発現強度順位の分布

ンドウ長さを 10 とすることにした．以上のことから，多少の余裕を見て，ウインドウ長さを 10 から 50 めで 5 ずつ変化させた場合のウインドウ類似度に対する影響を確認する．

調査は表 4.2 で示した転写制御領域における各塩基の出現確率に基づきランダムに 2 つのウインドウを生成し，このウインドウ間のウインドウ類似度平均値を各ウインドウ長さについて調べ，ウインドウ長さに対するウインドウ類似度の変化を調査する．各ウインドウ長さにおける計算回数は 10000 回とした．これは特にウインドウ長さが短い場合，繰り返し回数が少ないと，特異度閾値以上の長さ 6 の文字列が含まれていた場合と，そうでない場合に偏りが生じると正確なデータが得られないため，かなり多めに設定して調査を行うことにしたためである．得られた調査結果を図 5.10 に示す．ウインドウ長さに対するウインドウ類似度の平均値はほぼ一直線上に分布しており，線形の関係にあることがわかる．

5.3.2 ウインドウ長さの推定への影響に関する調査

これまで未知のバインディングサイトを推定するのに, 本研究で用いている方法では単一のウインドウ長さを用いて実験を行っていた．しかしながら，バインディングサイトの長さがそれぞれ異なる遺伝子に対し，ウインドウ長さ一定でかまわないのか，それともバインディングサイト長さごとに，最適なウインドウ長さが存在し，未知のバインディングサイトを推定するためにはウインドウ長さを変更し，繰り返し推定する必要があるのか十

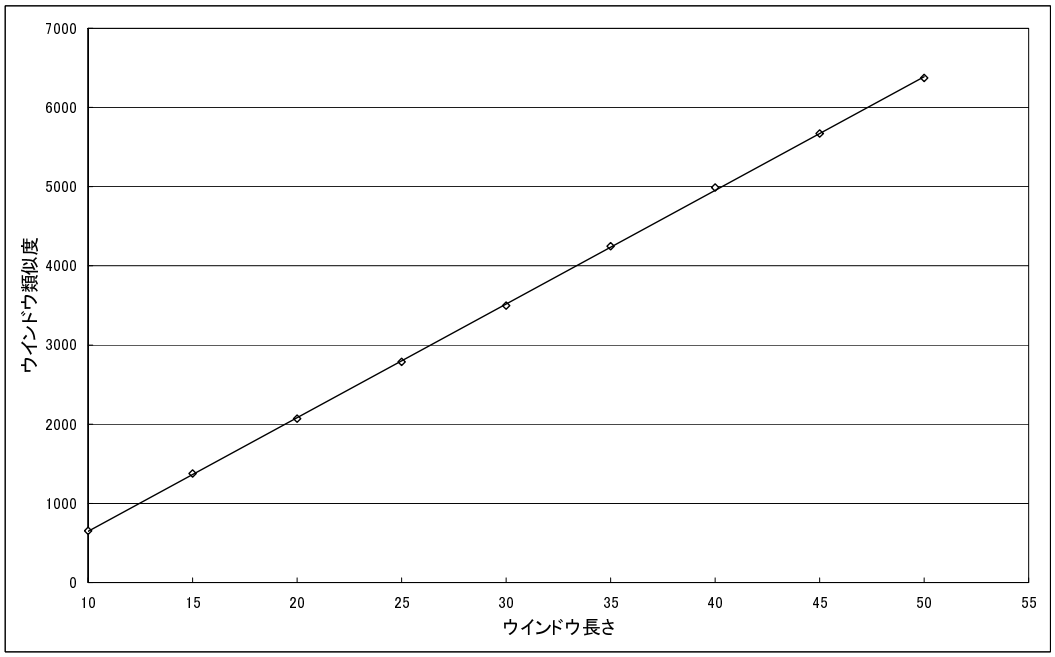

図 5.10: ウインドウ長さに対する平均ウインドウ類似度

分な検証を行ってこなかった．このため，一つの遺伝子に対し，ウインドウ長さを変化させた場合の推定結果について検証することにする．

調査対象の遺伝子として yneA を用いる．yneA のバインディングサイトに関する情報は表 5.8 に示したとおりであり，太字の部分がバインディングサイトである．この長さ 14 のバインディングサイトに対し，ウインドウ長さ l を 12，18，24 と変化させた場合の推定を行い比較する．

表 5.8: 遺伝子 yneA のバインディングサイトデータ

項目	データ
Binding factor	LexA
Regulation	Negative
Binding seq.	AAATG **CGAACAAACATTCC** TGTTG

バインディングサイト付近の推定結果のグラフを図 5.11 に示す．ウインドウ長さを長くするに従い若干ウインドウ類似度の値が増加していることと，位置にわずかに違いがみられるが，結果として全てのウインドウ長さにおいて推定が可能であった．このためウインドウ長さの推定結果への影響は小さいように思われる．

この結果は先のウインドウ長さとウインドウ類似度の関係が影響していると考えられる．ウインドウ長さと平均ウインドウ類似度の関係は線形であり，ウインドウ長さ 30 の

場合でも平均ウインドウ類似度は約 3500 である．これに対し，ピーク位置におけるウインドウ類似度閾値は 8000 であり，その差は非常に大きい．このため，ウインドウ長さに従いウインドウ類似度は若干変化するが，それは特異的文字列が一致した場合のウインドウ類似度の変化に比べると十分に小さく推定結果には殆ど影響を及ぼさないものと考えられる．

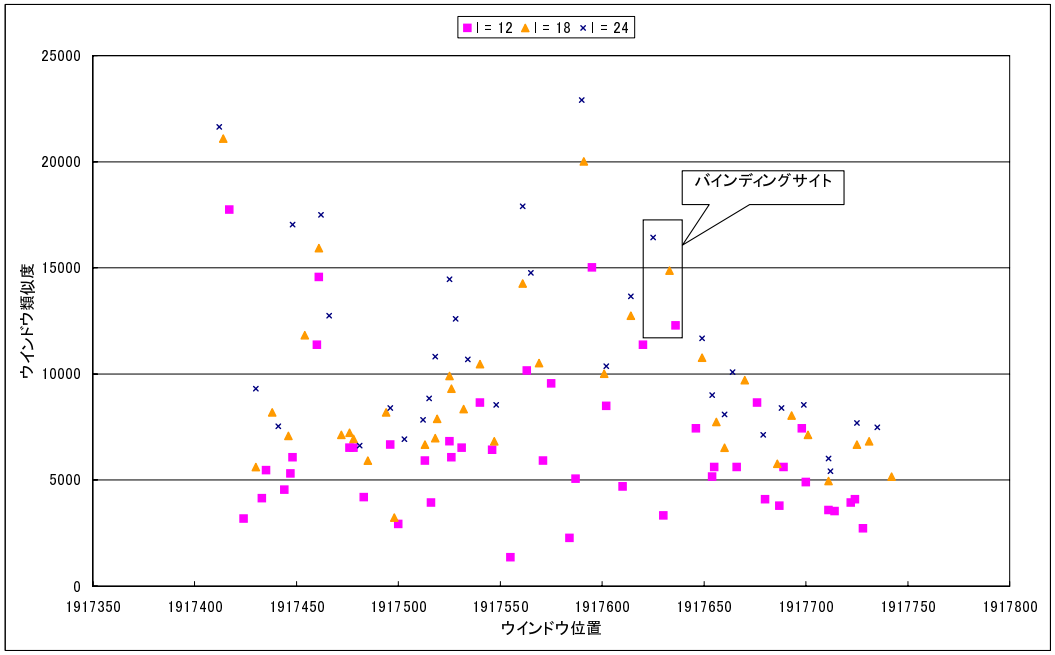

図 5.11: ウィンドウ長さを変えた場合の推定結果比較 (比較元：yneA)

第6章　まとめ

本研究では，既存手法における問題点の解決のため，手法の改良と各種パラメータに関する最適値の調査を行った．

6.1　ピーク位置について

既存手法におけるピーク位置の定義では，バインディングサイトを含むウインドウ位置がピーク位置として選択されない場合があり，正しく推定できない一因となっていた．これに対し本研究では，これまで利用していなかった比較元のウインドウ位置 i の変化と，位置 i に対する最大ウインドウ類似度をとる対象遺伝子側のウインドウ位置 j の関係を新たに発見し，これをピーク位置の決定に利用することで改良を行った．また，バインディングサイトが短い被制御遺伝子についても推定がおこなえるよう，ピーク位置の前提として設けていたウインドウ類似度の閾値を 15000 から 8000 にまで引き下げた．これらの変更により，これまでバインディングサイトを含むウインドウ位置であるにもかかわらずピーク位置として選択されていなかったウインドウをより正確にピーク位置として決定することが可能になった．

6.2　類似関係のグラフ構造を利用した絞込みについて

既存手法により得られた類似関係をグラフ上に表したとき，同一転写制御因子の被制御遺伝子がクリークを構成するという性質を利用して，被制御遺伝子の判別を行う方法を提案した．提案手法により，既存手法と比較して，ピーク位置の数を 1/3 程度にまで絞り込むことに成功した．これはバインディングサイトを特定するために大きな成果である．判別の条件として，最低でも三つの遺伝子でクリークを構成するという条件を設定したが，一つの転写制御因子の被制御因子の数が不明であることを考慮すれば妥当であると考えられる．

しかしながら，全体の推定結果としては依然としてバインディングサイトを特定するのに，十分な絞込みができているとはいえない．バインディングサイトを特定するためには，クリークによる判別を行う前に，さらに対象遺伝子を絞り込む必要がある．クリークによる判別はピーク位置を頂点と考えて実行しており，対象遺伝子が一つ減るだけでもピーク位置は十数個は減少し，減少したピーク位置とクリークを構成していたピーク位置

もかなりの数が減少すると考えられ，クリークによる判別は対象遺伝子の数が少ないほど効果的に働くと考えられる．

6.3 バインディングサイトの特定方法について

推定結果として得られたピーク位置の中からバインディングサイトである可能性が高い部分を特定する方法として，ランダム配列による推定結果との比較により特異的に高いウインドウ類似度をとる部分をバインディングサイトの可能性が高いと推定する方法を提案した．この方法はバインディングサイトの長さが長い遺伝子についてはかなり有効である．しかしながら，この方法ではバインディングサイトが短い場合には利用できないことため別の方法によりバインディングサイトの可能性が高い位置を特定する必要があると考えられる．

6.4 各種パラメータについて

6.4.1 ウインドウ類似度における移動幅について

同一転写因子のバインディングサイトであっても，塩基の置換，挿入，削除により遺伝子ごとに若干の違いが存在する．これに対応するためウインドウ類似度の定義において移動幅を用いているが，現在用いている移動幅は 3(7 箇所での比較) は広すぎる可能性があったため，移動幅とウインドウ類似度の関係，移動幅によるバインディングサイトにおけるウインドウ類似度の効果について調査を行い，最適な移動幅は 2(5 箇所での比較) であるという結論を得た．

6.4.2 マイクロアレイデータによる同一転写因子の絞込み

本研究の推定手法ではマイクロアレイによる遺伝子発現傾向の相関により，転写制御領域における文字列比較をおこなう対象となる遺伝子を絞り込んでいる．これまで相関の高い上位遺伝子 300 種を可能性の高い遺伝子としていたが，既知の被制御遺伝子の発現強度と相関係数について値と順位の分布を調査した結果，発現強度，相関係数ともに高い値に集中していうるわけではなかった．これに対し，順位の調査では発現強度，相関係数ともに高い順位の部分に同一転写因子の被制御遺伝子が集中していることが明らかになった．しかしながら，現在用いている上位 300 遺伝子では不十分であり，上位 400 遺伝子とすることで全体の約 40 %の遺伝子をカバーすることにした．上位 400 以降の遺伝子にも被制御遺伝子は多数含まれているが，全体に平均的に分布しており，被制御遺伝子の可能性が高い遺伝子を判別するという目的からもこれ以上順位を下げることは望ましくないと考えられる．相関データの精度はマイクロアレイデータの数を増やすことにより，ある程度

の改善されることが期待できることと，この段階で，遺伝子数を絞り込まなければバインディングサイトの特定が困難である点から考えて上位400というのは妥当であると考えられる．

6.4.3　ウインドウ長さについて

ウインドウ長さとウインドウ類似度の関係について調査した．調査は遺伝子制御領域における各塩基の出現確率を元にランダムに生成した文字列配列を用い，ウインドウ長さとウインドウ類似度平均値の関係は線形であるという結果が得られた．ウインドウ長さに比例して，平均ウインドウ類似度は高くなるため，バインディングサイトが含まれていても，周囲との差が小さくなることは確かであり，特異的な文字列の一致を発見することが困難になることが考えられる．しかし，実際の推定における設定は，既知のバインディングサイトの大部分は長さ30以下であるため，設定する最大ウインドウ長さを30程度と考えてよい．この場合，最大平均ウインドウ類似度は約3500，これに対し，特異的な文字列の類似が含まれていると判断する基準として用いているピーク位置の閾値は，長さ6以上の文字列の完全一致した場合の最低ウインドウ類似度として8000を用いている．平均類似度に比べて非常に高い類似度を最低値としていため，ウインドウ長さの変化による平均ウインドウ類似度の変化は推定結果に対し，殆ど影響を与えないと考えられる．実際，同一の遺伝子に対しウインドウ長さを変更して行った推定結果においても，バインディングサイトを正しく推定可能であることが確認できた．

このためウインドウ長さの設定は，バインディングサイト全体をそのウインドウ内に含むことが可能であることのみが条件であり，ウインドウ長さ30程度までならば推定結果に対する影響は皆無と考えられる．但し，実用の面から考えると，ウインドウ長さは計算時間の短縮のためには短いほうが望ましいが，未知のバインディングサイトを調べるためには，ある程度余裕があったほうが望ましいと考えられる．

6.5　今後の課題

本研究による改良により，従来手法よりもバインディングサイト候補を絞り込むことが可能になったが，比較対象となる遺伝子数が400と非常に多く，バインディングサイトを特定するほど十分に絞込めたとはいえない．今後さらに候補遺伝子を絞り込む方法として，オペロンに関する情報を利用することが有効と考えられる．

本研究で用いた枯草菌などの原核生物における転写では，複数の遺伝子がまとめて転写されるポリシストロン性転写があることが知られている．一つの転写単位の中に含まれる関連遺伝子群と転写の制御に関わる領域をまとめてオペロンと呼び，オペロンに含まれる遺伝子はオペロンの先頭部分に存在する制御領域で一括して発現を制御されている．このためオペロンが推定可能になれば，数種の遺伝子を一つの比較対象としてあつかえるため，遺伝子の数を大幅に減少させることが期待できる．また，これは生物学的な発現のメ

カニズムをより正確に再現した形となる上，事実上，比較対象となる文字列の数が減少するため，偶然の文字列の類似が減少しより正確にバインディングサイトをピーク位置として特定可能になると思われる．またクリークによる絞込みにおいても元となるピーク位置の総数が減少することで，より効果が期待できる，またオペロンに関する推定は，DNA塩基配列上での各遺伝子の間隔からある程度推定可能であると考えられる．

謝辞

本研究を行うにあたり，指導教官の平石 邦彦教授には多数の助言をいただき，深く感謝しております．また，日頃から多大なる議論と激励を頂きました平石研究室の皆様に厚くお礼申し上げます．

参考文献

[1] デービッド W. マウント, バイオインフォマティクス第2版, メディカル・サイエンス・インターナショナル, 2005.

[2] 田村 隆明, 山本 雅, 改定第2版 分子生物学イラストレイテッド第2版, 羊土社, 2003.

[3] 上田 智之, 遺伝子転写制御領域に含まれる特異的文字列の解析とDNAマイクロアレイデータを用いた遺伝子間の依存関係推定, 北陸先端科学技術大学院大学 修士論文, 2004.

[4] Nir Friedman, Inferring Cellular Networks Using Probabilistic Graphical Models, *Science*,303,799 - 805,2004.

[5] Tatsuya Akutsu Identification of genetic networks from a small number of gene expression patterns under the Boolean network model, *Proc. Pacific Symposium on Biocomputing*,4,17 - 28,1999.

[6] Chris J Needham, James R Bradford, Andrew J Bjlpitt, David R Westhead, Inference in Bayesian networks, *Nature Biotechnology*,24, 51 - 53,2006.

[7] Matthew J. Beal, Francesco Falciani, Zoubin Ghahramani, Claudia Rangel, David L. Wild, A Bayesian approach to reconstructing genetic regulatory networks with hidden factors, *Bioinformatics*,21, 349 - 356,2005.

[8] 北野宏明, システムバイオロジーの展開, シュプリンガー・フェアラーク東京株式会社, 2001.

[9] 松原 謙一, ゲノム機能 発現プロファイルとトランスクリプトーム, 中山書店, 2000.

[10] Makita Y, Nakao M, Ogasawara N, Nakai K., DBTBS: database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics, *Nucleic Acids Res.*,32,D75-77,2004.