修 士 論 文

転写制御領域の解析と破壊株データからの遺伝子の依存関係推定に関する研究

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

小倉 亨

2003 年 3 月

修 士 論 文

転写制御領域の解析と破壊株データからの遺伝子の依存関係推定に関する研究

指導教官　平石邦彦 助教授

審査委員主査　平石邦彦 助教授
審査委員　金子峰雄 教授
審査委員　浅野哲夫 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

11028 小倉 亨

提出年月: 2003 年 2 月

概 要

近年、ヒトを始めとする様々な生物のデオキシリボ核酸（DNA）上の塩基配列が決定した。そのため、研究の焦点は遺伝子や生体活動に関わる蛋白質などを明らかにしようとする動きに変わってきている。

DNA は遺伝情報を担っており、細胞内の核に存在する物質である。遺伝子は DNA 上ので生体機能に関わる蛋白質を生成する発現を行う部分の事を指す。遺伝子の発現は転写、翻訳の段階を経て蛋白質を生成する現象である。発現は遺伝子の上流部分にある転写制御領域と呼ばれる部位に転写活性化因子と呼ばれる蛋白質が結合することにより開始する。転写段階では遺伝子上の塩基配列と対になる配列を持つメッセンジャーリボ核酸（mRNA）が生成される。転写段階で生成された mRNA は核の外に移動し翻訳の段階で蛋白質を生成する元になる。翻訳段階では mRNA の塩基配列から対応したアミノ酸を結合させて蛋白質を生成する。また、上流部分にある別の転写制御領域に転写抑制因子と呼ばれる蛋白質が結合する事により転写は抑制される。

転写を活性化させる蛋白質や転写を抑制する蛋白質の事を転写因子と呼ぶ。また、転写因子を発現によって生成する遺伝子は調節遺伝子と呼ばれる。調節遺伝子が発現により生成する転写因子によって発現に影響を受ける遺伝子は調節遺伝子と依存関係を持つ。

調節遺伝子と依存関係を持つ遺伝子の上流部分には調節遺伝子が発現して生成した転写因子が結合する転写制御領域が存在し、その領域内の塩基配列は特異的である。本論文では DNA 配列を文字列として扱うため、転写制御領域内の塩基配列は転写制御領域の部分文字列として扱う。この様な部分文字列の事を本論文では文字列パターンと呼ぶ。ゆえに調節遺伝子と依存関係を持つ遺伝子は特定の文字列パターンを持つ。文字列パターンを持つ遺伝子とは転写制御領域にその文字列パターンを含む事を意味している。本論文では同一の文字列パターンを持つ遺伝子群は共通の転写因子によって影響を受けているかどうかを破壊株データを用いて検証する。この方法により遺伝子間の依存関係推定における転写制御領域の解析の有意性を示す事を目的とする。破壊株データとは標的となる遺伝子を破壊し各遺伝子の発現量を観測した DNA マイクロアレイデータの事である。本論文の目的は遺伝子間の依存関係推定に対して転写制御領域を考慮した推定方法の基礎的部分を担うものである。推定された遺伝子間の依存関係は細胞内の状態変化を定性的な解析に利用する事が出来る。また、細胞内の状態変化を解析する事は新薬開発など様々な利用法が考えられる。本論文では対象となる生物として DNA 上の塩基配列や各遺伝子の場所などが知られている枯草菌を用いる。

本論文の手法としては枯草菌の DNA 塩基配列を文字列として表現したデータから調節遺伝子が発現して生成する転写因子が結合する事が既知である文字列パターンを持つ遺伝子をパターンマッチングを用いて調査する。DNA は二つのストランドと呼ばれる塩基配列が互いに結合して二重螺旋構造を形成している。今回は各遺伝子の上流部分を両方のストランドから抜き出し調査を行った。遺伝子の上流部分を特定するため、各遺伝子の場

所、転写方向等のデータを利用して上流部分を範囲を決めて抜き出した。各遺伝子の上流部分を文字列として抜き出して、調べる対象となる文字列パターンについて全遺伝子の上流部分を調査した。

文字列パターンを持つ遺伝子群が調節遺伝子に影響を受けている事を調べるために破壊株データを用いて統計学的手法を行った。調節遺伝子と他の遺伝子との相関係数を破壊株データから求め、調節遺伝子と文字列パターンを持つ遺伝子との相関係数が高い場合にはその二つの遺伝子間には依存関係を持つ可能性がある。文字列パターンを持つ遺伝子群を標本とし、全遺伝子を母集団とした時に調節遺伝子との相関係数を求め、標本と母集団の相関係数の平均値に差が有意であるか調査した。

今回は文字列パターンを持つ依存関係既知の遺伝子群、依存関係既知の遺伝子と一律に発現に対して影響を受ける遺伝子群（オペロン)、文字列パターンを持つ全ての遺伝子群、文字列パターンを持つ依存関係未知の遺伝子群を標本とし調査を行った。結果としては文字列パターンを持つ依存関係既知の遺伝子群やそのオペロン群に対しては本研究の統計的手法において半数以上の調節遺伝子に対して標本と依存関係を持つ可能性があると判断する事が検証できた。これはDNAマイクロアレイデータを用いた本論文の解析手法で半数以上の調節遺伝子に対して既知の事実と一致する事を示している。しかし、文字列パターンを持つ全ての遺伝子群では依存関係を持つ可能性がある調節遺伝子は半数以下になり、文字列パターンを持つ依存関係未知の遺伝子群では1つの調節遺伝子のみに関して依存関係がある可能性を持つという結果になった。

文字列パターンを持つ依存関係未知の遺伝子群の調節遺伝子との相関係数を調べると文字列パターンを持つ依存関係既知の遺伝子群に近い相関係数を持つ遺伝子は存在するが、相関係数が低い遺伝子を多数含むため、依存関係未知の遺伝子群全体の平均値が低くなる事が分かった。

文字列パターンを持つ依存関係未知の遺伝子群の中に低い相関係数を含む理由としては、未知の転写因子によって遺伝子が影響を受けることと、依存関係を持たない遺伝子が存在することが挙げられる。

未知の転写因子によって影響を受ける場合にはその転写因子を生成する調節遺伝子の特定が必要になってくる。また、依存関係を調べる方法として調節遺伝子との相関係数ではなく、偏相関係数を用いる方法でこの様な問題を解決することが出来ると考えられる。

標本内に依存関係を持たない遺伝子を含む事に対しては、調節遺伝子との相関係数で調査するのではなく、既知に影響を受ける遺伝子との相関係数を求める必要がある。この様な相関係数を疑似相関係数と呼び、疑似相関係数が正に高いほど既知の遺伝子とデータ上での値の傾向が類似する事を意味する。この疑似相関係数に閾値を設けて依存関係の無い遺伝子を削除していく方法が有効であると考えられる。

目 次

第 1 章　遺伝子の依存関係推定について　1
1.1 はじめに 1
1.2 従来研究 4
1.3 本研究の目的について 6

第 2 章　対象となる菌類と使用するデータについて　8
2.1 枯草菌について 8
2.2 DNA 配列データについて 9
2.3 破壊株データについて 10

第 3 章　実験 1：パターンマッチングを用いた全遺伝子の上流部分の調査　12
3.1 DNA 配列データの上流部分について 12
3.2 文字列パターンについて 13
3.3 パターンマッチングについて 16

第 4 章　実験 2：統計的手法を用いた DNA マイクロアレイデータの解析　17
4.1 相関係数 17
4.2 平均値の検定 18

第 5 章　結果　21
5.1 文字列パターンを持つ遺伝子について 21
5.2 統計的手法による DNA マイクロアレイデータの解析結果 22
5.2.1 文字列パターンを持つ依存関係既知の遺伝子を標本にした場合 . . . 23
5.2.2 文字列パターンを持つ依存関係既知の遺伝子のオペロンを標本にした場合 26
5.2.3 文字列パターンを持つ全遺伝子を標本にした場合 29
5.2.4 文字列パターンを持つ依存関係未知の遺伝子を標本にした場合 . . . 30
5.3 各標本との比較 31

第 6 章　考察　33
6.1 他の未知の転写因子に影響を受けている場合 33
6.2 依存関係がない場合 34

6.2.1　文字列 p ターンを挟んで異なるストランドに遺伝子が存在する場合　35
6.2.2　今回使用した文字列パターンの情報の欠損していた場合　. 36
6.2.3　疑似相関係数を用いた依存関係未知の遺伝子の調査　. 37

第 7 章　おわりに　40

第1章　遺伝子の依存関係推定について

1.1　はじめに

近年、ヒトを始めとする様々な生物のデオキシリボ核酸（DNA）上のアデニン（A)、グアニン（G)、シトシン（C)、チミン（T）からなる塩基配列決定が完了した。そのため、研究の焦点は遺伝子や生体活動にかかわる蛋白質の機能などを明らかにしようとするポストゲノムの段階に移りつつある。

DNA は遺伝情報を担っていおり、核内の染色体を構成している物質である。遺伝子とは遺伝情報を決定する単位であり、DNA 上で発現という現象を行う部分の事を指す。

遺伝子の発現とは図 1.1 で示す様に転写段階 (transcription) でメッセンジャーリボ核酸（mRNA）を生成し翻訳 (translation) の段階を経て酵素等の生体機能に関わる蛋白質を生成する現象である。

転写とはリボ核酸ポリメラーゼ（RNA polymerase）と呼ばれる物質によって遺伝子部分の塩基配列に対応する mRNA を生成する（図 1.2a)。塩基配列の対応としては C → D、D → C、T → A となるが、RNA ではチミン (T) のかわりにウラシル（U）が使われるため A → U となる。生成された mRNA はスプライシングと呼ばれる行程により蛋白質を生成するのに必要な部分以外は切り放される (図 1.2b)。この様に生成された mRNA は核外に向かい発現の翻訳に使用される。

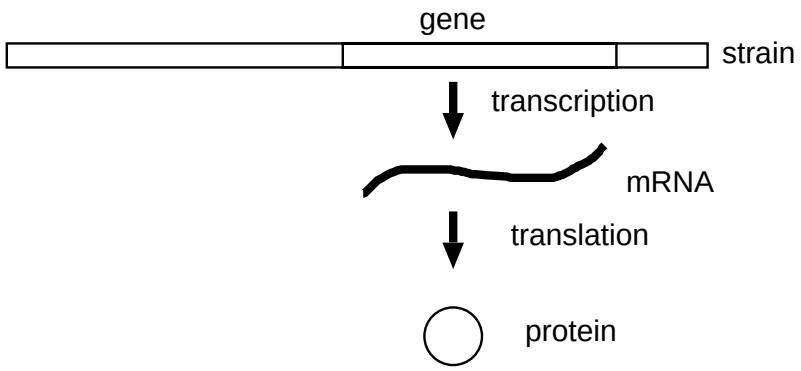

図 1.1: 遺伝子発現の過程

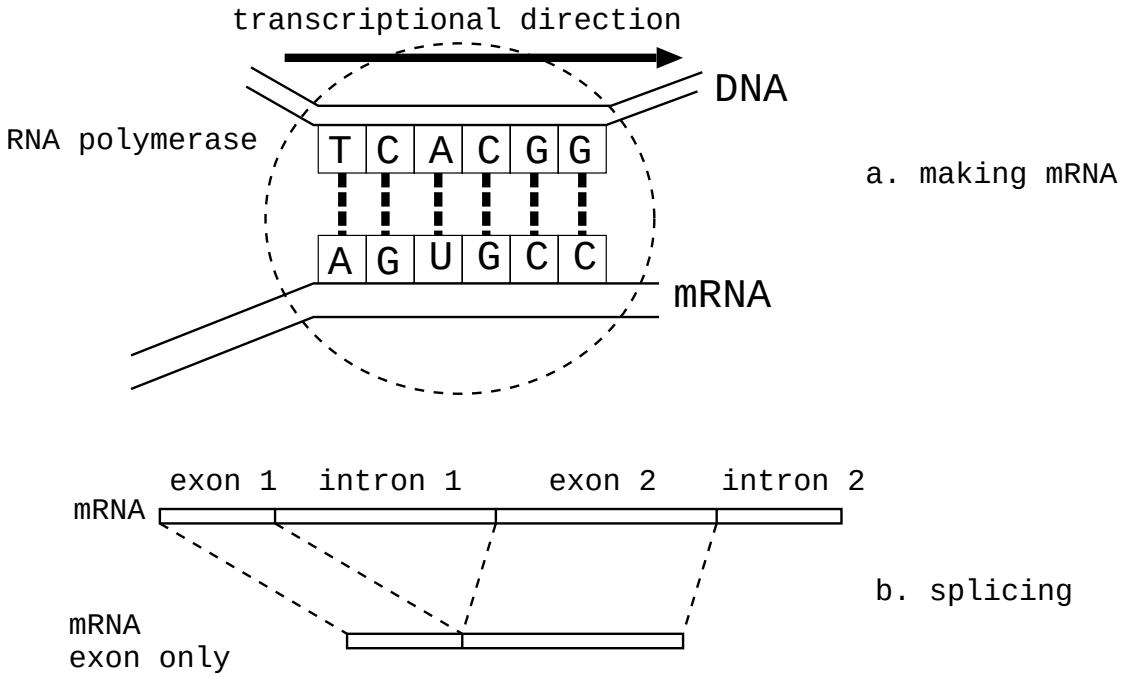

図 1.2: 転写段階

翻訳は転写段階で生成された mRNA から 3 つの塩基に対応するアミノ酸を表 1.1 の遺伝子暗号を元に結合させ、蛋白質を生成する [3]。3 つの塩基を 1 組とする単位をコドンと呼ぶ。mRNA をコドン毎にリボソームと呼ばれる物質が遺伝子暗号からアミノ酸を決定し、トランスファー RNA（tRNA）が対応するアミノ酸を輸送する。遺伝子暗号上での終止は終止コドンを意味しており、終止コドンは翻訳を終わらせるコドンの事を指す。運ばれてきたアミノ酸同士は互いに結合して蛋白質を形成する。

発現は遺伝子の上流部分にあるプロモータ（promoter）と呼ばれる部分に転写を活性化する蛋白質であるアクチベータ（activator）が結合して開始する（図 1.3 参照）。また、オペレータ（operator）と呼ばれる部位に転写を抑制する蛋白質であるリプレッサ（repressor）が結合する事により遺伝子の発現は抑制される。この様に遺伝子の発現を制御する蛋白質は転写因子と呼ばれ、転写因子を発現によって生成する遺伝子の事を調節遺伝子と呼ぶ。また、転写因子が結合する各部分は転写制御領域と呼ばれる。

他の遺伝子の転写制御領域に結合した転写因子によって発現に影響を受ける遺伝子もある。ある一つの転写制御領域に結合した転写因子によって一律に発現に影響を受ける遺伝子群の事をオペロンと呼ぶ。オペロンと呼ばれる遺伝子群の間は非常に狭い。図 1.4 の場合、遺伝子 A の上流部分にある転写制御領域に転写因子が結合する事により、遺伝子 A,B,C は一律にその発現に影響を受ける。また、オペロンは転写単位とも呼ばれる。

遺伝子が転写因子を発現により生成する事が生物学的に知られている。その遺伝子の発現より生成された転写因子が他の遺伝子の転写制御領域に結合することにより、その遺伝

1文字目	2文字目				3文字目
	U	C	A	G	
U	フェニルアラニン	セリン	チロシン	システイン	U
	フェニルアラニン	セリン	チロシン	システイン	C
	ロイシン	セリン	終止	終止	A
	ロイシン	セリン	終止	トリプトファン	G
C	ロイシン	プロリン	ヒスチジン	アルギニン	U
	ロイシン	プロリン	ヒスチジン	アルギニン	C
	ロイシン	プロリン	グルタミン	アルギニン	A
	ロイシン	プロリン	グルタミン	アルギニン	G
A	イソロイシン	トレオニン	アスパラギン	セリン	U
	イソロイシン	トレオニン	アスパラギン	セリン	C
	イソロイシン	トレオニン	リシン	アルギニン	A
	メチオニン	トレオニン	リシン	アルギニン	G
G	バリン	アラニン	アスパラギン酸	グリシン	U
	バリン	アラニン	アスパラギン酸	グリシン	C
	バリン	アラニン	グルタミン酸	グリシン	A
	バリン	アラニン	グルタミン酸	グリシン	G

表 1.1: 遺伝子暗号表

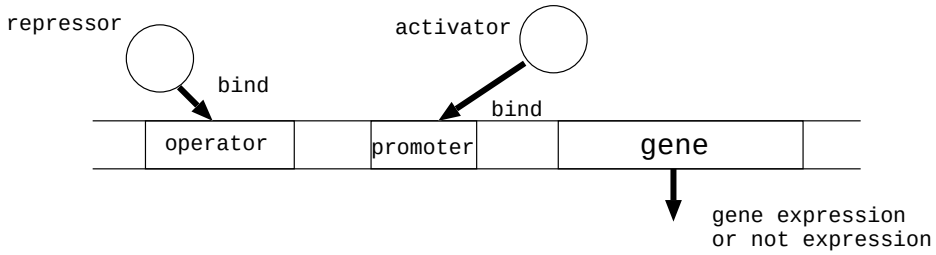

図 1.3: 転写制御領域と発現

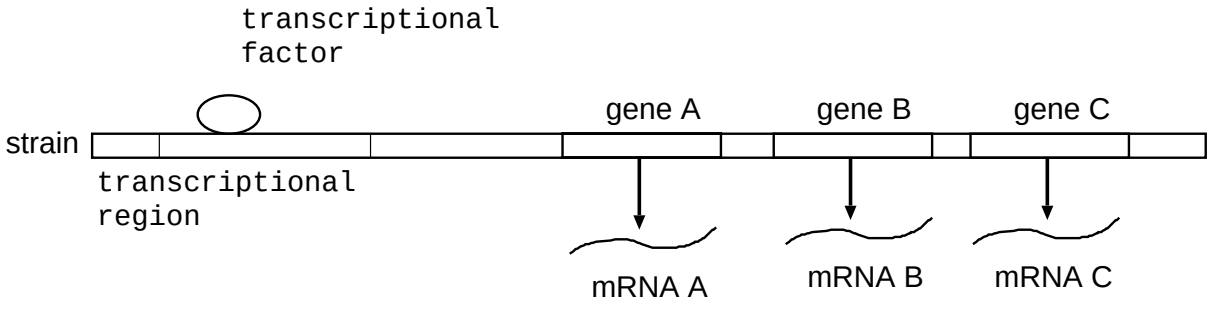

図 1.4: オペロンについて

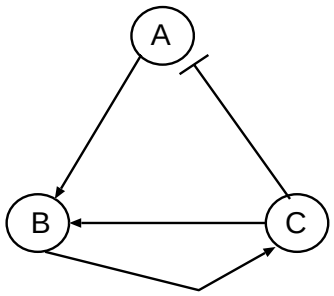

図 1.5: ブーリアンネットワークでの例

子の発現に対して活性、または抑制を行う。この様な遺伝子間の依存関係を解析することは蛋白質を起因とする癌等の病気の新薬開発に利用できると考えられているため、現在、様々な研究が行われている。

しかし、細胞内の転写因子の濃度とその転写因子が結合する転写制御領域を上流部分に含む遺伝子の mRNA の転写量や生成する蛋白質濃度を測定して依存関係を解析する研究は遺伝子学、分子生物学の分野で行われているが、全遺伝子の依存関係をこの方法で調べる場合には費用的にも時間的にも非常にコストがかかる。そのため、現在では大量の遺伝子の mRNA の転写量を同時に測定できる DNA マイクロアレイを用いて遺伝子間の依存関係を推定する研究が主に行われている。

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。従来研究では DNA マイクロアレイデータのみを用いて遺伝子間の依存関係を推定するトップダウン法と呼ばれる研究が主に行われてきた。代表的なトップダウン法としては、ブーリアンネットワーク、ベイジアンネットワーク、S-system と多階層有効グラフ等を利用して依存関係推定を行う研究が挙げられる。

ブーリアンネットワーク [1, 2] では単位時間毎に観測された DNA マイクロアレイデータでの各遺伝子の値から各遺伝子の状態を 0（発現していない）か 1（発現している）の状態に分ける。各遺伝子間の発現に関する依存関係は AND,OR,NOT 等の論理関数の形式で表現される。依存関係はある単位時間 t と $t+1$ の各遺伝子の状態遷移表を作成して遺伝子毎に各状態に対応する論理関数を推定していく。ブーリアンネットワークの特徴としては同期して動作する論理回路の様なモデルである。図 1.5 では A、B、C の 3 つの遺伝子が依存関係を持つ場合のブーリアンネットワークモデルである。各頂点を遺伝子、頂点を結ぶ各アークは依存関係を表している。A、B、C の各遺伝子は図 1.5 内の論理関数によってそれぞれの遺伝子間の依存関係が表現されている。

ベイジアンネットワーク [7] は確率変数間の依存関係を非循環な有効グラフで表現した

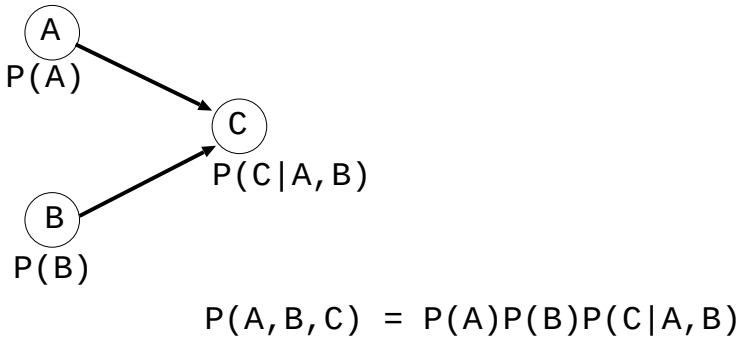

図 1.6: ベイジアンネットワークでの例

モデルである。各遺伝子の状態についてはDNAマイクロアレイデータの値から状態を決め、その状態になる確率を計算する。発現に関しての依存関係は条件付き確率を用いて表現する。全体としてはネットワークを構成する確率変数の同時確率分布になる。図1.6では3つの遺伝子間の依存関係をベイジアンネットワークを用いて表現している。遺伝子Cがある値になる確率は遺伝子A、Bとの条件付き確率となり遺伝子A、Bの確率変数が取りうる値に依存する。この場合遺伝子A、Bを遺伝子Cの親ノードと呼ぶ。このネットワーク全体の同時確率分布は$P(A,B,C) = P(A)P(B)P(C|A,B)$となる。

Friedmanらは各遺伝子の確率変数がとりえる値ををを-1（mRNA転写量が少ない）、0（mRNA転写量が変わらない）、1（mRNA転写量が多い）の3つの状態に分け、この手法を用いて酵母菌の遺伝子約800個について依存関係を推定した[7]。

また、DNAマイクロアレイデータから各遺伝子の状態を連続値として扱い遺伝子間の依存関係を推定する研究としてS-systemと多階層有効グラフ[5]があげられる。S-systemとはSavageauが提案したもので以下の微分方程式で表現される動的システムの記述法である。

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

上の式は遺伝子iが他の遺伝子から受ける影響を微分方程式で表したものである。α_i、β_iはそれぞれ遺伝子iの生成項、分解項にあたる。また、g_{ij}は遺伝子iに対して生成項に関与する他の遺伝子からの相互作用係数になり、h_{ij}は遺伝子iに対して分解項に関与する他の遺伝子からの相互作用係数になる。DNAマイクロアレイデータからこの式を満たす依存関係を表す各パラメータ$\alpha_i, \beta_j, g_{ij}, h_{ij}$を導出する。また、多階層有効グラフでは実験的手法を用いて遺伝子を破壊した場合のデータと破壊しない場合のデータの両方を比較し影響の有無を判断する。依存関係の向きがループ構造になるような遺伝子群を同値類として一つの遺伝子と見なし、遺伝子間の依存関係を構築していく。S-systemから求めたパラメータと多階層有効グラフで求めた依存関係の二つを統合することによって遺伝子間の依存関係とその強さを推定していく。

これら従来研究の問題点としては、

- 組み合わせ爆発：N 個の遺伝子から構成することができるネットワークの数は遺伝子が増えると膨大になる。その中から条件を満たすネットワークを探索するのは非常に計算量がかかる。
- 複数の準最適解：元のデータと同じような発現量を作り出す事の出来るネットワークは複数個存在する。最悪の場合ではその数は非常に大きな数になる。
- パラメータ探索：同じような構造のネットワークでも、遺伝子間の依存関係の強さなどのパラメータを発見することも、構造の検討と同時に行われる必要があり、このこと自体でも膨大な計算量が必要となる。
- データのノイズと信頼性：データにノイズなどの誤差が多く含まれており、実験手法などによって値が変わってしまうなどの信頼性の問題。

があげられる [1]。

現在、多くの研究ではデータのノイズの問題は扱っていない。理由としては、今後測定手法の精度が向上することが予測できるという事と、ノイズの問題は他の問題を解決する事と独立に取り組めることがあげられる。

そのため、ベイジアンネットワークを除く２つの手法は実際に生物から測定されたデータではなく、理想的なデータを用いて推定を行っている。現在、ブーリアンネットワークでは組み合わせ爆発の問題を解消する研究が中心に行われている。また、S-sytem ではパラメータ探索を遺伝的アルゴリズムを用い依存関係に関わるパラメータを導出している。

1.3 本研究の目的について

生物学的知識から依存関係を持つ遺伝子はある特異的な文字列パターンを転写制御領域内に持つことが知られている。文字列パターンとは転写制御領域を文字列として扱う場合に、転写制御領域の部分文字列の事を指す。また、文字列パターンを持つ遺伝子とは、その遺伝子の転写制御領域内に文字列パターンを持つ事を指す。しかし、特異的な文字列パターンを持つ遺伝子の発現量がその文字列パターンに結合する転写因子を生成する遺伝子の発現量に影響を受けているかは知られていない。本研究では調節遺伝子が発現して生成する転写因子が結合することが既知である文字列パターンを持つ遺伝子を調査し、破壊株データを用いて影響を受けている事を検証する。この事により依存関係推定における転写制御領域の解析に対する有意性を示す事を目的とする。本研究の目的は遺伝子間の依存関係推定において転写制御領域を考慮した推定方法に対する基礎的部分を担うものである。遺伝子間の依存関係推定は細胞内の状態変化を定性的な解析等に利用する事ができる。細胞内の状態変化の解析は新薬等の開発等の様々な利用法が考えられる。本研究では対象となる生物として全遺伝子の塩基配列と DNA 上での場所が既知である枯草菌を用いる。

本研究の手法としては DNA の塩基配列データから調節遺伝子毎に結合することが既知である文字列パターンを持つ遺伝子を調査し、破壊株データを用いて影響を受けている事

を統計学的手法から調査する。統計学的手法としては調節遺伝子と全遺伝子の相関係数を求めて、影響を受けていると考えられる遺伝子群の相関係数の平均値と全遺伝子の相関係数の平均値を比較し、影響を受けていると考えられる遺伝子群の平均値の方が高い場合には平均値の検定を用いて各々の平均値の差が有意である事を検証する。この手法により調節遺伝子との依存関係を検証する。本研究で使用するデータについては 2 章で説明する。調節遺伝子が発現して生成する転写因子が既知に結合する文字列パターンを上流部分に持つ遺伝子を調査する方法については 3 章で説明する。また、相関係数の導出、平均値の検定方法については 4 章で説明する。実験を行った結果については 5 章で説明する。本研究での結果に対する考察は 6 章で説明する。

第2章　対象となる菌類と使用するデータについて

本章では本研究で依存関係推定の対象となる菌類と使用したDNA塩基配列データと破壊株データについて説明する。

今回使用する枯草菌のDNA配列データはNational Center for Biotechnology Information (NCBI)[8]で公開されているデータを使用した。また、統計的手法に用いる破壊株データについて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学講座から99種類の破壊株データを使用した。

2.1　枯草菌について

枯草菌（Bacillus subtilis）は原核生物であり、好気性、毒性の無いグラム陽性の胞子形成菌である。土壌など自然界に広く分布し、環境条件に伴い内生胞子を形成して休眠する。食材として身近な納豆を作る納豆菌は枯草菌の一種である。

日本とヨーロッパの30以上の枯草菌研究室の国際共同研究により、1997年に枯草菌のゲノムの全塩基配列が決定され「Nature」誌に発表された。枯草菌のDNAは420万塩基配列を持ち、DNA上に4,100以上の蛋白質遺伝子と5つRNA遺伝子が存在する事が明らかになった。蛋白質遺伝子とは生体機能に関わる蛋白質を発現する遺伝子の事であり、RNA遺伝子とはmRNA以外のRNA（rRNA,5SRNA,tRNA)を発現によって生成する遺伝子の事である。rRNAは翻訳の段階でコドンを遺伝子暗号を用いて解読するリボソームの構成要素として存在するRNAであり、5SRNAもリボソームの構成要素である。tRNAはトランスファーRNAの事を指し、翻訳の段階でアミノ酸を輸送する物質である。蛋白質遺伝子の半数は機能が既知であったり、他の蛋白質との類似性から機能推定は可能だが残り半数については機能は未知である[6]。

現在、NCBIでBacillus subtilis Marburg 168株のDNA配列データが公開されており、Bacillus Subtilis Operon Reading Frames database(BSORF)[9]で転写地図等も公開されている。転写地図とはDNA上のどの部分にどの遺伝子が存在するかを示したものであり、転写方向も表している。また、BSORFでは各遺伝子のDNA配列等も参照することができる。

枯草菌は毒性が無い事と、外部からのDNAを取り込み、自身のDNA上での塩基配列の似た部分と取り替えるというDNA形質置換を性質に持つ。この性質を利用する事によ

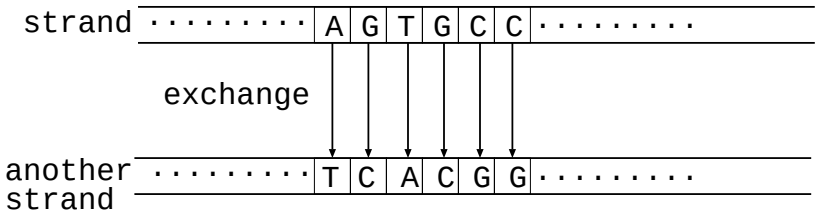

図 2.1: ストランドの変換について

り、遺伝子部分の DNA を実験的手法によって破壊し、発現機能を欠失させた破壊株を作成することが他の生物に比べ容易である。そのため、標的となる遺伝子を破壊した枯草菌と破壊しない枯草菌に対して様々な実験を行い、実験データを比較して破壊された遺伝子が発現して生成される蛋白質の機能の特定等が行われている。

2.2 DNA 配列データについて

デオキシリボ核酸（DNA）はアデニン (A)、グアニン (G)、シトシン (C)、チミン (T) の４つの塩基からなるストランド (strand) が二重螺旋構造をしている物質であり、細胞の核内に存在する。ストランドとは DNA の二重螺旋構造の片側部分を指す。両方のストランドは A と T、C と G が互いに結合して DNA の二重螺旋構造を構成している。

DNA 配列データは各塩基の頭文字 A,G,C,T で構成される文字列のデータである。NCBI で公開されている枯草菌の DNA 配列データは一つのストランドだけであるため転写制御領域を調べるにあたり、A を T、G を C に変換してもう一方のストランドを生成した（図 2.1)。DNA の構造上からストランド毎に転写方向は逆方向になる事が知られている。

今回、遺伝子の場所を特定するために NCBI からの各遺伝子の場所やストランドの種類を示したデータを用いた。枯草菌は転写地図が作成されているため、各遺伝子の場所、長さ、どちらのストランド上に遺伝子が存在するか等が既知である。表 2.1 はこの様なデータファイルの一部であり、Location はその遺伝子が DNA 上のどの部分に存在するかを表している。2 行めの遺伝子 dnaA は 410bp から 1750bp までに存在する事を意味する。bp は塩基配列 1 つを示す単位である。Strand は + の場合は NCBI で公開しているストランド上、– では NCBI で公開しているストランドのもう一方ストランド上に遺伝子があることを示している。Length はその遺伝子の長さ、PID は遺伝子が発現して生成される蛋白質の認識番号、Gene は遺伝子名、Synonym Code は遺伝子番号を表している。また、Synonym Code の Bsu は Bacillus Subtilis の略である。今回はこのようなデータを使用して各遺伝子の場所とストランドを特定し上流部分を調査した。

Location	Strand	Length	PID	Gene	Synonym Code
410..1750	+	446	16077069	dnaA	Bsu0001
1939..3075	+	378	16077070	dnaN	Bsu0002
3206..3421	+	71	16077071	yaaA	Bsu0003
3437..4549	+	370	16077072	recF	Bsu0004
4567..4725	+	52	16077073	yaaB	Bsu0005
4866..6782	+	638	16077074	gyrB	Bsu0006
6993..9458	+	821	16077075	gyrA	Bsu0007
14845..15792	−	315	16077076	yaaC	Bsu0008
15913..17379	+	488	16077077	guaB	Bsu0009

表 2.1: 遺伝子の場所、長さ、ストランド等を記すファイルの内容の一部

2.3 破壊株データについて

破壊株とは実験的手法で標的となる遺伝子を破壊した株のことを指す。また、遺伝子を破壊していない株の事を野生株と呼ぶ。また、株とは遺伝的形質が同じ生物の別の個体の事を指す。本研究で破壊株データは破壊株、野生株の各発現量を比較するために観測したDNA マイクロアレイデータの事を指す。DNA マイクロアレイデータとはマイクロアレイ法で測定されたデータである。マイクロアレイ法はスライドガラス上に数千から数万の遺伝子の塩基配列（mRNA の鋳型）を高密度に配列させる。次に二つの対象となる生物の mRNA 調節した後に生物毎に mRNA に蛍光色を付着させる。そして付着させておいた鋳型に結合させ、余分な mRNA は洗浄する。蛍光色を付着させているため、画像としてデータを保存し画像処理の技術を使用してそれぞれの生物の mRNA の蛍光強度を測定する [4]。図 2.2 は以上の手順を図にしたものである。本研究ではマイクロアレイ法で観測された蛍光強度を発現量とする。現在、DNA マイクロアレイ技術の支流は Cy5, Cy3 の二つの蛍光色を用いて主に野生株の測定対象と同生物の破壊株の両方の遺伝子の発現量を測定する方法である。

実際に使用したデータは実験による実データから実験条件等によって生じた誤差を補正し、各遺伝子毎に野生株と破壊株の発現量の比を計算した補正データを使用した。

破壊株 X における遺伝子 i の発現量の比の値は

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

となる。補正データにおいて対象となる調節遺伝子 X を破壊したデータ上ではその調節遺伝子の影響を受ける事が既知である遺伝子 i に関して受ける影響に対して約 8 割の遺伝子が $ratio_i^X$ と受ける影響に一致する。この発現量の比を全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。そのため、今回は対数をとることにより正規化した値を用いた。この様な変換は DNA マイクロアレイを用いて解析を行

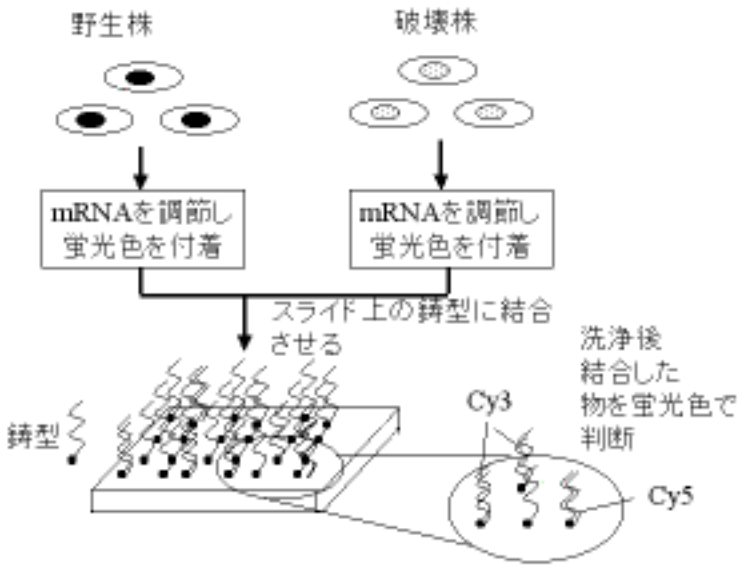

図 2.2: マイクロアレイ法

う時によく用いられる。今回使用したデータは 2spot/1gene で観測されたデータである。すなわち、ある破壊株のデータ内で一つの遺伝子に対して二つの実験値を持つデータである。二つの値を独立した実験データとして扱っても良いため、今回は正規化した値の絶対値が大きい方のデータを使用した。$ratio_i^X$ に対し今回は閾値 t を設け以下のような値を取るデータのみを使用するという方法を用いた。

$$\log(ratio_i^X) > \log(t)$$

$$\log(ratio_i^X) < -\log(t)$$

閾値を設けて使用するデータを選択する理由としては生物学的に調節遺伝子の発現に影響を受ける遺伝子数は少ない事とデータ誤差による値の分散を考慮する事が挙げられる。今回は標的となる遺伝子を破壊したデータと破壊しないデータの各発現量の比を用いるが、標的となる遺伝子が調節遺伝子である場合には少数の遺伝子にしか影響を与えないため、データ上の多数の遺伝子に関しては 1 になる。しかし、データの誤差により影響を受けていない遺伝子の値は 1 にはならずに 1 の周辺にばらついた値となる。そのため、今回は 1 の周辺の値に関しては閾値 t を設けて除外し、影響を受けていると判断できるデータを各遺伝子毎に使用するデータとして選択した。

今回、統計的手法において閾値外のデータを除外したものとしないもの両方を使用した。 以後、閾値を設けて使用するデータを選択した場合を「データ選択」、データの選択をしない場合を「データ非選択」とする。

以上のデータを使い既知に転写因子が結合する文字列パターンを持つ遺伝子を調査し、結合する転写因子を生成する調節遺伝子との依存関係の可能性の有無を調査した。

第3章　実験1：パターンマッチングを用いた全遺伝子の上流部分の調査

本研究ではBacillus subtilis Marburg 168株のDNA配列データに対し全遺伝子の上流部分に既知の文字列パターンを含むかどうかをパターンマッチングを用いて調査した。本研究で文字列パターンとは転写制御領域を文字列として扱った場合の部分文字列のことを指す。また、文字列パターンを持つとは転写制御領域内にその文字列パターンを含むことを意味している。仮にある遺伝子の上流部分に転写因子と結合する事が既知である文字列パターンを含む場合には、その遺伝子は文字列パターンを持つとして調節遺伝子と依存関係を持つ候補となる。

DNAを構成する物質で塩基部分であるアデニン、グアニン、シトシン、チミンをそれぞれA,G,C,Tの文字として扱い、上流部分を文字列として扱う。また上流部分の文字列に対して文字列パターンとのマッチングを行った。

今回、全遺伝子の上流部分に各文字列パターンが存在するか以下の手順で調査した。

1. DNA配列データと各遺伝子の場所を記したデータから全遺伝子の上流部分を両方のストランドから抜き出す。
2. 調べる対象となる文字列パターンを決定する。
3. 全遺伝子の上流部分に対してパターンマッチングを行う。
4. 対象となる文字列パターンを変え3を行う。

以下に各手順について説明する。

3.1　DNA配列データの上流部分について

今回は全遺伝子に対して対象となる文字列パターンの調査を両方のストランドを用いて行った。上流部分を抜き出すためにDNA配列データと各遺伝子の場所と、どのストランド上に遺伝子が存在するかを記したデータを用いた。遺伝子の開始点と転写方向により上流部分を特定した。

また上流部分として抜き出す範囲としては転写開始点を0としたときに-500bp までとした（図3.1 a)。既知の文字列パターンが-500bp以上の上流部分に存在しなかったため、調査する上流部分の範囲を-500bpと決めた。遺伝子間が-500bpに満たない場合はその範囲を上流部分として調査し（図3.1 b)、今回使用した既知の文字列パターンの最小の長さ

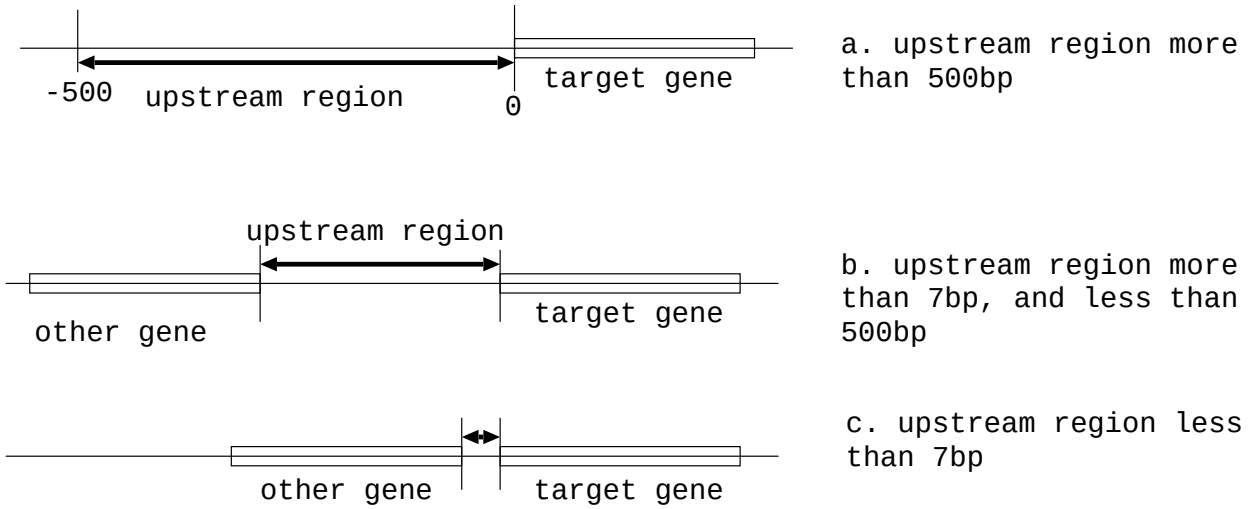

図 3.1: 各遺伝子の上流部分の調査について

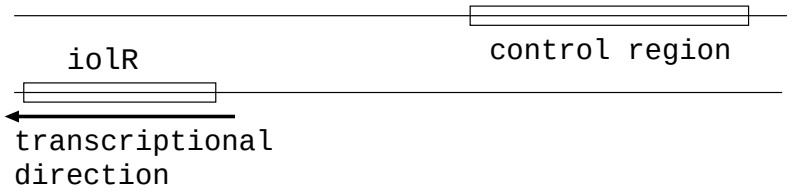

図 3.2: iolR の転写制御領域

が 7 bp なため、それ以下の遺伝子間には転写制御領域は存在せず、別の遺伝子のオペロンの一つであるとし、調査は行わない（図 3.1 c)。

DNA は二つのストランドから構成されるため、文字列パターンの調査は遺伝子の上流部分を両方のストランドから調査しなければならない。転写因子 IolR を発現する遺伝子 iolR は自身の転写制御領域にも結合し発現を抑制するが、iolR あるストランドとは別のストランドの上流部分に転写制御領域が存在する事が生物学的に知られているからである [13]。（図 3.2 参照)。

今回、以上の方法で両方のストランドから上流部分を抜き出した。また、抜き出した上流部分に対して文字列パターンを含むかどうかパターンマッチングを用いて調査を行った。

3.2 文字列パターンについて

ある調節遺伝子から発現される転写因子が結合する転写制御領域の文字列パターンは既知に影響を受ける遺伝子毎に多少異なる場合がある。これは各調節遺伝子が発現して生成する転写因子の立体構造の違いが原因である。

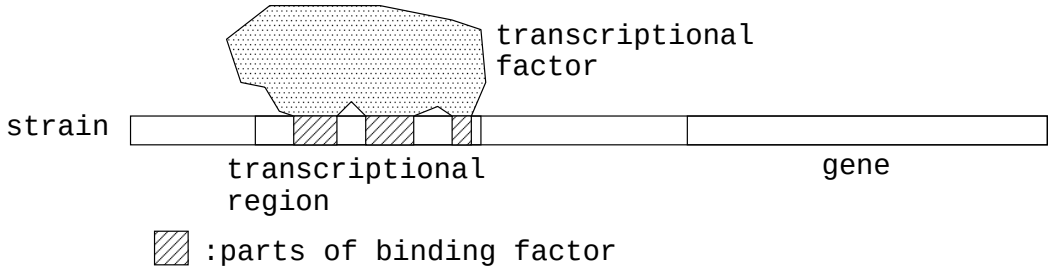

図 3.3: 転写因子の構造による結合する部分と転写制御領域

表 3.1 に調節遺伝子 glnR が発現して生成する転写因子 GlnR に既知に影響を受ける遺伝子、影響、結合する文字列パターンを示す。Negative は発現を抑制することを意味する。

遺伝子	制御	文字列パターン
nasA	Negative	TGTCANNNNNNNTTACA
nasB	Negative	TGTAANNNNNNNTGACA
ureA-P3	Negative	TGTAANNNNNNNTAACA
ureA-P3	Negative	TGTGANNNNNNNTAACA
glnR	Negative	TGTTANNNNNNNTTACA
glnR	Negative	TGACANNNNNNNTAACA

表 3.1: glnR が結合して発現に影響を与える既知の遺伝子とその転写制御領域

表 3.1 での 4 行目と 5 行目の ureA-P3 は遺伝子 ureA が変異した遺伝子である。また、glnR は依存関係が既知である遺伝子全てに対して負の影響を与えている。また各遺伝子の文字列パターンの中の N については A、G、T、C のどの配列が当てはまっても良いことを表している。これは図 3.3 の様に転写因子の立体構造によるものだと考えられている。

この部分以外の部分に転写因子は結合する。これは DNA の二重螺旋構造に起因していると考えられる。また各遺伝子の文字列パターンの一意な部分は各々の遺伝子によって異なるが比較的類似したパターンになっている。

遺伝子の転写制御領域内に転写因子が結合する幾つかの特定の文字列パターンを一つでも含んだ時に本研究では文字列パターンを持つとする。また、文字列パターンを持たないとは転写因子が結合する全ての特定の文字列パターンを転写制御領域に含まない事を意味する。

ある一つの転写因子が結合する事が既知である文字列パターンを持つ遺伝子の中には同じパターンであるにも関わらず異なる影響を受けるものも存在する。転写因子に既知に影響を受ける遺伝子でもこの様な場合があるため、今回は文字列パターンを持つ依存関係未知の遺伝子についてはその文字列パターンから発現に対して受ける制御を決定しない。

今回使用する文字列パターンは転写因子を生成し他の遺伝子の転写制御領域に結合する事が既知である調節遺伝子 16 種類に対してそれぞれ転写因子に結合する既知の文字列パターンを DBTBS[10] より決定した。

表 3.2 に今回使用した既知に転写制御領域に結合する転写因子を発現する 16 種類の調節遺伝子名と発現して生成される転写因子の働きを示す [14]。

調節遺伝子名	発現物質の働き
araR	transcriptional repressor of the arabinose operon(araABDLMNPQ)
ccpA	transcriptional regulator involved in carbon catabolite control
comA	two-component response regulator[ComP] or late competence genes surfactin production.
degU	pleiotropic regulator involved in various post-exponential phase responses makes a two component system with DegS kinase
deoR	transcriptional repressor of the dra/nupC/pdp operon(deoxyribonucleoside)
gerE	transcriptional regulator required for expression of late spore coat genes
glnR	transcriptional repressor involved in the expression of the phosphotransferase gene (glnA)
iolR	transcriptional repressor of the myo-inositol catablism operon (iolABCDEFGHIJ/iolRS)
phoP	two-component response regulator[PhoR] involved in phosphate regulation(phoA,phoB,phoD,resABCDE)
purR	transcriptional repressor of the purine operon (purEKBCLQFMNHD)
rocR	transcriptional activator of arginine utilization operons(rocABC,rocDEF)
senS	transcriptional regulator of extracellular enzyme genes(amyE,aprE,nprE)
spo0A	two-component response regulator[KinC] central for the initiation of sporulation(spo0A,abrB,kinA,kinC,spoIIA,spoIIE,spoIIG)(part of phosphorelay:Spo0B〜P→Spo0A〜P)
spoIIID	transcriptional regulator of sigE-and sigK-dependent genes
tnrA	transcriptional pleiotropic regulator incolced in global nitrogen regulation(expression of nrgAB, nasB, gapP, ureABC,glnRA)
xylR	trabscriptional repressor of the xylose operon (xylAB)

表 3.2: 今回使用した調節遺伝子とその働き

今回使用した 16 種類の調節遺伝子の中にも文字列パターンが GlnR での nasA の様に N の部分が存在するものもある。そのため、今回のパターンマッチングでは N の部分が既知である文字列パターンについては N の部分長さを考慮して行った。

3.3　パターンマッチングについて

今回は −500bp で抜き出した上流部分の文字列内に対する文字列パターンの調査については従来のアルゴリズムでのパターンマッチングを行った。パターンを両方のストランドにおける文字列に対してマッチングを行う。文字列にパターンがマッチした場合、その遺伝子が依存関係未知であれば、その遺伝子はその特定の文字列パターンを持つとし、対象となる調節遺伝子との依存関係の候補となる。

以上の方法で 16 種類の転写因子が結合する既知の文字列パターンを用いて全遺伝子の上流部分をパターンマッチングを用いて転写因子を生成する遺伝子と依存関係のある遺伝子の候補を調査した。

第4章　実験2：統計的手法を用いたDNAマイクロアレイデータの解析

本章では前章で転写因子が結合することが既知である文字列パターンを持つ遺伝子群に対して調節遺伝子との依存関係を持つの可能性をDNAマイクロアレイデータを用いて調査する統計的手法について説明する。

本研究では遺伝子間に依存関係があることを調べる方法の一つとしてDNAマイクロアレイデータを用いて相関係数を求める。16種類の調節遺伝子と他の遺伝子との相関係数を求めた後、平均値の検定を用いて遺伝子間の依存関係について調べるする。

4.1　相関係数

本研究では調節遺伝子と他の遺伝子との依存関係調べるためにをDNAマイクロアレイデータを用いて相関係数を求めた。調節遺伝子 x と他の遺伝子 y の n 種類のDNAマイクロアレイデータ上の値をそれぞれ x_i、$y_i(i = 1, \ldots, n)$ とすると相関係数は以下の式から計算される。

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

相関係数 r_{xy} は $-1 \leq r_{xy} \leq 1$ の範囲の値を取る。r_{xy} が 0 より大きいの時に正の相関、0未満の時に負の相関、0の時相関がないと判断される。また、$r_{xy} = 1$ の時に非常に強い正の相関、$r_{xy} = -1$ の時非常に強い負の相関と判断される。相関係数では符合は相関の向きであり、値が相関の強さを表す。図4.1は強い正の相関、強い負の相関の時のx,yの値をプロットしたものを表している。データにノイズ等の誤差がなく、他の転写因子に影響せず、発現量に対して時間を考慮しない場合、ある転写因子を生成する遺伝子の発現量とパターンに結合し発現を活性化される事が既知である遺伝子の発現量は高い正の相関

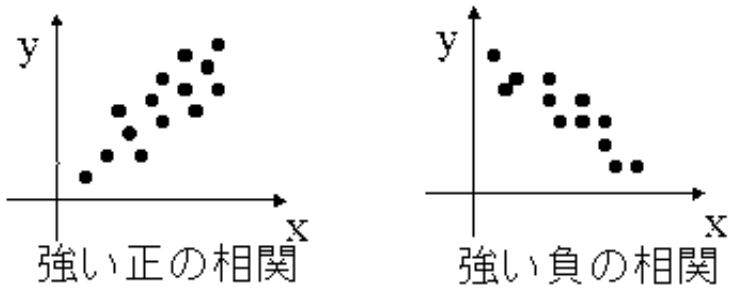

図 4.1: 強い正の相関と負の相関の場合について

があると考えられる。また、既知に抑制される遺伝子の発現量とは高い負の相関があると考えられる。

調節遺伝子毎に既知に影響を受ける遺伝子との相関係数の符合と受ける影響は約 0.83 の割合で一致した。これは転写因子が転写制御領域に結合し、発現を活性化させた場合には相関係数は正の値になり、逆に発現を抑制した場合には負の値の相関係数に高い確率でなる事を示している。

4.2 平均値の検定

平均値の検定は、標本として取り出した集団の値の平均値が母集団の平均値と異なる事を検定する方法である。検定は以下の手順で行う。

1. 仮定の設定
2. 統計量を求める
3. 確率を求める
4. 判定

それぞれの手順について説明する。

手順 1 仮定の設定

「母集団の平均値と標本の平均値とは異なる」という仮説は違いの程度を特定しないと検定できない。今回は以下のような仮定を行う。

- 帰無仮説 H_0:「母集団の平均値と標本の平均値は等しい」
- 対立仮説 H_1:「母集団の平均値と標本の平均値は異なる」

検定結果が帰無仮説 H_0 を棄却する判定した場合に対立仮説 H_1 を採択し母集団の平均値は標本の平均値と異なると判定する。

手順 2 統計量を求める

母集団の平均値を μ、標準偏差を σ としたときに無作為に取り出した標本の平均値の分布から対象となる標本の平均値がどれだけ偏っているかを調べる統計量 z を計算する。母集団のデータ数が少ない場合には正規分布に従う時、今回の様に扱うデータ数が約 4000 と大きい場合は母集団がどのような分布でも無作為に取り出した n 個の標本の平均値の分布は

$$E(\bar{x}) = \mu$$

$$Var(\bar{x}) = \frac{\sigma^2}{\sqrt{n}}$$

に従う正規分布になる（図 4.2)。

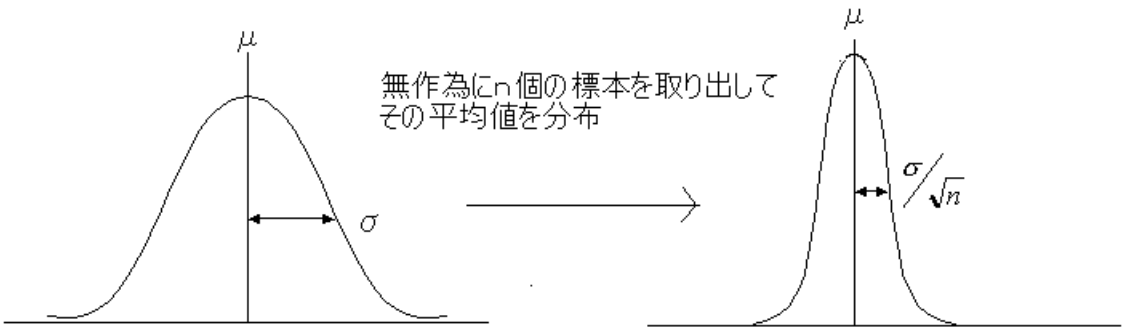

図 4.2: 無作為に取り出した標本の平均値の分布

対象となる標本の標本数 n、平均 $\bar{x}$ の場合、平均値の分布上での $\bar{x}$ の値を標準化すると、

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{n}}$$

となる。標準化とは正規分布を標準正規分布に変換する方法である。X が平均 μ'、標準偏差 σ' の正規分布に従う場合、z' は標準正規分布に従う。

$$z' = \frac{X - \mu'}{\sigma'}$$

標準化した z を用いて標本の平均値 $\bar{x}$ の偏り度を表す確率 P を計算する。

手順 3 確率を求める

手順 2 で求めた統計量 z を用いて $\bar{x}$ の偏り度の確率 P を計算する。確率 P は標準正規分布の両側確率から計算される。P についての式を以下に記す。

$$P = 2 \times (1 - \text{標準正規分布上での } z \text{ の値})$$

この P の値から帰無仮説を棄却できるかどうか判断する。

手順 4:判定

$P \leq \alpha$ の場合に H_0 は棄却され、標本の平均値が母集団の平均値とは偏ると判定する。今回は優位水準 α は 0.05 として検定を行った。

相関係数と平均値の検定を用いて調節遺伝子と調節遺伝子が生成する転写因子が結合する特定の文字列パターンを持つ遺伝子群との関係を調べる。

仮に標本である遺伝子群が負の相関係数のみの場合には母集団と平均値を比べる時に平均値の絶対値を取り、差を比較する。これは相関係数の平均値の強さを比べるためである。また、標本に正と負の両方の相関係数を含む場合には相関係数を全て絶対値に変換して平均値を計算して、同様に相関係数を絶対値に変換した母集団の平均値と比較する。これは標本内の相関の強さの平均値と母集団の相関の強さの平均値を比べるためである。

この様な標本の相関係数の平均値が全遺伝子群である母集団の平均値よりも高く、各々の平均値が異なると判定された場合にはその標本となる遺伝子群は全遺伝子群の中でも高い相関係数を持つ遺伝子を集めた標本である。ゆえに、その標本となった遺伝子群は調節遺伝子と相関関係を持つ可能性がある。

今回、以下に該当する遺伝子群を標本とした。

- 文字列パターンを持つ依存関係既知の遺伝子群、
- 文字列パターンを持つ依存関係既知の遺伝子のオペロン
- 文字列パターンを持つ全ての遺伝子群
- 文字列パターンを持つ依存関係未知の遺伝子群

これら標本において高い相関関係を持つ場合、標本は文字列パターンを持つため調節遺伝子と標本である遺伝子群との間に依存関係を持つ可能性があると判断する。

以上の標本で母集団との平均値について調べ、標本の平均値が高い場合に検定を行う。

第5章　結果

5.1　文字列パターンを持つ遺伝子について

16 種類の調節遺伝子を対象に各文字列パターンを持つ遺伝子を調べた結果を表 5.1 に記す。

調節遺伝子名	文字列パターンを持つ遺伝子数	文字列パターンを持つ依存関係未知の遺伝子数
araR	3	0
ccpA	21	7
comA	5	3
degU	208	201
deoR	1	0
gerE	31	23
glnR	30	28
iolR	2	0
phoP	11	5
purR	9	2
rocR	3	0
senS	2	1
spo0A	1519	1510
spoIIID	76	71
tnrA	26	20
xylR	2	0

表 5.1: 各調節遺伝子での文字列パターンを持つ遺伝子数

調節遺伝子 araR, deoR, iolR, rocR, xylR に関しては文字列パターンを持つ依存関係未知の遺伝子は見つからなかった。これらの遺伝子が発現して生成する転写因子が結合する文字列パターンは特異的である事を意味している。また、degU、spo0A に関しては、文字列パターンを持つ遺伝子数が非常に多い。degU に関しては発現して生成される DegU が結合する文字列パターン内に一意である配列が少なく、DNA の二重螺旋構造により A,G,T,C

の全ての配列でも許す転写因子が結合しない部分であるNが多いため多数の遺伝子の上流部分に存在することが理由に上げられる。また、spo0Aに関しては文字列パターン自体が短い事が理由にあげられる。

5.2 統計的手法によるDNAマイクロアレイデータの解析結果

本節では4章で紹介した統計的手法を用いてDNAマイクロアレイデータの解析結果について説明する。

本研究では調節遺伝子と他の遺伝子の関係を調べるためにDNAマイクロデータの値から相関係数を求めた。生物学的に調節遺伝子によって発現量に影響を受ける遺伝子は少数であるため相関係数が低い値となる遺伝子は多い。そのため調節遺伝子に対して全遺伝子の相関係数の平均値は非常に低くなる。ゆえに標本となる遺伝子群の相関係数の平均値が母集団である全遺伝子の平均値と同じ、もしくは低い場合にはその標本である遺伝子群は調節遺伝子と相関関係を持たない事になる。

また、標本である遺伝子群の平均値が母集団である全遺伝子の平均値と比べ高い場合でもその差が非常に小さいものであった場合には全遺伝子に対して標本の方が調節遺伝子と相関関係を持つとは言い切れない。そのため、標本と母集団の平均値の差が有意である事を調べなければならない。

ゆえに標本の平均値が母集団の平均値より高い場合には平均値の検定を行い、各々の平均値に差があると判定された場合にはその差が有意であるため、標本の平均値は母集団の平均値より高いため標本となる遺伝子群は調節遺伝子は相関関係を持つ可能性がある。今回は標本を文字列パターンを持つ遺伝子群とするため、標本と調節遺伝子が相関を持つ可能性がある場合、生物学的知識から調節遺伝子が発現して生成される転写因子はパターン部分に結合して発現に影響を与えるため、標本となる遺伝子群には依存関係を持つ可能性があると判断する。

以上の方法により遺伝子間の依存関係推定において文字列パターンを持つ遺伝子が調節遺伝子と依存関係を持ち、発現量に影響を受けることを検証する。そのため、文字列パターンを持つ遺伝子群に対して依存関係を持つ可能性の有無を調節遺伝子毎に調査した。

文字列パターンを持つ依存関係既知の遺伝子群やオペロンを標本にした理由としては、データでのノイズや転写が始まる前に測定した等の要因から今回の手法で依存関係が既知である遺伝子群がDNAマイクロアレイデータの解析結果で事実と一致する調節遺伝子の数を調べるためである。また、依存関係が既知である遺伝子群やオペロン群の調査結果は文字列パターンを持つ全ての遺伝子群や文字列パターンを持つ依存関係未知の遺伝子群との結果の比較に用いる。

文字列パターンを持つ依存関係既知の遺伝子群とオペロンを含んだ遺伝子群は調節遺伝子が発現によって生成した転写因子が各遺伝子の発現に与える影響が既知であるため、影

響毎に分けて依存関係を調査出来るが、文字列パターンを持つ依存関係未知の遺伝子群に対しては影響毎に分けることが出来ない。この様な場合、調節遺伝子と他の遺伝子の相関係数を絶対値に変換する。相関係数を絶対値に変換する理由としては、標本を影響毎に分けられないため、標本内の遺伝子の相関係数は正の値、負の値共に存在する。そのため、非常に高い正の相関係数を持つ遺伝子と非常に高い負の相関係数を持つ遺伝子が存在した場合に平均値を計算すると低い値になってしまう。故に絶対値に変換することにより、正、負の相関に関わらず値の高さで依存関係を持つ可能性の判断を行うためである。

また平均値の差についても影響が負である標本の場合には平均値の絶対値と母集団の平均値の絶対値で以下の式によって計算する。

$$差 = |標本の平均値| - |母集団の平均値|$$

これは母集団に対して標本が高い相関係数を持つことを調べるため、正、負の相関で判断するのではなく、値の高さで標本と母集団を比較するためである。

以下に各標本での調査結果を記す。

5.2.1 文字列パターンを持つ依存関係既知の遺伝子を標本にした場合

各調節遺伝子毎に標本を以下の様に分ける。

- 調節遺伝子から受ける影響が活性である遺伝子
- 調節遺伝子から受ける影響が抑制である遺伝子
- 依存関係が既知である全ての遺伝子

また、母集団は全遺伝子となる。

文字列パターンを持つ依存関係既知の遺伝子の発現量は各調節遺伝子の発現量に影響を受ける。そのため、データ上のノイズ等の誤差が無い場合調節遺伝子と依存関係既知の遺伝子との相関係数は依存関係を持たない遺伝子の相関係数よりも高い値を示すと考えられる。

また、標本が依存関係が既知な全ての遺伝子である場合では正に高い相関係数を持つ遺伝子と負に高い相関係数を持つ遺伝子を含んだ場合に平均値は低い値を示してしまうため、標本、母集団共に相関係数の絶対値を取った値で調査した。

各調節遺伝子毎に標本と母集団の平均値の差を調べ、標本の方が高い相関係数である場合に平均値の検定を行った。この場合、母集団の平均値と標本の平均値が異なると判定出来れば標本は調節遺伝子と依存関係を持つ可能性があると判断する。

また、DNA マイクロアレイデータは全遺伝子に対して測定されていないため、今回のデータで測定されていない文字列パターンを持つ依存関係既知の遺伝子は対象から除外した。また、前章の調査結果で senS は文字列パターンを持つ依存関係既知の遺伝子が senS のみであったため対象から除外した。

調節 遺伝子名	平均値の差 （活性）	平均値の差 （抑制）	平均値の差 (全体)	判定 （活性）	判定 （抑制）	判定 （全体）
araR	-	0.32	0.19	-	○	○
ccpA	0.30	0.25	0.096	×	○	○
comA	0.67	-	0.51	○	-	○
degU	0.15	0.04	0.02	×	×	×
deoR	-	0.15	0.01	-	×	×
gerE	0.53	-	0.43	○	-	○
glnR	-	0.26	0.13	-	○	×
iolR	-	0.27	0.12	-	×	×
phoP	0.25	0.09	0.098	○	×	○
purR	-	0.31	0.17	-	○	○
rocR	0.73	-	0.60	○	-	○
spo0A	0.24	0.04	0.19	○	×	○
spoIIID	0.40	0.17	0.19	○	×	○
tnrA	0.30	0.17	0.16	○	×	○
xylR	-	0.05	−0.05	-	×	検定しない

表 5.2: 標本を依存関係既知の遺伝子群とした場合のデータ選択での平均値の差と検定結果

表 5.2 はデータにデータを選択した場合、表 5.3 はデータを選択しない場合での調節遺伝子毎に標本と母集団の平均値の差と検定結果を示したものある。表の中の-は標本に該当する遺伝子が存在しないことを表している。また○は標本の平均値が母集団の平均値と異なる検定結果を意味し、×は標本の平均値は母集団の平均値と同じという検定結果を表したものである。

表 5.2、表 5.3 からデータ選択、データ非選択のどちらの場合でも標本が依存関係既知の遺伝子全体の時の xylR 以外では標本の平均値は母集団の平均値よりも高い。その様な調節遺伝子に対して平均値の検定を行った。標本と母集団の平均値が異なると判定された調節遺伝子に関しては依存関係既知の遺伝子群は他の遺伝子に比べ高い相関係数を持つと言える。この場合、既知の事実と DNA マイクロアレイデータでの解析結果が一致したと言える。

また、表 5.4 に標本が既知の事実と DNA マイクロアレイデータの解析結果が一致した調節遺伝子の割合をデータ別、標本別に割合を用いて示す。標本 i での○の数を Hit_i、全調節遺伝子数を All、標本に該当する遺伝子が存在しない調節遺伝子数を $Nothing$ とすると、割合は以下の式で求める。

$$\text{割合} = \frac{Hit_i}{All - Nothing}$$

調節 遺伝子名	平均値の差 （活性）	平均値の差 （抑制）	平均値の差 (全体)	判定 （活性）	判定 （抑制）	判定 （全体）
araR	-	0.20	0.10	-	○	×
ccpA	0.22	0.18	0.07	×	○	○
comA	0.56	-	0.45	○	-	○
degU	0.12	0.04	0.02	×	×	×
deoR	-	0.18	0.06	-	×	×
gerE	0.50	-	0.43	○	-	○
glnR	-	0.24	0.14	-	○	×
iolR	-	0.28	0.17	-	×	×
phoP	0.17	0.07	0.07	○	×	○
purR	-	0.23	0.16	-	○	○
rocR	0.62	-	0.53	○	-	○
spo0A	0.20	0.001	0.008	○	×	○
spoIIID	0.36	0.15	0.15	○	×	○
tnrA	0.29	0.15	0.14	○	×	○
xylR	-	0.04	−0.4	-	×	検定しない

表 5.3: 標本を依存関係既知の遺伝子群とした場合のデータ非選択での平均値の差と検定結果

	活性	抑制	全体
データ選択	0.78	0.33	0.667
データ非選択	0.78	0.33	0.60

表 5.4: DNA マイクロアレイデータの解析結果と既知の事実が一致した調節遺伝子の割合

結果としてはデータ選択、非選択ともに半数以上の調節遺伝子に関して今回使用したDNA マイクロアレイデータの解析結果と依存関係既知であり影響が活性である遺伝子群対して半数以上の調節遺伝子に対して既知の事実と一致した。

表 5.4 から文字列パターンを持つ依存関係既知の遺伝子群を標本にした場合にはデータ選択、データ非選択共に大きな差は無い。また、調節遺伝子の半数以上で依存関係を持つ可能性があるという結果になり、DNA マイクロアレイデータの解析結果と既知の事実が一致した。

5.2.2 文字列パターンを持つ依存関係既知の遺伝子のオペロンを標本にした場合

前節と同様にオペロンに対しても依存関係の調査を行った。また標本は依存関係既知の遺伝子群同様に以下の様に分けた。

- 調節遺伝子から受ける影響が活性であるオペロン
- 調節遺伝子から受ける影響が抑制であるオペロン
- 依存関係が既知である全てのオペロン

オペロンは一つの転写制御領域に結合した転写因子によって発現に影響を受ける遺伝子群の事であり、その文字列パターンを持つ遺伝子と同等の相関係数を示すと考えられるためである。

既知のオペロンについては B.Subtilis Operon Table [11] を参照した。今回も senS は B.Subtilis Operon Table でオペロンが見つからなかったため検定対象から除外した。今回も DNA マイクロアレイデータ上に無い遺伝子に関しては除外した。標本と母集団の平均値の差と平均値の検定結果を表 5.5、表 5.6 に示す。 表 5.5、表 5.6 共に標本を依存関係既知の遺伝子群の場合と同様に-は対象となる遺伝子が無い、○は平均値が異なるという検定結果、×は平均値は同じであるという検定結果を表している。

標本をオペロンにした場合にはデータ選択、データ非選択共に抑制で標本が依存関係を持つ可能性を持つ調節遺伝子の数が増えた。オペロンのみ依存関係を持つ可能性のある調節遺伝子に関しては文字列パターンを持つ依存関係既知の遺伝子より高い相関係数を持つオペロンが存在する事が言える。同時にデータ選択、非選択共に調節遺伝子 spo0A における抑制の影響をうけるオペロンを標本にした場合については標本の平均値よりも母集団の平均値の方が高くなってしまった。これは、文字列パターンを持つ依存関係既知の遺伝子より低い相関係数を持つオペロンが存在する事を意味している。

依存関係既知の遺伝子同様に標本であるオペロンに対して依存関係を持つ可能性がある調節遺伝子の割合を表 5.7 に示す。オペロンを含んだ事により標本に高い相関係数を持つ遺伝子が増えたため、依存関係をもつ可能性がある遺伝子が増えた。そのため、全体として割合はパターンを持つ依存関係既知の遺伝子群よりも増えた。

調節遺伝子名	平均値の差（活性）	平均値の差（抑制）	平均値の差（全体）	判定（活性）	判定（抑制）	判定（全体）
araR	-	0.22	0.14	-	○	○
ccpA	0.30	0.23	0.08	×	○	○
comA	0.68	-	0.52	○	-	○
degU	0.07	0.037	0.006	×	×	×
deoR	-	0.18	0.04	-	×	×
gerE	0.53	-	0.43	○	-	○
glnR	-	0.16	0.03	-	○	×
iolR	-	0.27	0.12	-	×	×
phoP	0.25	0.17	0.11	○	○	○
purR	-	0.29	0.15	-	○	○
rocR	0.72	-	0.59	○	-	○
spo0A	0.24	−0.004	0.07	○	検定しない	○
spoIIID	0.40	0.17	0.18	○	×	○
tnrA	0.24	0.23	0.12	○	○	○
xylR	-	0.07	−0.03	-	×	検定しない

表 5.5: データ選択での標本をオペロンとした時の母集団の平均値の差と検定結果

調節遺伝子名	平均値の差（活性）	平均値の差（抑制）	平均値の差（全体）	判定（活性）	判定（抑制）	判定（全体）
araR	-	0.17	0.11	-	○	○
ccpA	0.22	0.16	0.06	×	○	○
comA	0.56	-	0.45	○	-	○
degU	0.06	0.04	0.005	×	×	×
deoR	-	0.18	0.07	-	○	×
gerE	0.50	-	0.43	○	-	○
glnR	-	0.14	0.04	-	○	×
iolR	-	0.28	0.17	-	×	×
phoP	0.17	0.13	0.08	○	○	○
purR	-	0.20	0.14	-	○	○
rocR	0.63	-	0.53	○	-	○
spo0A	0.20	−0.019	0.10	○	×	○
spoIIID	0.36	0.15	0.16	○	×	○
tnrA	0.23	0.19	0.12	○	○	○
xylR	-	0.06	−0.025	-	×	検定しない

表 5.6: データ非選択での標本をオペロンとした時の母集団の平均値の差と検定結果

	活性	抑制	絶対値
データ選択	0.778	0.417	0.667
データ非選択	0.778	0.583	0.667

表 5.7: 標本をオペロンとした場合で依存関係を持つ可能性があると判断された調節遺伝子の割合

以上のDNAマイクロアレイデータの解析結果から文字列パターンを持つ既知の依存関係を持つ遺伝子のオペロン群は、半数以上の調節遺伝子に対して既知の事実と一致することが分かった。

5.2.3 文字列パターンを持つ全遺伝子を標本にした場合

標本をパターンを持つ全遺伝子群にした場合について依存関係の調査を行った。文字列パターンを持つ全遺伝子を標本にする場合には依存関係未知の遺伝子を含むため、依存関係既知の遺伝子群やそのオペロン群の様に活性、抑制に分けることが出来ない。そのため、調節遺伝子毎に全遺伝子に対して相関係数の値を絶対値に変換する。変換した母集団に対して平均値、標準偏差を計算した後に標本の平均値を再計算してから調査した。

表5.8に文字列パターンを持つ全遺伝子を標本にした場合の平均値の差、平均値の検定結果を記す。表5.8での○は平均値が異なるという検定結果を表しており、×は平均値が

調節遺伝子名	平均値の差（データ選択）	平均値の差（データ非選択）	判定（データ選択）	判定（データ非選択）
araR	0.188	0.108	○	×
ccpA	0.064	0.053	○	○
comA	0.286	0.253	○	○
degU	0.009	0.006	×	×
deoR	0.010	0.006	×	×
gerE	0.155	0.145	○	○
glnR	0.005	0.005	×	×
iolR	0.118	0.165	×	×
phoP	0.092	0.080	○	○
purR	0.147	0.133	○	○
rocR	0.598	0.529	○	○
senS	-0.088	-0.085	検定しない	検定しない
spo0A	-0.002	-0.002	×	×
spoIIID	0.010	0.015	×	×
tnrA	0.026	0.023	×	×
xylR	-0.053	-0.040	検定しない	検定しない

表 5.8: 標本を文字列パターンを持つ全ての遺伝子とした場合の平均値の差と検定結果

同じであるという検定結果を表している。

調節遺伝子毎に標本をパターンを持つ全ての遺伝子群にした場合、多くの調節遺伝子で平均値の差は低くなった。また、標本の平均値が母集団の平均値よりも低くなる調節遺伝

子の数も増えた。

表 5.9 にパターンを持つ全遺伝子を標本にした場合、標本が依存関係を持つ可能性があると判断された調節遺伝子の割合を示す。データ選択、データ非選択共に依存関係を持つと判断された調節遺伝子割合は変わらない。以上の結果から標本を文字列パターンを持つ

データ選択	データ非選択
0.4375	0.375

表 5.9: 文字列パターンを持つ全ての遺伝子群と依存関係を持つ可能性のある調節遺伝子の割合

全遺伝子した場合、依存関係既知の遺伝子群やオペロン群を標本にした場合よりも平均値が低くなる事が分かった。この事には文字列パターンを持つ依存関係未知の遺伝子群の中で低い相関係数を持つ遺伝子が存在することを意味している。

5.2.4 文字列パターンを持つ依存関係未知の遺伝子を標本にした場合

文字列パターンを持ち依存関係が未知である遺伝子に対して依存関係の調査を行った。文字列パターンを持つ全ての遺伝子群を標本にした時と同様に調節遺伝子毎に全遺伝子との相関係数を絶対値に変換して、母集団の平均値、標準偏差と標本の平均値を計算した。また、文字列パターンを持つ依存関係未知の遺伝子を持たない調節遺伝子に関しては除外した。表 5.10 に標本と母集団の平均値の差と平均値の検定結果を記す。文字列パターン

調節遺伝子名	平均値の差（データ選択）	平均値の差（データ非選択）	判定（データ選択）	判定（データ非選択）
ccpA	-0.001	0.003	検定しない	×
comA	0.062	0.059	×	×
degU	0.008	0.005	×	×
gerE	0.052	0.042	○	○
glnR	-0.004	-0.004	検定しない	検定しない
phoP	0.085	0.090	×	○
purR	-0.042	-0.042	検定しない	検定しない
senS	-0.088	-0.085	検定しない	検定しない
spo0A	0.001	0.002	×	×
spoIIID	0.0001	0.007	×	×
tnrA	-0.014	-0.016	検定しない	検定しない

表 5.10: 文字列パターンを持つ依存関係未知の遺伝子の平均値の差の検定結果

を持つ依存関係未知の遺伝子群を標本にした場合、半数近くの調節遺伝子に関して標本の平均値が母集団の平均値より低い値になった。また、データ選択では 1 つ、データ非選択では 2 つの調節遺伝子に対して標本である依存関係未知の遺伝子群と依存関係を持つ可能性があると判断する結果になった。ゆえに文字列パターンを持つ依存関係未知の遺伝子群との依存関係を持つ可能性のある調節遺伝子数が、依存関係既知の遺伝子群やオペロン群よりも少ないという結果になった。

5.3 各標本との比較

本節では各標本での平均値の検定結果を比較する。表 5.11 はデータ選択の場合、表 5.12 はデータ非選択の場合の調節遺伝子と他の遺伝子との相関係数を絶対値で変換した時の調査結果を示す。

調節遺伝子 xylR と senS については標本となる遺伝子を持つ場合において標本の平均値が母集団の平均値より低いため検定を行わなかったので、ここでは削除した。

	依存関係既知	オペロン	パターンを持つ全ての遺伝子	パターンを持つ依存関係未知
araR	○	○	○	-
ccpA	○	○	○	検定しない
comA	○	○	○	×
degU	×	×	×	×
deoR	×	×	×	-
gerE	○	○	○	○
glnR	×	×	×	検定しない
iolR	×	×	×	-
phoP	○	○	○	×
purR	○	○	○	検定しない
rocR	○	○	○	-
spo0A	○	○	×	×
spoIIID	○	○	×	×
tnrA	○	○	×	検定しない

表 5.11: データ選択での各標本の検定結果

表 5.11、表 5.12 では文字列パターンを持つ依存関係未知の遺伝子群以外の標本で依存関係を持つ可能性のある調節遺伝子が多いことがわかる。これは、標本が依存関係既知の遺伝子群やオペロン群よりも依存関係未知の遺伝子群の方が平均値が低い事を表している。しかし、依存関係未知の遺伝子群の全ての遺伝子が相関が低いわけではなく、既知の

	依存関係既知	オペロン	パターンを持つ全ての遺伝子	パターンを持つ依存関係未知
araR	×	○	×	-
ccpA	○	○	○	×
comA	○	○	○	×
degU	×	×	×	×
deoR	×	×	×	-
gerE	○	○	○	○
glnR	○	×	×	検定しない
iolR	×	×	×	-
phoP	○	○	○	○
purR	○	○	○	検定しない
rocR	○	○	○	-
spo0A	○	○	×	×
spoIIID	○	○	×	×
tnrA	○	○	×	検定しない

表 5.12: データ非選択での各標本の検定結果

遺伝子群よりも高い相関係数を持つ遺伝子も存在する。ゆえに、依存関係未知の遺伝子群に相関係数の低い遺伝子が多数存在している事が考えられる。次の章で文字列パターンを持つ依存関係未知の遺伝子群内に低い相関係数を含む原因について考察する。また、表 5.11 と表 5.12 を比較すると、データ選択、データ非選択では結果に大きな違いはない。これはデータを選択した場合に全体として相関係数が高くなり母集団の相関係数も高くなるため、結果としてはデータ非選択での場合と違いが無くなることを示している。今回の統計手法では二つのデータに大きな違いは生まれなかった。しかし、データ選択により相関係数が高くなる遺伝子や低くなる遺伝子も存在するため、別の手法によっては二つのデータに大きな違いが生まれる可能性があると考えられる。

第6章　考察

本研究ではDNA配列データからパターンマッチングを用いることにより文字列パターンを持つ遺伝子を調査した。調査後に文字列パターンを持つ遺伝子に関しては調節遺伝子と依存関係を持つ遺伝子の候補とした。その後、DNAマイクロアレイデータから調節遺伝子と他の遺伝子との相関係数を求めて、全遺伝子を母集団、転写因子が結合する文字列パターンを持つ遺伝子群を標本として、標本の相関係数の平均値が母集団の相関係数の平均値より高い場合には平均値の検定を行い、平均値が異なると判定された場合には標本は相関係数の高い遺伝子を集めたものである。その様な場合、文字列パターンを持つ遺伝子群を標本としているため、調節遺伝子と標本である遺伝子群との間に依存関係を持つ可能性があると判断した。

標本を文字列パターンを持つ既知の遺伝子群やオペロン群にした場合は半数以上の調節遺伝子が各標本である遺伝子群と依存関係を持つ可能性があるという結果になった。しかし、文字列パターンを持つ依存関係未知の遺伝子群では依存関係を持つ可能性のある調節遺伝子はデータ選択では1つ、データ非選択では2つのみであった。これは、文字列パターンを持つ依存関係既知の遺伝子群を標本にした場合やそのオペロン群を標本にした場合と大きく異なる。これは、依存関係未知の遺伝子群では平均値が低いことが原因であると考えられる。この様な遺伝子群を調査した結果、依存関係既知の遺伝子群の相関係数の平均値以上の相関係数を持つ遺伝子も存在する事がわかった。しかし、依存関係未知の遺伝子群の中に調節遺伝子との相関係数が低い遺伝子が多数あるため、平均値は低い値になってしまう。故に、依存関係未知の遺伝子群では調節遺伝子と依存関係を持つ可能性が無いという結果になる調節遺伝子が増えたと考えられる。

調節遺伝子との相関係数が低い依存関係未知の遺伝子に対して以下の事が考えられる。

- 他の未知の転写因子に影響を受けているため相関係数が低い
- 調節遺伝子との依存関係は無いため相関係数が低い

調節遺伝子との相関係数が低い文字列パターン持つ依存関係未知の遺伝子群について考察する。

6.1　他の未知の転写因子に影響を受けている場合

文字列パターンを持つ遺伝子の発現に影響を与える未知の転写因子がある場合を図6.1を用いて説明する。文字列パターンを持つ遺伝子に対してその転写因子が存在しなくとも

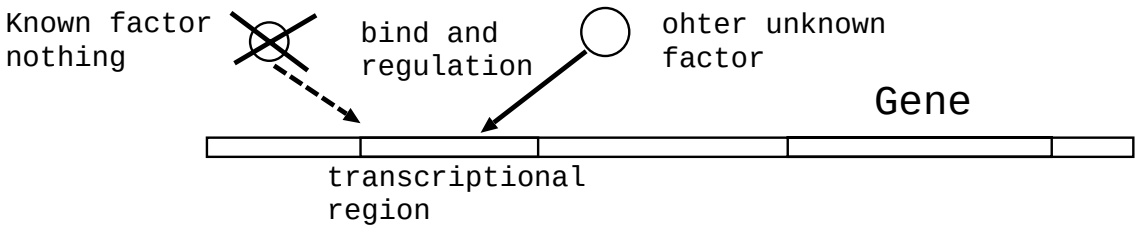

図 6.1: 未知の転写因子による遺伝子の発現への影響

別の未知な転写因子が細胞内に存在した場合、遺伝子の発現は未知な転写因子の影響を受けるため、影響を受けている遺伝子に対しても低い相関係数になる。この様な場合には相関係数のみで影響を判断することは出来ない。この事は依存関係既知の遺伝子群にもありうると考えられる。

この問題に対しては他の調節遺伝子の発現量を固定して調節遺伝子と依存関係未知の遺伝子とのを求める偏相関係数を用いる事により解決できると考えられる。しかし、全遺伝子に関しては逆行列を作成できず偏相関係数は計算出来ないため、全遺伝子をクラスタリング等を行い分割して関係式を構成している変数を取り除いて偏相関係数を求める必要がある [17] また、未知の転写因子を生成する調節遺伝子の候補を探し出す方法もある。未知の転写因子を生成する遺伝子の候補を探し出す手段としては対象となる調節遺伝子の遺伝子部分の DNA 配列と類似する配列を持つ遺伝子の調査や、転写後に生成される mRNA のコドンからアミノ酸配列を用いて類似する遺伝子の調査を行う。この方法によって類似する遺伝子を発見した場合はこの遺伝子の発現量を固定して調節遺伝子と依存関係未知の遺伝子の偏相関係数を求めて依存関係を調べる方法により解決できると考えられる。

6.2 依存関係がない場合

本節では依存関係の無い遺伝子を依存関係のある遺伝子の候補にしてしまう場合について説明する。

生物学的には文字列パターンを持つ遺伝子は転写因子が結合する事により発現に影響を受ける。実際に DNA マイクロアレイデータを用いた依存関係の調査結果では依存関係が既知である遺伝子に対しては半数近くの調節遺伝子において DNA マイクロアレイデータの解析結果が事実と一致した。

しかし、多くの調節遺伝子に関して文字列パターンを持つ依存関係未知の遺伝子群では依存関係を持つ可能性のある事を示すことは出来なかった。今回の DNA マイクロアレイデータを用いた依存関係の調査は各標本の相関係数の平均値を用いて行った。そのため、文字列パターンを持つ依存関係未知の遺伝子群に相関係数が低い遺伝子が多数あるために平均値は低くなり依存関係を持つ可能性が無いという結果になったと考えられる。各転

写因子を生成する遺伝子に対して依存関係未知の遺伝子群を調べると依存関係既知の遺伝子群の平均値以上の相関係数を持つ依存関係未知の遺伝子は存在する。しかし、パターンマッチングによって選びだされた依存関係未知の遺伝子の中には調節遺伝子と依存関係を持たない可能性のある遺伝子も含んでいると考えられる。

これら依存関係の無い遺伝子を候補として含んでしまう原因として以下の事が考えられる。

- 文字列パターンを挟んで異なるストランドに遺伝子が存在する。
- 今回使用した文字列パターンの情報の欠損

以上の事に関して考察を行う。

6.2.1 文字列 p ターンを挟んで異なるストランドに遺伝子が存在する場合

文字列パターンを挟んで異なるストランド上に遺伝子が存在する場合、今回の調査方法では両方とも調節遺伝子と依存関係を持つ遺伝子の候補としている。図 6.2 の様な場合において、遺伝子 A と B の間の文字列パターンに対して調節遺伝子との間に起りうる依存関係は４つあげられる。

- 遺伝子 A のみが依存関係を持つ
- 遺伝子 B のみが依存関係を持つ
- 遺伝子 A,B の両方が依存関係を持つ
- 遺伝子 A,B の両方とも依存関係を持たない。

この場合において、遺伝子 A のみが依存関係を持つ場合、B は調節遺伝子と依存関係を持たない遺伝子であるが、今回の調査方法では依存関係のある遺伝子の候補としてしまう。また遺伝子 B のみが依存関係を持つ場合にも同様の事が考えられる。

遺伝子 A,B が調節遺伝子と依存関係を持つ事を調査する方法としては調節遺伝子の DNA マイクロアレイデータの値を参考に依存関係があるか調べる方法がある。調節遺伝子を破壊した DNA マイクロアレイデータ上で遺伝子 A の発現比が 1 に近く、遺伝子 B の発現比が 1 より遥かに大きい、もしくは小さい時に遺伝子 B のみが調節遺伝子との依存関係を持つ事になる。また、調節遺伝子と A、B の相関係数を比較することにより相関係数の高い方を選択する方法もある。

しかし、データ誤差や他の転写因子等の影響により発現比が遺伝子 A、B 共に 1 に近い値場合や A、B 共に相関が低い場合も考えられ、調節遺伝子との依存関係を持つ可能性は調べられない。また、仮に A や B が調節遺伝子と相関係数が高い遺伝子であっても A や B が調節遺伝子の発現に影響を与えているとも考えられる。

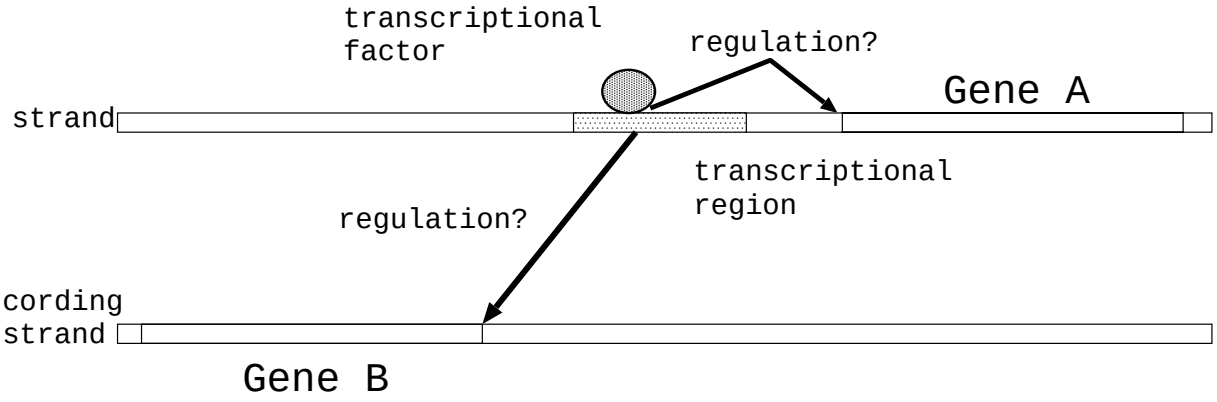

図 6.2: ストランド別に文字列パターンを挟む遺伝子について

DNA マイクロアレイデータ上の値が 1 に近い値であったり、相関係数が A、B 共に低い場合には生物学的実験によって依存関係を明らかにしていく方法が良いと考えられる。遺伝子 A と文字列パターンを含む部分の DNA を切取り、調節遺伝子が生成する転写因子に結合して発現するかどうか調べる。IolR を生成する調節遺伝子 iolR はこの方法と似た実験で転写制御領域を特定した [13]。同じく遺伝子 B にも同様の実験を行い調べていく方法によって互いの依存関係を決定する方法を提案する。

6.2.2 今回使用した文字列パターンの情報の欠損していた場合

今回使用した各調節遺伝子の文字列パターンが転写因子が結合する文字列パターン全てを表しているとは限らない。今回使用した文字列パターン以外の上流部分に一意的な配列が存在し、その部分も含めて文字列パターンとしている場合がある。

今回は既知の文字列パターンを調べたが、図 6.3 の様にパターンが因子と結合する一部分であった場合は他の部分を調べる必要がある。転写因子 PurR を発現する purR は purR Box と呼ばれる特異的な塩基配列に結合する事が知られていたが、その後の実験等の結果から実際には他の上流部分にも一意的な配列を持つことが明らかになった [16]。

パターンと他の一意的な配列が存在する上流部分を持つ遺伝子のみに影響を与えている場合、今回の文字列パターンのみで依存関係の候補を探し出す今回の方法では影響を与えていない遺伝子も候補としてしまう。

この様な場合において解決方法としては、既知に依存関係を持つ遺伝子の上流部分で文字列パターン以外に一意的、もしくは類似する配列部分を文字列パターンの上流と下流部分から範囲を決めて調査する必要がある。多くの文字列パターンにおいて転写因子の立体構造上、結合する部分は数ヵ所にある。依存関係既知の遺伝子の上流部分でその様な部分配列が発見された場合、その部分配列も含めた文字列パターン用いて依存関係のある候補となる遺伝子を調査する。

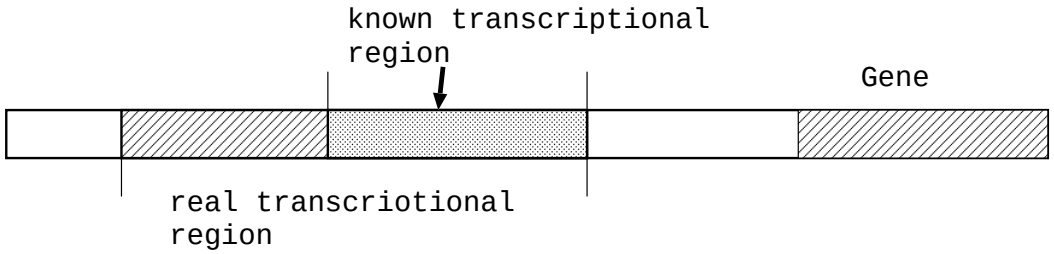

図 6.3: 他の上流部分に一意的な転写因子が結合する配列を持つ場合

6.2.3 疑似相関係数を用いた依存関係未知の遺伝子の調査

文字列パターンを持つ依存関係未知の遺伝子群の相関係数に対して依存関係既知の遺伝子群の相関係数の平均値以上になる遺伝子を取り出しても、実際に調節遺伝子との間に依存関係があるとは限らない。

この様な場合、文字列パターンを持つ依存関係既知の遺伝子群と依存関係未知の遺伝子群との相関係数を計算する方法を提案する。もし、相関係数が正の値で高い場合には依存関係未知の遺伝子は依存関係既知の遺伝子と各データ上で類似した値を持つと考えられる。この様な相関係数は調節遺伝子と依存関係未知の遺伝子の相関係数に対して疑似相関係数と呼ばれる。

ある閾値を設けて疑似相関係数がそれ以上になる場合に遺伝子に対して転写因子を生成する遺伝子との相関係数を調べる。もし相関係数が高い場合にはその依存関係未知の遺伝子は調節遺伝子と依存関係を持つ可能性があると考えられる。

今回、疑似相関係数の閾値を依存関係既知の全ての遺伝子同士の相関係数の平均値として閾値以上の疑似相関を持ち、調節遺伝子との相関係数が依存関係既知の全ての遺伝子以上の依存関係未知の遺伝子が存在するか調査した。依存関係が既知である遺伝子全てが今回使用した文字列パターンを持つわけでは無く、疑似相関を利用する場合は依存関係が既知な全ての遺伝子での平均値を用いることにする。調査結果としては調節遺伝子 ccpA, comA, degU, spo0A, spoIIID に関して上記の条件を満たす遺伝子が存在した。表 6.1 に疑似相関係数と調節遺伝子との相関係数が高い依存関係未知の遺伝子数を示す。その様な遺伝子群を標本とした時の平均値と母集団の平均値の差と平均値の検定結果について表 6.2 に示す。調節遺伝子 ccpA 以外はこの標本に対して調節遺伝子は依存関係を持つ可能性がある事が示せた。また、comA、spoIIID ではデータ選択、非選択共に疑似相関係数と相関係数が高い遺伝子は全て同じ遺伝子であった。表 6.3 に調節遺伝子 ccpA、comA、spoIIID での疑似相関係数と相関係数が高い遺伝子数と発現して生成する蛋白質の機能的分類と同じ機能的分類である依存関係既知の遺伝子を記す。これは既知の依存関係を持つ遺伝子群が発現して生成する蛋白質が同じ機能的分類である場合が多い事から疑似相関

	データ選択	データ非選択
ccpA	1	0
comA	1	1
degU	39	30
spo0A	61	58
spoIIID	7	7

表 6.1: 疑似相関係数と相関係数が高い遺伝子数

調節遺伝子	平均値の差（データ選択）	平均値の差（データ非選択）	判定（データ選択）	判定（データ非選択）
ccpA	0.17	-	×	-
comA	0.23	0.19	○	○
degU	0.13	0.11	○	○
spo0A	0.137	0.123	○	○
spoIIID	0.256	0.230	○	○

表 6.2: 疑似相関係数と相関係数が高い遺伝子での平均値の差と検定

係数と相関係数の高い遺伝子の機能的分類の中に依存関係既知の遺伝子群を含むかを調べたものである。

調節遺伝子 ccpA で選ばれた遺伝子 ytkA は発現して生成する蛋白質の機能がわかっていないため、既知に依存関係を持つ機能未知の遺伝子 yxjC と同じ分類であるとはまだ分からない。しかし、調節遺伝子 comA で選ばれた遺伝子 rapA や spoIIID で選ばれた dacF は既知の依存関係を持つ遺伝子と同じ機能的分類になる事が分かった。この様な遺伝子は依存関係を持つ可能性が高い事を意味する。

以上の様な疑似相関係数を用いる方法で相関係数の低い遺伝子を取り除く方法が有効であると考えられる。

調節遺伝子	相関係数と疑似相関係数が高い遺伝子	蛋白質の機能的分類	同じ機能的分類の依存関係既知の遺伝子
ccpA	ytkA	In other organisms	yxjC
comA	rapA	Sporulation	rapE, rapC
spoIIID	yabP	In other organisms	無し
spoIIID	ybaN	Miscellaneous	無し
spoIIID	ymxG	Metabolism of Amino Acids and Related Molecules	無し
spoIIID	yndM	In other organisms	無し
spoIIID	dacF	Cell Wall	spoVD
spoIIID	nucB	Metabolism of nucleotides and Nucleic Acids	無し
spoIIID	ywcB	In other organisms	無し

表 6.3: 疑似相関係数と相関係数が高い遺伝子の機能的分類について

第7章　おわりに

今回、16 種類の調節遺伝子に関して調節遺伝子毎に文字列パターンを持つ遺伝子を全遺伝子の上流部分をパターンマッチングによって調査し、調節遺伝子と既知の依存関係を持つ遺伝子との関係について DNA マイクロアレイデータから統計的手法を用いて調節遺伝子との依存関係について調査した。

標本が文字列パターンを持つ依存関係既知の遺伝子群とそのオペロン群である場合には半数を越える調節遺伝子に対しては標本の相関係数の平均値は全遺伝子の相関係数の平均値よりも高く、平均値の検定でも母集団と異なると判定された。これは半数以上の調節遺伝子に対して DNA マイクロアレイデータの解析結果と既知の事実が一致したことを表している。

しかし、標本を文字列パターンを持つ全遺伝子に対しては標本の平均値が母集団の平均値よりも高く、平均値の検定で全遺伝子との平均値と異なる調節遺伝子は半数近くになり、依存関係未知の遺伝子群を標本にした場合では 1 つの調節遺伝子のみであった。これは依存関係未知の遺伝子群の平均値は低い事を表している。しかし、各標本に対して依存関係を持つ可能性がある調節遺伝子は皆無ではないため、文字列パターンを持つ遺伝子群は調節遺伝子に影響を受けていることが検証された。この事は文字列パターンを持つ依存関係未知の遺伝子群の中には依存関係を持つ遺伝子も含んでいると考えられるため、転写制御領域の解析は遺伝子の依存関係に有意であると考えられる。

以上の事から文字列パターンを持つ依存関係未知の遺伝子群は相関係数の低い遺伝子群が多数存在する事が分かった。依存関係未知の遺伝子群に相関係数の低い遺伝子が存在する理由については未知の転写因子に遺伝子の発現が影響を受ける事と、依存関係が無い遺伝子が依存関係未知の遺伝子群に存在することがあげられる。この事は文字列パターンのみからでは遺伝子間の依存関係を推定できないことを表している。

未知の転写因子に影響を受ける遺伝子に対しては、偏相関係数を用いて遺伝子間の依存関係を推定する方法が有効であると考えられる。

また、依存関係が無い遺伝子に関しては、依存関係未知の遺伝子群内の遺伝子に対して疑似相関係数を用いて依存関係のある遺伝子を見つけ出す事によって解決する事が出来ると考えられる。

以上の方法は転写制御領域が未知の遺伝子に対してデータ等で影響を受けている遺伝子群の上流部分を調べ転写制御領域を明らかにして依存関係を求めていく事に重要な役割を果たすと考えられる。

謝辞

最後に研究を進めるにあたり、指導教官であり、北陸先端科学技術大学院大学情報科学研究科の平石 邦彦助教授には、終始様々な助言、ご指導頂きました。また、同研究の宋 少秋助手、高島 康裕助手にも終始様々な助言、ご指導頂きました。

ここに深く感謝の意を表します。

参考文献

[1] 北野宏明 『システムバイオロジー 生命をシステムとして理解する』（秀潤社 2001）

[2] 北野宏明『システムバイオロジーの展開～生物学の新しいアプローチ～』（シュプリンガー・フェアラーク東京社 2001）

[3] NTT「人体」プロジェクト 驚異の小宇宙・人体 III 遺伝子・DNA １生命の暗号を暗号を解読せよ（日本放送研究会 1999)

[4] 村松正明, 那波宏之 監修 DNA マイクロアレイと最新 PCR 法

[5] 松原謙一, 榊 佳之 ポストシークエンスのゲノム科学６ ゲノム情報生物学 BioInformatics と Information Biology （中山書店 2001）

[6] 松原謙一, 榊 佳之 ポストシークエンスのゲノム科学２ 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）

[7] N. Friedman, M. Linial, I. Nachman, D. Pe'er "Using bayesian Network to Analyze Expression Data," *Journal of Computational Biology*, vol. 95, no. 25, pp.14863-14868, 1998

[8] http://www.ncbi.nlh.gov/

[9] http://bacillus.genome.ad.jp/

[10] http://dbtbs.hgc.jp/

[11] http://www.cib.nig.ac.jp/dda/taitoh/bsub.operon.html

[12] 市原 清志 バイオサイエンスの統計学 （南江堂 1990）

[13] Ken-Ichi Yoshida,Tsukasa Shibayama,Daiki Aoyama and Yasutaro Fujita "Interaction of a Repressor and its Binding Sites for Regulation of the Bacillus subtilis iol Divergon" *Journal of Molecular Biology* vol. 285, Issue 3, oo, 917-929, January, 1999

[14] F. Kunst, N. Ogasawara, etc "The complete genome sequence of the Gram-positive bacterium Bacillus subtilis" *NATURE* vol. 390, no. 20, pp. 249-267, November 1997

[15] http://133.100.212.50/ bc1/Biochem/Genome.htm

[16] HANS H. SAXILD, KATJA BRUNSTED, KARIN I. NIELSEN, HANNE JARMER "Definition of Bacillus subtilis PuR Operator Using Genetic and Bioinfomatic Tools and Expanision of the PurR Regulon with glyA, guaC, pbuG, xpt-pbuX, yqh-folD, and pbuO" *Journal of Bacteriology* vol. 183, no. 21, pp. 6175-6183, Nov. 2001

[17] 日本品質管理学会、テクノメトリックス研究会 グラフィカルモデリングの実際 （日科技連出版者 1999）