NAIST-IS-MT0851144

修士論文

Escherichia coli の Fur 結合領域における コンセンサス配列探索

小川 祐樹

2010 年 2 月 4 日

奈良先端科学技術大学院大学
情報科学研究科 情報生命科学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
修士(工学)授与の要件として提出した修士論文である。

小川 祐樹

審査委員：	金谷 重彦	教授	(主指導教員)
	湊 小太郎	教授	(副指導教員)
	Md.Altaf-Ul-Amin	准教授	(副指導教員)
	中村 建介	特任准教授	(副指導教員)
	高橋 弘喜	助教	(副指導教員)

Escherichia coli の Fur 結合領域における コンセンサス配列探索*

小川 祐樹

内容梗概

コンセンサス配列とは、ある特定の機能を持つ遺伝子領域中に高頻度に見られる塩基配列パターンのことである。コンセンサス配列を特定することにより、遺伝子上に存在する別の未知コンセンサス配列の予測、さらにはその遺伝子の機能予測を可能にすると期待されている。しかし、コンセンサス配列にはゆらぎが大きく、その法則性も曖昧である。コンセンサス配列を探索する手段は様々な統計学的指標を用いて考案されているが、正確に予測できるものはまだ確立されていない。

コンセンサス配列を持つものの一つとして、真正細菌の FUR-box が知られている。真正細菌にとって、鉄は生存に必要不可欠な要素のひとつである。細菌生体内の鉄の恒常性は siderophore と呼ばれる三価鉄輸送キレートによって保たれている。Fur (Fe up-take regulator)はsiderophore コード領域上流の DNA 領域と結合することで、siderophore の転写を厳密に制御している。この Fur が結合する DNA 領域を FUR-box と呼ぶ。

本研究では、実際の実験データを用いた新たなアプローチでコンセンサス配列を探索する新規アルゴリズムを構築した。このアルゴリズムを使用するデータモデルとして、真正細菌 *Escherichia coli* の W3110 株とO157 Sakai 株のデータを用いた。本研究の手順として、まず *Escherichia coli* W3110 株と O157 Sakai 株において Fur をプローブタンパクとした ChIP-on-chip データを用い、Fur 結合領域を選出した。さらに、ChIP-on-chip のシグナル値、塩基配列の出現頻度による p値、そして重み行列を考慮した新規探索アルゴリズムを作成し、コンセンサス配列を推定した。そしてEM アルゴリズムを用いた既存配列探索モデルである MEME と比較評価することで、その有用性を検証した。その結果、現段階での精度はMEMEに劣るものの、今後の改善次第ではさらに良いアルゴリズムになることが分かった。

キーワード:

コンセンサス配列、Fur、FUR-box、ChIP-on-chip、p値、重み行列

*奈良先端科学技術大学院大学 情報科学研究科 情報生命科学専攻 修士論文

NAIST-IS-MT0851144 2010年2月4日

Searching the consensus sequence on the Fur-binding domain of *Escherichia coli* *

Yuki Ogawa

Abstract

A consensus sequence is a conserved nucleotide pattern which is usually found at meaningful sites across DNA and often works as a binding site for a protein or a protein complex or other bio-molecules associated to certain function. Determination of the consensus sequence is expected to lead to identify similar consensus sequence at other places on the genome DNA, furthermore the function prediction of the related gene. Most of the defined consensus sequences, however, have no clear rule of consensus pattern. To discover consensus sequence, many researchers have suggested methods with various statistical measures; nevertheless all-powerful method has not been conceived yet. The "FUR-box" of bacteria is known as one of consensus sequences associated with iron transportation. Iron is one of essential elements for survival of most bacteria. In bacteria's cell, iron III chelator called "siderophore" strictly maintains the iron homeostasis. Fur (Fe uptake regulator) closely controls the transcription of siderophore by binding the upstream of the DNA sites encoding siderophore synthesis. The genome region on which Fur binds is called FUR-box. In this study, I propose a new algorithm to find consensus sequence based on ChIP-on-chip data and weight matrix. First, I elected several genome regions expected to contain Fur-binding domain by using the ChIP-on-chip experimental data, in which Fur is used as the probe protein, of *Escherichia coli* str. K-12 substr. W3110 and O157:H7 str. Sakai. Secondly, I built a new algorithm for searching consensus sequence as follows; extracting short motifs based on χ^2 test from ChIP-on-chip experimental data, estimating consensus sequence of FUR-box based on the position weight matrix. Lastly, I investigated the efficiency of the new algorithm by comparing with the results from MEME; it is the existing tool with EM algorithm. From comparative analysis, it has become clear that new algorithm has lower precision than MEME, still it is a novel strategy and expected to be a better algorithm depending on some improvements.

Keywords:

consensus sequence, Fur, FUR-box, ChIP-on-chip, p-value, weight matrix

*Master's Thesis, Department of Information Bioinformatics and Genomics, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT0851144, February 4, 2010.

目次

第1章　　研究背景

1.1　コンセンサス配列探索について

コンピュータとインターネットの世界的普及、そして DNA シーケンス技術のめざましい発展によって、日々膨大な種の生物の塩基配列がデータベースに蓄積されている (Miyazaki, 2008)。これにより、この塩基配列から新たな生物学的な意味を探求する試みがなされている。その中でも特に重要なものの一つは、コンセンサス配列の探索である。コンセンサス配列とは、ある特定の機能を持つ遺伝子領域に共通して見られる配列パターンのことである。タンパクや代謝物質など、生物を形づくる要素の多くは、DNA 領域周辺に転写制御タンパクが結合し、 DNA 転写を調節することでその産生をコントロールしている。そして、転写制御タンパクが結合する部分にはコンセンサスな配列が多く見受けられる。それゆえにコンセンサス配列を特定することは、ゲノム上に存在する別の未知コンセンサス配列を予測すること、さらにはその部分の遺伝子機能を推定することにつながると期待されている。

その一方で課題もある。一概にコンセンサス配列といっても、そのパターンは多岐にわたる。ある程度の共通性はあっても、そこに存在するゆらぎは様々であり、その規則も曖昧なものが多い。また、コンセンサス配列はふつう数塩基から多くても数十塩基程である。数万、数十万塩基もあるゲノム配列の中からコンセンサスな配列を人の手のみで見つけることは困難である。しかし、コンピュータの高速な演算能力を利用することによって、短期間に膨大な計算を行うことが可能になった。それによって、コンセンサス配列を探索する研究アプローチが急速に進むことになった。

1.2　Fur について

鉄は多くの生物において、生存に必要不可欠な要素である。そのため、生体内における鉄分は厳密な制御機構によってその恒常性が保たれている (Stojiljkovic *et al.*, 1994)。真正細菌においては、siderophore と呼ばれる三価鉄イオン輸送キレートがその役割を担っている。siderophore の転写を制御しているレギュレータを Fur (Fe uptake regulator) と呼ぶ。Fur は真正細菌に幅広く保存されていることが知られている (Panina *et al.*, 2001)。Fur は鉄イオンの多い環境下で siderophore がコードされている遺伝子領域の上流に結合し、その転写を抑制する。逆に鉄イオンが少ない環境下では、Fur が外れることによって siderophore の転写が進む (Braun and Hantke, 1991)。Fur にはいくつかのホモログが知られており、それらは二価の金属イオン輸送タンパクのコード領域上流に結合することで、その転写を制御している。Zur (Zinc uptake regulator、亜鉛の制御) 、

PerR (peroxide regulon repressor、鉄やマンガンの制御) などがそれにあたる (Bsat *et al.*, 1998, Gaballa and Helmann, 1998)。また、Fur は病原性細菌において、毒性物質を生成する遺伝子の制御にも関わっているとされる (Calderwood and Mekalanos, 1987)。

Fur が結合する DNA 領域は FUR-box と呼ばれており、コンセンサスな配列を持つことが分かっている。*Escherichia coli* においては 19 塩基のコンセンサス配列 (GATAATGATATACATTATC) が報告されている (de Lorenzo *et al.*, 1987)。ただしこれは DNase I footprinting を用いた部分的な解析によるものであり (de Lorenzo *et al.*, 1988; Griggs and Konisky, 1989) 、網羅的に解析された研究成果はまだ報告されていない。また、鉄の恒常性保持の機構と毒性物質の生成の機構が異なることから、それらを制御する Fur、またそれが結合する FUR-boxのコンセンサス配列も似て非なるものである可能性がある。

1.3 研究目的

1.1 でも述べたように、コンセンサス配列(モチーフとも呼ぶ)を予測することは、その遺伝子機能を知るのに非常に有意義である。そのため、正規表現や相互情報量、重み行列など、様々な統計的指標を用いたモチーフのモデル化が提案され、そこからコンセンサス配列を探索するアルゴリズムが考案されてきた (Tsai *et al.*, 2006; Halpern *et al.*, 2007; Robin *et al.*, 2007; Miyazaki, 2008)。しかし、その規則性のないゆらぎが原因で、一意にコンセンサス配列を同定するアルゴリズムの確率にはまだ至っていない。また 1.2 より、*Escherichia coli* にとって重要な要素である鉄を制御する Fur、そして FUR-box のコンセンサス配列について、網羅的な研究を行った例はまだない。

そこで本研究では、ChIP-on-chip のシグナル値、塩基配列の出現頻度による p 値、重み行列を組み合わせた新たな統計的モチーフ探索アルゴリズムを作成し、*Escherichia coli* W3110 株と *Escherichia coli* O157 Sakai 株の Fur をプローブタンパクとした ChIP-on-chip のデータを用いて、*Escherichia coli* に共通する、もしくは W3110 株と O157 Sakai 株とで異なる FUR-box のコンセンサス配列を情報学的、統計学的に推定する。また、EM アルゴリズム (Dempster *et al.*, 1977) を用いた既存のコンセンサス配列検索ソフトウェアである MEME (Bailey and Elkan, 1994)で配列検索を行い、その結果と比較検討することで、この新規アルゴリズムの有用性を検証する。

第2章　　研究手法

2.1　データセット

2.1.1　*Escherichia coli* W3110 株と O157 Sakai 株のゲノム配列データ

本研究を行うにあたって、NCBI (National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/) に登録されている2種の真正細菌 *Escherichia coli* (大腸菌)のゲノム配列データを用いた。

・*Escherichia coli* str. K-12 substr. W3110
2009 年 8 月 16 日に取得、AP009048.1 、GI:85674274、4,646,332 bp

・*Escherichia coli* O157:H7 str. Sakai chromosome
2009 年 8 月 18 日に取得、BA000007.2 、GI:47118301、5,498,450 bp

2.1.2　Fur をプローブにした ChIP-on-chip 実験データ

本研究では、Fur をプローブタンパクとした *Escherichia coli* W3110 株 (総プローブ数 236,714) と O157 Sakai 株の ChIP-on-chip 実験データ(総プローブ数 127,269) を、各株 2 実験ずつ計 4 つのデータを用いて行った。なおこの実験データは、W3110 株については本学情報科学研究科システム細胞学講座より、O157 Sakai 株については大阪大学医学研究科感染防御学分野より提供されたものである。ここで ChIP-on-chip とは、タイリングアレイとクロマチン免疫沈降法を組み合わせた実験手法である。この手法を用いることにより、プローブタンパクが結合する DNA の領域を調べることができる (Buck and Lieb, 2004)。

・ChIP-on-chip 実験のワークフロー

(1) 大腸菌細胞をホルムアルデヒドで処理することにより、プローブタンパク(Fur)を標的 DNA 領域 (FUR-box) に結合(クロスリンク)させる。
(2) 細胞を溶解させて DNA を抽出し、超音波または酵素処理により DNA を断片化させる。クロスリンク DNA、通常の二本鎖 DNA ともに 1 kbp 以下のサイズに断片化される。
(3) 免疫沈降を行う。プローブタンパクに特異的に結合する抗体タンパクを用い、プローブと抗体の間に多数の架橋構造を形成させることで大きな構造体を作

り、不溶化、沈殿させる。クロスリンクしていない DNA 鎖は溶液中に残る。

(4) 遠心分離を行い、沈殿した DNA-プローブタンパク-抗体複合体を回収する。

(5) 洗浄後に PCR を行い、(4)で得た複合体の DNA のみを一本鎖で増幅させる。その後、ビオチン処理を施す。

(6) 元のゲノムデータからデザインされた、標的生物の全ゲノムをカバーする 25 mer ずつのオリゴヌクレオチドプローブをスライドガラス上に配置する。このとき、元のゲノム転写産物に完全に一致するパーフェクトマッチ配列、25 mer のうちの中央の 1 塩基のみ異なるミスマッチ配列を作成する。パーフェクトマッチ配列とミスマッチ配列におけるシグナル強度の差を考慮することにより、クロスハイブリダイゼーション由来のノイズを取り除くことができる。(ただし、今回の実験ではミスマッチ配列のデータは使用していない。)

(7) ビオチン処理した標的 DNA とプローブをクロスハイブリダイゼーションさせ、シグナル強度を計測する。

(8) これとは別に、特定のプローブタンパクとクロスリンクさせない DNA を用いて、コントロール実験のシグナル値を計測しておく。この値を考慮することにより、配列に特異的なノイズを取り除くことができる。

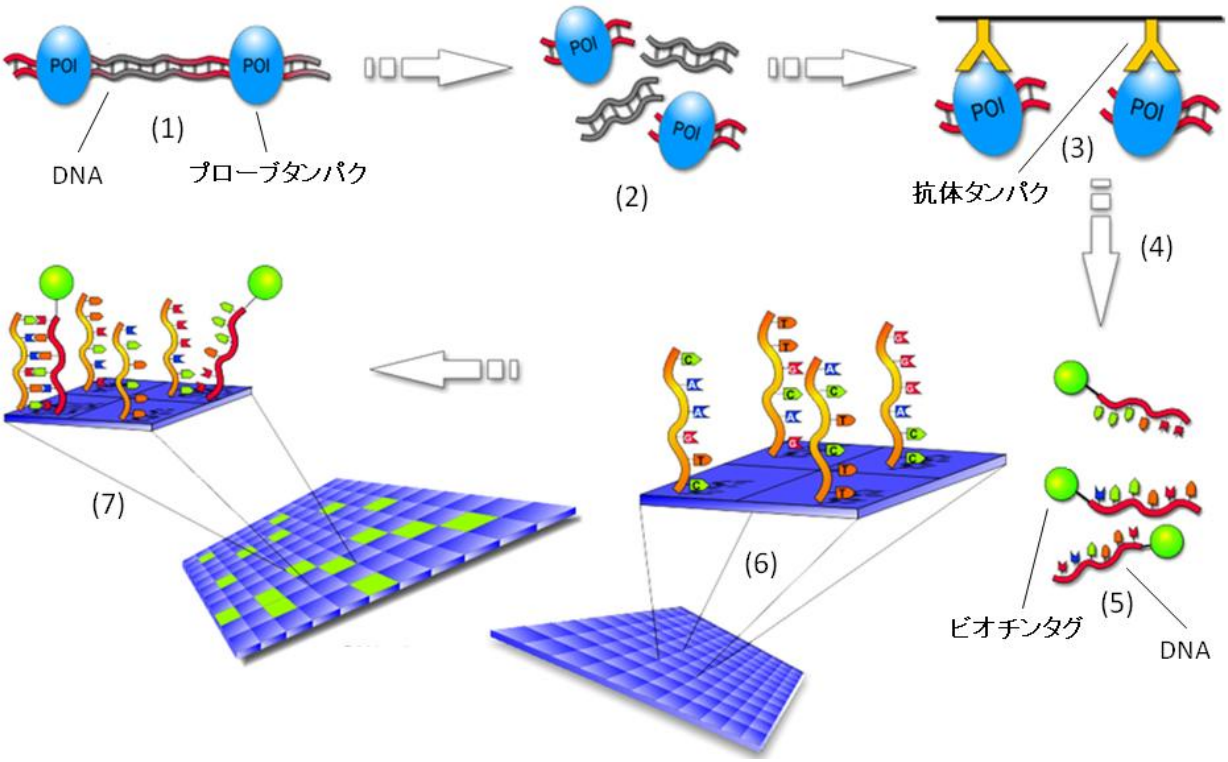


Figure 1. ChIP-on-chip の実験手順概要(Hentrich, 2007 を改変).

2.2 研究手法

2.2.1 ChIP-on-chip 実験データの成形

2.1.2 の ChIP-on-chip 実験データにおいて、シグナル値の単純移動平均(Simple Moving Average)を取った。単純移動平均の式を以下に示す。

$$T_n = \frac{1}{2k} \sum_{j=-k+1}^{k} y_{n+j}$$

DNA の位置y_nを基準とした前 $k-1$ bp、後ろ k bp ずつ、計 $2k$ bp の移動平均値T_n

この式において、本研究では k = 500 bp とした。つまり、y_nを基準とした前 499 bp、後 500 bp ずつ、計 1000 bp の移動平均を求めた。ここから求められた T_nをゲノム上の y_n項の位置にプロットし、n を 500, 600, 700, …, と 100 bp ずつ移動させて移動平均を求めていった。これにより、極端な値を示すシグナルノイズの影響を除去した。

2.2.2 FUR-box 候補領域の決定

2.1.2 の ChIP-on-chip 実験データでは、Fur タンパクが結合する DNA 領域、つまり FUR-box 領域で高いシグナル値を示す。FUR-box 候補領域を決定するにあたって、以下の条件を設定した。

(1) 全シグナル値の上位 5 %に含まれる連続領域(5 個以上)
(2) 各株の ChIP-on-chip 実験データにおいて、2 枚のアレイデータともに(1)を満たす領域

上記の条件に当てはまるものを、本研究での FUR-box 候補領域に決定した。さらに、NCBI でこの候補領域周辺の遺伝子を検索することにより、各 FUR-box 候補領域が転写制御を行っている可能性のある遺伝子を設定した。このとき、各 FUR-box 候補領域中で最も高いシグナル値を示す場所の前後 500 bp 以内に転写開始点が含まれる遺伝子をピックアップした。詳細を補足資料 I に示す。

2.2.3 FUR-box 候補領域周辺遺伝子の BLAST 解析

2.2.2 で決定した各株の FUR-box 候補領域周辺遺伝子を用いて BLAST 検索を行った。このことにより、W3110 株、O157 Sakai 株に共通して存在する遺伝子、もしくは各株に特異的な遺伝子を分け、それぞれの領域で違ったコンセンサス配列を見つけることへの手助けとした。

手順として、まず各株の FUR-box 候補領域周辺遺伝子をクエリ、お互いの全遺伝子配列をデータベースとして BLAST 検索を行った(W3110 株 FUR-box 候補領域周辺遺伝子をクエリ → O157 Sakai 株全遺伝子配列をデータベース、O157 Sakai 株 FUR-box 候補領域周辺遺伝子をクエリ → W3110 株全遺伝子配列をデータベース)。その検索結果の中で、

(1) identity が 60 %以上
(2) 一致した塩基配列の長さが、検索に使った遺伝子(クエリ、データベース)の 60 %以上
(3) (1), (2) の条件で双方向トップヒットの組み合わせ

を共通の遺伝子と推定した。このデータを用いて、W3110 株と O157 Sakai 株の FUR-box 候補領域周辺遺伝子を、共通、または特異的なものにカテゴリ分けした。

2.2.4 FUR-box 候補領域における高頻度配列の検出

2.2.2 で決定した FUR-box 候補領域におけるコンセンサス配列を見つけるために、配列の出現頻度を求めた。なお、W3110 株と O157 Sakai 株に共通、または特異的なコンセンサス配列を検出するために、2.2.3 で分けられたカテゴリごと (カテゴリ 1 については W3110 株と O157 Sakai 株ともに) に出現頻度を求めた。まず、W3110 株と O157 Sakai 株における各 FUR-box 候補領域の中で、最もシグナル値の高い塩基位置をピークとした。次に、このピークを中心に前後 250 bp ずつ、計 501 bp の配列を抽出した。このとき、対象 DNA 配列の相補鎖配列も同時に取得した。この 501 bp の配列を 5 bp、もしくは 8 bp ずつの単位で読み取っていき、$4^5 = 1024$ 種類(5 bp)、または $4^8 = 65536$ 種類(8 bp)の配列パターンの頻度をそれぞれカウントした。これを用いて、各株の全ての FUR-box 候補領域における配列パターンカウントの合計を求めた。詳細を補足資料 I に示す。

2.2.5　不完全ガンマ関数を用いた χ^2 検定

2.2.4 で求めたカウントデータを用いて、χ^2 仮説検定による高頻度配列の選出を行った。2.2.4 のカウントデータとは別に、W3110 株と O157 Sakai 株の全ゲノムにおけるカウントデータを作成し、これらを使って各 FUR-box 候補領域における配列出現確率の p 値を求めた。p 値を求めるにあたって、本研究では不完全ガンマ関数を用いた χ^2 検定を用いた。これより求められたデータのうちで、p 値が低いもの、すなわち有意とされるものをコンセンサス配列候補と推定した。不完全ガンマ関数の数式を補足資料 III に示す。

Table 1. χ^2 検定に用いた Contingency Table.
配列パターンは各々の塩基配列、ピーク領域はピークの前後 250 bp ずつの領域を指す。この 2×2 分割表を元に、不完全ガンマ関数を用いた χ^2 検定を行った。

	ピーク領域	非ピーク領域	
配列パターン	a	b	a+b
非配列パターン	c	d	c+d
	a+c	b+d	N

a: ピーク領域中に存在する配列パターン A の数
b: ピーク領域でない領域に存在する配列パターン A の数
c: ピーク領域中に存在する A 以外の配列パターンの数
d: ピーク領域でない領域に存在する A 以外の配列パターンの数
N: 配列パターンの全数

2.2.6　コンセンサス配列を探索する新規アルゴリズム

コンセンサス配列を探索する新規アルゴリズムとして、ChIP-on-chip 実験データのシグナル値、配列の出現頻度による p 値、そして重み行列を考慮したものを作成した。その詳細を以下に示す。また、大まかな概要を Figure 2 に示す。

Step 1.　2.2.5 で求めた各カテゴリの FUR-box 候補領域内で、最も p 値の低い配列(8 配列)をモチーフ探索用配列として入力する。

Step 2.　モチーフ探索用配列を 1〜8 bp 目に含む合計 19 bp の配列を各 FUR-box 候補領域から一つずつ抽出する。このとき、モチーフ探索用配列からの塩基置換を 1 塩基だけ認める(そのため、いくつかの配列の組み合わせ候補が選出される。このときは、FUR-box 候補領域の中心位置に近いものを採用する)。また、相補鎖配列も考慮する。

Step 3.　抽出した配列において、各位置での塩基の出現頻度を数える。

Step 4.　配列の出現頻度を列ごとに合計した重み行列を作成する。

Step 5.　Step 2 において、次にモチーフ探索用配列を 2〜9 bp 目に含む候補を抽出し、重み行列を計算する。これを順々に 12〜19 bp 目のものまで繰り返し行う。

Step 6.　Step 5 で選出した候補ごとに以下の式を計算し、スコアの最も高い組み合わせ候補を採用する。これは、塩基の出現頻度がどれくらい偏っているかを求めるものである。

$$Score = \sum_{j=1}^{S} \mathrm{argmax}\left\{w(i,j) - \frac{N}{4}\right\}^2 \ , \ i = \mathrm{A, C, G, T}$$

i は塩基、*j* は重み行列の位置、*N* は用いる FUR-box 候補領域の数を表す。

Step 7.　重み行列の各位置で最も出現頻度の高い塩基を繋げていき、これをコンセンサス配列とする。

さらにこのアルゴリズムを用いて、各カテゴリにおける FUR-box 候補領域におけるコンセンサス配列を推定した。推定されたコンセンサス配列の表示には WebLogo (Crooks *et al.*, 2004) を用いた。

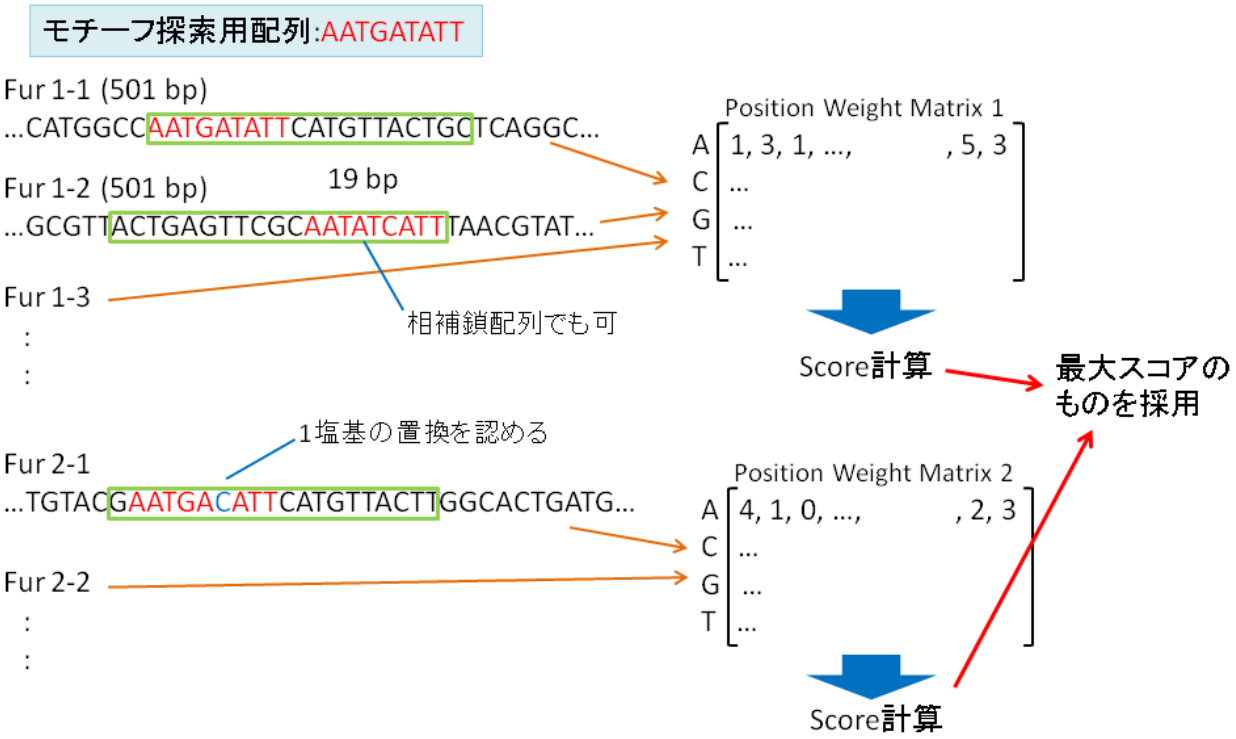


Figure 2. 新規アルゴリズムの手順概要.

2.2.7 既存アルゴリズム MEME によるコンセンサス配列探索

2.2.6 のアルゴリズムとは別に、EM アルゴリズムを用いたコンセンサス配列探索プログラム MEME でコンセンサス配列の探索を行った。

EM アルゴリズムは期待値最大化法とも呼ばれ、最尤法に基づいた期待値計算 (E-step) と期待値最大化 (M-step) を交互に繰り返すことにより、そのモデルパラメータの尤度を最大化させる。EM アルゴリズムの詳細を補足資料 IV に示す。

なお MEME で探索を行うにあたり、2.2.6 のアルゴリズムと容易に比較できるよう、コンセンサス配列の配列長さを 19 bp に、また各 FUR-box 候補領域にコンセンサス配列が一つだけあると仮定して、コンセンサス配列の分布に関するオプションを OOPS (one occurrence per sequence) に設定した。推定されたコンセンサス配列の表示には WebLogo を用いた。

2.2.8　統計的スコアによる新規アルゴリズムと MEME との比較

2.2.6 で作成した新規アルゴリズムの性能を検証するために、新規アルゴリズムと MEME、それぞれのアルゴリズムを用いて推定したコンセンサス配列のデータを比較した。比較の基準として、位置特異的スコア行列に基づいた、以下のスコアを用いた。

$$E = \sum_{j=1}^{S} \sum_{i=\mathrm{A,C,G,T}} \left\{ w(i,j) - \frac{N}{4} \right\}^2$$

j, S は FUR-box 候補領域の数、*i* は A, C, G, T、
N は用いた FUR-box 候補領域の数を表す。

上記の式ではまず、推定されたコンセンサス配列における各列での塩基の出現頻度を数える。そこから統計的な出現頻度を差し引くことにより、塩基の出現頻度の偏りを数値で表すことができる。このスコアをコンセンサス配列全体で求めたものが上記の式となる。ここから求めたスコアを各アルゴリズム、各カテゴリからのコンセンサス配列で比較し、新規アルゴリズムから推定したコンセンサス配列を評価した。

第3章　　結果と考察

3.1　ChIP-on-chip 実験データの成形

2.1.2 の ChIP-on-chip 実験データにおいて、2.2.1 で述べた移動平均を取ることにより、シグナルノイズを除去することができた。その図を Figure 3-A, B, C, D に示す。

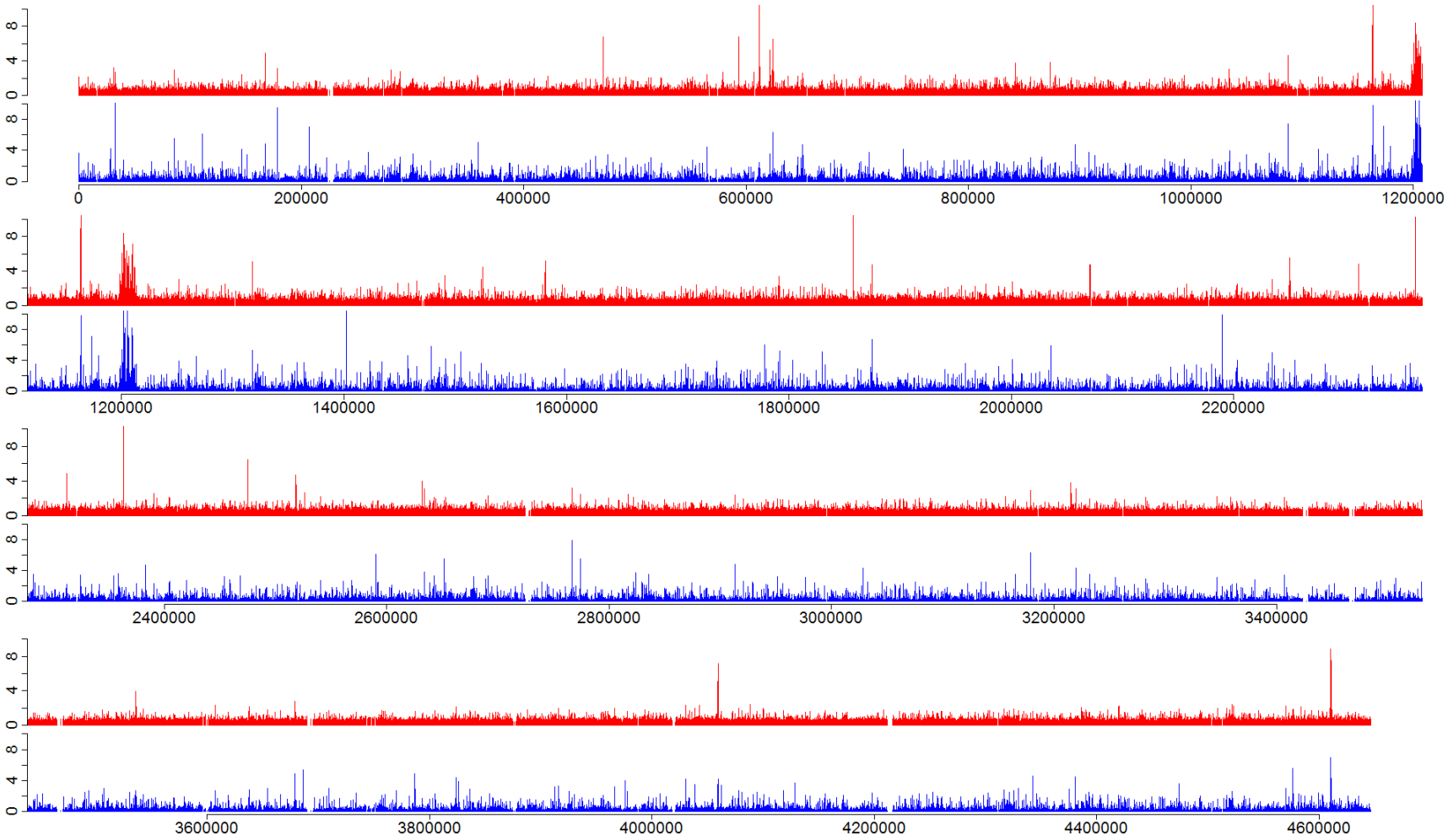


Figure 3-A. *Escherichia coli* W3110 株における補正前のシグナル値.
横軸はゲノムのポジション、縦軸はシグナル値の大きさを表す。
この図には 2 実験のデータを載せている(赤色と青色のプロット)。

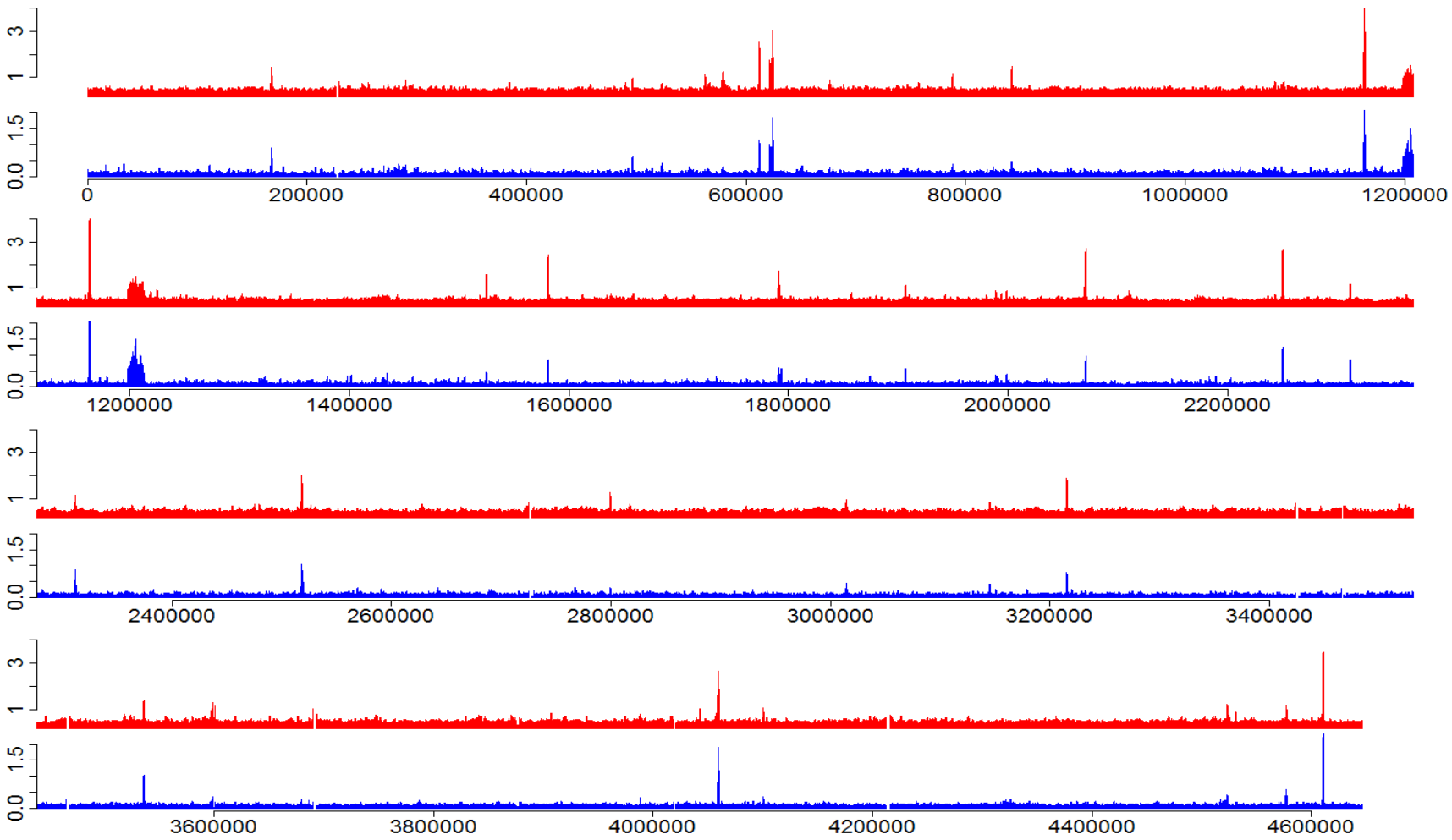


Figure 3-B. *Escherichia coli* W3110 株における補正後のシグナル値.
横軸はゲノムのポジション、縦軸はシグナル値の大きさを表す。
この図には 2 実験のデータを載せている(赤色と青色のプロット)。

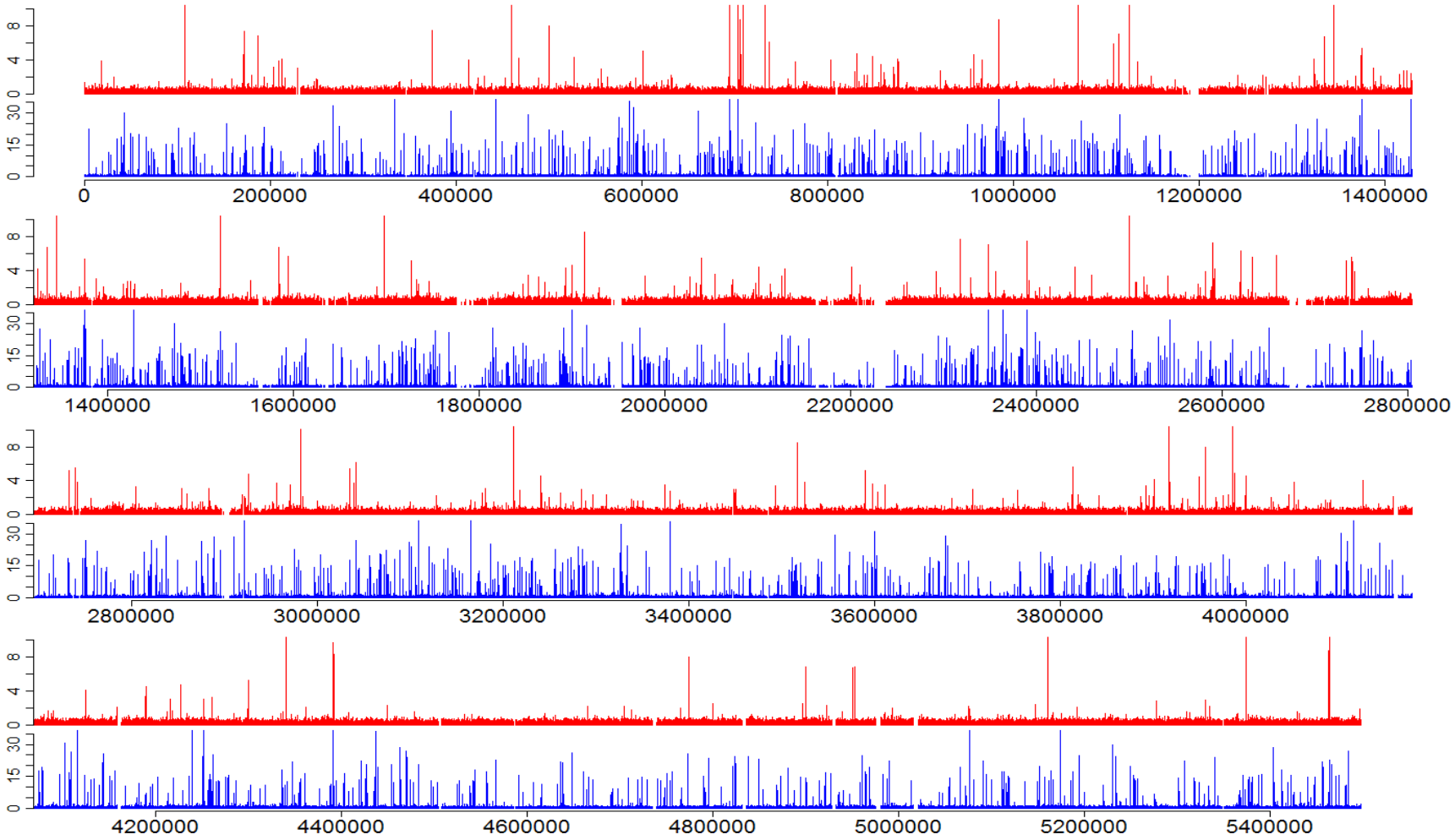


Figure 3-C. *Escherichia coli* O157 Sakai 株における補正前のシグナル値.
横軸はゲノムのポジション、縦軸はシグナル値の大きさを表す。
この図には 2 実験のデータを載せている(赤色と青色のプロット)。

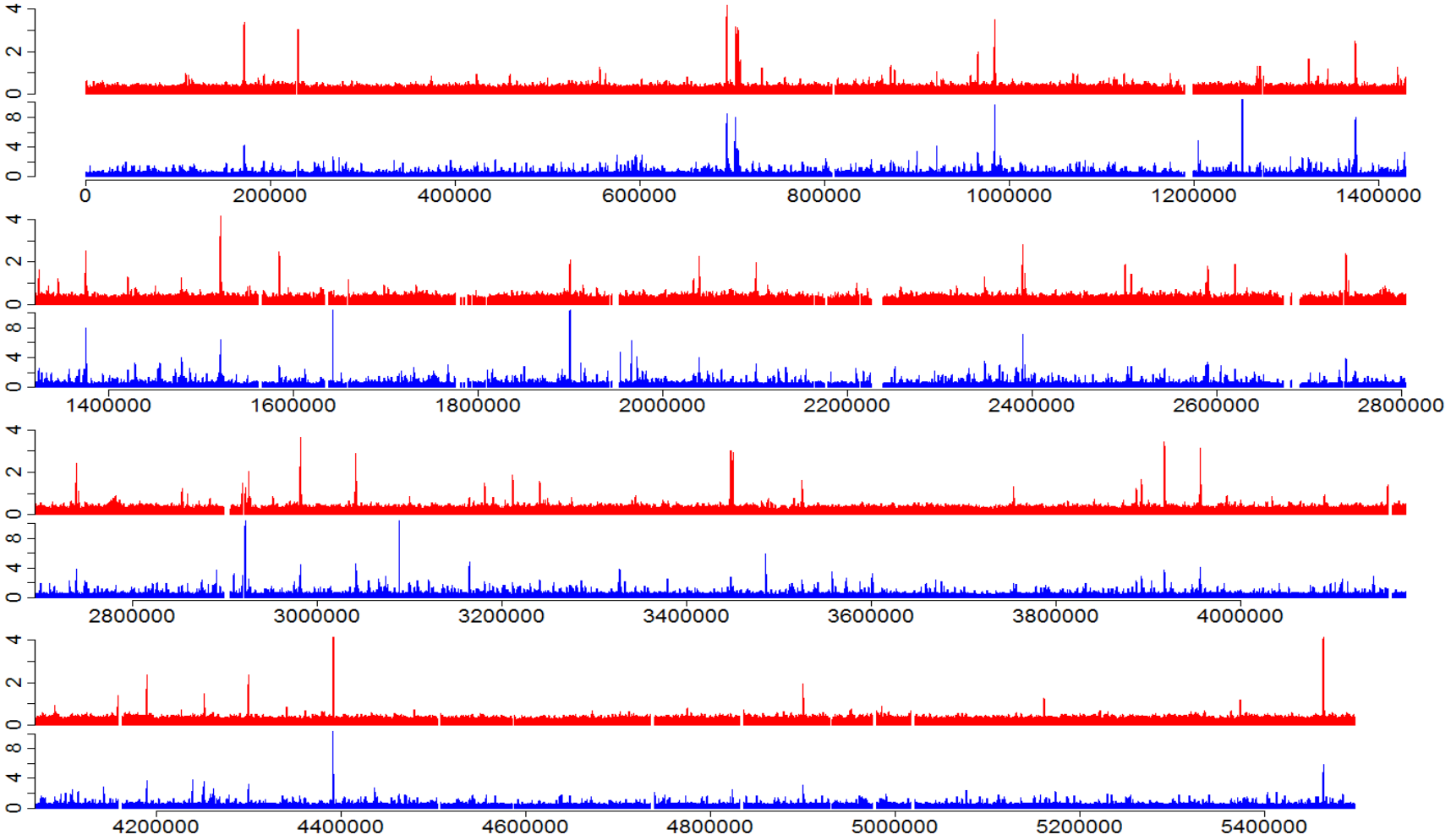


Figure 3-D. *Escherichia coli* O157 Sakai 株における補正後のシグナル値.
横軸はゲノムのポジション、縦軸はシグナル値の大きさを表す。
この図には 2 実験のデータを載せている(赤色と青色のプロット)。

3.2　FUR-box 候補領域の決定

3.1 で作成したデータを用いて、2.2.2 の条件を満たす FUR-box 候補領域を決定した。上位 5 %を満たすシグナル値の分布のヒストグラムを Figure 4 に示す。W3110 株においては 27 箇所、O157 Sakai 株においては 53 箇所決定できた。その図を Table 2, Figure 5-A, B に示す。また、各々の FUR-box 候補領域が転写制御を行っている可能性のある遺伝子を補足資料 II に示す。

この結果において、W3110 株は 27 箇所、O157 Sakai 株は 53 箇所と、各株によって FUR-box 候補領域の数に差が見られた。この理由として、W3110 株と O157 Sakai 株の種としての違いが挙げられる。W3110 株は野生株であり、O157 Sakai 株は野生株に様々なファージが取り込まれ、また遺伝子の組み換えが起こることによって生まれた亜種である。全塩基配列数においても、W3110 株は 4,646,332 bp、O157 Sakai 株は 5,498,450 bp と、およそ 100 万塩基もの差がある。この 100 万塩基のなかに、今回数に差が出た FUR-box 候補領域が含まれているものと推定される。

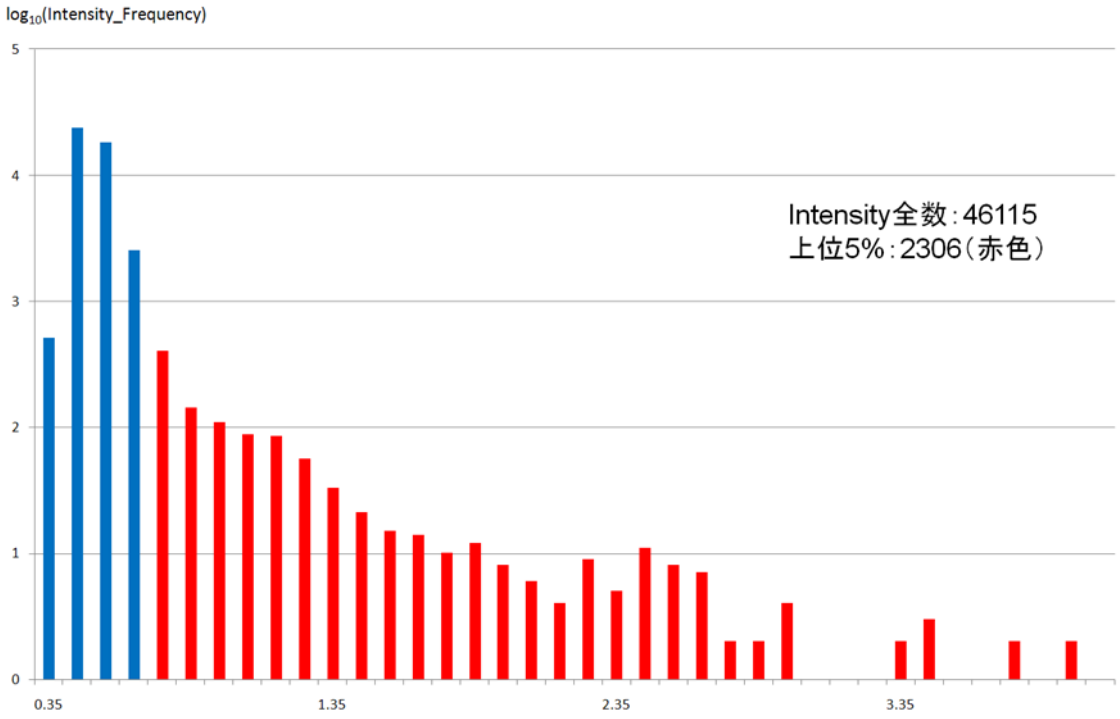


Figure 4. *Escherichia coli* W3110 株の ChIP-on-chip シグナル値分布ヒストグラム.
横軸はシグナルの対数値、縦軸はそのシグナル値を示したものの個数(対数で表示)を示す。赤色のものが上位 5 %に含まれるものである。

Table 2. W3110 株と O157 Sakai 株における FUR-box 候補領域の選定.
各株の ChIP-on-chip 実験データにおいて、2.2.2 の条件を満たした領域の数を示す。実験 1 は赤色のプロットのもの、実験 2 は青色のプロットのものである。実験 1, 2 に共通して確認できた領域を共通のものとし、これを FUR-box 候補領域とした。

	W3110 株	O157 株
実験 1(赤色)	40	62
実験 2(青色)	31	56
共通	27	53

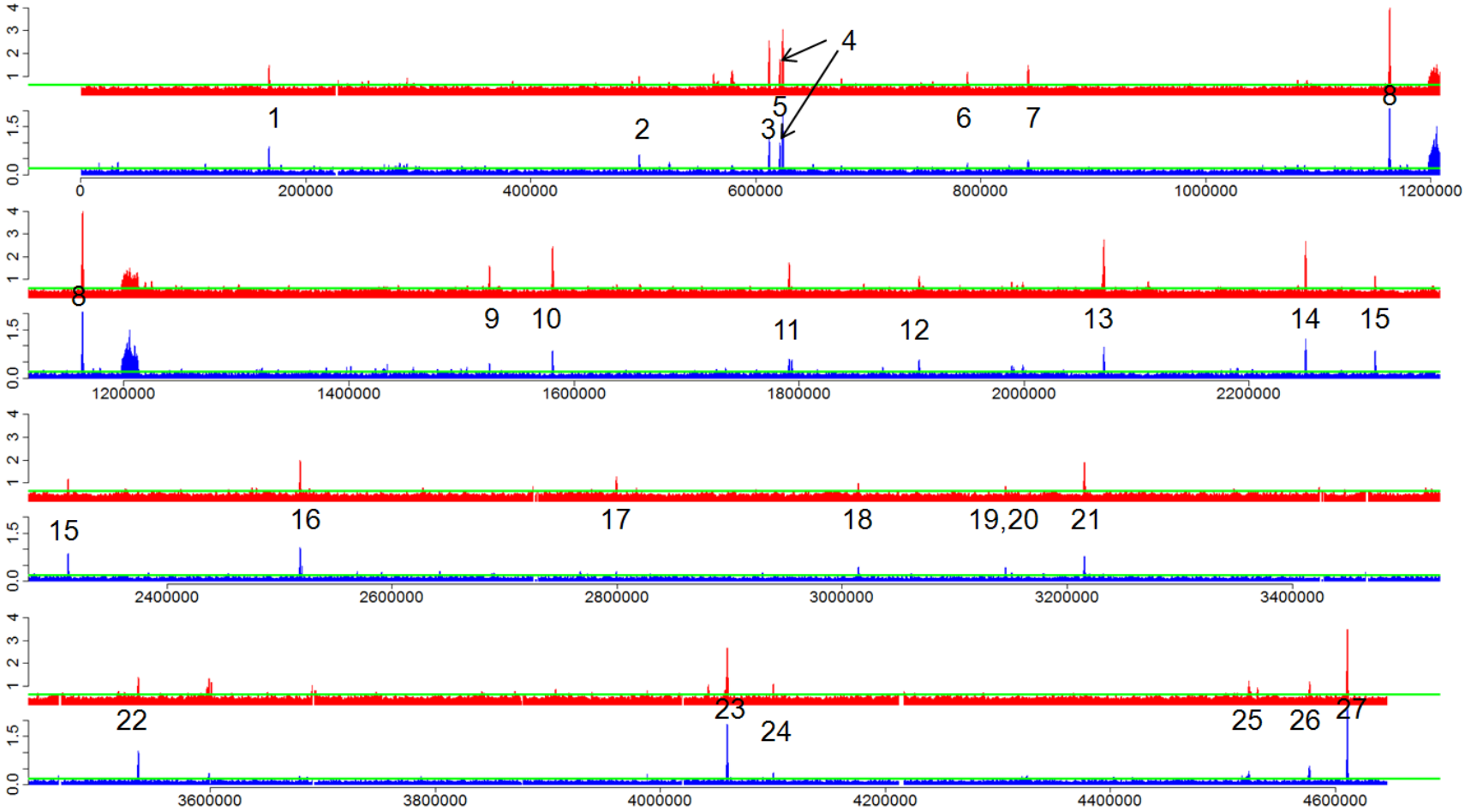


Figure 5-A. *Escherichia coli* W3110 株における FUR-box 候補領域.

横軸、縦軸の値は Figure 2-A, 2-B に同じ。

図中の緑線は上位 5 %のボーダーライン、数字は FUR-box 候補領域の ID を示す。

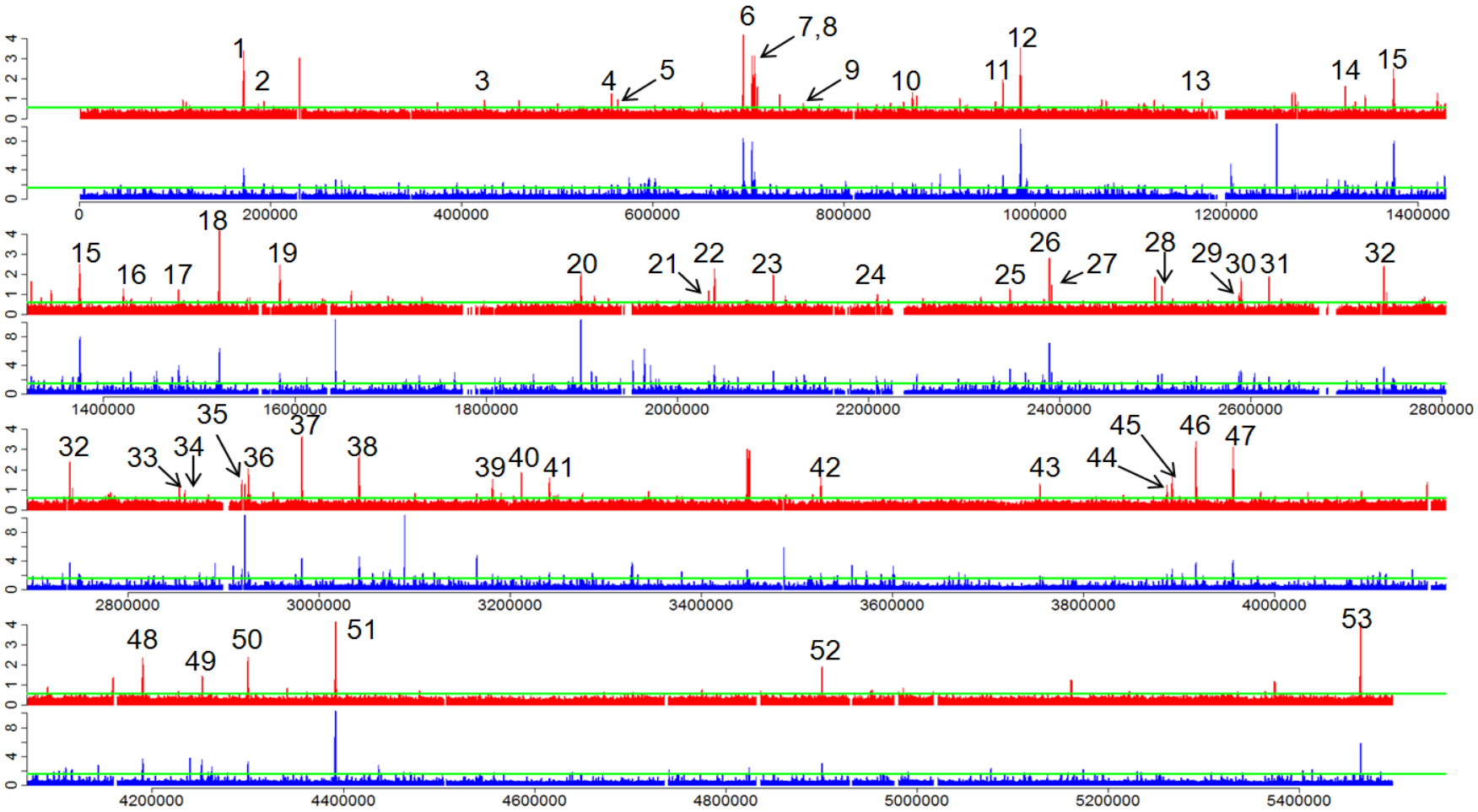


Figure 5-B. *Escherichia coli* O157 Sakai 株における FUR-box 候補領域.

横軸、縦軸の値は Figure 2-A, 2-B に同じ。

図中の緑線は上位 5 %のボーダーライン、数字は FUR-box 候補領域の ID を示す。

3.3 FUR-box 候補領域周辺遺伝子の BLAST 解析

3.2 において選抜した FUR-box 候補領域周辺の遺伝子を、お互いの全遺伝子配列をデータベースとして BLAST 検索にかけた。その結果、W3110 株と O157 Sakai 株とで共通の FUR-box 候補領域周辺遺伝子、互いの別の遺伝子と一致したもの、そして各株に特異的な遺伝子にカテゴリ分けすることができた。その図を Figure 6 に示す。この結果に対する考察を行う前提として、O157 Sakai 株は、マクロファージによる水平伝播などによって W3110 株の遺伝子に変異が入った種であることを考慮しなければならない。このことから、W3110 株中の FUR-box や Fur は O157 Sakai 株にも保存されている可能性が高いということ、そして O157 Sakai 株においては特異的な遺伝子がいくつかあるだろうということである。

上記のことを踏まえた考察として、まずカテゴリ 1 について、これらの遺伝子群は各株に共通に保存された遺伝子であり、またこれらは Fur をプローブとした ChIP-on-chip 実験のシグナルピーク値から選抜されたものであるから、FUR-box による転写制御が成されている遺伝子であると推定される。いくつかの論文(Fecker and Braun, 1983; Grunberg-Manago *et al.*, 1985; Earhart, 1987; Sauer *et al.*, 1990; Van Hove *et al.*, 1990; Braun and Hantke, 1991; Klebba *et al.*, 1993; Panina *et al.*, 2001)で Fur によって制御されていると報告されている遺伝子が複数あることからも、この可能性が高い。また、FUR-box 候補領域によっては複数の遺伝子を含むものもあることから、一つの FUR-box が複数の遺伝子の転写制御を行っていることも推定される。次にカテゴリ 2 と 4 について、これらはお互いの FUR-box 候補に選ばれなかった候補領域の周辺遺伝子に一致したものである。このカテゴリにも、Fur によって制御されていると報告されている遺伝子がいくつかあることから、実際には FUR-box に制御されているとしても、本研究で FUR-box 候補領域を選抜した条件にたまたま含まれなかったのかもしれない。もしくはその逆もあり得る。最後にカテゴリ 3 と 5 についてであるが、これらは各株に特異的な遺伝子群である。特に O157 Sakai 株においては数が多く存在する。その中でも、ECs2974 はベロ毒素遺伝子であることが分かった。これは *S.dysenteria* 1 が産生する志賀毒素と同一の毒素であることが報告されており(O'Brien *et al.*, 1982)、これを持つことが、O157 Sakai 株が病原性大腸菌である理由となっている。O157 Sakai 株には多くのマクロファージ由来の水平伝播が起こっていることを考えると、カテゴリ 5 の遺伝子群はそれに由来する変異であることが推察される。

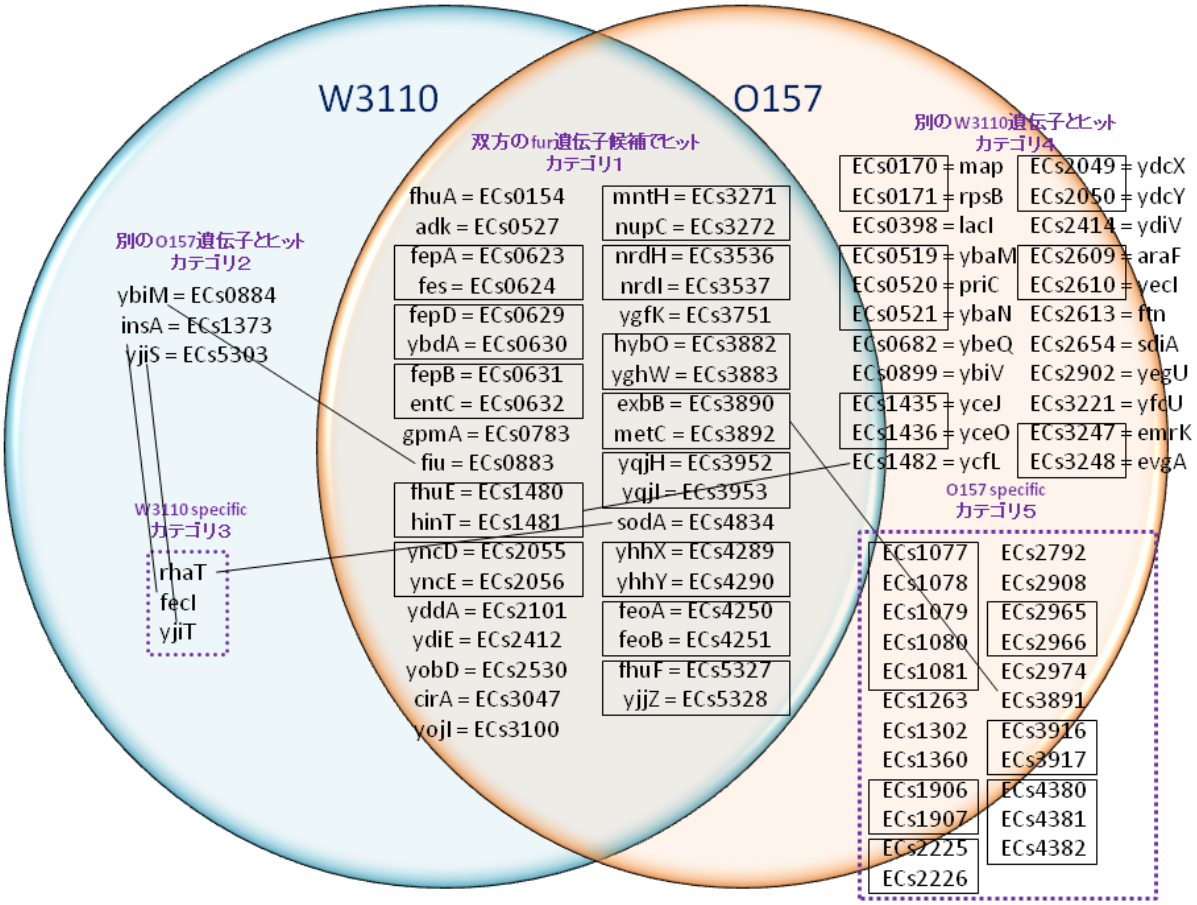


Figure 6. *Escherichia coli* W3110 株と O157 Sakai 株における BLAST 解析結果.

・カテゴリ 1：W3110 株と O157 Sakai 株に共通の FUR-box 候補領域周辺遺伝子(24 組)
・カテゴリ 2：W3110 株において、O157 Sakai 株の FUR-box 候補領域でない遺伝子とヒットした遺伝子(3 組)
・カテゴリ 3：W3110 株に特異的な遺伝子(4 個)
・カテゴリ 4：O157 Sakai 株において、W3110 株の FUR-box 候補領域でない遺伝子とヒットした遺伝子(22 組)
・カテゴリ 5：O157 Sakai 株に特異的な遺伝子(23 個)

を示す。なお、黒枠で囲まれた遺伝子、または直線で結ばれた遺伝子は、同じ FUR-box 候補領域を持つ周辺遺伝子であることを示す。

3.4 FUR-box 候補領域における高頻度配列の検出

3.2, 3.3 において決定した各カテゴリの FUR-box 候補領域について、各領域の塩基出現頻度を調べ、そこから配列出現確率の p 値を求めた。各カテゴリ、各株における p 値の低い上位 10 配列を Table 3-A~F に示す。

まずカテゴリ 1、5 塩基カウントのデータにおいて、各株ともに p 値の低い配列、つまり高頻度に存在する配列にはある程度の共通性が見られる。DNase I footprinting を用いた研究により報告されている FUR-box コンセンサス配列中に含まれる塩基配列が、上位のランクに来ていることが分かる(AATGA, TCATT, ATAAT, ATTAT, ATGAT など)。8 塩基カウントのデータにおいても、各株の配列に共通性が見られ、かつ W3110 株で最も p 値が低く、O157 Sakai 株でも 4 番目の配列である AATGATAA が、報告されている Fur のコンセンサス配列中に含まれている。その他にも、完全一致、もしくは 1 つか 2 つの塩基が違うだけの配列が多く見られる。これらのことから、3.2 において決めた FUR-box 候補領域には、上記の配列を含んだコンセンサス配列が存在する可能性が高いと推定される。カテゴリ 2~5 においては、カテゴリ 1 とは異なる配列が多く見られる。しかしその中でも、例えばカテゴリ 2 とカテゴリ 4 の配列にはある程度の共通性が見られる。さらに、各株に特異的な遺伝子を制御していると推定されるカテゴリ 3 とカテゴリ 5 では、さらに他と異なる。次章 3.5 では、各カテゴリのコンセンサス配列を詳細に求める。

Table 3-A. カテゴリ 1 の W3110 株における FUR-box 候補領域の p 値上位 10 配列.
それぞれ左のセルが塩基配列、右のセルがその配列の p 値を表す。

5 塩基でカウント		8 塩基でカウント	
AATGA	1.26E-23	AATGATAA	1.18E-58
TCATT	1.26E-23	TTATCATT	1.18E-58
ATAAT	1.90E-18	ATCTAGTA	3.40E-44
ATTAT	1.90E-18	TACTAGAT	3.40E-44
ATATA	1.44E-13	AAATGATA	1.33E-37
TATAT	1.44E-13	TATCATTT	1.33E-37
ATGAT	9.24E-13	AATGAGAA	2.97E-37
ATCAT	9.24E-13	TTCTCATT	2.97E-37
TAATG	5.42E-11	ATTATCAT	5.39E-34
CATTA	5.42E-11	ATGATAAT	5.39E-34

Table 3-B. カテゴリ 1 の O157 Sakai 株における FUR-box 候補領域の p 値上位 10 配列.
それぞれ左のセルが塩基配列、右のセルがその配列の p 値を表す。

5 塩基でカウント		8 塩基でカウント	
AATGA	4.02E-20	AATGAGAA	1.66E-32
TCATT	4.02E-20	TTCTCATT	1.66E-32
ATAAT	5.86E-15	CTTGCAAG	1.76E-32
ATTAT	5.86E-15	AATGATAA	1.16E-30
ATGAT	1.45E-11	TTATCATT	1.16E-30
ATCAT	1.45E-11	AATGATTA	2.09E-29
ATATA	4.96E-11	TAATCATT	2.09E-29
TATAT	4.96E-11	TTAATTAA	2.95E-28
TATCA	4.49E-09	ATATATCT	3.74E-27
TGATA	4.49E-09	AGATATAT	3.74E-27

Table 3-C. カテゴリ 2 における FUR-box 候補領域の p 値上位 10 配列.
それぞれ左のセルが塩基配列、右のセルがその配列の p 値を表す。

5 塩基でカウント		8 塩基でカウント	
TTAGA	1.25E-06	GCTCCAAG	2.38E-58
TCTAA	1.25E-06	CTTGGAGC	2.38E-58
TATAA	6.22E-06	ACCATCTA	1.22E-55
TTATA	6.22E-06	TAGATGGT	1.22E-55
AAAAT	1.86E-05	TCCAAGTG	3.28E-50
ATTTT	1.86E-05	CACTTGGA	3.28E-50
TTTGA	3.09E-05	ACTTGGAG	4.22E-44
TCAAA	3.09E-05	TCTAAGTG	4.22E-44
AAATG	3.81E-05	CACTTAGA	4.22E-44
CATTT	3.81E-05	CTCCAAGT	4.22E-44

Table 3-D. カテゴリ 3 における FUR-box 候補領域の p 値上位 10 配列.
それぞれ左のセルが塩基配列、右のセルがその配列の p 値を表す。

5 塩基でカウント		8 塩基でカウント	
TTAGA	1.25E-06	TCTAAGTG	4.22E-44
TCTAA	1.25E-06	CACTTAGA	4.22E-44
AAATG	3.72E-06	ATTGGAAA	1.15E-43
CATTT	3.72E-06	TTTCCAAT	1.15E-43
TATAA	6.22E-06	GAGATGTG	9.43E-42
TTATA	6.22E-06	CACATCTC	9.43E-42
AAAAT	1.86E-05	ACACAAGT	1.39E-41
ATTTT	1.86E-05	ACTTGTGT	1.39E-41
ATGAG	2.24E-05	AGGGTATA	2.17E-37
CTCAT	2.24E-05	TATACCCT	2.17E-37

Table 3-E. カテゴリ 4 における FUR-box 候補領域の p 値上位 10 配列.
それぞれ左のセルが塩基配列、右のセルがその配列の p 値を表す。

5 塩基でカウント		8 塩基でカウント	
AATGA	1.90E-17	GGACCTAG	3.58E-163
TCATT	1.90E-17	CTAGGTCC	3.58E-163
TATAA	6.94E-12	TCGGAGGG	1.99E-41
TTATA	6.94E-12	CCCTCCGA	1.99E-41
ATAAT	6.84E-09	ACTAGGTC	9.55E-34
ATTAT	6.84E-09	ACCTAGTA	9.55E-34
AATAA	8.47E-08	TACTAGGT	9.55E-34
TTATT	8.47E-08	GACCTAGT	9.55E-34
ATATA	1.07E-07	ACACATAG	9.06E-22
TATAT	1.07E-07	CTATGTGT	9.06E-22

Table 3-F. カテゴリ 5 における FUR-box 候補領域の p 値上位 10 配列.
それぞれ左のセルが塩基配列、右のセルがその配列の p 値を表す。

5 塩基でカウント		8 塩基でカウント	
AAAAA	9.30E-17	ATCTAGGT	1.15E-63
TTTTT	9.30E-17	ACCTAGAT	1.15E-63
TGAGA	1.65E-15	ATAATTAT	7.72E-46
TCTCA	1.65E-15	GAGAGCTG	5.45E-41
AAAAT	4.70E-13	CAGCTCTC	5.45E-41
ATTTT	4.70E-13	AAGCTAGA	9.80E-39
AAATA	1.00E-12	TCTAGCTT	9.80E-39
TATTT	1.00E-12	ACTAATAG	1.01E-38
ATAAT	4.13E-11	CTATTAGT	1.01E-38
ATTAT	4.13E-11	AATGAGAA	1.27E-34

3.5 新規アルゴリズムと MEME によるコンセンサス配列の探索

2.2.6 で作成した新規アルゴリズム、また 3.4 で得られたデータを用いてコンセンサス配列の推定を行った。ここから推定されたコンセンサス配列を、WebLogo を利用して視覚的に表したものを Figure 7-A~F の各上段に示す。

まずカテゴリ 1 について、各株におけるコンセンサス配列は共通している部分が多いことが分かる。また、DNase I footprinting を用いた研究により報告されているコンセンサス配列とも類似している。ここから、このカテゴリのものは W3110 株と O157 株における共通の FUR-box である可能性が高く、また Figure 7-A にある配列が高く保存されていることが推察される。他のカテゴリのものについては、どのカテゴリにおいてもカテゴリ 1 のものとは異なる配列となっている。カテゴリ 2 とカテゴリ 3 は、元のデータ配列 3 つのうち 2 つが同じのものであるため、ある程度類似した配列が見られる。次に O157 Sakai 株に特異的なカテゴリであるカテゴリ 5 において、独自の配列が推定された。カテゴリ 1 のものとは異なる配列であるにも関わらずこれらが FUR-box であると推定されていることから、これらが制御している Fur には、カテゴリ 1 の Fur とは別の働きがあるのかもしれない。

推定されたコンセンサス配列全体についてだが、取れてきた配列はモチーフ探索用配列として入力した 8 配列が主であり、その周辺配列はあまり取れていない。詳細は 3.6 で述べる。

次に、3.4 で得られたデータを用いて、MEME によるコンセンサス配列の探索を行ったここから推定されたコンセンサス配列を、WebLogo を利用して視覚的に表したものを Figure 7-A~F の各下段に示す。

カテゴリ 1 について、どちらの株においても、ほぼ同じ配列が推定された。また、3 新規アルゴリズムから推定したコンセンサス配列とも類似している。これらを総合すると、”AATGAGAAT-ATTATCA”の配列(“-“は不明な塩基を表す)がコンセンサス配列である可能性が非常に高いと言える。カテゴリ 2 とカテゴリ 3 について、MEME においても新規アルゴリズム同様に類似した配列が推定された。しかしながら、各アルゴリズムでは推定された配列に相違がある。どちらの方が精度が高いのかは、次章で述べることとする。このことはカテゴリ 4、カテゴリ 5 についても同様である。

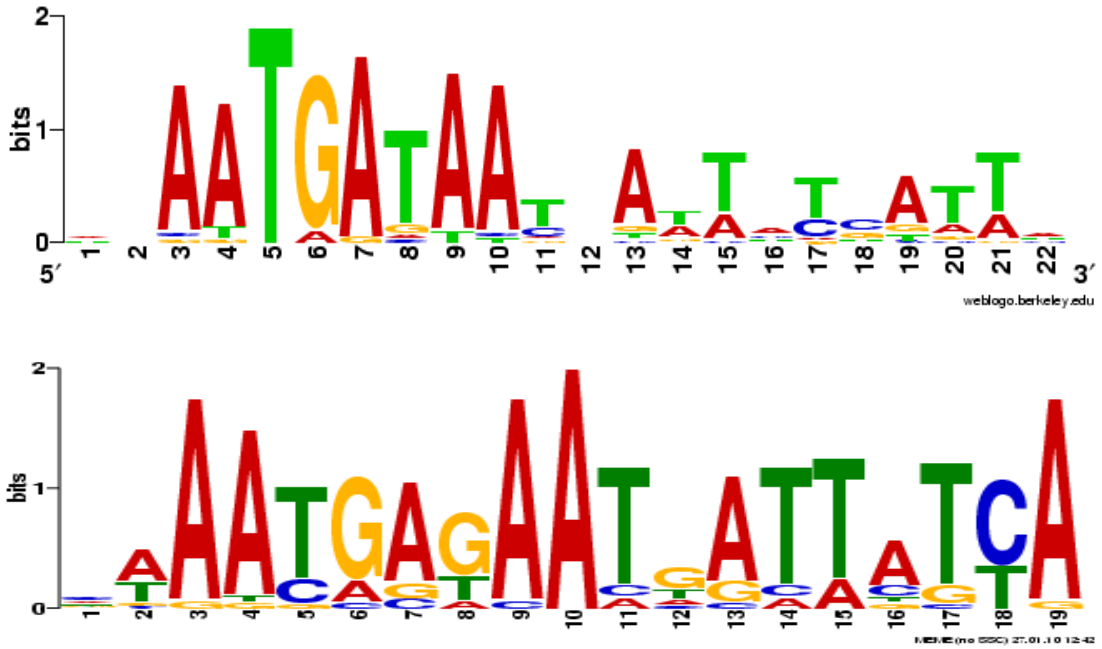


Figure 7-A. カテゴリ 1 の W3110 株で推定されたコンセンサス配列.
上段が新規アルゴリズム、下段が MEME によって推定されたコンセンサス配列である。

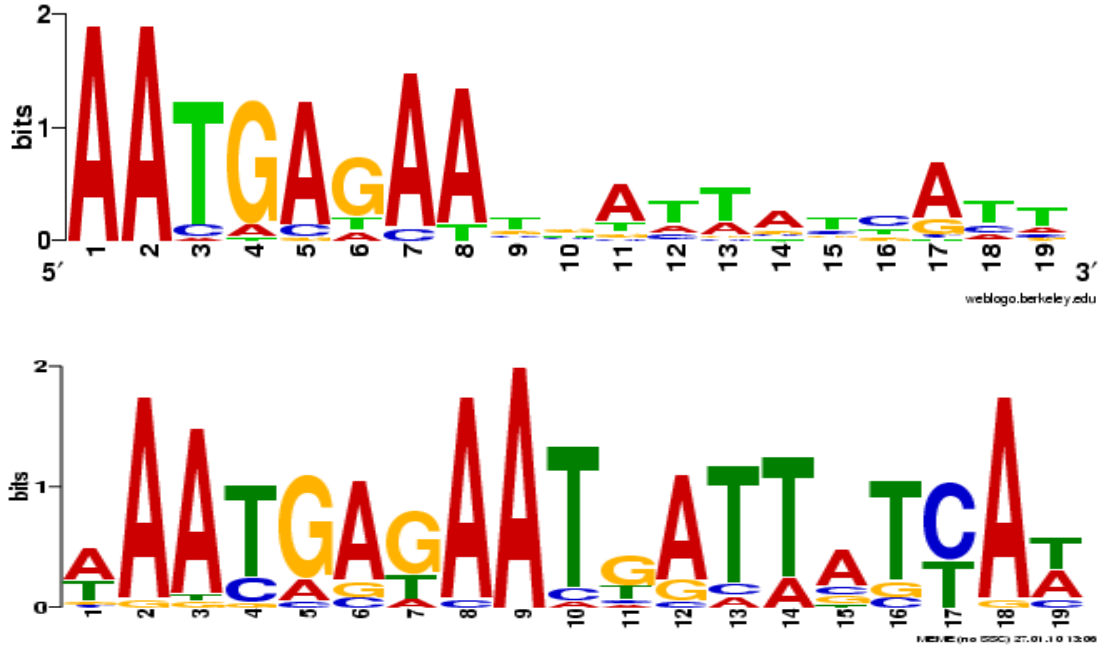


Figure 7-B. カテゴリ 1 の O157 株で推定されたコンセンサス配列.
上段が新規アルゴリズム、下段が MEME によって推定されたコンセンサス配列である。

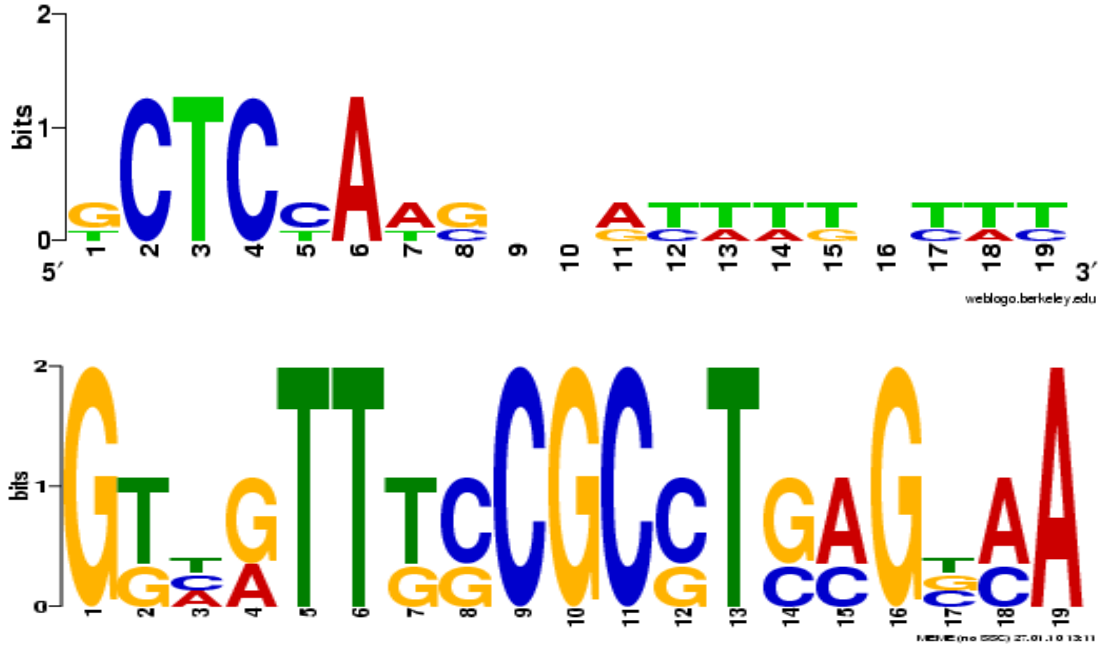


Figure 7-C. カテゴリ 2 で推定されたコンセンサス配列.
上段が新規アルゴリズム、下段が MEME によって推定されたコンセンサス配列である。

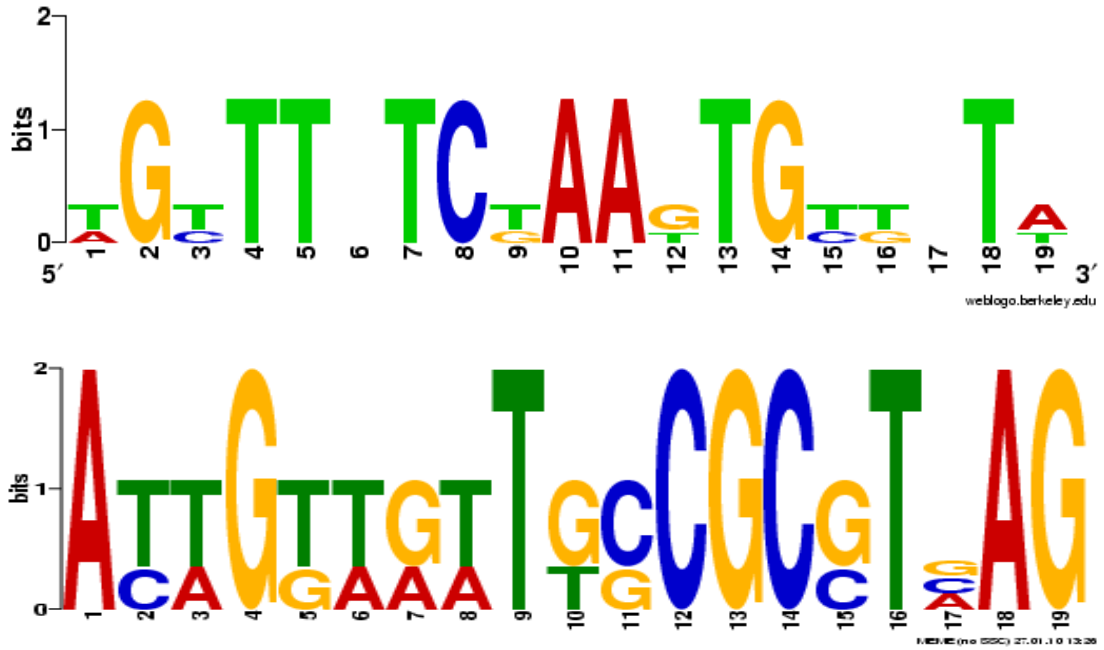


Figure 7-D. カテゴリ 3 で推定されたコンセンサス配列.
上段が新規アルゴリズム、下段が MEME によって推定されたコンセンサス配列である。

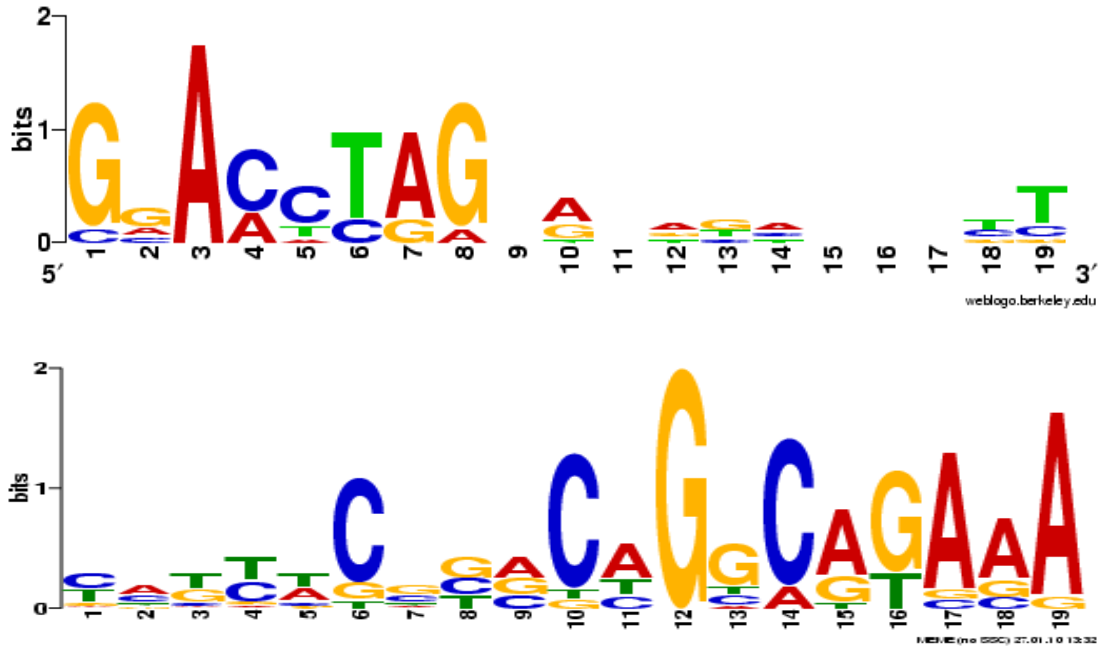


Figure 7-E. カテゴリ 4 で推定されたコンセンサス配列.
上段が新規アルゴリズム、下段が MEME によって推定されたコンセンサス配列である。

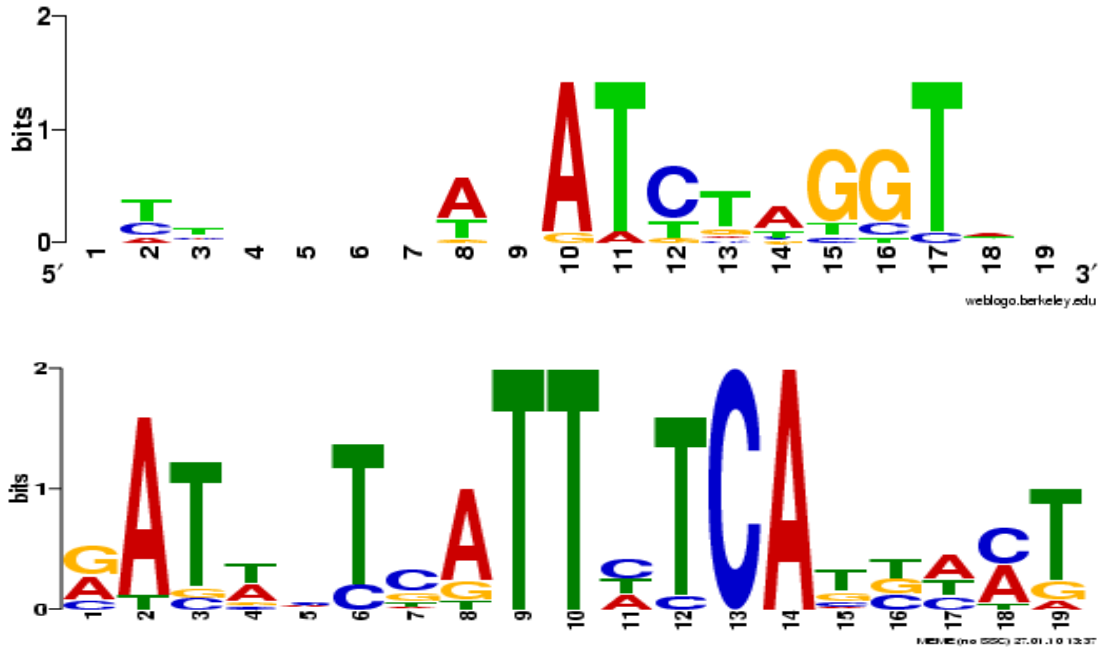


Figure 7-F. カテゴリ 5 で推定されたコンセンサス配列.
上段が新規アルゴリズム、下段が MEME によって推定されたコンセンサス配列である。

3.6　統計的スコアによる新規アルゴリズムと MEME との比較

3.5 で推定された 2 つのアルゴリズムからのコンセンサス配列について、その精度を検証するために、各々の配列のスコアを求めた。各スコアをまとめた表を Table 4 に示す。結果を見てみると、全てのカテゴリにおいて、新規アルゴリズムより MEME の方がスコアが高い。カテゴリ 3 やカテゴリ 4 においては、2 倍ほども水を空けられてしまっている。この理由のひとつとして、新規アルゴリズムでは始めにモチーフ探索用配列を入力し、そしてその配列が結果に大きく反映されることが挙げられる。今回の解析では最も p 値の低い配列をモチーフ探索用配列として入力した。しかし、例えば p 値がほぼ同じの配列が複数あったとしても、結果に反映されるのは 1 つだけである。それに対して、MEME では全ての配列の中から最も出現期待値の高いコンセンサス配列を探索する。こういった配列探索方法の違いがスコアの差に現れたと推察した。

Table 4. 各アルゴリズムによるコンセンサス配列のスコア.
NEW は新規アルゴリズムを表す。

	Figure	PWM	MEME
カテゴリ 1_W3110	6-A	4120	4524.00
カテゴリ 1_O157	6-B	3634	4574.00
カテゴリ 2	6-C	62.25	84.25
カテゴリ 3	6-D	88.25	158.08
カテゴリ 4	6-E	531.25	1204.25
カテゴリ 5	6-F	704.25	1168.25

第4章　　結論

本研究により、*Escherichia coli* の W3110 株と O157 株における FUR-box の推定に成功した。W3110 株では 27 箇所、O157 Sakai 株では 53 箇所の FUR-box を推定し、BLAST の結果を用いて 5 つのカテゴリに分けた。このカテゴリ別に推定されたコンセンサス配列により、各株に共通なもの、あるいは特異的な FUR-box が存在することが示唆された。その中でも、カテゴリ 1 の各株共通 FUR-box、そしてカテゴリ 5 の O157 Sakai 株に特異的な FUR-box が推定されたことは重要である。カテゴリ 1 の FUR-box について、これが転写制御する遺伝子は、生体内での鉄の恒常性を保つという *Escherichia coli* の生存に必要な働きを持つものであるものと推測される。一方、カテゴリ 5 における FUR-box が転写制御する遺伝子の中に、W3110 株にはない性質である毒性物質の産生に関わっているものがあると推定されたことから、その他の FUR-box も、毒性物質の産生、あるいは他の O157 Sakai 株に特有な機能を持つ遺伝子を制御している可能性が高い。

また、本研究で作成したコンセンサス配列探索の新規アルゴリズムについて、Chip-on-chip のシグナル値、p 値、重み行列を利用することで、コンセンサス配列を求めることが可能であると分かった。統計情報のみからコンセンサス配列を推定する一般的な探索アルゴリズムとは違い、実際の実験から得たデータを用いるという点で、その信頼性が高いといえる。その反面、3.6 でも述べたように、このアルゴリズムはモチーフ探索用配列として入力する塩基配列に結果が大きく左右されてしまうことも分かった。そのため今回のアルゴリズムの比較では、全ての配列から期待値を最尤化する MEME に比べてスコアが低い結果となった。p 値が顕著に低い配列が存在すれば良いが、コンセンサス配列探索においては常にそうとは限らない。これらの課題を解決する方法として、モチーフ探索用配列として入力する配列を複数設定できるようにすることが挙げられる。p 値の低い配列を複数考慮したコンセンサス配列を推定できるようにすれば、計算時間は長くなっても、より精度の高いコンセンサス配列が推定できるようになることが期待できる。

謝辞

本研究に取り組むにあたり、主指導教官として適切なご指導と最適な研究の場を与えてくださった情報科学研究科 比較ゲノム学講座 金谷重彦 教授に深く感謝致します。

本研究全般に対し、研究方針や考え方など、適切な御指導と助言を頂きました、高橋弘喜 助教に厚く御礼を申し上げます。

また、本研究を進めるにあたって貴重な実験データを提供していただいた、情報科学研究科システム細胞学講座 大島拓 助教には、心から感謝を申し上げます。

そして、研究生活の様々な面でお世話になりました、Md. Altaf-Ul-Amin 准教授、中村建介 特任准教授をはじめとする比較ゲノム学講座の皆様に深く感謝の意を表します。

最後に、長期間の学生生活を支えてくださった家族に心から御礼を申し上げます。

参考文献

1. Aparicio, O., Geisberg, J. V., and Struhl, K. (2004) Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo, *Curr Protoc Cell Biol Chapter 17*, Unit 17 17.
2. Baichoo, N., and Helmann, J. D. (2002) Recognition of DNA by Fur: a reinterpretation of the Fur box consensus sequence, *J Bacteriol 184*, 5826-5832.
3. Bailey, T. L., and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proc Int Conf Intell Syst Mol Biol 2*, 28-36.
4. Bailey, T. L., and Elkan, C. (1995) Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization, *Mach Learn 21*, 51-80.
5. Bailey, T. L., and Elkan, C. (1995) The value of prior knowledge in discovering motifs with MEME, *Proc Int Conf Intell Syst Mol Biol 3*, 21-29.
6. Braun, V., and Hantke, K. (1991) Genetics of bacterial iron transport, In *CRC Handbook of Microbial Iron Chelate*, pp 107-130, CRC Press Inc.
7. Bsat, N., Herbig, A., Casillas-Martinez, L., Setlow, P., and Helmann, J. D. (1998) Bacillus subtilis contains multiple Fur homologues: identification of the iron uptake (Fur) and peroxide regulon (PerR) repressors, *Mol Microbiol 29*, 189-198.
8. Buck, M. J., and Lieb, J. D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments, *Genomics 83*, 349-360.
9. Calderwood, S. B., and Mekalanos, J. J. (1987) Iron regulation of Shiga-like toxin expression in *Escherichia coli* is mediated by the fur locus, *J Bacteriol 169*, 4759-4764.
10. Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004) WebLogo: a sequence logo generator, *Genome Res 14*, 1188-1190.
11. de Lorenzo, V., Giovannini, F., Herrero, M., and Neilands, J. B. (1988) Metal ion regulation of gene expression. Fur repressor-operator interaction at the promoter region of the aerobactin system of pColV-K30, *J Mol Biol 203*, 875-884.
12. de Lorenzo, V., Wee, S., Herrero, M., and Neilands, J. B. (1987) Operator sequences of the aerobactin operon of plasmid ColV-K30 binding the ferric uptake regulation (fur) repressor, *J Bacteriol 169*, 2624-2630.
13. Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) Maximum likelihood from incomplete data using the EM algorithm., *J. of Royal StatisticalSociety Series B **39***, 1-38.
14. Earhart, C. (1987) Ferrienterobactin transport in *Escherichia coli*, In *Iron Transport in Microbes, Plants and Animals*, pp 67-84, VCH Verlagsgesellschaft.

15. Fecker, L., and Braun, V. (1983) Cloning and expression of the fhu genes involved in iron(III)-hydroxamate uptake by *Escherichia coli*, *J Bacteriol 156*, 1301-1314.
16. Gaballa, A., and Helmann, J. D. (1998) Identification of a zinc-specific metalloregulatory protein, Zur, controlling zinc transport operons in *Bacillus subtilis*, *J Bacteriol 180*, 5815-5821.
17. Griggs, D. W., and Konisky, J. (1989) Mechanism for iron-regulated transcription of the *Escherichia coli* cir gene: metal-dependent binding of fur protein to the promoters, *J Bacteriol 171*, 1048-1054.
18. Grunberg-Manago, M., Hershey, J. B., Plumbridge, J. A., Sacerdot, C., Springer, M., Fayat, G., Lestienne, P., Mayaux, J. F., and Blanquet, S. (1985) Regulation of gene expression of translation components in *Escherichia coli*: initiation factors and aminoacyl tRNA synthetases, *Curr Top Cell Regul 26*, 503-520.
19. Halpern, D., Chiapello, H., Schbath, S., Robin, S., Hennequet-Antier, C., Gruss, A., and El Karoui, M. (2007) Identification of DNA motifs implicated in maintenance of bacterial core genomes by predictive modeling, *PLoS Genet 3*, 1614-1621.
20. Hentrich, T. (2007) Workflow overview of the wet-lab portion of a ChIP-on-chip experiment. http://en.wikipedia.org/wiki/File:ChIP-on-chip_wet-lab.png
21. Klebba, P. E., Rutz, J. M., Liu, J., and Murphy, C. K. (1993) Mechanisms of TonB-catalyzed iron transport through the enteric bacterial cell envelope, *J Bioenerg Biomembr 25*, 603-611.
22. Kukreti, P., Singh, K., and Modak, M. J. (2007) Identification of R/KRRY motif in Klenow Fragment of *E-coli* DNA polymerase I: Its role in coordinating polymerase and exonuclease activity, *Faseb J 21*, A656-A657.
23. Miyazaki, S. (2008) Bio-database literacy and its application with cis-regulatory modules to find novel drug target proteins, *Yakugaku Zasshi **128***, 1525-1535.
24. O'Brien, A. D., and LaVeck, G. D. (1983) Purification and characterization of a Shigella dysenteriae 1-like toxin produced by *Escherichia coli*, *Infect Immun 40*, 675-683.
25. O'Brien, A. D., LaVeck, G. D., Thompson, M. R., and Formal, S. B. (1982) Production of *Shigella dysenteriae* type 1-like cytotoxin by *Escherichia coli*, *J Infect Dis 146*, 763-769.
26. O'Brien, A. O., Lively, T. A., Chen, M. E., Rothman, S. W., and Formal, S. B. (1983) *Escherichia coli* O157:H7 strains associated with haemorrhagic colitis in the United States produce a *Shigella dysenteriae* 1 (SHIGA) like cytotoxin, *Lancet 1*, 702.
27. Panina, E. M., Mironov, A. A., and Gelfand, M. S. (2001) Comparative analysis of FUR regulons in gamma-proteobacteria, *Nucleic Acids Res 29*, 5195-5206.

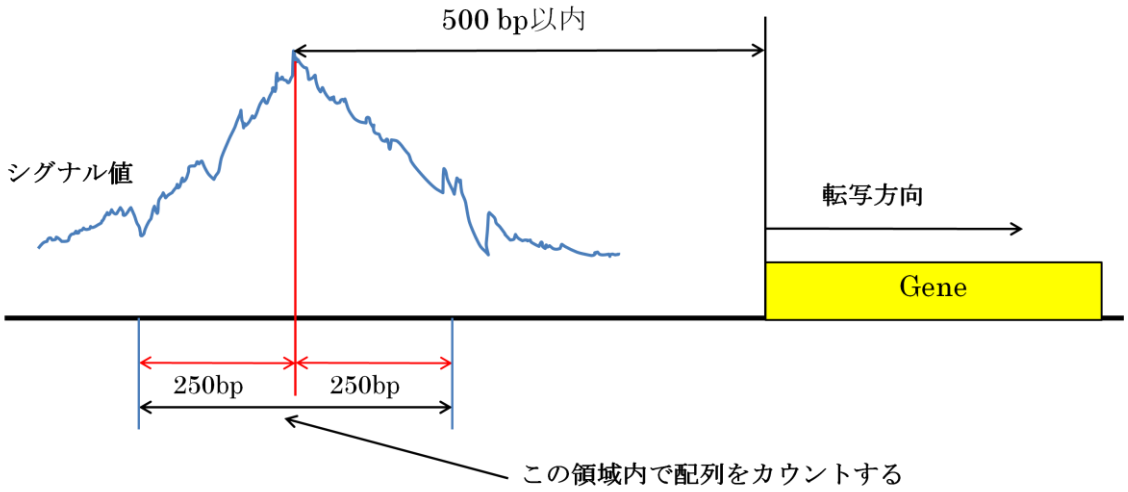
28. Robin, S., Schbath, S., and Vandewalle, V. (2007) Statistical tests to compare motif count exceptionalities, *BMC Bioinformatics 8*, 84.
29. Sauer, M., Hantke, K., and Braun, V. (1990) Sequence of the fhuE outer-membrane receptor gene of *Escherichia coli* K12 and properties of mutants, *Mol Microbiol 4*, 427-437.
30. Schaffer, S., Hantke, K., and Braun, V. (1985) Nucleotide sequence of the iron regulatory gene fur, *Mol Gen Genet 200*, 110-113.
31. Sharov, A. A., and Ko, M. S. (2009) Exhaustive search for over-represented DNA sequence motifs with CisFinder, *DNA Res 16*, 261-273.
32. Stojiljkovic, I., Baumler, A. J., and Hantke, K. (1994) Fur regulon in gram-negative bacteria. Identification and characterization of new iron-regulated *Escherichia coli* genes by a fur titration assay, *J Mol Biol 236*, 531-545.
33. Tsai, H. K., Huang, G. T., Chou, M. Y., Lu, H. H., and Li, W. H. (2006) Method for identifying transcription factor binding sites in yeast, *Bioinformatics 22*, 1675-1681.
34. Van Hove, B., Staudenmaier, H., and Braun, V. (1990) Novel two-component transmembrane transcription control: regulation of iron dicitrate transport in *Escherichia coli* K-12, *J Bacteriol 172*, 6749-6758.
35. 奥村晴彦, 首藤一幸, 杉浦方紀, 土村展之, 津留和生, 細田隆之, 松井吉光, and 光成滋生 (2003) *Javaによるアルゴリズム事典*, 技術評論社.
36. 株式会社バイオマトリックス研究所. Affymetrix®-GeneChip® 3' Expression Array.
37. 丸山修, and 阿久津達也 (2007) *バイオインフォマティクス -配列データ解析と構造予測-*, Vol. 4, 朝倉書店.

補足資料

I. FUR-box 候補領域の決定とその領域における高頻度配列の検出

2.2.2 と 2.2.3 における FUR-box 候補領域の決定、またその領域における高頻度配列の検出の詳細を示す。FUR-box 候補領域の条件として、全シグナル値の上位 5 % に含まれる連続領域を対象とした。各株の ChIP-on-chip 実験データにおいて、2 枚のアレイデータともにこの条件を満たす領域を候補領域とした。次に各 FUR-box 候補領域中で最も高いシグナル値を示す場所の前後 500 bp に含まれる遺伝子をピックアップした。

さらに、各 FUR-box 候補領域の中で、最もシグナル値の高い塩基位置をピークとした。このピークを中心に前後 250 bp ずつ、計 501 bp の配列を取得し、この領域内で塩基配列をカウントした。



Supplementary Figure 1. FUR-box 候補領域における高頻度配列の検出.

II. FUR-box 候補領域周辺の遺伝子

2.2.2 において、FUR-box 候補領域のシグナルピーク位置から前後 500 bp 以内に転写開始点が含まれる遺伝子を列挙した。W3110 株におけるものを Supplementary Table 1-A に、O157 Sakai 株におけるものを Supplementary Table 1-B に示す。コード領域は、その遺伝子の全ゲノム中における読み位置を示す。ピークの前後 500 bp 以内に遺伝子が存在しない場合は N.A.と示している。

Supplementary Table 1-A. W3110 株の FUR-box 候補領域周辺遺伝子.
ID は Figure 4-A と同じもの、ピーク位置は移動平均をプロットしたゲノム上の位置、コード領域は候補遺伝子が存在するゲノム上の読み位置を表す。

ID	ピーク位置	候補遺伝子	コード領域	候補遺伝子	コード領域
1	167400	fhuA	167484..169727		
2	496400	adk	496399..497043		
3	611950	fes	612038..613162	fepA	609477..611717
4	621400	fepD	620408..621412	ybdA	621523..622773
5	624000	entC	624108..625283	fepB	622777..623733
6	788100	gpmA	787265..788017		
7	842050	fiu	839671..841953	ybiM	842218..842478
8	1163250	hinT	1163462..1163821	fhuE	1160939..1163128
9	1524850	yncE	1525021..1526082	yncD	1522677..1524779
10	1581000	yddA	1579371..1581056		
11	1791300	ydiE	1791327..1791518		
12	1906700	yobD	1906515..1906973		
13	2071050	N.A.			
14	2250300	cirA	2248112..2250103		
15	2312000	yojI	2310306..2311949		
16	2518200	mntH	2516914..2518152	nupC	2518488..2519690
17	2799300	nrdI	2799621..2800031	nrdH	2799379..2799624
18	3014450	ygfK	3014716..3017814		
19	3145200	yghW	3145106..3145393	hybO	3143799..3144917
20	3150650	metC	3150892..3152079	exbB	3149906..3150640
21	3215300	yqjI	3215435..3216058	yqjH	3214383..3215147
22	3535900	rhaT	3536156..3537190	sodA	3535251..3535871
23	4059400	yhhY	4058789..4059277	yhhX	4059610..4060647
24	4100350	feoA	4100026..4100253	feoB	4097688..4100009
25	4522950	insA	4523207..4523482	fecI	4522394..4522915
26	4576900	yjiT	4577094..4578611	yjiS	4576431..4576595
27	4610600	fhuF	4609555..4610343	yjjZ	4610484..4610720

Supplementary Table 1-B. O157 Sakai 株の FUR-box 候補領域周辺遺伝子.
ID は Figure 4-B と同じもの、ピーク位置は移動平均をプロットしたゲノム上の位置、コード領域は候補遺伝子が存在するゲノム上の読み位置を表す。

ID	ピーク位置	遺伝子	コード領域	遺伝子	コード領域
1	171700	ECs0154	171806..174049		
2	193000	ECs0170	192198..192992	ECs0171	193360..194085
3	423850	ECs0398	423204..424286		
4	556600	ECs0520	556180..556707	ECs0521	556777..557154
5	563100	ECs0527	563070..563714		
6	694400	ECs0624	694471..695595	ECs0623	691910..694150
7	703900	ECs0629	702841..703845	ECs0630	703956..705206
8	706300	ECs0632	706355..707530	ECs0631	705210..706166
9	757100	ECs0682	756929..757516		
10	871650	ECs0783	870715..871467		
11	965900	ECs0883	963700..965982		
12	984400	ECs0899	983445..984260		
13	1174550	ECs1080	1174631..1174843	ECs1079	1174449..1174586
14	1324150	ECs1263	1324307..1325137		
15	1374900	ECs1302	1374948..1375244		
16	1420350	ECs1360	1418370..1420460		
17	1478350	ECs1435	1477621..1478187	ECs1436	1478448..1478588
18	1520900	ECs1481	1520992..1521351	ECs1482	1521354..1521731
19	1584250	N.A.			
20	1899200	ECs1906	1898703..1899365	ECs1907	1899362..1899637
21	2033000	ECs2050	2033411..2033644	ECs2049	2033104..2033325
22	2038950	ECs2055	2036726..2038828	ECs2056	2039070..2040131
23	2100600	ECs2101	2098901..2100586		
24	2209450	ECs2226	2209505..2209876	ECs2225	2208933..2209355
25	2348000	N.A.			
26	2389300	ECs2412	2389327..2389518		
27	2391700	ECs2414	2391021..2391734		
28	2506900	ECs2530	2506630..2507088		
29	2587750	ECs2610	2588077..2588580	ECs2609	2586291..2587280
30	2589800	ECs2613	2589868..2590365	ECs2612	2589374..2589697

31	2619100	ECs2654	2618381..2619103		
32	2739300	ECs2792	2739356..2741503		
33	2853800	ECs2902	2853628..2854632		
34	2859650	ECs2908	2859168..2859407		
35	2919350	ECs2965	2919200..2919385	ECs2966	2919138..2919605
36	2926000	ECs2974	2924769..2925716		
37	2982000	ECs3047	2979771..2981750		
38	3041700	ECs3100	3039959..3041602		
39	3181300	ECs3221	3178812..3181451		
40	3211500	ECs3247	3210313..3211476	ECs3248	3211892..3212506
41	3241000	ECs3271	3239601..3240839	ECs3272	3241175..3242377
42	3524900	ECs3537	3525115..3525525	ECs3536	3524873..3525118
43	3754050	ECs3751	3754211..3757309		
44	3886900	ECs3882	3885497..3886615	ECs3883	3886804..3887091
45	3892300	ECs3890	3891607..3892341	ECs3892	3892593..3893780
46	3917250	ECs3916	3916210..3917019	ECs3917	3917393..3919534
47	3956200	ECs3952	3955258..3956022	ECs3953	3956310..3956933
48	4190300	N.A.			
49	4252200	ECs4251	4252582..4254903	ECs4250	4252338..4252565
50	4300200	ECs4290	4300393..4300881	ECs4289	4299022..4300059
51	4391900	ECs4382	4392129..4393043	ECs4380	4389464..4391446
52	4900700	ECs4834	4900770..4901390		
53	5464100	ECs5328	5464164..5464400	ECs5327	5463235..5464023

III. 不完全ガンマ関数を用いた χ^2 分布検定

自由度 ν において、ある確率変数 $z_1, z_2, \ldots, z_\nu$ が独立に標準正規分布に従うとすると、ν と $\chi^2 = z_1^2 + z_2^2 + \ldots + z_\nu^2$ を用いて、不完全ガンマ関数による χ^2 分布の累積確率関数は

$$P\left(\frac{\nu}{2}, \frac{\chi^2}{2}\right) = \frac{\gamma\left(\frac{\nu}{2}, \frac{\chi^2}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} = \frac{\int_0^{\frac{\chi^2}{2}} e^{-t} t^{\frac{\nu}{2}-1} dt}{\int_0^{\infty} e^{-t} t^{\frac{\nu}{2}-1} dt}$$

で表される (奥村 *et al.*, 2003)。本研究では、これを用いて仮説検定を行った。

IV. EM アルゴリズム

EM アルゴリズムは観測データが不完全な場合に利用され、最尤法を用いて確率モデルのパラメータを推定する。これは反復法の一種であり、期待値計算ステップ(E-step)と最大化ステップ(M-step)を交互に行うことで、そのパラメータの期待値の最大値を求めるものである。

コンセンサス配列探索における EM アルゴリズムについて、ここでは丸山, 阿久津 (2007)のものを参考にした。推定したいコンセンサス配列が、長さ L の DNA 配列 S_i 上の変数 u_i から始まる、長さ K の非観測データ(未知の配列)であるとする。このとき、EM アルゴリズムでは尤度関数 $L(\theta_0, \Theta \mid S_1, \ldots, S_n)$ を最大化する。E-step では、$u_i = j$ は配列 S_i 上のコンセンサス配列が S_i の第 j 番目から始まる事象を示す。$\theta = (\theta_0, \Theta)$ と $\bar{\theta} = (\overline{\theta_0}, \overline{\Theta})$をそれぞれモデルパラメータとするとき、

$$Q(\theta \mid \bar{\theta}) = \sum_{i=1}^{n} \sum_{j=1}^{L-K+1} P(u_i = j \mid S_i, \bar{\theta}) \log P(S_i, u_i = j \mid \theta)$$

の式を計算する。これは、モデルパラメータを $\bar{\theta}$ と仮定したとき、配列 S_i 上でのコンセンサス配列の始まりの位置 u_i の分布が $P(u_i \mid S_i, \bar{\theta})$ で与えられた条件下における $\log P(S_i, u_i = j \mid \theta)$ の期待値を計算することと同義である。M-step においては、上記の式 $Q(\theta \mid \bar{\theta})$ が θ の関数となっていることから、θ を動かすことにより、$Q(\theta \mid \bar{\theta})$ を最大化する。