JAIST Repository
https://dspace.jaist.ac.jp/

Title	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
Author(s)	上田，智之
Citation	
Issue Date	2004-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1805
Rights	
Description	Supervisor:平石 邦彦，情報科学研究科，修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

修 士 論 文

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
情報科学研究科情報システム学専攻

上田 智之

2004年3月

修 士 論 文

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

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　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

210006 上田 智之

提出年月: 2004年2月

Copyright © 2004 by Tomoyuki Ueda

目 次

第 1 章　遺伝子間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 5
1.3 本研究の目的および特色 7

第 2 章　使用する菌類と各種データの紹介 9
2.1 枯草菌について 9
2.2 ＤＮＡ塩基配列データについて 10
2.3 ＤＮＡマイクロアレイデータについて 11

第 3 章　実験１：転写制御領域に含まれる文字列の統計的調査 14
3.1 統計調査について 14
3.2 文字列の長さと各長さの文字列の出現頻度に関する調査 15
3.3 各長さの文字列の出現頻度 16

第 4 章　実験２：文字列の統計的特異度 20
4.1 特異性について 20
4.2 特異度の定義 20
4.3 長さ６の文字列の特異度 21
4.3.1 文字列の特異度 21
4.3.2 特異度で表現された転写制御領域 21
4.4 特異的文字列の候補 23

第 5 章　実験３：転写制御領域の類似性に関する調査 25
5.1 類似性について 25
5.2 類似度の定義 25
5.3 従来研究と比較した特色 27
5.4 制御関係既知の遺伝子群についての調査 29

第 6 章　考察 33
6.1 転写制御領域について 33
6.1.1 転写制御領域の定義 33
6.1.2 転写制御領域の開始位置 33

6.1.3 転写単位 34
6.2 特異的文字列について 34
6.2.1 特異性 34
6.2.2 文字列の長さ 34
6.2.3 文字列の出現頻度 35
6.2.4 統計的特異度の定義 35
6.2.5 閾値 35
6.3 類似性について 36
6.3.1 文字列の類似度 36
6.3.2 局所的な領域の類似度 36
6.3.3 遺伝子間の類似度 37
6.3.4 計算量 37
6.4 遺伝子間の制御関係推定について 38
6.4.1 転写制御領域の類似性による推定 38
6.4.2 ＤＮＡマイクロアレイデータによる推定 38
6.4.3 転写制御領域の類似性とＤＮＡマイクロアレイデータによる推定 . 38

図 目 次

1.1 遺伝子の発現 1
1.2 転写制御 2
1.3 特異的結合 2
1.4 遺伝子間の依存関係 3
1.5 遺伝子間ネットワーク 4

2.1 相補配列の生成 11
2.2 ＤＮＡマイクロアレイ法 12
2.3 破壊株データの例 (CcpA 破壊株) 13

3.1 転写制御領域の長さ 15
3.2 長さ６の文字列の出現遺伝子数 16
3.3 長さ８の文字列の出現遺伝子数 17
3.4 長さ１０の文字列の出現遺伝子数 17
3.5 出伝遺伝子数と既知の特異的文字列 18
3.6 長い文字列を構成する長さ６の部分文字列 19

4.1 文字列の特異度と出現頻度 22
4.2 glpQ の転写制御領域に含まれる特異的文字列 22
4.3 cotC の転写制御領域に含まれる特異的文字列 23
4.4 特異的文字列の候補 24

5.1 各遺伝子から取り出す転写制御領域 26
5.2 従来手法による依存関係推定 28
5.3 本研究による依存関係推定 28
5.4 局所的な領域に含まれる特異的文字列の評価 29
5.5 glpQ 30
5.6 ackA 31
5.7 araE 31
5.8 ccpA 32

表 目 次

1.1 転写制御に影響を与える原因 3
1.2 標準ＩＵＢ／ＩＵＰＡＣ核酸略号 6

2.1 枯草菌遺伝子の位置情報 10
2.2 遺伝子間の依存関係に関する情報 10

4.1 各文字の出現頻度と出現確率 21

5.1 文字列の類似度スコア 27

第1章　遺伝子間の依存関係推定について

1.1　はじめに

近年、ヒトや猿、大腸菌など様々な生物において、それらの生物学的な違いの特徴を示すＤＮＡ（デオキシリボ核酸）の塩基配列が決定されている。そして、今後のゲノムの研究は、ＤＮＡが持つ情報の構造に関する研究から情報の持つ機能に関する研究に移りつつある。ＤＮＡは、４つの塩基（Ａ：アデニン、Ｇ：グアニン、Ｔ：チミン、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、遺伝情報を担う染色体の主成分物質である。遺伝子とはＤＮＡ中の機能的な役割を持つ領域を指し、発現によって生体機能に関わる蛋白質を生成するための設計図としての役割を持っている。遺伝子の発現（manifestation）は、転写（transcription）と翻訳（translation）に分けられる。まず、転写段階ではＲＮＡポリメラーゼが転写開始位置に結合し、遺伝子下流に移動するときに遺伝子の塩基配列をｍＲＮＡにコピーされる。そして、ｍＲＮＡは蛋白質の合成に必要のない部位を切り離され、翻訳段階でそれを基に蛋白質が合成される。(図 1.1)

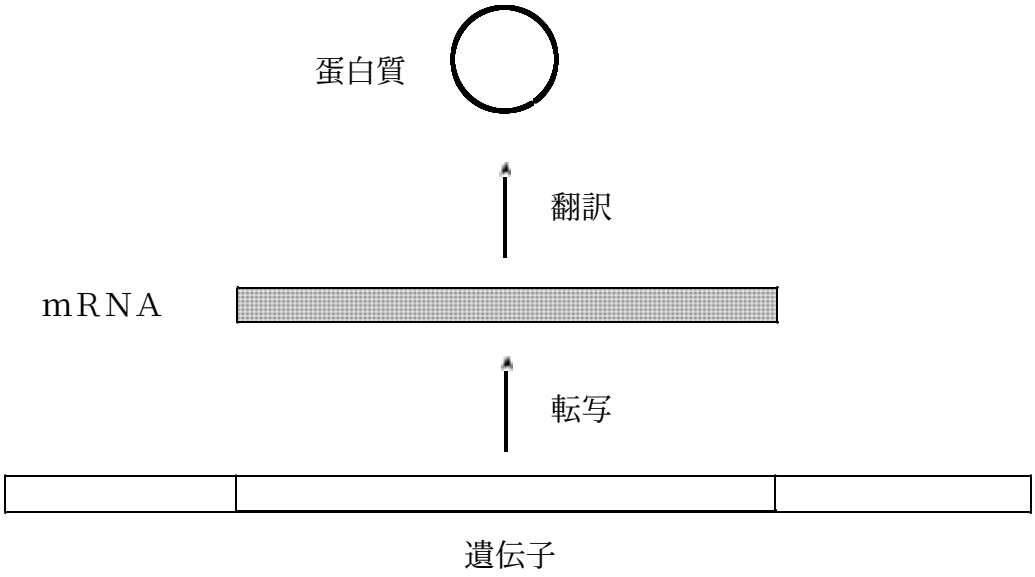

図 1.1: 遺伝子の発現

このように生成された蛋白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、他の機能として自身や他の遺伝子の転写に影響を与える役割を持つ場合

もある。(図 1.2)　このとき、影響を与える蛋白質を転写制御因子と呼ぶ。

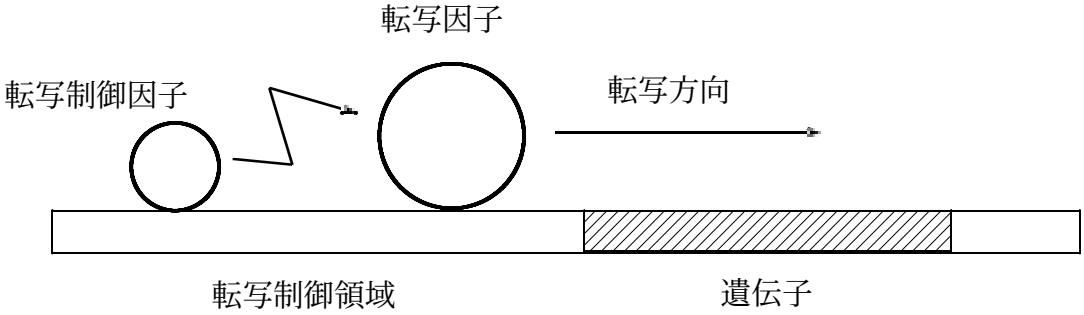

図 1.2: 転写制御

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写の制御に関与する領域に特異的に結合し、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。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蛋白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２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６～十数塩基であることが知られている。また、１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。（図 1.3）

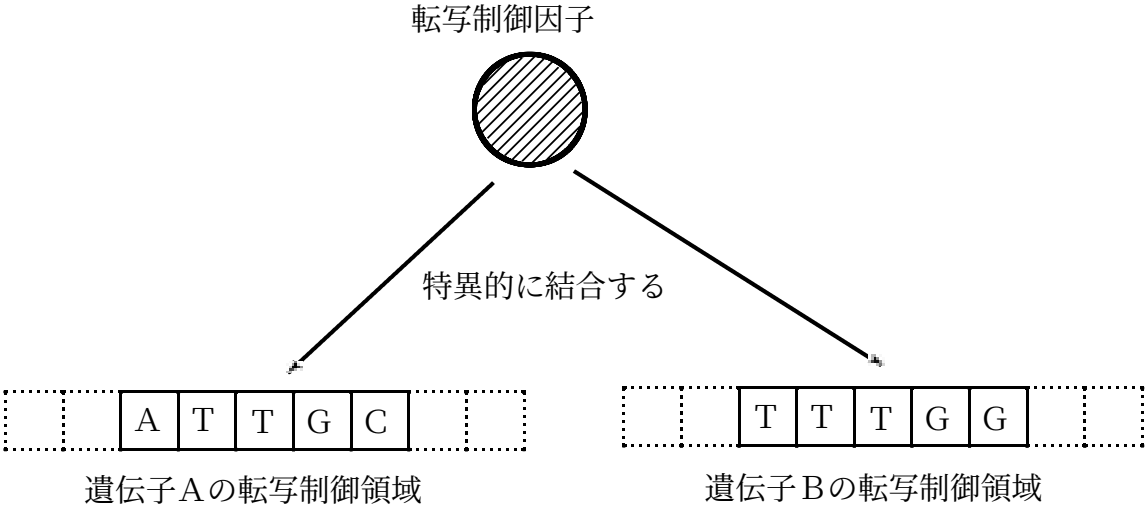

図 1.3: 特異的結合

そのような転写制御因子が結合する塩基配列は、人工的に生成でき、局所的な結合においては生物学的な実験によって確かめることができる。しかし、全体における転写制御因子の結合は、表 1.1 の理由 [3] により、その再現性は必ずしも示されない。

表 1.1: 転写制御に影響を与える原因

	主な原因
1	転写制御因子のリン酸化などによる修飾
2	複数の転写制御因子が同一の文字列を認識
3	コンセンサス配列の周りの配列が影響
4	細胞種
5	細胞の外界刺激
6	ゆらぎ

遺伝子間の依存関係とは、ある遺伝子Ａが生成する蛋白質がある遺伝子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の発現を制御することを指し、遺伝子Ａと遺伝子Ｂは依存関係にあるという。（図 1.4）　また、このとき、遺伝子Ａを調節遺伝子と呼び、遺伝子Ｂを被調節遺伝子と呼ぶ。

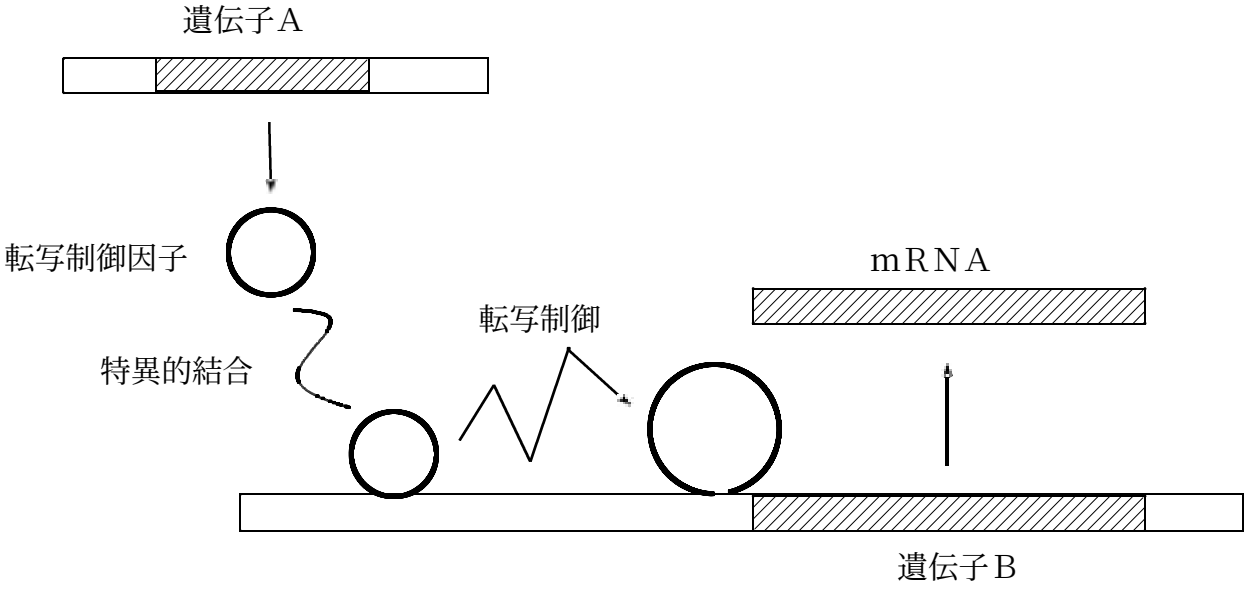

図 1.4: 遺伝子間の依存関係

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る。遺伝子Ａが遺伝子Ｂと依存関係にあり、遺伝子Ｂと遺伝子Ｃが依存関係にあるとき、遺伝子Ａと遺伝子Ｃは間接的な依存関係にある。このような遺伝子間の依存関係を表現したのが、遺伝子間ネットワークである。（図 1.5）各ノードは遺伝子を表し、ブランチは遺伝子間の依存関係を表す。図では、遺伝子Ａから遺伝子Ｂ、Ｄへ向かう矢印は、遺伝子Ａ

が調節遺伝子であり、遺伝子Ｂと遺伝子Ｄが被調節遺伝子であることを表している。

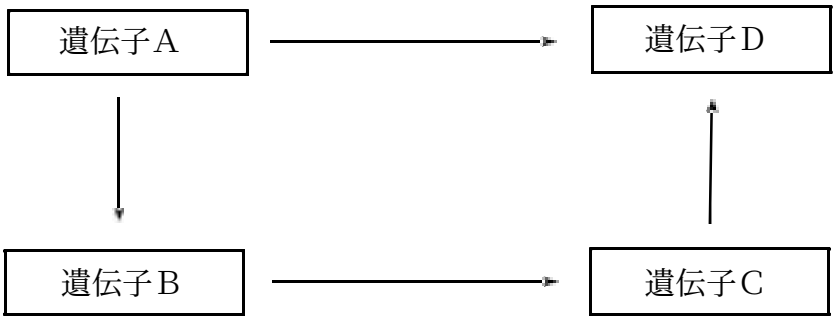

図 1.5: 遺伝子間ネットワーク

本研究で取り扱う文字 w とは、ＤＮＡの各塩基を表す A,T,G,C である。

$$w \in \{A,T,G,C\}$$

従って、文字列 s は、文字 w の繰り返しとして表現できる。

$$s = w^{|s|} \qquad \text{例：} s = AGGC$$

ただし、$|s|$：文字列ｓに含まれる文字数　、$|s| \geq 0$ であり、文字列の中の i 番目の文字は、$s[i]$ と表される。上の例では、$s[1] = A, s[2] = G, s[3] = G, s[4] = C, |s| = 4$ である。長さがゼロの文字列は空文字列といい、ϵ で表す。文字列パターンとは、文字からなる正規表現である。

$$\text{例：} TGT(A \cup G) = \{TGTAA, TGTGA\}$$
$$ACA(T^2 \cup T^4) = \{ACATT, ACATTTT\}$$

また、$A \cup G = R$（プリン、purine）のような核酸の略号は、標準ＩＵＢ／ＩＵＰＡＣ核酸略号を用いる。（表 1.2）

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、ＤＮＡマイクロアレイデータのみを用いたトップダウン法と呼ばれる手法が一般的である。代表的なものに、ブーリアンネットワーク [4] やベイジアンネットワーク [5]、Ｓ－ＳＹＳＴＥＭ [6] 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、閾値の決定法や近年のＤＮＡマイクロアレイデータの精度向上に対して情報の過剰な損失が問題となっており、また、推定のためには多量のＤＮＡマイ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、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り、いくつかの成果を挙げている。しかし、これらの手法はＤＮＡマイクロアレイデータの精度に強く依存しており、また、推定のための情報量不足が問題となっている [4]。このような従来手法における問題を解決する方法として、ＤＮＡマイクロアレイデータの他に生物学的な情報を加えて遺伝子間の依存関係を推定する手法が注目されている [10]。生物学的な情報を与えるプロモーターシーケンスに関する研究には、コンセンサスシーケンスやモチーフなどがある [7]。これらは依存関係が既知である遺伝子群を対象に、それらの転写制御領域において類似する配列を発見する手法であり、依存関係既知の遺伝子群の構造的特徴と類似する配列をもつ遺伝子を共通の制御を受ける遺伝子として推定することに役立つ。このように構造的な類似性によって機能的な類似性を推定する手法は、既知の遺伝子のデータベースを参照して、未知の生物における遺伝子の発見することなどに役立てられており、今日の遺伝子発見の基礎となっている。ここでは、モチーフと呼ばれる手法 [8] を紹介する。

表 1.2: 標準ＩＵＢ／ＩＵＰＡＣ核酸略号

略号	核酸（Nucleic Acid）
A	アデニン（Adenine）
C	シトシン（Cytosine）
G	グアニン（Guanine）
T	チミン（Thymine）
U	ウラシル（Uracil）
M	A or C（アミノ、amino）
R	A or G（プリン、purine）
W	A or T（弱い、weak）
S	C or G（強い、strong）
Y	C or T（ピリオジン、pyrimidine）
K	G or T（ケト、keto）
V	A or C or G
H	A or C or T
D	A or G or T
B	C or G or T
N	A or C or G or T

モチーフと呼ばれるこの手法は、依存関係が既知である遺伝子を対象としており、共通の転写制御因子から制御を受ける遺伝子群の転写制御領域を統計的に特徴付けた文字列を出力とする。この手法の特徴は、文字列の開始点を各遺伝子の転写開始位置とし、各位置における塩基の出現頻度を調べ、情報量（Information Content）によって評価し、モチーフとなる文字列を決定することである。また、この手法の主要な定義を以下に示す。文字列の長さをＳ、最小空白数を G_1、最大空白数を G_2 とすると、文字列の組み合わせは、

$$4^S(\mathrm{S}-1)(G_1-G_1+1)$$

となる。また、ｎ個のシーケンスにおいて、文字列中の位置ｋにｉ番目の文字が出現する数を t_{ik} とすると、

$$\sum_{k=1}^{S}\sum_{i=1}^{4} t_{ik} = nS$$

という関係が成り立ち、位置ｋでの文字ｉの出現する確率 p_{ik} を

$$P_{ik} = \frac{t_{ik}+1}{n+4}$$

とすると、

$$\sum_{k=1}^{S}\sum_{i=1}^{4} p_{ik} = S$$

という関係が成り立つ。情報量ＩＣは、

$$IC = \sum_{k=1}^{S}\sum_{i=1}^{4} p_{ik}\log\left(\frac{p_{ik}}{P_i}\right)$$

と定義される。これにより、ある位置で最も多く出現した文字が評価され、モチーフとなる文字列に採用される。

1.3　本研究の目的および特色

本研究では、はじめに全遺伝子の転写制御領域を対象として、それに含まれる文字列を解析し、特異的文字列の候補を抜き出す。そして、各遺伝子の局所的な領域に含まれる特異的文字列の候補の類似性を比較し、各遺伝子間の構造的類似性を算出する。また、ＤＮＡマイクロアレイデータを用いることで、生物学的な構造類似性と動的な生体反応の両面から、遺伝子間の制御関係を推定することを目的としている。また、この目的は、ＤＮＡの構造的特徴から機能的な特徴を定量的に評価するアプローチであり、機能面のみを評価するＤＮＡマイクロアレイデータに対して、構造的な因果関係を付与することを示している。以下に本研究の特色を述べる。

1. アプローチ
   従来研究によって、転写制御領域に特異的文字列パターンを持つ遺伝子は共通の制御を受けることの有意性は知られており、転写制御領域の解析は遺伝子間の依存関係推定において有効な情報を与えると考えられる。また、ＤＮＡ構造の生物学的な特徴を用いるため、ＤＮＡマイクロアレイデータのみを用いる手法に比べ、推定の信頼性が高くなり、依存関係に因果関係を与えることができることが挙げられる。

2. 統計解析に基づく特異的文字列集合
   転写制御領域に含まれる特異的文字列の解析では、転写制御領域に含まれる文字列の統計的な解析に基づく結果を用いて、特異的文字列の候補を得る。また、そのような特異的文字列の集合のみを対象として類似性を評価するため、制御関係に関係がないと考えられる文字列を取り除くことができる。

3. 推定された遺伝子群
   全遺伝子を対象とすることで確定的な類似性を示すことができる。また、既知の結合文字列の長さが限られた範囲に多いことから、転写制御領域の局所的な領域について類似性を評価する。

第2章　使用する菌類と各種データの紹介

本章では本研究で対象とする菌類とそのＤＮＡ配列データとＤＮＡマイクロアレイデータ（破壊株データ）について説明する。対象とする菌類は枯草菌とし、ＮＣＢＩで公開されている全塩基配列や遺伝子の位置などの構造的なデータを使用する。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９９種類の破壊株データを使用する

2.1　枯草菌について

枯草菌（こそうきん、Bacillus subtilis）は、大腸菌と並んで分子生物学の代表的な研究対象である [9]。枯草菌は、原核生物であり、毒性の無いグラム陽性、胞子形成菌であり、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。身近なところでは、納豆菌がその一種である。その特徴は分子遺伝解析が非常に進んでいることと、胞子形成、蛋白分泌、抗生物質生産、DNA 形質転換をおこなうことである。これらの特徴を生かして、染色体複製の遺伝的制御、遺伝子内微細マッピング、転写制御因子シグマカスケード、相互に入り組んだシグナル伝達系、糖代謝のカタボライト抑制機構など多くの分子遺伝学の研究が行われてきた。また、国際的な共同研究により、１９９７年には、全塩基配列が決定され、「Nature」誌に発表されたことで、引き続き遺伝子破壊による機能未知遺伝子群の解析が国際協力体制で進みつつある。枯草菌はこのような歴史的背景をもつため、大腸菌と並ぶ代表的な研究対象として扱われている。。DNA の塩基数は４２１万４８１０塩基対あり、約４１００のタンパク質コーディング遺伝子を含む。

本研究で使用する枯草菌は、全塩基配列が決定されており、遺伝子の位置等の情報も BSORF : Bacillus subtilis open reading frame （http://bacillus.genome.ad.jp/）で公開されている。また、依存関係が既知である遺伝子群や転写制御因子が結合する DNA 塩基配列などのデータは DBTBS : database of transcriptional regulation in Bacillus subtilis（http://dbtbs.hgc.jp/）で公開されている。表 2.1 は、BSORF で公開されているデータの一部である。Location はその遺伝子が DNA 上の位置であり、単位は bp である。bp は、base pair の略で１つの塩基を示している。Strand は、２つある DNA 塩基配列のどちらかを示しており、公開されている DNA 塩基配列側をプラスとし、その相補配列側をマイナスとしている。Length は遺伝子の長さ。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、遺伝子の名前。Synonym Code は、遺伝子の通し番号である。

表 2.2 は、DBTBS で公開されているデータの一部である。Transcription factor は転写

表 2.1: 枯草菌遺伝子の位置情報

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	guaB	Bsu0006

制御因子を生成し影響を与える調節因子名である。Consensus seq は、影響を受ける各遺伝子転写制御領域の結合サイトのコンセンサスシーケンスである。Regulated gene は、被調節遺伝子名。Operon は遺伝子が属するオペロン。Sigma は、影響を受けるシグマ因子名。Regulation は、転写制御の種類（活性または抑制）。Absolute は DNA 塩基配列での位置。Binding seq は、転写制御因子が結合する領域及び配列である。

表 2.2: 遺伝子間の依存関係に関する情報

Transcription factor : PhoP

Consensus seq : TT(A/T/C)ACA-N3- to N7-TT(A/T/C)ACA

Gene	Operon	Sigma	Regulation	Absolute position	Binding seq
glpQ	glpTQ	SigA	Positive	233918..233993	TTAATAGTTTTTCCAACA
phoA	ND	ND	Positive	1017962..1018025	TTTTCATTTCCTTACAA
phoB	ND	ND	Positive	621245..621303	AATCCAATCTTTAAAAT
phoD	ND	ND	Positive	283462..283511	TTACAATCAGTTCACA

2.2 ＤＮＡ塩基配列データについて

ＤＮＡ（デオキシリボ核酸）とは、染色体の主成分であり、４つの塩基文字（A、T、G、C）から構成される１つの塩基配列とその相補配列が互いに結合し、二重らせん構造をとっている物質である。ＤＮＡ塩基配列データとは、そのような配列の片側の塩基配列を表したデータであり、このデータから相補配列を生成することができる。ある塩基配列

に対する相補配列は、互いの塩基に対応する塩基の置き換えによって生成することができ、各塩基の対応関係はA⇔T、G⇔Cである。(図 2.1)

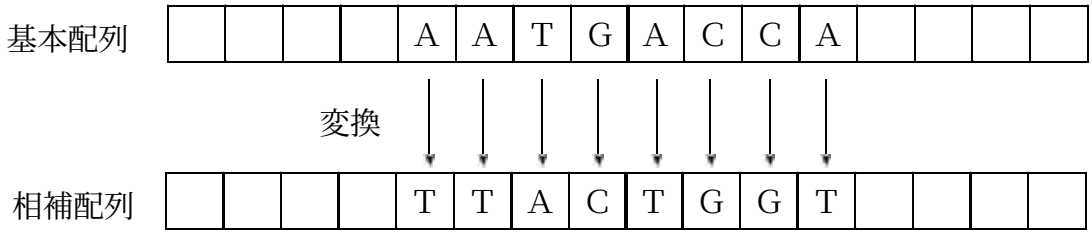

図 2.1: 相補配列の生成

2.3 ＤＮＡマイクロアレイデータについて

ＤＮＡマイクロアレイデータとは、ＤＮＡマイクロアレイ技術によって得られた遺伝子の発現データである。ＤＮＡマイクロアレイ技術とは、同時に数千から数万の遺伝子の発現量を測定できる技術であり、現在、遺伝子研究に広く用いられている。この技術で観測される遺伝子の発現量とは、遺伝子が発現して生成された蛋白質の量を測定することではなく、遺伝子の転写によって生成されたｍＲＮＡの量である。基本原理は、２本の DNA 塩基配列が互いに相補的な場合にハイブリダイズすることである。つまり、遺伝子が発現して生成されたｍ RNA とそれと一致する用意した DNA 塩基配列（プローブ）とがハイブリダイズすることである。観測原理は、はじめに遺伝子が発現して生成されたｍ RNA とそれと一致する DNA 塩基配列にそれぞれ蛍光色素 Cy-3,Cy-5 を付着させる。そして、遺伝子の発現量に応じて色の強度が変わることを利用して発現量を測定する。（図 2.2）

今回使用する破壊株データとは、ＤＮＡマイクロアレイデータの一種で、破壊株と野生株の発現量の比を観測したデータである。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いう。また、野生株とは遺伝子を破壊していない株をいう。株とは、遺伝的形質が同じ生物の別の個体をいう。破壊株データは、破壊した遺伝子が調節遺伝子であった場合にその遺伝子から影響を受ける遺伝子の発現量を観測することに用いられる。また、野生株データでは遺伝子を破壊していない通常の発現量を観測する。この二つのデータの比を得ることにより、破壊した遺伝子からの影響を観測することができる。つまり、遺伝子間の依存関係を観測できる。したがって、下式は破壊株Ｘにおける遺伝子 i の発現量である。

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

発現量の比は、遺伝子間に依存関係がない場合、１となり、被調節遺伝子が負の制御を受ける場合は１より大きくなり、正の制御を受ける場合は、１よりも小さくなる。しかし、

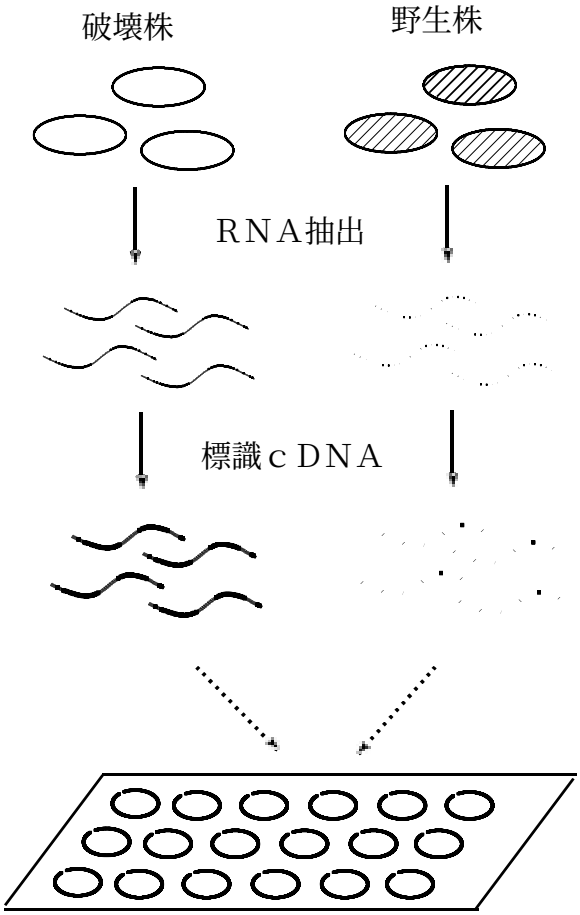

図 2.2: ＤＮＡマイクロアレイ法

実験条件等によって誤差が生じるため、それらの誤差を補正する必要がある。したがって、本研究では実験データによる実データを補正したデータを使用する。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ことが既知である遺伝子 i に関して約８割の遺伝子が $ratio_i^X$ と受ける影響が一致する。この発現量の比を全ての破壊株に対して１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、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。

図 2.3 は CcpA 遺伝子を破壊した環境下での各遺伝子の発現量比をプロットしたものである。

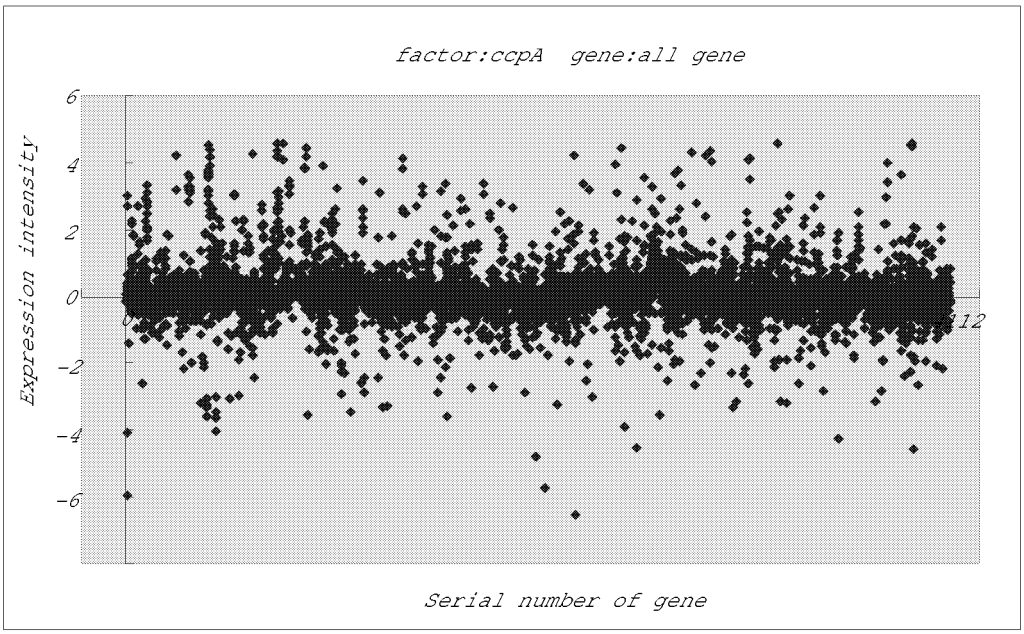

図 2.3: 破壊株データの例 (CcpA 破壊株)

第3章　実験１：転写制御領域に含まれる文字列の統計的調査

本研究では、全遺伝子転写制御領域を対象として、そこに含まれる文字列の調査を行ない、特異的文字列の特異性の統計的な尺度を考える。このような統計調査は、「転写制御領域に含まれる特異的文字列はある転写制御因子の特異的結合を表す」という生物学の基本的な考えに基づくものである。また、その考えを裏付ける結果として、依存関係既知の遺伝子群の転写制御領域には類似した配列が多く見つかっており、転写制御領域の統計的な調査の有意性を示している。その具体的な従来研究の成果例として、多くの遺伝子において転写開始位置から上流 25 ～30 塩基に転写開始因子が結合する TATA ボックスが発見されている。

3.1　統計調査について

統計的方法とは、標本を用いて母集団に関する情報を引き出すことであり、今回の調査での標本は、既知の特異的文字列の集合であり、母集団は既知および未知の特異的文字列の集合である。また、引き出したい情報とは特異的文字列集合の特異性である。従って、調査の目的は既知の特異的文字列集合の特異性を転写制御領域に含まれる全ての文字列集合との対比によって引き出し、それを利用して未知の特異的文字列の候補を得ることである。ここで扱う文字列とは、DNA を構成する４つの塩基文字 $w(A,G,T,C)$ であり、それらを組み合わせたものが文字列 S である。従って、文字列の長さを L とすると、文字列 S の組み合わせの数 C は、$C = 4^L$ である。生物学的に文字列の長さは転写制御因子の認識・結合能力を反映すると考えられる。転写制御因子の１つの結合ドメインが認識する文字列の長さは 5 ～ 10 数文字程度であり、複合体を形成している転写制御因子では複数の結合ドメインをもつことが知られている。このようなことから特異的文字列の長さは 5 ～ 数 10 文字程度と推測される。また、特異的文字列は少なくとも制御関係がある遺伝子に含まれていなければならない。そのため、多くの遺伝子に含まれる文字列は特異的文字列ではない可能性が高いと考えられる。逆に１つの遺伝子にしか含まれない文字列も文字の置換等を許さない限りにおいては、特異的文字列とは成り難い。従って、文字列の出現頻度は、共通の転写制御因子から影響を受ける可能性のある遺伝子数を表すと考えられる。また、従来の研究により、転写制御因子は特異的文字列を認識して結合するため、共通の特異的文字列をもつ遺伝子群は共通の制御を受けることが統計的に有意であること

が示されており、既知の特異的文字列の出現数は影響を与える遺伝子数を表すと言い換えることができる。以上のことより、本章では、はじめに文字列の長さと各長さの文字列の出現頻度に関する調査を行なう。

3.2 文字列の長さと各長さの文字列の出現頻度に関する調査

はじめに、各遺伝子の転写制御領域を用意する。遺伝子の転写制御領域は、一般に遺伝子の上流にあるが、その領域の長さや位置などは明確には決められていない。そのため、本研究では転写制御領域の開始位置を遺伝子の開始位置とし、終了位置は開始点からの距離 529 までとする。転写制御領域の長さ 0 以上のとき、その遺伝子は転写制御領域を持つとする。また、その長さは最低 30 とし、最大は 529 とする。このような定義に基づいて各遺伝子の転写制御領域を取り出したとき、その長さの分布は図 3.1 のようになる。この結果、転写制御領域を持つ遺伝子数は 3610 、持たない遺伝子数は 503 となる。

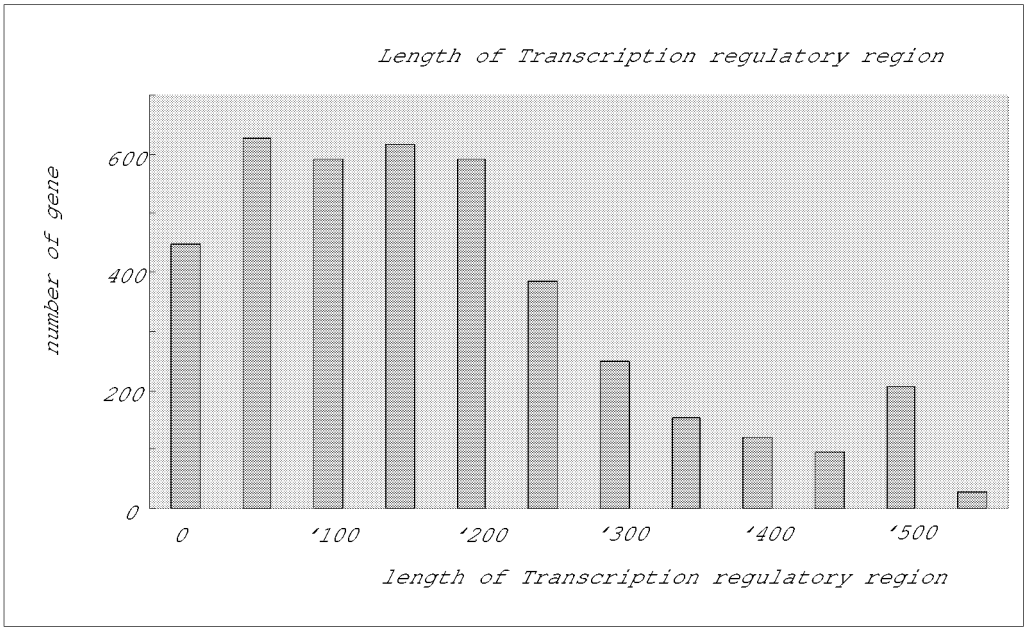

図 3.1: 転写制御領域の長さ

3.3 各長さの文字列の出現頻度

次に、各遺伝子の転写制御領域に含まれる各長さの文字列の出現頻度を調査する。調査する全遺伝子の転写制御領域に含まれる文字の合計は、2533596 である。ここで、遺伝子 i の転写制御領域の長さを $L_t[i]$ とし、全遺伝子転写制御領域から取り出す長さ L_S の文字列の数 M は、 $M = \sum_{i=0}^{4113}(L_t[i] - L_S + 1)$ となり、文字列の長さに反比例する。一方、文字列の種類数は指数的に増加するため、文字列の出現確率は低くなる。調査では取り出した文字列数 M を文字列の種類数で割った値が 1 以上となる長さ 10 までを調査する。図 3.2〜3.4 に各長さの文字列の出現頻度を示す。

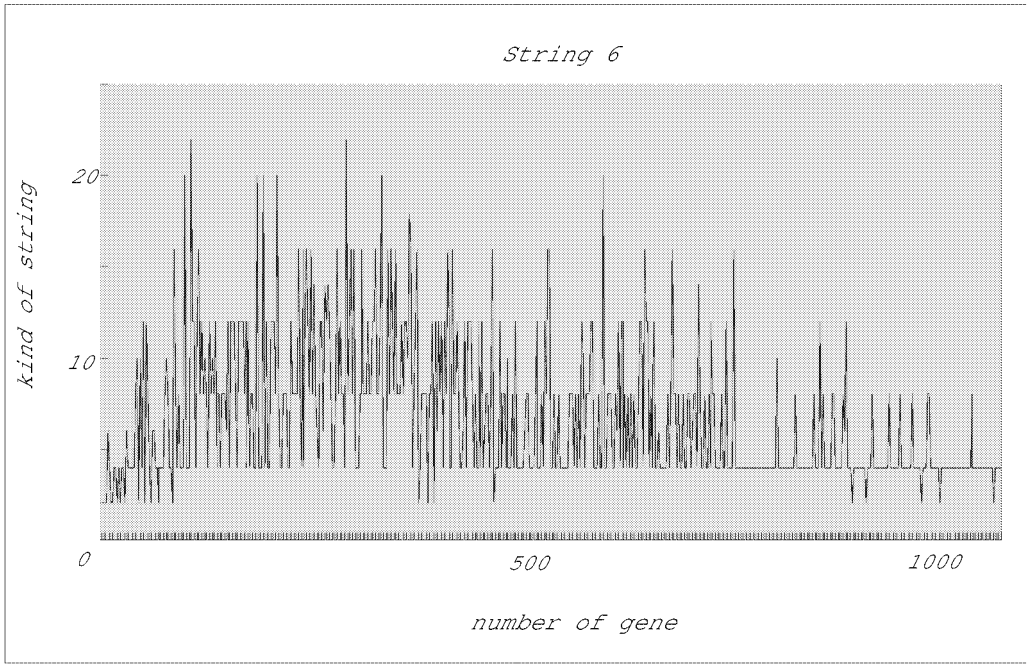

図 3.2: 長さ 6 の文字列の出現遺伝子数

これまでの調査結果から、転写制御領域に含まれる文字列数は、取り出す文字列の長さに反比例し、また全遺伝子転写制御領域内での出現回数も文字列の長さに対して指数的に減少することが確認できる。長さ 10 の文字列では全遺伝子中での出現頻度が 1 である文字列の種類数が最大であり、出現回数を特異性の尺度とした場合、特異的文字列の候補となる文字列数が多いという結果を導く。(図 3.5)　また、出現回数が少ない文字列集合に既知の結合文字列が含まれるとは限らず、出現回数が多い文字列集合にも既知の結合文字列は含まれる。長さ 6 の文字列に関しては、既知の特異的文字列は全体の 90[%] に含まれている。従って、文字列の出現回数だけからみて特異性文字列の特異性は確認できないため、異なる統計的特徴を考える。また、一般に長い文字列は短い部分文字列で構成されることや転写制御因子が結合する文字列の長さの最低は 5, 6 文字であることから、以後の調

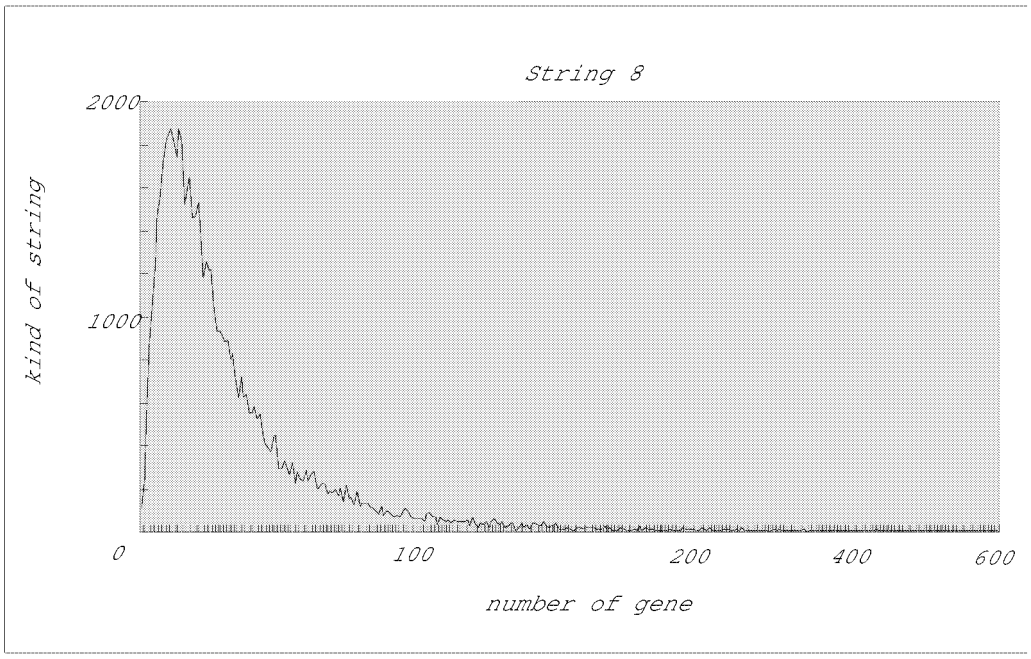

図 3.3: 長さ８の文字列の出現遺伝子数

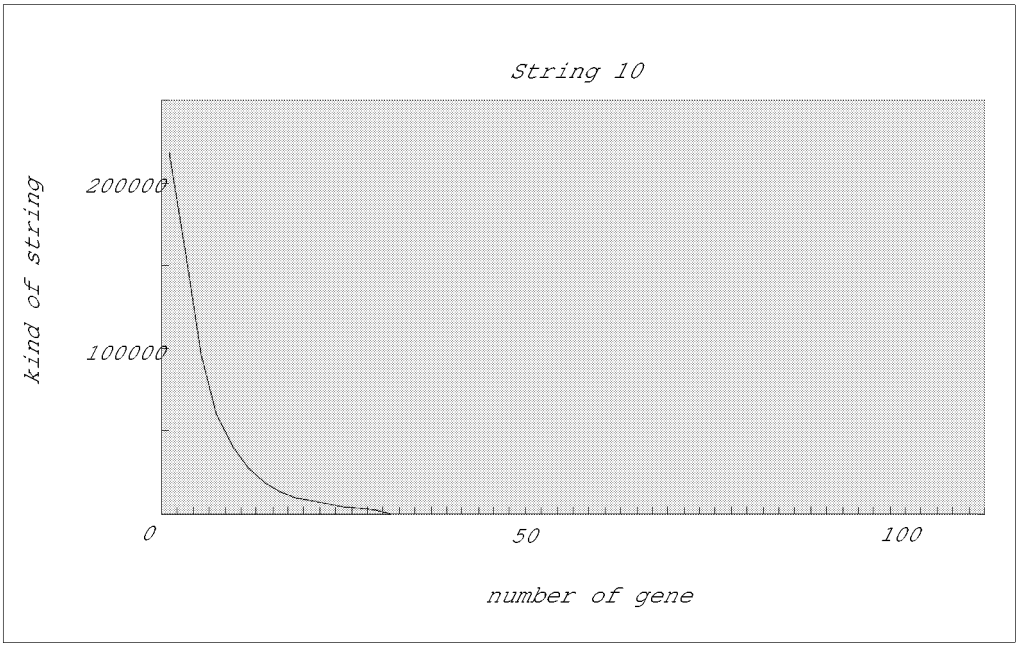

図 3.4: 長さ１０の文字列の出現遺伝子数

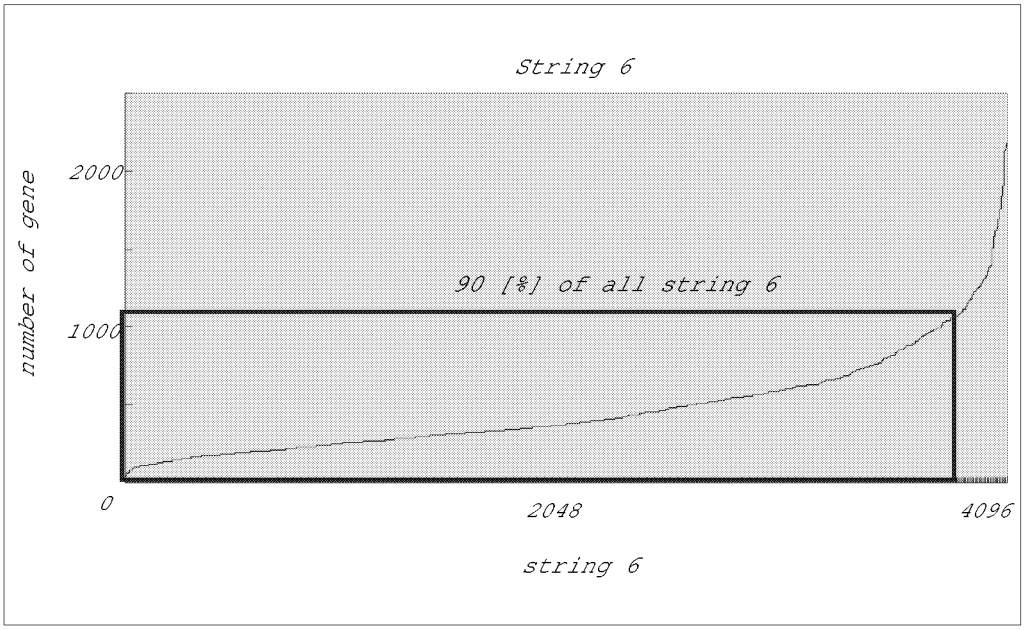

図 3.5: 出伝遺伝子数と既知の特異的文字列

査では短い部分文字列を対象とし、長い文字列は部分文字列の組み合わせとして表現する。(図 3.6)

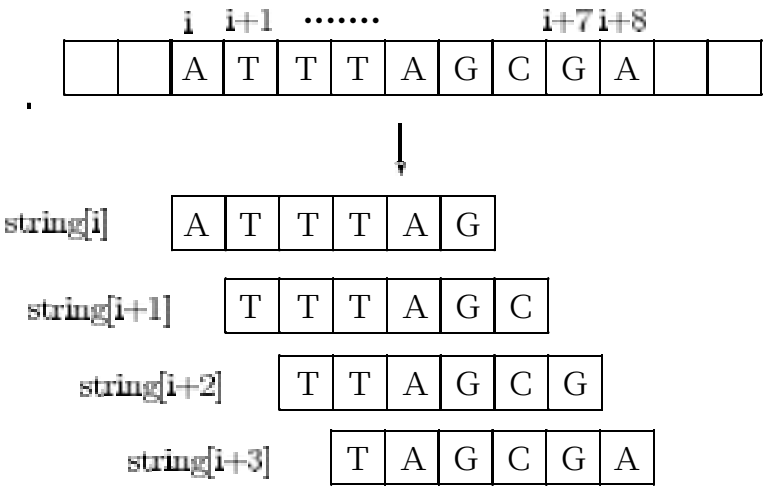

図 3.6: 長い文字列を構成する長さ 6 の部分文字列

第4章　実験２：文字列の統計的特異度

4.1　特異性について

遺伝子の転写制御領域に含まれる文字列の統計的な調査により、長さの短い部分文字列を対象に異なる統計的側面から特異性を調査する必要があることが分かった。従って、本節では、はじめに特異度を定義し、転写制御領域に含まれる長さ 6 の文字列 4096 種類に対して、各文字列の特異度を表す。一般に DNA 中に含まれる塩基文字の割合には偏りがある。今回調査した転写制御領域に含まれる各塩基文字の割合は、$A+T=63.3[\%]$ 、$G+C=36.7[\%]$ である。ここで、文字がランダムに出現すると仮定すると、転写制御領域から取り出す文字列に各文字が含まれる割合も同様になると考えられ、A と T を多く含む文字列の出現頻度が多くなると予想される。従って、本章では各文字の出現頻度を考慮して文字列の期待値を定義し、実際の出現頻度と期待値との差を調査する。

4.2　特異度の定義

遺伝子転写制御領域に含まれる各塩基 A,G,T,C の出現回数を w_A,w_T,w_G,w_C とすると、それらの合計 w_{total} は、

$$w_{total}=w_A+w_T+w_G+w_C$$

となる。
このとき、各文字の出現確率を p_A,p_T,p_G,p_C とする。

$$p_A=w_A/w_{total},p_T=w_T/w_{total},p_G=w_G/w_{total},p_C=w_C/w_{total}$$

また、相補配列も対象としていることから、

$$w_A=w_T,w_G=w_C,p_A=p_T,p_G=p_C$$

である。従って、 A,T と G,C のそれぞれの出現確率 p,q をとすれば、

$$p=(p_A+p_T)/2$$
$$q=(p_G+p_C)/2$$

表 4.1: 各文字の出現頻度と出現確率

	A	T	G	C
w	825270	825270	477618	477618
p	0.3167	0.3167	0.1833	0.1833

となる。実際の測定値は表 4.1 であるため、$p = 0.6334$, $q = 0.3666$ となる。

ここで以下の仮定をおく。
文字列の出現がランダムならば、A および T が k 回、G および C が $n-k$ 回出現する長さ n の文字列 s の出現確率 P_s は、

$$P_s = p^k q^{n-k}$$

となる。取り出した遺伝子転写制御領域の合計 L をとする。文字列の出現がランダムならば、文字列 s または逆文字列 s^{-1} が、転写制御領域または相補配列に含まれる個数の期待値 E_s は、

$$E_s = 4LP_s$$

である。
文字列の統計的特異度 D_s を以下の式で表現する。

$$D_s = (O_s - E_s)/E_s$$

ただし、O_s は文字列 s の実際の出現頻度である。

4.3 長さ 6 の文字列の特異度

4.3.1 文字列の特異度

長さ 6 の文字列の出現頻度と特異度の調査結果を示す。(図 4.1)
長さ 6 の文字列の種類数は 4096 である。特異度がマイナスの文字列数は 2686（65.6[%]）であり、プラスの文字列数は 1410（34.4[%]）である。また、特異度の最大値は 5.24 であり、最小値は −0.76 である。出現頻度で高い値を示す文字列には AAAAAA、TTTTTT であり、それらは連続的に出現しているため、高い出現頻度を示している。

4.3.2 特異度で表現された転写制御領域

既知の特異的文字列をもつ遺伝子の転写制御領域を特異度を用いて表現する。(図 4.2,4.3) 図中に示される四角で囲まれた領域は、実際の転写制御因子が結合する文字列である。

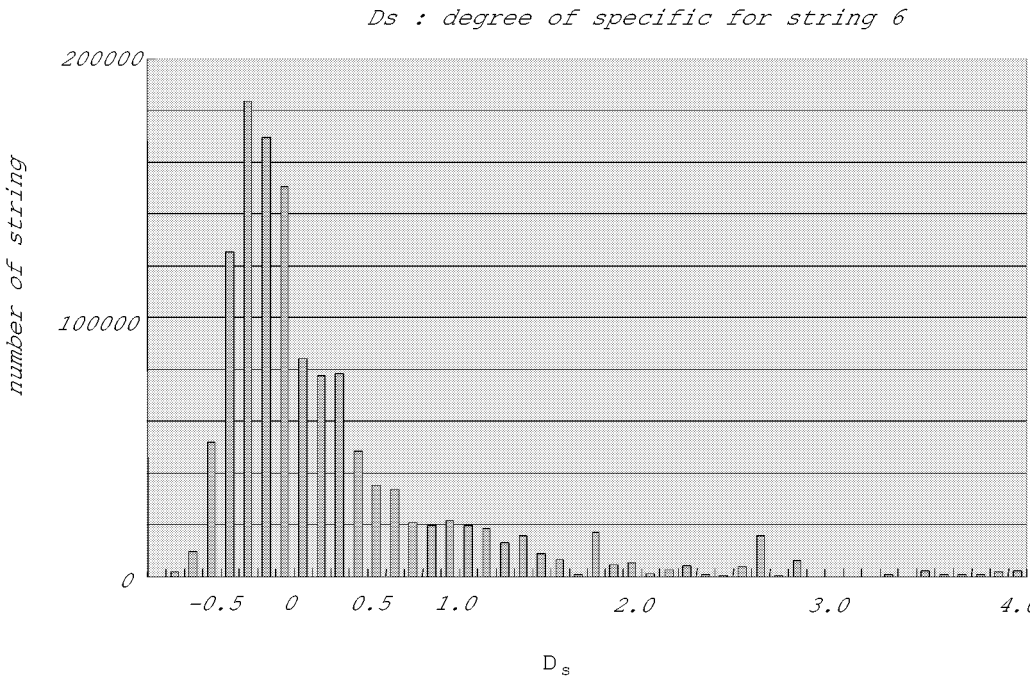

図 4.1: 文字列の特異度と出現頻度

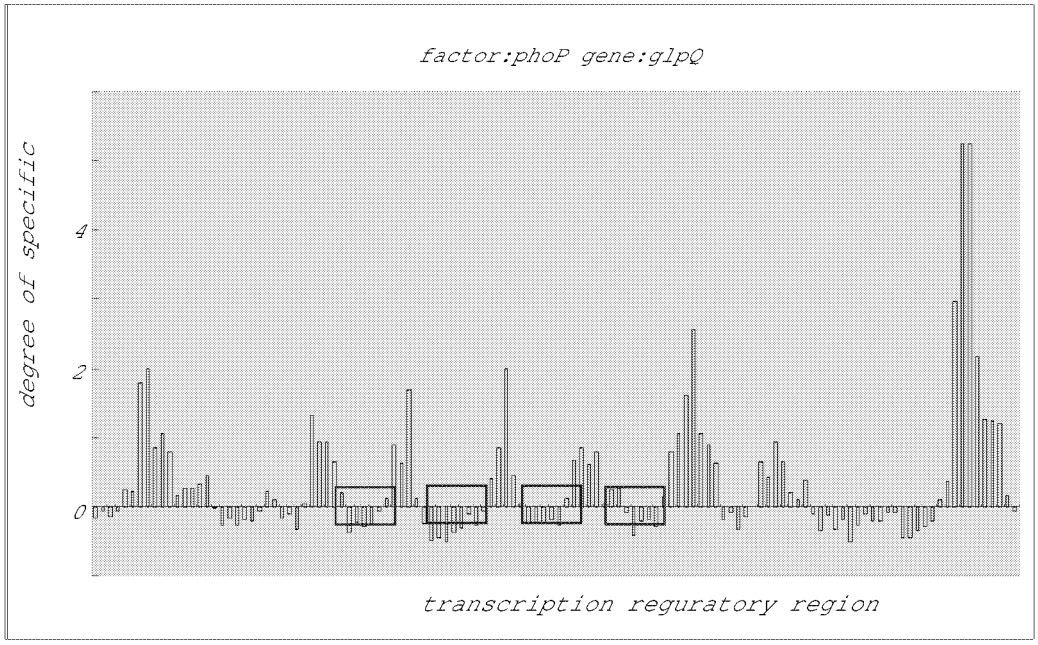

図 4.2: glpQ の転写制御領域に含まれる特異的文字列

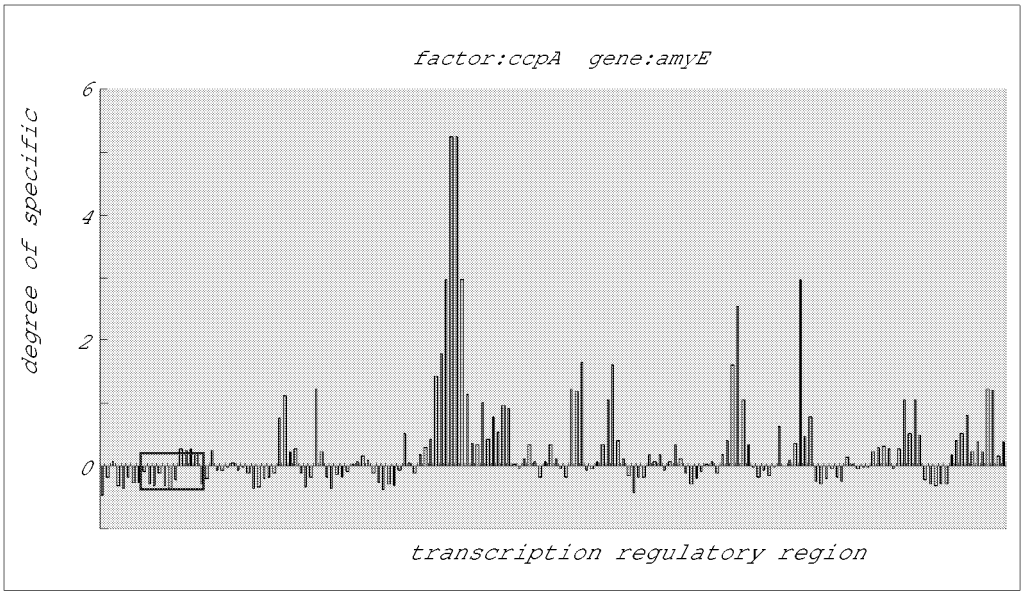

図 4.3: cotC の転写制御領域に含まれる特異的文字列

4.4 特異的文字列の候補

特異度で表現された転写制御領域について見てみると、既知の特異的文字列が特異度のマイナス領域に多くみられることが分かった。従って、ここでは既知の特異的文字列が定義した特異度に対してどのような値をとるかを調査する。調査対象となる既知の特異的文字列は、長さ 6 の文字列について調査していることから、それ以上の長さの文字列とし、定義した転写制御領域に含まれいていた１９８の文字列である。そして、これらの文字列について調査した。結果、既知の特異的文字列は特異度のマイナス領域に多く現れるという統計的特徴を得た。そして、ある特異度以下の文字列がカバーする既知の特異的文字列の割合は、図 4.4 のようになることが分かった。例えば、−0.1 以下では、既知の特異的文字列全体の 90[%] をカバーする。 また、特異度が −0.1 以下の長さ 6 の文字列の数は、2294 であり、文字列の種類数で見ると全体の半数近くに絞ることができる。そのため、本研究における特異的文字列の候補を特異度 −0.1 以下の文字列集合とし、以後の調査はそれらを対象に行なうこととする。

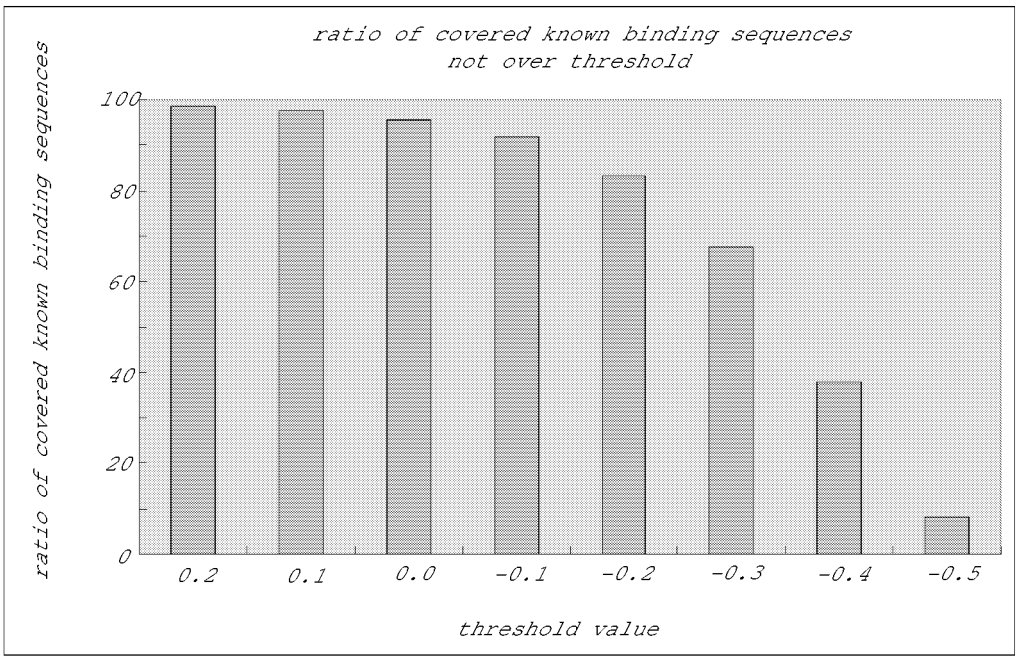

図 4.4: 特異的文字列の候補

第5章　実験３：転写制御領域の類似性に関する調査

5.1　類似性について

本章では、類似度を定義し、各遺伝子転写制御領域の類似性を評価する。従来研究による類似性は２つの文字列にアラインメント、３つ以上の文字列に対するマルチプルアラインメントなどがあり、文字列内において文字の置換、削除、挿入を行なう。本研究では、類似性を局所的な領域に含まれる長さ６の特異的文字列の候補どうしを比較することを特色としている。これを類似度に実装するためには様々な方法があるが、従来研究で評価される点を含み、さらに従来研究では評価されない点も評価するような類似度を定義する。このような類似度を本実験で適用することは、生物学的な特異性を広く評価することでもあると考えられる。また、実験を繰り返し行なうことで適切な条件を得ることができる。

5.2　類似度の定義

本節では、本研究で使用する類似度を定義する。

1. 各遺伝子の転写制御領域
   各遺伝子 g がもつ転写制御領域を $t(g,d)$ とする。
   ただし d は転写制御領域の４つの方向を指し、 $d=(0,1,2,3)$ である。(図 5.1)（転写制御因子の結合は文字列の方向に依存しないと考えられているため、逆方向の配列も定義する。）

2. ウィンドウ
   ウィンドウ W を連続する局所的な転写制御領域の長さ $L \geq 6$ とする。今回の実験では転写制御領域の長さの最低が３０であるため、$L=30$ とする。従って、ある遺伝子 g の長さ L の転写制御領域に含まれるウィンドウ L の数は、

   $$W_{total} = \sum_{d=0}^{3}(t(g,d)-29)$$

   となる。また、遺伝子 g の転写制御領域の i 番目の文字から始まるウィンドウを $w_g^i[k]$ とし、そこに含まれる特異度が閾値以下の文字列の集合を $S_g^i[k]$ とする。

方向

開始位置 → 終了位置

配列０	A	T	T	T	C	G	G	C	A	T	G	C	T	C	C	C
配列１	T	A	A	A	G	C	C	G	T	A	C	G	A	G	G	G

（配列０の相補配列）

配列２	C	C	C	T	C	G	T	A	C	G	G	C	T	T	T	A

（配列０の逆配列）

配列３	G	G	G	A	G	C	A	T	G	C	C	G	A	A	A	T

（配列２の相補配列）

図 5.1: 各遺伝子から取り出す転写制御領域

3. 文字列に対する類似度
ある２つの特異的文字列 s, s' の類似度を $\delta(s, s')$ とする。２つの文字列の類似度 $\delta(s, s')$ は、一致する文字数を k としたとき、長さ 6 のランダムな２つの文字列において k 文字が一致する確率の逆数とする。ただし、ランダムな文字列集合において $\delta(s, s')$ の平均値 m を引く。実際のデータにおいて X 軸を文字列対数、Y 軸を類似度の合計としたグラフを描くとほぼ直線となる。したがって、回帰分析により得られた直線が傾きが 80.3 であり、 m の実測値になる。
従って、類似度 $\delta(s, s')$ は、

$$\delta(s, s') = \frac{1}{(1/4)^k (3/4)^{6-k}} - 80.3$$

となり、各の値 k に対する類似度の値は、表 5.1 となる。

4. ウィンドウに対する類似度
遺伝子 A と遺伝子 B のウィンドウ間の類似度は、各ウィンドウに含まれる特異的文字列に対して計算する。
ある遺伝子 A のある位置 k のウィンドウ $W_A^{30}[k]$ に含まれる文字列集合と遺伝子 B のある位置 k' のウィンドウ $W_B^{30}[k']$ に含まれる文字列集合は、それぞれ $S_A^{30}[k]$ と $S_B^{30}[k']$ であり、それぞれの集合に含まれる文字列数を n, n' とし、２つのウィンドウの類似度を

$$\delta(W_A^{30}[k], W_B^{30}[k']) = \sum_{i=0}^{n-1} max_{j \in 0,k'}(\delta(s[i], s'[j]))$$

とする。

表 5.1: 文字列の類似度スコア

k	$\delta(s, s')$
0	5.619
1	16.86
2	50.57
3	151.7
4	455.1
5	1365
6	4096

5. 遺伝子間の類似度
   遺伝子 A と遺伝子 B の類似度は、それらの転写制御領域に含まれる各ウィンドウの類似度の最大値とし、

$$\delta(gene[A], gene[B]) = max_{k \in gene[A], k' \in gene[B]}(\delta(W_A^{30}[k], W_B^{30}[k']))$$

   とする。

5.3 従来研究と比較した特色

上で定義した類似性は、従来研究における類似性とは異なる構造を評価する。従来研究では、ある 2 つの文字列 s, s' における文字の挿入、削除、置換の回数に依存した類似性評価を行なう。このような文字列比較の問題は以前から研究されており、遺伝子研究に対しても多くの実績が挙げられている。従来手法を依存関係推定において適用する場合、まず、依存関係既知の遺伝子群から転写制御領域の類似性により文字列パターンを得て、次にそのような文字列パターンと類似した文字列をもつ未知の遺伝子を発見する。この方法によって導かれる未知の遺伝子群と他の遺伝子群との関係は図 5.2 のようになる。

従来の手法は、依存関係既知の遺伝子群で見られる特徴を文字列パターンとする。つまり、転写制御因子の特異的結合文字列を文字の削除、挿入、置換のみを用いて、直列に表現するため、文字列の縮退などは考慮されない。したがって、そのような変化が生じた遺伝子の転写制御領域からは、類似した特徴を見つけられないため、対象から外される。本研究での類似性は、はじめに特異的文字列の候補を取り出す。(図 5.3)

次に、各遺伝子転写制御領域に含まれる閾値以下の文字列に対して、その類似性を評価する。したがって、図 5.4 のような集合において類似性を評価する。

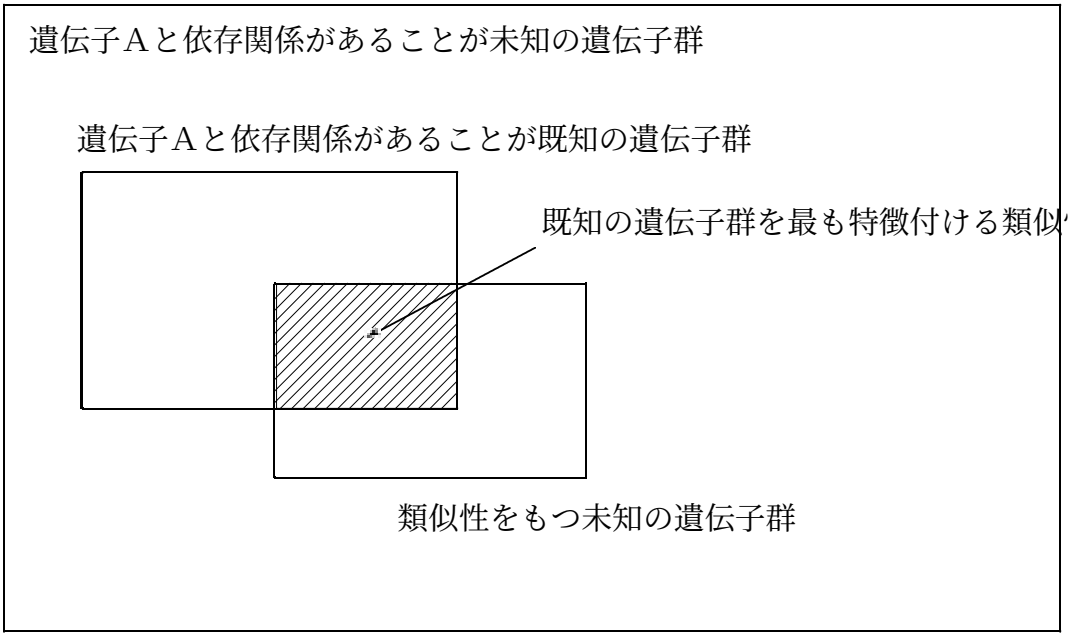

図 5.2: 従来手法による依存関係推定

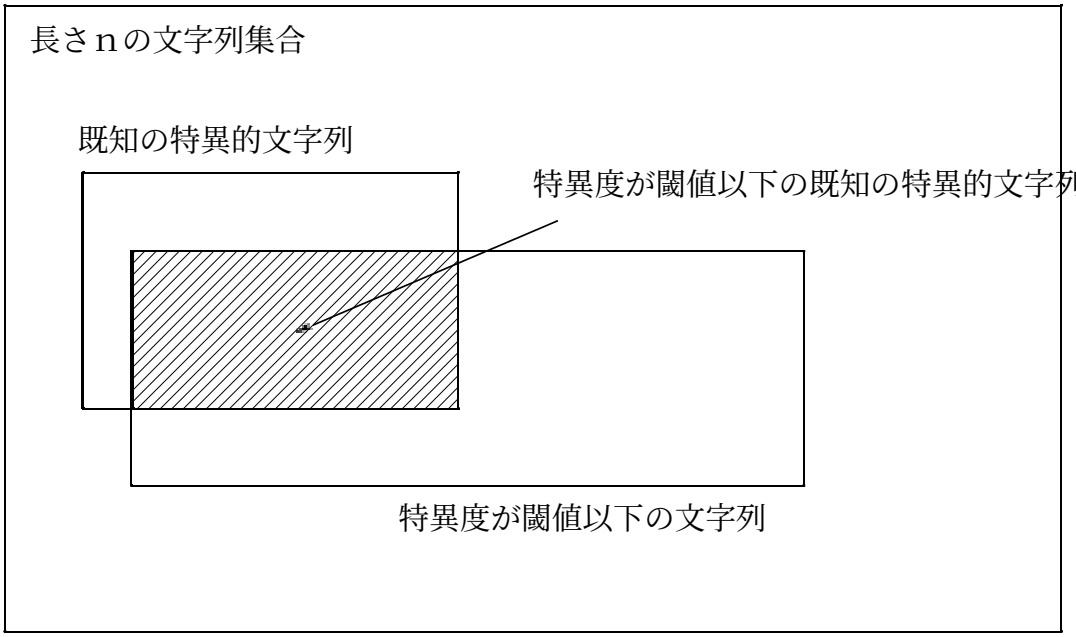

図 5.3: 本研究による依存関係推定

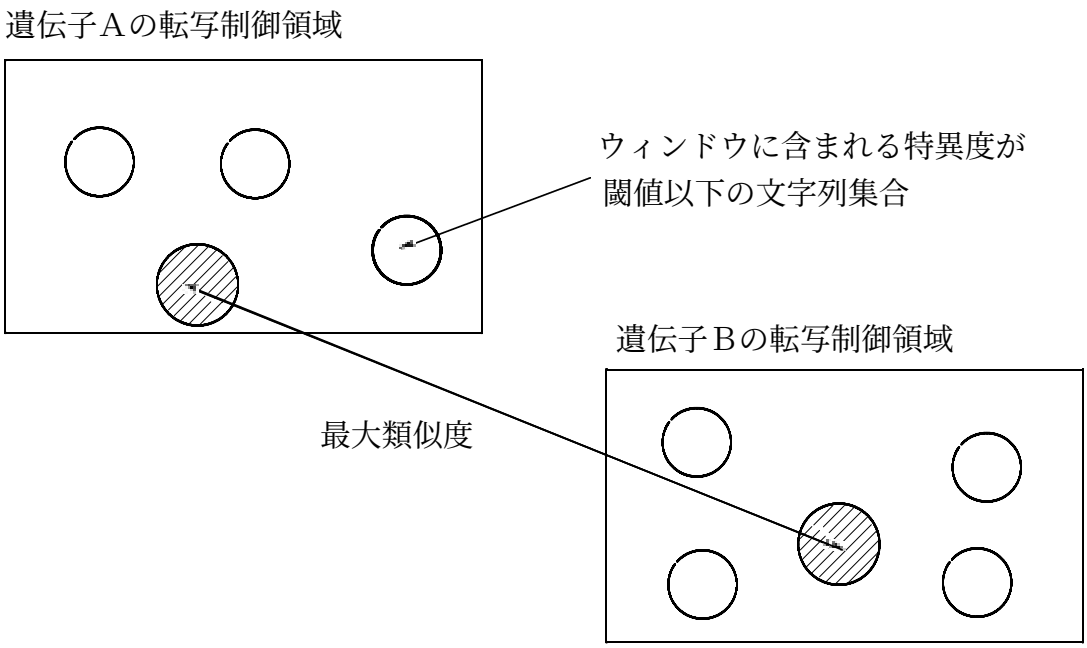

図 5.4: 局所的な領域に含まれる特異的文字列の評価

5.4 制御関係既知の遺伝子群についての調査

定義された類似度を用いて各遺伝子の転写制御領域の類似度を計算する。今回の調査では、依存関係が既知である遺伝子群を対象として行なう。ある転写制御因子から共通の制御を受ける遺伝子群では、類似した文字列をもつと考えられる。したがって、共通の制御を受ける遺伝子群の間での類似度は、関係がない遺伝子間の類似度と比較して、高くなると考えられる。

図 5.5 は転写制御因子 phoP によって制御される遺伝子群の１つである glpQ と全遺伝子との類似性を比較した結果と DNA マイクロアレイデータとの結果を示した図である。四角の実線で囲われた領域が制御関係がありそうな遺伝子群であり、三角点は転写制御因子 phoP から制御を受けることが既知である遺伝子群である。スコアについて見てみると、９遺伝子のうち、６遺伝子は平均スコア 20598 より高いスコアを示しており、類似性が高いと考えられる。しかし、高いスコアを示す遺伝子は他にも多くあり、実際に制御関係がある遺伝子とそれらの遺伝子と区別する方法が必要である。また、高いスコアを示した遺伝子（phoB,tagD,tagA,phoD,ykoL,resD）に対する glpQ の転写制御領域の位置は、既知の結合領域ではなく、σ因子（sigA）が結合する領域を含む領域であった。他の遺伝子（phoA,tuaA,yjbC）に対する位置については、既知の結合領域を含む領域が評価された。σ因子は他の多くの遺伝子（σ因子 sigA では、１３９遺伝子）の制御にも関与しているため、類似度の調査において評価されやすいと考えられる。

次に、転写制御因子 ccpA に影響を受ける遺伝子群についての調査結果を図 5.6,5.7 に

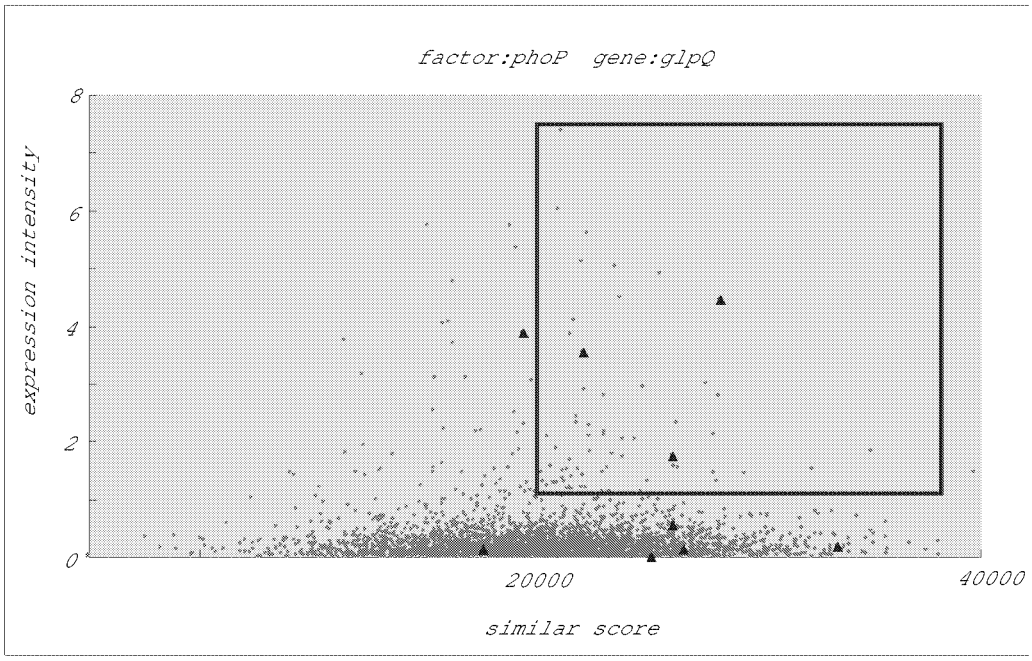

図 5.5: glpQ

示す。

CcpA から影響を受けることが既知である遺伝子は４０あり、図以外のスコアにおいてもほぼ同様にバラついた分布を示した。各遺伝子について高いスコアを示した位置を見てみると、既知の結合領域が含まれている遺伝子は少ない (４０遺伝子中３遺伝子)。また、window の幅を 15 、20 と変えて調査を行なった結果も同様な分布を示た。しかし、制御関係が既知である遺伝子以外の遺伝子では、比較される遺伝子によってスコアの順位が大きく変動する。従って、調査した遺伝子群において各遺伝子のスコアを合計してみた結果を図 5.8 に示す。

個々の遺伝子に対するスコアと比較した場合、制御関係が既知である遺伝子群のスコアが高くなっていることが分かる。このような分布になる原因として制御関係が既知である遺伝子の間にはバラつきの少ない類似性があると考えられる。つまり、この分布において低いスコアを示す遺伝子は、既知のある遺伝子についてのみ高い類似性を持ち、それ以外では低い類似性を持つと考えられる。今回調査した制御関係が既知である遺伝子群には、共通の類似性が保たれていることが知られていることから考えると、１つの遺伝子に対してのみ高い類似性を示すことが制御関係の推定に役立つとは言えず、あるグループにおいて共通の類似性が保たれていることが、制御関係の推定に役立つ情報であると考えることができる。

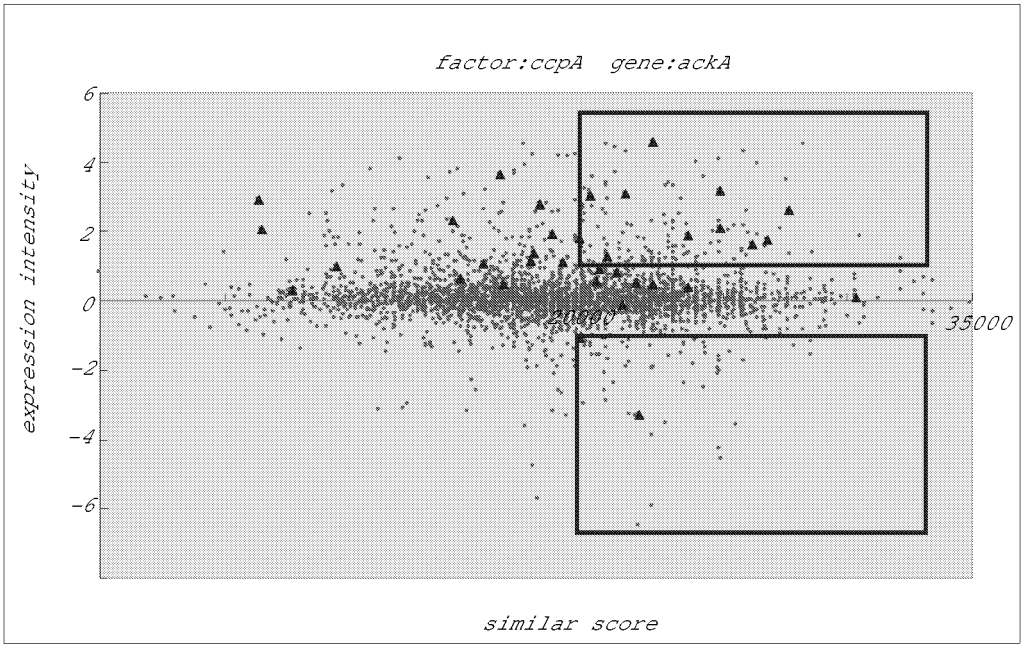

図 5.6: ackA

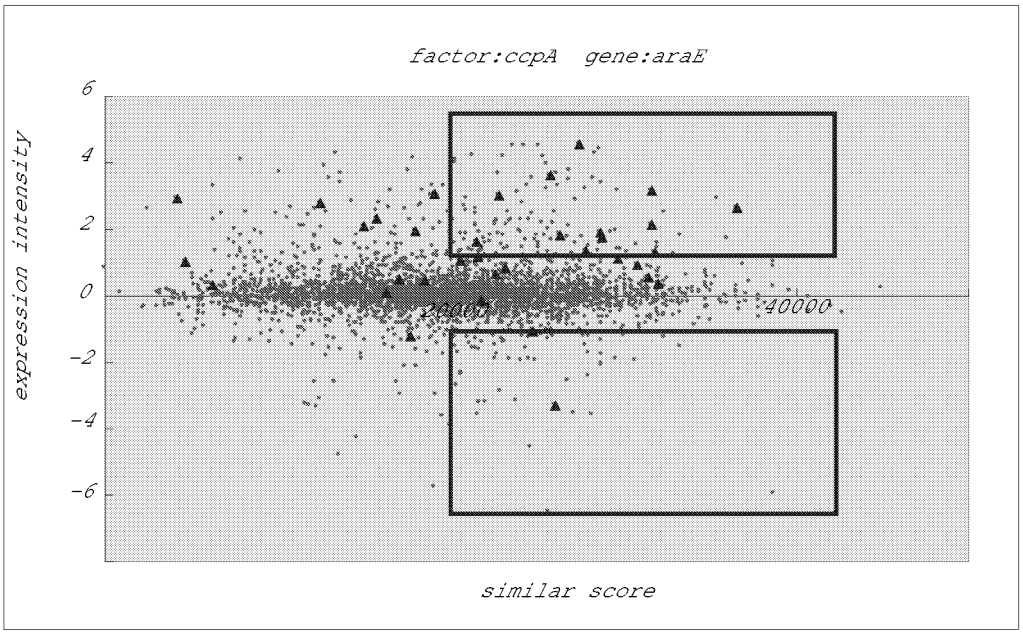

図 5.7: araE

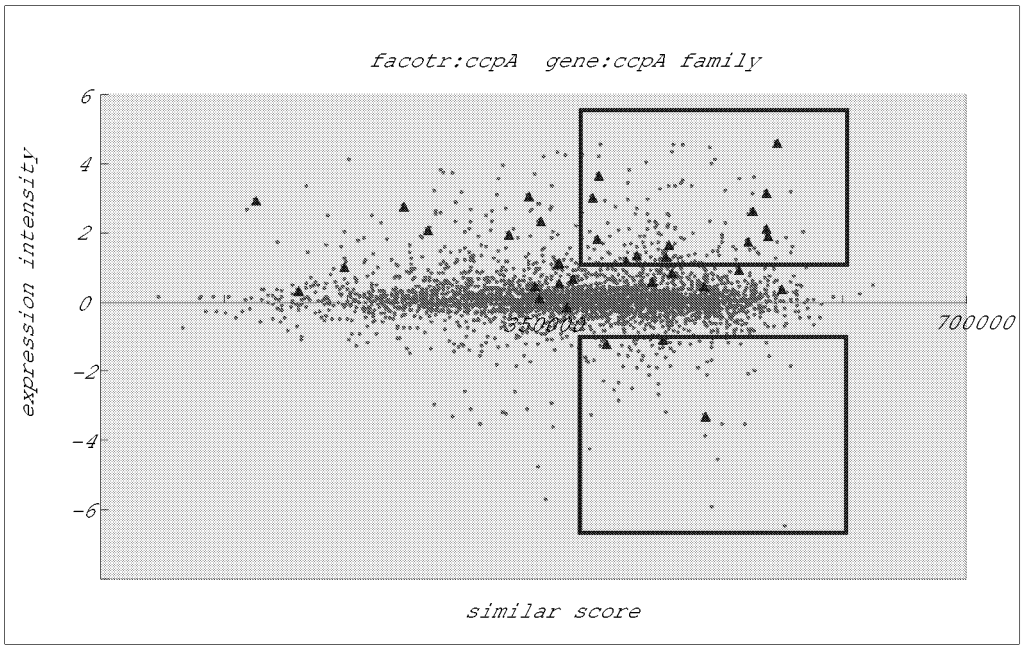
facotr:ccpA gene:ccpA family
expression intensity
6
4
2
0
-2
-4
-6
550000
700000
similar score

図 5.8: ccpA

第6章　考察

本研究では、はじめに全遺伝子転写制御領域を対象として統計的な調査を行ない、転写制御領域に含まれる文字列がもつ特異性を解析した。調査の結果、長さ 6 の文字列については、定義した特異度において既知の特異的文字列パターンがマイナスに多く含まれることが分かった。また、その割合は特異度が -0.1 において、既知の特異的文字列の 90[%] をカバーすることが分かった。そのため、類似性の評価では、特異度が閾値（ -0.1 ）以下の文字列を対象として、また、既知の特異的文字列の長さが 30 以下がほとんどであったことから、転写制御領域の局所的な領域について類似性を評価した。

6.1　転写制御領域について

6.1.1　転写制御領域の定義

本研究では、転写制御領域を遺伝子の開始位置から最大 500 までの上流領域とした。このように定義した理由は一般には転写制御領域は遺伝子の上流に存在するという抽象的な性質しか分かっていないためである。つまり、現在までに見つかっている転写制御因子の結合する領域の多くがそのような場所であるということである。そのため、明確な位置や範囲は規定できない。また、例外的に遺伝子の遥か上流に転写に関わる領域が存在する。例えば、転写に関わるエンサンサーやサイレンサーと呼ばれる領域が存在し、それらは遺伝子から数千塩基離れた場所に見つかっており、上流にも下流にも見つかっている。従って、そのような稀なケースを含めて考えることは転写に関わらない多くの場所を含んでしまうため、それらの領域を研究対象として扱うことは難しい。また、従来研究においても転写制御領域を遺伝子の上流 500[bp] 程度までとすることが一般的である。

6.1.2　転写制御領域の開始位置

転写制御領域の開始位置は、転写開始位置から上流であると考えることが自然に思われる。しかし、本研究では転写開始位置の明確な情報は公開されていないため、遺伝子の開始位置から上流とした。一般に遺伝子の転写は、遺伝子配列以外の余計な配列を含んで転写される。そして転写後、スプライシングによって余分な部位が切り取られる。従って、遺伝子の開始位置よりも上流を転写範囲に含むことが知られているが、その範囲は知られていない。

6.1.3　転写単位

遺伝子の転写に関しては、複数の短い遺伝子が同時に転写されることがあり、それら複数の遺伝子を１つの転写単位として扱うことがある。そのような転写単位はオペロンと呼ばれる。従って、実際の転写に関わるのは先頭に位置する遺伝子の転写制御領域であると考えられるため、それらを考慮した転写制御領域を検討する必要がある。しかし、オペロンを考慮する場合、その転写単位に含まれる遺伝子を特定しなければならず、より専門的な情報が必要である。

6.2　特異的文字列について

6.2.1　特異性

本研究では、統計的特異度を定義し、その特異度が閾値（ -0.1 ）以下の文字列を特異的文字列の候補とした。　これまでの特異的文字列の発見は、制御関係が既知である遺伝子群において発見されているため、発見された文字列は制御関係に深く関わると考えられる。本研究ではそのように発見された文字列の統計的な側面を調査し、制御関係推定にフィードバックしている。このような統計的特徴を制御関係推定に利用する手法では、その統計的特徴の捉え方が重要であることは直感的に理解できる。その捉え方の重要性とは、統計的特徴が生物学的な性質を説明できなければならないということである。そのため、ある統計的な尺度がもつ意味は常に重要である。今回の調査では、文字列の長さと出現頻度が統計的な尺度である。最も重要なことは、特異的文字列は転写制御因子の結合能力を表すということである。従って、その能力を表すために文字列の長さ、影響を受ける遺伝子数、類似度を使用した。文字列の長さは転写制御因子の物理的構造、文字列の出現頻度は転写制御因子から影響を受ける遺伝子数、類似度は転写制御因子が持つ文字列の認識能力をそれぞれ表すと考えられる。

6.2.2　文字列の長さ

物理的構造は複雑であり、転写制御因子によって異なるため、個々の特徴を個々の文字列の長さに対応させることは難しい。特に全遺伝子を対象とした調査の場合は特定の転写制御因子を想定していないため、文字列の長さは既知の転写制御因子が持つ特徴を全て含んでいることが望ましいと思われる。従って、異なる転写制御因子の特徴を含んだ文字列の長さは、曖昧さを含むものとなってしまう。そのため、より厳密な物理的構造を特徴付けるためにクラス分けなどを検討する必要があると考えられる。

6.2.3　文字列の出現頻度

文字列の出現頻度は、それらの文字列に結合する転写制御因子から影響を受ける遺伝子数を表すと考えられるが、それは類似度と大きく関係する。転写制御因子からみた場合、異なる文字列であっても同じ文字列として認識し結合するため、そのような文字列をもつ遺伝子が真の遺伝子数である。しかしながら、この節では厳密な文字列のみの出現頻度を扱い、類似度に関しては次節で扱うこととする。従来研究によって同じ文字列パターンを持つ遺伝子であっても影響を受けない場合があることが知られている。その原因は非常に沢山あり、複雑である。また、一方ではそのような文字列を持つ遺伝子群が共通の制御を受けると判断することは統計的に有意であることも知られている。これらの理由により、文字列の出現頻度だけでは影響を受ける遺伝子群と判断できず、また調査の結果、既知の特異的文字列は出現頻度によらないことが示され、統計的特異度を定義するに至った。

6.2.4　統計的特異度の定義

統計的特異度の定義については、はじめに文字列の出現確率は文字の個数のみに依存し、それを確率変数とみなすと２項分布となり、χ^2 検定によりに特異性を判定することが目的であった。しかし、文字列の出現頻度と期待値との差の二乗と期待値との比では、有意に差はないという結果であった。また、既知の特異的文字列を含む遺伝子の転写制御領域を χ^2 値で表すと、既知の特異的文字列の多くが定義された特異度の谷に含まれていた。有意に差がないという結果は、出現頻度と期待値との差がプラスでもマイナスでも二乗によって同等に扱われてしまうためであり、また、既知の特異的文字列が谷に含まれていることをより詳しく知るために、差の二乗から乗数を外し出現頻度が期待値よりも多いか少ないかを表せるようにした。その結果、既知の特異的文字列の多くが定義された特異度のマイナスに含まれていた。ここで想定した出現確率は、全遺伝子転写制御領域に含まれる塩基文字の割合の逆数を一文字の出現確率として、ある長さの文字列をそれらの組み合わせの積で表した。従って、この出現確率は文字列に含まれている文字の個数のみに依存する。文字列の出現確率を想定するにあたって、これ以外の要素を取り入れることは現在のところ、相応しくないと思われる。

6.2.5　閾値

閾値（－０．１）については、そのような値によって文字列の集合を分けることで、既知の特異性を保つ集合が得られると考えた。値の選定では、既知の特異的文字列を全て含むような値では特異的文字列の候補数が非常に多くなる。しかし、それでは特異性の低い文字列を多く含むことになるため、文字列を半数近い２２９４を含み、かつ、既知の特異的文字列の９０％をカバーすることで、その特異性を保つように値を選定した。

6.3 類似性について

本研究では、転写制御領域の局所的な領域について、そこに含まれる特異的文字列の候補を対象として類似度の計算を行なった。遺伝子に関する類似性の研究はこれまでにも盛んに行なわれているため、同様の手法を用いることも可能である。しかし、本研究では長さ 6 の文字列において特異的文字列の候補となる文字列集合を選択し、それらを対象とする。そのため、類似性を比較するのは長さ 6 の文字列どうしであり、また、局所的な領域に含まれるそれらの文字列どうしの比較の合計がその領域の類似性を表すため、従来手法を直接的には適用できない。また、転写制御因子は文字列の類似性を認識し結合すること以外に、領域の立体構造を認識し結合すること、つまり、らせん構造の角度を認識し結合することが生物学的に知られている。そのため、今後は本研究での問題点と合わせて立体構造を類似性に反映させることを検討する。

6.3.1 文字列の類似度

文字列の類似度については、一致する文字数をｋとしたとき、長さ 6 のランダムな２つの文字列においてｋ文字が一致する確率の逆数とした。類似度の計算では、文字の削除、挿入、置換操作による方法が一般的であり、それらの操作の回数や得られた文字列における文字の一致数でスコアが決まる。しかし、本研究において削除、挿入を行なう場合、隣り合う文字列に影響を及ぼす。もし隣の文字列が特異的文字列の候補でない場合、空白となってしまう。したがって、長さ 6 の文字列の比較において削除、挿入操作は考慮しなかった。

6.3.2 局所的な領域の類似度

局所的な領域の類似度については、ウィンドウの幅を 30 として、その領域に含まれる特異的文字列の候補どうしの類似度を計算した。ウィンドウの幅については、既知の特異的文字列のほとんどが 30 文字以下であることと、転写制御領域の長さの最小が 30 であるため、その幅を最大とし、それによって全ての遺伝子について偏りのない類似度を計算できると考えた。しかし、既知の特異的文字列がウィンドウの幅よりも短い場合、それらを含む領域の類似度は類似度の低い領域を多く含んでしまうので、領域全体のスコアは低くでてしまう。従って、ウィンドウの幅は今回使用したよりも短くすることも検討しなければならない。ウィンドウ内の文字列の比較は、ウィンドウ A 内のある１つの文字列に対してウィンドウ B 内の全ての文字列の類似度を計算し、その最大値をウィンドウ A のある１つの文字列の類似度とした。そのため、従来研究よりも遥かに広い類似性を含むことになり、高いスコアを示すウィンドウであっても、人から見た２つの文字列は類似しているようには見えない場合が多くなる。これは従来研究によってカバーされない特異性を含むように定義したためであるが、今後改善する必要がある。改善方法としては生物学的

な結果を示すことできる条件を発見することがあると考えられる。つまり、遺伝子 A のある文字列と遺伝子 B のある文字列の類似度、距離、順序等の条件をスコアに反映させることである。

6.3.3 遺伝子間の類似度

遺伝子間の類似度は、各遺伝子のウィンドウの比較において最大を示した類似度とした。本研究では局所的な類似性を評価するため、遺伝子間の類似度は多数あることになる。従って、通常それらの中から選択する場合、その最大値を選択することが正しいと考えられる。しかし、ｇｌｐ Q と他の遺伝子との比較においては、特異的文字列が含まれている領域は最大とならず、σ因子が結合する領域で最大を示した。従って、類似度の最大のみを遺伝子間の類似度とすることには問題がある。このような結果を回避するためには、２つの遺伝子間での類似度だけではなく、複数の遺伝子を対象として共通の位置での類似度を合計したものを評価することで、既知の特異的文字列を含む遺伝子群のそれらを含む領域を高く評価できると考えられる。

6.3.4 計算量

本研究では４１１３の遺伝子を対象として各遺伝子転写制御領域に含まれる文字列どうしの類似度を計算した。本研究で用いた手法によって表現される遺伝子間の類似度に対する計算量は、各遺伝子の転写制御領域の長さ L_t、取り出す文字列の長さ L_s、ウィンドウの長さ L_W に依存する。従って、１つの遺伝子間の類似度の計算量は

$$(L_W - L_s + 1)^2 \times (L_t - L_W + 1)^2$$

となる。また、各遺伝子と全ての遺伝子を比較することから、遺伝子の比較回数は、4113^2 である。従って、全ての遺伝子についての類似度の計算量は、

$$4113^2 \times (L_W - L_s + 1)^2 \times (L_t - L_W + 1)^2$$

となる。各遺伝子の転写制御領域の長さを $L_t = 500$ であると仮定するとし、文字列の長さ $L_s = 6$、ウィンドウの長さ $L_W = 30$ の条件の下での全遺伝子間の計算量 O は、

$$4113^2(30 - 6 + 1)^2(500 - 30 + 1)^2 = 2.3 \times 10^{15}$$

となる。このように全ての遺伝子を対象とした類似性の計算には非常に多くの時間が掛かると考えられる。そのため類似性の計算するにあたっては類似性の定義、アルゴリズム、計算結果の評価方法を十分に検討しなければならない。

6.4　遺伝子間の制御関係推定について

本研究では、遺伝子転写制御領域に含まれる特異的文字列の類似性とＤＮＡマイクロアレイデータを用いて遺伝子間の制御関係を推定する手法を検討した。これまでにも遺伝子間の制御関係推定に関する様々な研究が行なわれている。本節では本研究で扱った２つのデータを用いた手法の有効性について検討する。

6.4.1　転写制御領域の類似性による推定

遺伝子間の制御関係に大きく関わる転写制御領域に関する研究は以前から行なわれており、多くの成果を残している。代表的な研究には先にも述べたコンセンサスやモチーフといった研究があり、遺伝子研究に広く用いられている。これらの手法は配列の類似性に関する研究であり、類似した配列の発見に役立つ。しかしながら、遺伝子間に制御関係があることが既知である遺伝子群の転写制御領域であっても、類似性が見られない配列もある。また、ある集合の類似性を評価した結果得られた配列は、配列の集合においてそれぞれの配列と距離がある。さらにそのような配列を持つ遺伝子であっても制御関係に関わらない遺伝子も存在する。このような理由の背景には複雑な生物のシステムが存在する。そのため、現在のような局所的な生物システムを表現する手法は多くの誤りを含むことは避けられないと考えられる。このように生物のシステムを直接的に観測できない立場においては限界がある。つまり、情報科学における遺伝子研究では、情報の不足や生物学的実験が行なえないなどの制約がある。したがって、入手できる情報から考えられる情報を取り出し、遺伝子の理解に役立てられることが大きな使命であると考えられる。実際、類似性研究を用いた遺伝子発見や相同性検索は現在の遺伝子研究に大きく貢献している。従って、情報科学の立場から行なう遺伝子研究の有意性はあると考えられる。

6.4.2　ＤＮＡマイクロアレイデータによる推定

遺伝子の発現を定量的に観測したデータを用いることは、生物学的な説明をともなっている。しかしながら、複雑な遺伝子間の依存関係においては、様々な要因が作用することが知られている。そのため、単独のデータからの推定は非常に難しいと言われている。このようなことから、ＤＮＡマイクロアレイデータを用いた推定のためにデータマイニングを応用した研究が盛んに行なわれている。

6.4.3　転写制御領域の類似性とＤＮＡマイクロアレイデータによる推定

現在、ＤＮＡマイクロアレイデータ以外の生物学的情報を付加した推定方法が注目されている。そのような情報としては、蛋白質間相互作用、蛋白質－ＤＮＡ相互作用、プロモーター領域に含まれる共通配列などがある。これらはいずれも遺伝子の発現に関する情

報を含んでいる。そのため、遺伝子間の依存関係推定にとって重要な情報を与えると考えられる。

謝辞

本研究を行うにあたり、多大なるご指導・御鞭撻を賜りました平石 邦彦 教授に深く感謝の意を表します。また、本研究をまとめるにあたり、有益な御助言を頂きました宋 少秋助手、土井 洋文 セレスター・レキシコ・サイエンシス株式会社代表取締役社長 様に心より感謝します。

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諸先輩方、平石研究室の皆様に厚くお礼申し上げます。

最後に、大学院での貴重な研究生活を与えて頂き、暖かく見守ってくれた両親、兄妹、そして友人に心から感謝致します。

参考文献

[1] 松原謙一　編集 “ゲノム機能”, 中山書店,2000

[2] 小倉亨 “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　修士論文,2003

[3] 堀越正美　編著 “遺伝子発現”, 中外医学社,2001

[4] 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

[5] 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

[6] 高木利久　冨田勝　編集 “ゲノム情報生物学”, 中山書店,2000

[7] Ann Vanet,Laurent Marsan,Mrie-France Sagot, “Promoter sequences and algorithmical methods for identifying them”,1999 Editions scientifiques et medicales Elsevier SAS, Microbiol.150(1999),779-799

[8] Yishai M.Fraenkel,Yael Mandel,Devorah Friedberg and Hanah Mrgalit “Identification of common motifs in unaligned DNA sequence : application to Escherichia coli Lrp regulon”,CABIOS vol.11 no.4(1995) 379-387

[9] F.kunst, N.Ogasawara, and other researchers “The complete genome sequence of the Gram-positive bacterium Bacillus sutilis”、Nature vol.390 、November 1997

[10] Jaak Vilo, Alvis Brazma, Inge Jonassen, Alan Robinson, Esko Ukkonen ‘Mining for putative regulatory elements in the yeast genome using gene expression data”、American Association for Artifical Intelligence、2000