東京大学大学院新領域創成科学研究科
情報生命科学専攻

平成 18 年度

修士論文

微生物をターゲットとしたメタゲノム解析からの遺伝子予測と系統分類

2007 年 3 月提出
指導教員　高木　利久　教授

56909　朴　重鎬

要旨

従来の微生物ゲノムに対する研究は個々の生物種を実験室で単離、培養して行われていた。しかし実際の環境で生息する 99%以上の微生物が現在の技術では培養できないとされていて、またそれらは環境中で密接に相互作用しあい生態系の中で重大な役割を担っている。そういった中から培養という過程を経ずに環境中に存在する微生物集団の DNA をまとめてシーケンスして解析を行う環境ゲノム（メタゲノム）解析が近年注目を浴びている。2004 年頃から大規模な環境をターゲットとしたメタゲノム解析の論文が相次いで発表され、現在では様々な環境に対するメタゲノムプロジェクトが完了しまた進行中である。これらの解析から得られるゲノムデータは数千に上る生物種が混在しているので、シーケンスされてくる量からはそのほとんどがアセンブルできない one-pass の配列（一回のシーケンスにより得られる配列、平均長は 700 bp）となっている。このような大規模データからの遺伝子発見は大きなテーマの 1 つであるが、残念な事に既存の手法はメタゲノムデータには適用できない。なぜなら既存の手法は予測を行う上で生物種の情報が既知である必要があるが、メタゲノム配列は様々な生物種が混在したデータなので生物種の情報を得る事ができないためである。またそれ以外にも、新規配列を多く含んだこれらのゲノム断片がそれぞれどの生物種なのかを知る事も同時に重要なテーマとなっている。よって我々はメタゲノム配列からの遺伝子予測と系統分類の 2 つの問題に着目して本研究に取り組んだ。

本研究では与えられたゲノムの GC%からコドン使用頻度を推定するアプローチにより遺伝子予測を実現した。また他にもいくつかの指標を取り入れ、結果として原核生物の断片ゲノムに対して感度 95%、特異性 90%での高い予測精度を達

成した。さらには実際のメタゲノムデータに本手法を適用して、約 40 万個の新規遺伝子を予測することができた。断片ゲノムからの系統分類ではゲノム上にある遺伝子領域のコドン使用頻度に注目して分類を行った。結果として既知の生物種の断片配列には Genus 以上の系統レベルで F 値（感度と特異性の調和平均）90%以上の分類精度を達成でき、また未知の生物種に関しても Genus では F 値 75%、Domain では F 値 87%の分類精度を達成できた。メタゲノム断片を対象として遺伝子予測を行った研究は今までになく、また系統分類に関してもコドン使用頻度に注目してメタゲノム断片の分類を試みた研究は未だない。そういった点で本研究の重要性は高いと考えられる。

目次

1. はじめに

メタゲノム解析とは環境中に生息する微生物集団のゲノムをまとめてシーケンスし、そこに存在する集団や遺伝子をまとめて同定しようとする技術である。この解析により現在生息している微生物全体の99%以上とも推定されている難培養性の微生物にもアクセスが可能となったので、近年行われはじめたメタゲノムプロジェクトからは大量の新規配列が発見された[1-5]。また多くの環境には数千にのぼる微生物種が生息しているので、得られるゲノムデータのほとんどは解析でシーケンスされてくる量からはアセンブルできない[6]。例えば Sargasso Sea の全リード配列の 50%が、また Minnesota soil のほぼ 100%が one-pass（一回のシーケンスにより得られる配列、平均長は 700 bp）の状態である（表 1.1）。現在、このような多くの生物種が混在した断片配列の解釈はバイオインフォマティクスの分野で新たな問題として注目を浴びている。その中でも遺伝子発見の問題は現在最も重要なテーマの１つと考えられているが、残念な事に既存の手法はメタゲノムデータには適用できない。なぜなら既存の手法は予測を行う上で生物種の情報が既知である必要があるが、メタゲノム配列は様々な生物種がミックスしたデータなので生物種の情報を得る事ができないためである。またそれ以外にも、新規配列を多く含んだこれらのゲノム断片がそれぞれどの生物種なのかを知る事も同時に重要なテーマとなっている。よって我々はメタゲノム配列からの遺伝子予測と系統分類の 2 つの問題に着目して本研究に取り組んだ。

表 1.1 現在データが公開されているメタゲノムプロジェクト

環境	種の数（推定値）	コンティグ配列の数	コンティグ配列の平均長(bp)
Acid mine biofilm [1]	5	2,455	4,180
Sargasso sea [2]	1,800	811,372	1,006
Minnesota soil [3]	3,000	149,139	1,093
Whale falls [3]	150	116,464	1,008

種の数の推定値は[6]を参照した

２．メタゲノム配列からの遺伝子予測

２．１ 背景

コンピュータアルゴリズムを駆使したゲノム配列からの遺伝子予測手法の研究は長い歴史を持ち[7,8]、現在では多くの生物種に対する予測方法が提案されている。その中でも微生物に対する遺伝子予測の研究は盛んに行われていて、現在では高い精度を誇る遺伝子予測ツールが数多く存在する[9-14]。しかしこれらのツールで用いている手法は予測に必要な生物種固有のパラメータを既存の遺伝子情報を用いて推定しているので、予測を行うには与えられた配列の生物種をあらかじめ知っている必要がある。しかしながらメタゲノムデータから得られる配列には様々な生物種が混在しているので、それらがどの生物種かを特定することができない。よってこの手法は適用できない事になる。

一方で最近では与えられたゲノムの生物種の情報が未知でも多くの遺伝子を含む程のゲノム長があれば、自分自身のゲノムを学習して遺伝子予測を行う手法が提案されている[15,16]。しかしメタゲノム配列のほとんどはアセンブルできない one-pass の状態で、その中に含まれる遺伝子の数は１つか２つ程度である。よってこの手法もメタゲノム断片には適用ができない事になる。

このような状況で現在遺伝子を予測する唯一の方法として BLAST などによ

る相同性検索があり、この方法によりデータベース中にある既知の遺伝子と相同性を示す遺伝子を予測することができる。しかしメタゲノムデータは難培養性の新規生物種を含んでいるので多くのゲノムデータが相同性を示さない。現在ゲノムが読まれている原核生物の遺伝子密度は 80~90%と非常に高い事を考慮すると、これら相同性を示さないゲノムデータ中には多くの新規遺伝子が含まれていると考えられる。よってこのような状況の中で新たな遺伝予測手法の必要性は高い。

本研究では、メタゲノム配列からの遺伝子予測に取り組むにあたってゲノム GC%とコドン使用頻度の関係に注目した。一般にコドン使用頻度と GC%には強い相関がある事が知られており[17]、またこの相関関係を利用して遺伝子予測の可能性を検証した研究もある[18]。よって本手法では与えられた生物種未知の断片に対して GC%からコドン使用頻度を推定する事を試みた。そしてそのコドン使用頻度をパラメータとして遺伝子予測を行った。ここで予測精度の向上を目的として今回我々はモノコドン使用頻度を遺伝子予測のパラメータとして利用するのではなく、１つ前のコドンとの条件付確率を見るダイコドン使用頻度をパラメータとして利用する事にした。これはダイコドン使用頻度にもまた GC%との強い相関が見られたためである。

また真正細菌（Bacteria）と古細菌（Archaea）の間ではコドン使用頻度に有意な差がある事が知られている[19]。そう考えた場合、同じような GC%でも Bacteria と Archaea では異なるコドン使用頻度の傾向が見られるはずである。よって本研究では Bacteria と Archaea の２種類のモデルを用意して GC%からダイコドン使用頻度の推定を行った。またこの２種類のモデルは Bacteria と Archaea のダイコドン使用頻度の中からより適切なものを選ぶプロセスにより、与えられた断片配列の生物種ドメインを推定する役割も果たした（以降、この手法をドメ

イン分類と呼ぶことにする)。

さらに本研究では遺伝子予測の精度を向上させるためにダイコドン使用頻度以外にも新たな指標の導入を検討した。そして原核生物の遺伝子に見られる規則性を考察した結果、以下の３つの指標を導入することにした。１つ目は ORF 長の分布の利用である。これは現在知られている原核生物の遺伝子長の確率分布を元に予測の際の遺伝子候補を評価するものである。２つ目は最も上流に観測されるスタートコドンから真のスタートコドンまでの距離の分布の利用である。これは遺伝子候補のスタートコドンから最も上流にあるスタートコドンまでの距離を、既知の遺伝子から得られた確率分布を用いて評価するものである。３つ目は隣接する ORF 間の距離の分布である。これは隣接する ORF 間の距離を３パターンの向き(Tandem、Head-to-head、Tail-to-tail)を考慮して評価するものである。

以上、上述した４つの指標をそれぞれモデル化して、それらを統合する形で実際の予測モデルを構築した。そして評価としてはまず、本手法と既存の遺伝子予測手法の性能比較のため完全長ゲノムを用いてテストを行った。次に今回の目的である断片配列からの予測性能をテストした。さらに実際のメタゲノムデータからどの程度新規遺伝子を予測できるのかを検証するために、Sargasso Sea データへ本手法の適用を試みた。

２．２ 利用したデータ

遺伝子予測モデルを構築し、またその予測性能を評価するために NCBIftp サイト [20]にある Bacteria と Archaea の完全長ゲノムデータと遺伝子のアノテーションリストを利用した。そしてデータの偏りによる過学習、過評価を避けるために同じ属の中から１つの種だけを選び、これらの中からトレーニングデータとテストデータをそれぞれ用意した。実際にはトレーニングデータとして Bacteria116 種と Archaea15 種、またテストデータとして Bacteria9 種と Archaea3 種を今回利用した。

本研究は GC%からコドン使用頻度を推定するアプローチにより遺伝子予測を試みた。よってこの手法により予測が十分に行えるかを検証するために、既存の手法との性能比較も交えてまず完全長ゲノムをテストデータとして用い性能評価を行った。

次にメタゲノム配列を想定した断片ゲノムに対する予測性能の評価では、テストデータから得られた 700 bp の断片ゲノムを用いた。断片ゲノムは完全長ゲノムから 700 bp の断片を重複を許す形でランダムに抽出し、その断片を（ゲノム全長／700）個分用意する形で作成した。この断片セットを各生物種で作成し、それらを断片配列のテストデータとした。

さらに実際のメタゲノムデータへの評価として Venter Institute[21]にある Sargasso Sea のコンティグデータと遺伝子のアノテーションリストを利用した。

２．３ 手法

本手法はメタゲノム配列の特徴である生物種未知の短い断片から遺伝子を予測することを目的とした。そしてそのための指標としてダイコドン使用頻度、ORF 長の分布、最も上流にあるスタートコドンからの距離の分布、隣接する ORF 間の距離の分布の４つを今回用いた。それぞれの指標はすべて確率モデルであり実際の評価の際にはそれらのバックグラウンドの確率との対数オッズスコアを用いて、ゲノム中の各 ORF 候補に対して評価を行った。最初に４つの指標に対する確率モデルの構築方法を説明し、次にそれらのモデルを組み合わせて行った実際の予測の手順について説明する。

２．３．１　確率モデルの構築

２．３．１．１　コドンモデル

ゲノム上において、コドンとして使用される読み枠とそれ以外の読み枠の塩基組成には有意な違いが見られる。よって遺伝子の配列情報を学習して得られるパラメータは予測を行う上で最も重要である。しかし今回我々は生物種未知の配列をターゲットとしているので、このような既存の遺伝子を用いたパラメータの推定方法がとれない。そこで本手法は GC%とコドン使用頻度の相関に注目して、与えられた配列の GC%からコドン使用頻度を推定する方法を用いた。そして予

測精度をさらに向上させるために、今回は１つ前のコドンとの条件付確率を見たダイコドン使用頻度を用いた。

まずGC%からダイコドン使用頻度を推定するモデルを作るために、トレーニングデータを用いて GC%とダイコドン使用頻度の関係に対して回帰分析を行った。ここで求めたダイコドン使用頻度は３種類のストップコドンを除いた内部コドンに対する、１つ前のコドンが j のときのコドン i の頻度 $f_{i,j}(1 \le i, j \le 61)$ である。そして回帰を行うにあたって得られた GC%とダイコドン使用頻度をプロットした散布図を観察したところ、その多くがゆるやかな曲線を描いていた。そこですべてのダイコドン使用頻度とGC%の回帰に対する決定係数を求めたところ、ロジスティック回帰の方が直線回帰より当てはまりが良いという結果が得られた（図 2.1)。よって本手法はロジスティック回帰を用いてダイコドン使用頻度の推定を行った。さらに Bacteria と Archaea との間ではコドン使用頻度に有意な差があるので[19]、Bacteria と Archaea のデータを分けてそれぞれ独立に回帰を行ってモデルを構築した（図 2.2)。結果、得られた回帰方程式によって $f_{i,j}$ は GC%のシグモイド関数

$$f_{i,j}(x) = \frac{f_{i,j}^{\max} - f_{i,j}^{\min}}{1 + e^{-ax+b}} + f_{i,j}^{\min}, \qquad (1)$$

として表す事ができる。このとき x は GC%である。そして $f_{i.j}^{\max}$、$f_{i,j}^{\min}$ はデータから観測されるダイコドン使用頻度の最大値、最小値である（これは $f_{i,j}^{\min} \le f_{i,j}(x) \le f_{i,j}^{\max}$ となるような制約を加えた)。このときの回帰変数 a、b は

非線形最小二乗法によって求めた。

これらの推定されたダイコドン使用頻度を用いる事により、与えられた配列内の遺伝子らしい領域を評価することができる。しかしコドンとして使用されない読み枠（バックグラウンド）の頻度も同時に見てやる事で、より厳密な評価が可能となる。よってダイコドンのバックグラウンドの頻度に対しても同様の方法でGC%からの推定を行った。このとき $f_{i,j}(x)$ とバックグラウンドの頻度 $n_{i,j}(x)$ の対数オッズ比からスコアは

$$S_{i,j}(x) = \log \frac{f_{i,j}(x)}{n_{i,j}(x)}, \qquad (2)$$

となる。よってダイコドンに対してはGC%から61×61のスコア行列がBacteriaとArchaeaの２種類について作成される。

次に、ダイコドンのスコア以外に遺伝子領域を評価するためには先頭の内部コドン（モノコドン）とスタートコドン、そしてストップコドンを評価するためのスコアが必要になる。

モノコドン使用頻度はGC%と強い相関が見られ、またBacteriaとArchaeaの間で有意な差が見られた。よってモノコドンに関してもダイコドンのときと同様の方法で、GC%から作成される1×61のスコア行列を２種類用意した。

原核生物ではATG、GTG、TTG、CTGの４種類のスタートコドンが使われている。その中でもATGは生物種共通で最も良く使われているコドンであり、またCTGに関してはほとんど使われていない(1%以下)。よって今回は３つのスタートコドンについてGC%との関連を見たが強い相関は見られなかった。しかし、

バックグラウンドの頻度を見ると GC%との間に強い相関が見られた。よってスタートコドンに関しても同様の方法で、GC%から作成される 1×3 のスコア行列を用意した。

次にストップコドンは TAA、TAG、TGA の 3 パターンが使われている。そしてストップコドンに関しては、ダイコドンのときと同様に 1 つ前のコドンとの条件付確率の使用頻度を用いた。そしてストップコドンの頻度を観察したところ、3 つとも GC%との間に強い相関が見られた。よってストップコドンに関しても同様の方法で、GC%から作成される 61×3 のスコア行列を用意した。

結果として、与えられた配列の GC%から 4 種類のコドン使用頻度を推定してスコア行列を生成するモデルを構築した（図 2.3)。

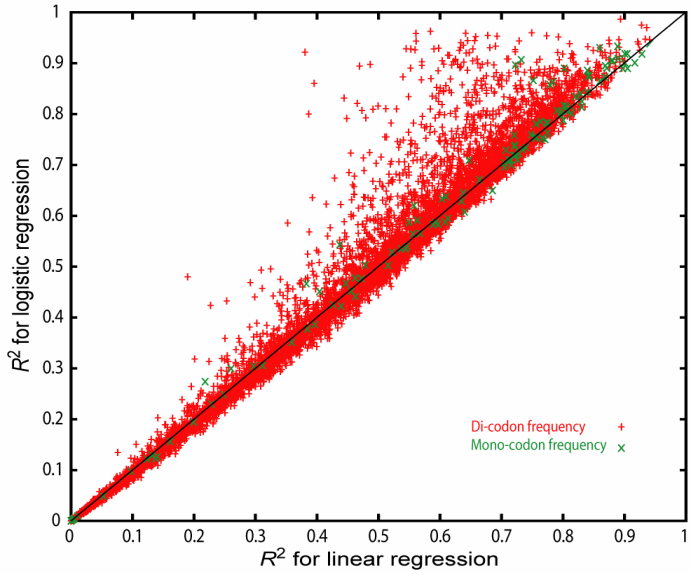


図 2.1 モノコドンとダイコドン使用頻度に対する、線形回帰とロジスティック回帰の決定係数(R^2)の比較。R^2 は 0 から 1 の範囲の値を示し、1 に近いほど回帰の当てはまりが良いことをあらわす。今回のモデルでは、ロジスティック回帰を用いた事が良い選択である事が言える。

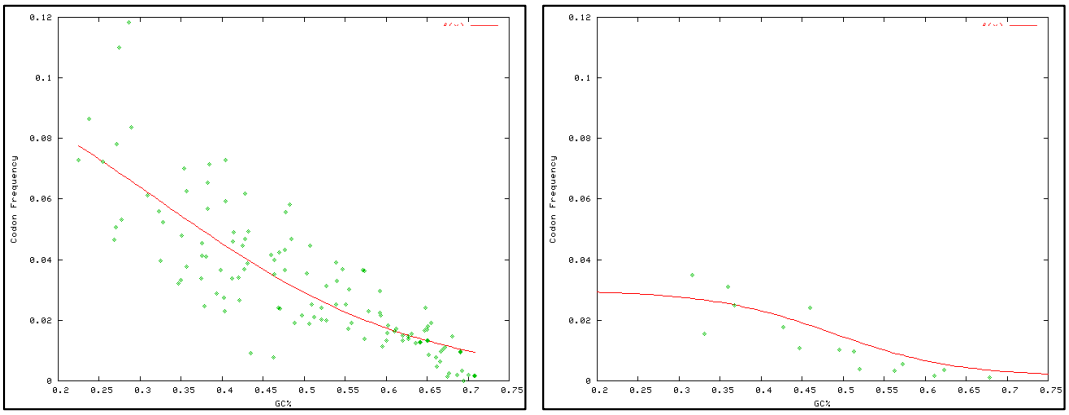


図 2.2　ダイコドン使用頻度（縦軸）と GC%（横軸）の関係に対するロジスティック回帰の例。図はダイコドン GTC-AAT と GC%の関係を Bacteria（左図）と Archaea(右図)データで行ったものである。２つのデータの分布には有意な違いが見られた。

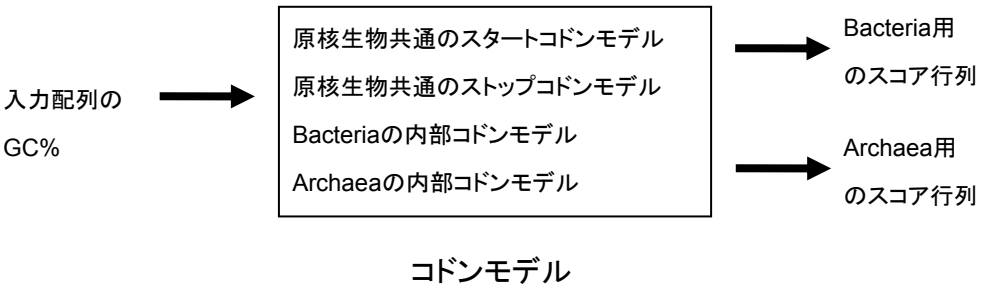


図 2.3 構築したコドンモデル。入力配列の GC%から Bacteria、Archaea２種類のスコア行列を生成する。実際の予測では、入力配列のドメインは未知なので２つのモデルを両方適用することになる。

２．３．１．２　ORF 長の分布に対するモデル

ORF 長は真の ORF と偽 ORF を区別する上で重要な指標の一つである（ここで偽 ORF とはスタートコドンからはじまりストップコドンで終わるが遺伝子ではないものであり、ヒトゲノムなどで使われる偽遺伝子（pseudogene）とは異なることに注意されたい)。もし塩基配列がゲノム上でランダムに並んでいれば、ストップコドン（TAA、TAG、TGA）は 3/64 の確率で出現する事になる。よって長い ORF 候補が見つかれば、それは真の ORF らしいという事になる。実際原核生物の平均 ORF 長は 950bp であり、これは生物種また GC%によらずほぼ一定であった（図 2.4.a)。一方、偽 ORF 長の分布（バックグラウンド）は真の ORF と比べて短く GC%により大きな違いがあった（図 2.4.b)。よって ORF 長に関しては１つの分布を、また偽 ORF に関しては GC%ごとに得られた分布をそれぞれ内挿補間することにより、予測中に出てくる ORF 候補の長さを評価するモデルを構築した。評価方法は、もし ORF 候補の長さが l ならばそのときの対数オッズスコアは

$$Score(l) = \log \frac{P(l)}{B(l)}, \qquad (3)$$

となる。ここで $P(l)$ は真の ORF の分布（図 2.4.a）から得られた確率であり、$B(l)$ は偽 ORF の分布（図 2.4.b）から得られた確率である。B に関してはゲノム GC% によって違う分布を用意した。

ところでメタゲノムデータから出てくる断片配列は、同時に断片 ORF を数多

く含んでいる（実際、今回生成した 700bp の断片配列テストデータの 92%が断片化した ORF であった)。そしてもし予測中に現れた断片化 ORF の長さが l ならば、その ORF の真の長さは l よりも長いはずである。よってこのような ORF を正当に評価するために、分布の上側確率 $P(x \geq l)$ 、$B(x \geq l)$ を用いてスコアを導出した。

$$Score(x \geq l) = \log \frac{P(x \geq l)}{B(x \geq l)}, \qquad (4)$$

２．３．１．３　最も上流に観測されるスタートコドンから真のスタートまでの距離の分布に対するモデル

一般に遺伝子は最も上流にあるスタートコドンを用いているが、いくつかの遺伝子については必ずしもそうではないことが知られている [12,22,26]。そしてスタートコドンのとる位置はゲノムの GC%に相関がある事もまた知られている [23]。よってスタートコドンのとる位置の情報をモデル化することは、遺伝子予測を行う上で重要な指標の１つとなる。そこで本研究では最も上流にあるスタートコドンから真のスタートコドンまでの距離が、実際にどのような分布になっているのかを各 GC%について調べた（図 2.4.c)。図によると実際に多くの ORF が最も上流に真のスタートコドンを使っていて、またその頻度は GC%によって有意な差が観測された。一方偽 ORF に関しても同様に分布を調べると真の ORF とは異なる傾向を示し、これらの分布もまた GC%に関して有意な差が見られた。

よって、真の ORF の分布 P と偽 ORF の分布 B を内挿補間することによりスタートコドンの距離に関するモデルを構築した。スコアの導出方法は式(3)と同様であり、P と B はそれぞれゲノム GC%によって異なる分布を用意した。また最も上流にあるスタートコドンが観測できないもの関しては、式(4)のように上側確率を用いてスコアを導出した。

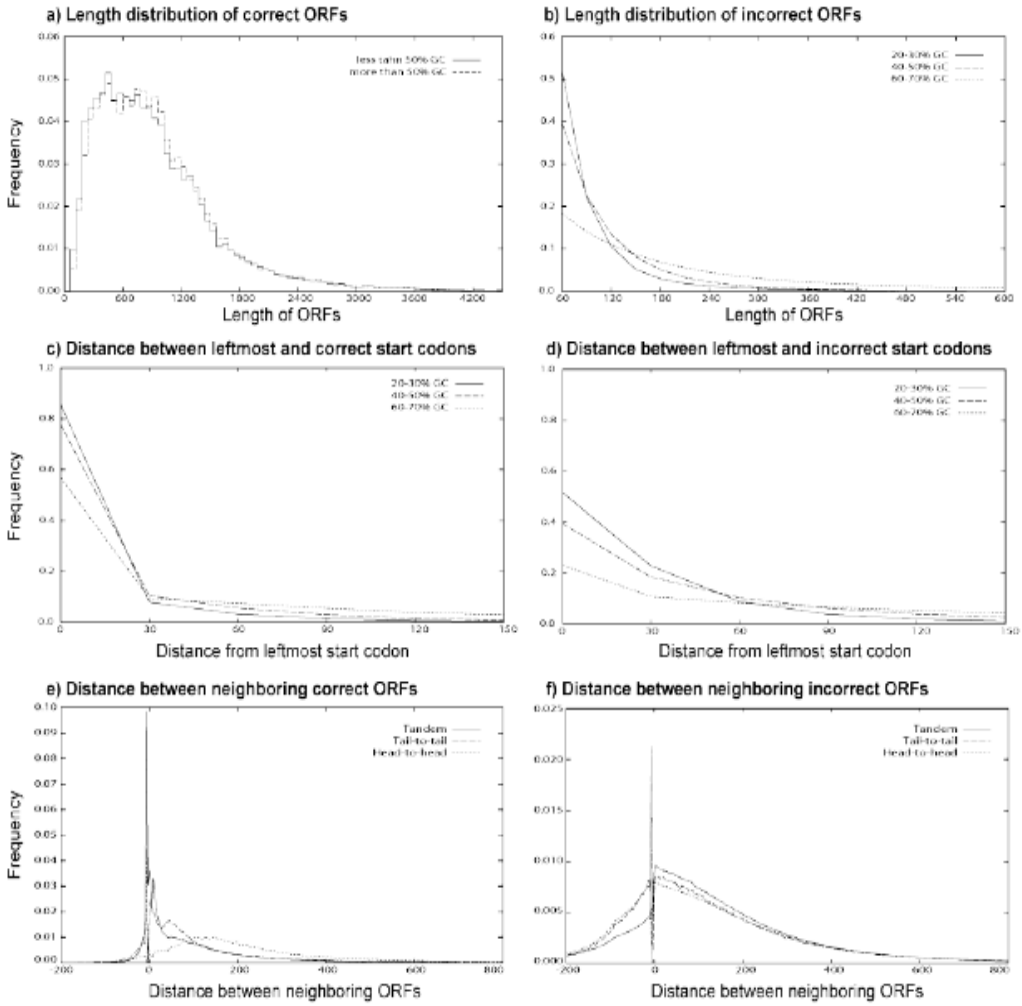
２．３．１．４　隣接する ORF 間の向きを考慮した距離の分布に対するモデル

原核生物は共通してゲノム上に遺伝子が密に並んでいてゲノム全体の 85～90%が遺伝子領域であり、またオーバーラップ遺伝子も多く存在する。よってこれら遺伝子の並び方の規則性をつかみモデル化する事は、遺伝子予測を行う上で重要な指標となりうる。そこで本研究ではゲノム上の隣接する ORF 間の向きと距離という指標に注目してモデルの構築を目指した。

まず遺伝子間の向きは Tandem、Head-to-head、Tail-to-tail の３パターンがあり、それぞれの存在比は 70%、15%、15%であった（一方でバックグラウンドは 50%、25%、25%となる）。バックグラウンドと比較しても Tandem の構造をとる遺伝子が極めて多いが、これは原核生物がオペロン構造を持つためである。次にこれら３パターンについて、それぞれ ORF 間の距離がどのような分布を形成しているのかを調べた（図 2.4.e）。図から読み取れるように、Tandem 構造は他の向きに比べてきわめて ORF 間の距離が短い事がわかる。これはオペロンの転写

のために、ORF 同士がコンパクトにまとめられているためだと考えられる。さらに Tandem と Head-to-head の分布では、プロモータ領域の存在もまた確認された。Head-to-head では理論的には２つのプロモータ領域を含んでいるので、その分布のピークが遠くにあったと考えられる。一方 Tail-to-tail の分布は他とは異なる分布を形成したが、これにはターミネータの存在が考えられる。

偽 ORF に関しても同様の分布を調べたところ、真の ORF に比べて有意な違いが観測された。また向きの違いによって分布に大きな違いは見られなかった。以上よりこれらの情報を予測に活用するために、真の ORF の分布 P と偽 ORF の分布 B を内挿補間することにより ORF 間の距離に関するモデルを構築した。スコアの導出方法は式(3)と同様であり、P と B はそれぞれ ORF 間の向きによって異なる分布を用意した。また ORF 間の距離が完全に観測できないものに関しては式(4)のように上側確率を用いてスコアを導出した。



図が見づらい場合はこちら

図 2.4 今回の予測モデルに取り入れた 3 つの指標の分布。(a)は ORF 長の分布を(b)はそのバックグラウンドの分布を表す。(c)は最も上流にあるスタートコドンから真のスタートコドンまでの距離の分布を、(d)はそのバックグラウンドを表す。(e)は隣接する ORF 間の向き(Tandem、Head-to-head、Tail-to-tail)を考慮した距離の分布を、(f)はそのバックグラウンドの分布を表す。(a)～(d)に関してはゲノム GC%で、(e)、(f)は向きの場合分けでそれぞれ分布を描いた。ORF 間の距離(e)(f)の負の値は隣接する ORF がオーバーラップしていることを表す。

２．３．２　予測の手順

今回我々は予測モデルとして、コドンモデル、ORF 長の分布に対するモデル、最も上流に観測されるスタートコドンから真のスタートコドンまでの距離の分布に対するモデル、ORF 間の向きを考慮した距離の分布に対するモデルの４つを用意した。さらにコドンモデルでは Bacteria と Archaea の２種類を用意したが、実際に予測を行う際には与えられた配列がどちらのドメインに属するかは未知なので、２通りのモデルを適用し最終的なスコアの高い方を正解のモデルとして採択した（これが今回提案したドメイン分類の手法でもある)。そして採択されたモデルの予測遺伝子とドメインを出力結果とした。実際の手順を以下に示す。

１.まず与えられた入力配列の GC%から 3 つのモデルのスコア行列を作成する。このとき、コドンモデルから作られたスコア行列を A、ORF 長のモデルからのものを B、最も上流にあるスタートコドンからの距離のモデルのものを C とする。

２．次に入力配列から、考えられるすべての ORF 構造を抽出する。このとき断片配列を考慮して、スタートコドンやストップコドンの欠如したものでも ORF として可能性のあるすべての構造を抽出してくる。また同じストップコドンに対して複数のスタートコドンを持つ ORF が考えられるが、それらもすべて抽出する。

３．抽出されたすべての ORF 候補を 5'末端から順に $g = \{g_1, g_2, \cdots, g_N\}$ とする。

任意の ORF 候補 g_i 中のコドンを z_j とすると、長さ l のコドン長を持つ ORF 候補は $g_i = (z_1, z_2, \cdots, z_l)$ となる。そして各 g_i に対してスコア行列 A、B、C を用いてスコア付けを行いその合計スコア $s(g_i)$ を求める。このとき、最も上流のスタートコドンから g_i のスタートコドンまでの距離を n とすると、$s(g_i)$ は

$$s(g_i) = A(z_1) + \sum_{k=2}^{l} A(z_{k-1}, z_k) + B(l) + C(n), \qquad (5)$$

と計算される。ここで z_1 がスタートコドンの場合、$A(z_1)$ はスタートコドンのスコア行列から得られたスコアであり、z_1 が内部コドンの場合、$A(z_1)$ はモノコドンのスコア行列から得られたスコアである。また z_l がストップコドンの場合、$A(z_{l-1}z_l)$ はストップコドンのスコア行列から得られたスコアであり、z_1 が内部コドンの場合、$A(z_{l-1}z_l)$ はダイコドンのスコア行列から得られたスコアである。g_i が断片 ORF の場合、$B(l)$ は分布の上側確率から得られたスコアになり、また最も上流にあるスタートコドンが観測されない場合、$C(n)$ は分布の上側確率から得られたスコアとなる。

４． g_i の向きを $r_i \in$ {直鎖,相補鎖}とする。そのとき $g = \{g_1, g_2, \cdots, g_N\}$ 中の同じストップコドンを共有しない２つの ORF 候補 g_i, g_j の向きが $r_i, r_j \in$ {Tandem, Head-to-head, Tail-to-tail}でその間の距離を $d_{i,j}(r_i, r_j)$ とすると、向きと距離に対するスコア $t(r_i, r_j)$ と $u(d_{i,j}(r_i, r_j))$ はそれぞれ

$$t(r_i, r_j) = \log \frac{p(r_i, r_j)}{n(r_i, r_j)}, \qquad (6)$$

$$u(d_{i,j}(r_i, r_j)) = D(d_{i,j}(r_i, r_j)), \qquad (7)$$

となる。ここで $p(r_i, r_j)$ は（Tandem, Head-to-head, Tail-to-tail）=（0.7, 0.15, 0.15）のいずれかから選ばれた確率であり、$n(r_i, r_j)$ はそれらのバックグラウンドの確率（Tandem, Head-to-head, Tail-to-tail）=（0.5, 0.25, 0.25）である。また D は隣接する ORF 間の向きを考慮した距離の分布から得られたスコア行列である。

そして $g = \{g_1, g_2, \cdots, g_N\}$ に対して $t(r_i, r_j)$ と $u(d_{i,j}(r_i, r_j))$、そして手順 3 で求めた $s(g_i)$ の 3 つのスコアの組み合わせから、ゲノム上で最適な ORF の並び $o = \{o_1, o_2, \cdots, o_M\}(M \le N)$ を決定する。このとき o のスコア $v(o)$ は

$$v(o) = s(o_1) + u(d_1) + \sum_{k=2}^{M} \left\{ s(o_k) + t(r_{k-1}, r_k) + u(d_{k-1,k}(r_{k-1}, r_k)) \right\} + u(d_M), \qquad (8)$$

となる。ここで $u(d_1)$ は 5’末端から o_1 までの距離に対する ORF 間の距離の分布の上側確率から導出されたスコアであり、$u(d_M)$ は o_M から 3’末端までの距離に対する ORF 間の距離の分布の上側確率から導出されたスコアである。

５．１～４までの手順を Bacteria モデルと Archaea モデルに対してそれぞれ行い、$v(o)$ の高い方を正解のモデルとする（図 2.5）。

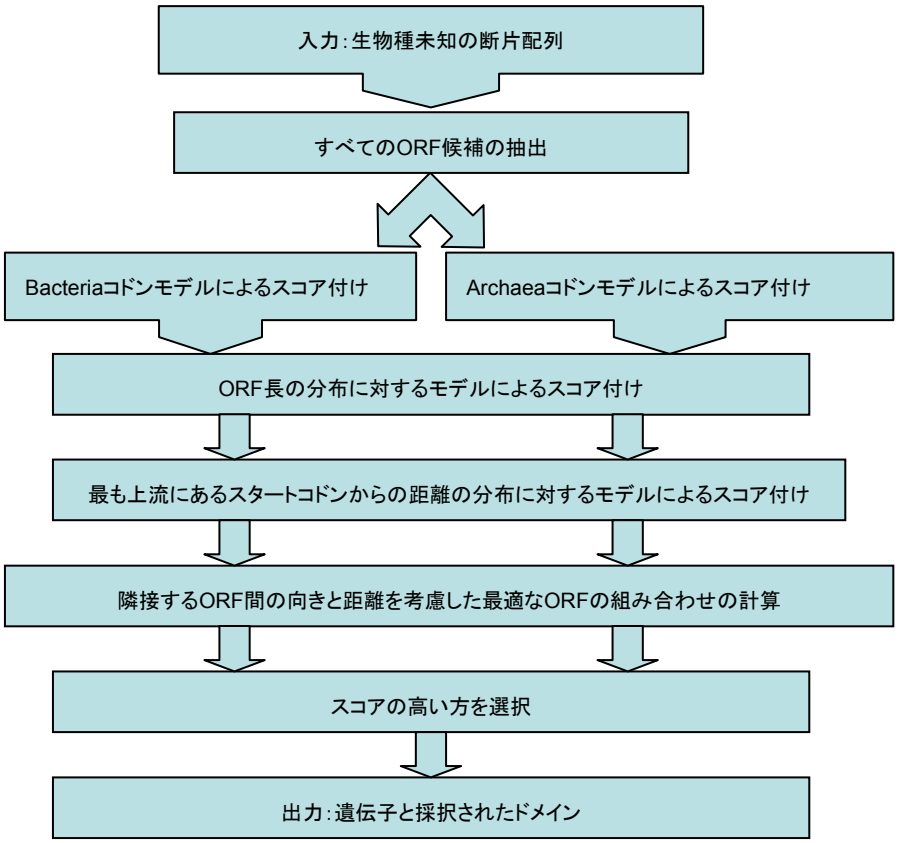


図 2.5 本手法の遺伝子予測の手順

２．４　結果と考察

２．４．１　完全長ゲノム対する予測結果

断片ゲノムをテストする前に今回構築した予測モデルの性能を検証するため、まずは完全長ゲノムに対してテストを行った。今回用いた手法は GC%からの回帰でコドン使用頻度の推定を行っているが、この手法が一般に良く使われている既存の遺伝子情報を学習する手法と比べて、どの程度精度に差があるかを既存のツール（GeneMark.hmm）[10]を用いて比較を行った（表 2.1）。GeneMark.hmm は既存の遺伝子情報の学習によりパラメータを推定し、隠れマルコフモデルを用いて予測を行っているツールである。よって今回のテスタデータに対してそれぞれ別の予測モデルを用意する必要がある。一方本手法は GC%からパラメータの推定を行っているので、１つのモデルですべてのテストデータに対して予測を行える。評価にはテストデータ１３種の完全長ゲノムを用いた。そして今回評価の指標として感度（Sn）と特異性（Sp）を以下のように定義した。

$$\mathrm{Sn} = \frac{TruePositive}{TruePositive + FalseNegative}, \qquad (9)$$

$$\mathrm{Sp} = \frac{TruePositive}{TruePositive + FalsePositive}, \qquad (10)$$

ここで*TruePositive*は遺伝子の読み枠とストップコドン位置が正しく予測できた

数である。また Exact match は*TruePositive*の中でスタートコドンとストップコドン両方を正しく予測できた割合である。

表 2.1 が示すように本手法は様々な GC%を持つテストデータに対して一様に高い精度での予測ができた。そして GeneMark.hmm との比較では Sn と Sp、Exact match 共に同程度の精度を達成した。一般に本手法で用いたゲノム GC%からダイコドン使用頻度を近似する手法は、GeneMark.hmm で用いている遺伝子をピンポイントで学習する手法よりも予測精度は劣ると考えられる。しかしながら今回同程度の予測精度を達成できたのは、微生物ゲノムはおおよそのコドン使用頻度が推定できればゲノム上で遺伝子領域と非遺伝子領域を見分けるには十分であるためだと思われる。よって本手法が用いたコドン使用頻度を回帰する方法は、遺伝子予測に対して十分適用できるという事を示せた。

しかし、各生物種の予測精度を細かく見たときには両者の精度に差が見られた。その中でも特に Exact match の割合において有意な差が見られたが、これは GeneMark.hmm が RBS モデルを取り入れているためである。特に *B.subtillis* ではその影響が顕著に見られた。しかし一方で *B.pseudomallei* などでは本手法の方が Exact match の割合が約 10%近く高く、全体の平均を見ても 2.2%しか差は無い。この理由は今回我々が取り入れた ORF 長のモデル、スタートコドンの距離のモデル、ORF 間の距離のモデルが有効に働いたためだと思われる。実際 2.4.3 の検証ではそのような結果が得られた（表 2.3）。

また両者の手法を比較すると、共通してうまく予測ができない生物種も存在した。例えば *W.endosymbiont* の Sp は極端に低いが、これは多くの遺伝子が偽遺伝子化しているために他の生物種に比べて遺伝子密度が低く、そのために今回構築した予測モデルでは擬陽性を多く発生したからである。また *C.tepidum* では

Sn が低いがこのゲノムは Archaea や真核生物由来の外来遺伝子を多く持っているために[24]、この生物種が使っているコドン使用頻度からではこれらの遺伝子を予測できなったからである。

今回完全長ゲノムのテストデータはすべて正しいドメインに分類されたが、トレーニングデータでは 116 種中 5 種の Bacteria と 15 種中 1 種の Archaea は正しく分類されなかった。そしてそれらの生物種を調べたところ、Bacteria は高熱菌や共生細菌のものから構成されていた。一般にコドン使用頻度は系統関係だけでなく生息環境にも依存するといわれているので、その原因が今回の分類結果として表われたと考えられる。

表 2.1 完全長ゲノムに対する遺伝子予測結果と比較

		本手法			GeneMark.hmm		
テストデータ	GC%	Exact match	Sn	Sp	Exact match	Sn	Sp
Bacteria							
B.aphidicola	26.3	88.5%	99.6%	94.6%	88.6%	99.8%	95.6%
P.marinus	31.2	86.6%	95.9%	94.9%	87.7%	97.1%	95.7%
W.endosymbiont	34.2	76.7%	95.2%	75.9%	85.7%	98.9%	75.0%
H.pylori	39.2	74.2%	96.3%	96.3%	88.2%	98.3%	95.1%
B.subtilis	43.5	61.8%	94.0%	96.9%	86.2%	97.9%	95.3%
E.coli	50.8	72.3%	94.7%	97.3%	74.3%	97.2%	96.8%
C.tepidum	56.5	59.9%	82.4%	95.1%	58.1%	84.1%	93.4%
C.jeikeium	61.4	70.0%	94.6%	97.0%	72.3%	95.5%	97.7%
B.pseudomallei chr.1	67.9	71.9%	97.4%	93.9%	61.0%	96.6%	95.0%
B.pseudomallei chr.2	68.5	70.3%	97.7%	91.5%	62.1%	95.9%	89.4%
Archaea							
M.jannaschii	31.3	69.9%	98.4%	96.1%	63.1%	98.9%	95.4%
A.fulgidus	48.6	73.5%	96.2%	94.8%	72.0%	96.9%	94.0%
N.pharaonis	63.4	80.7%	96.9%	97.7%	84.8%	95.8%	98.8%
平均		73.5%	95.3%	94.0%	75.7%	96.4%	93.6%

各テストデータの生物種名は上から順位に、*Buchnera aphidicola str. APS*、*Prochlorococcus marinus str. MIT 9312*、*Wolbachia endosymbiont strain TRS*、*Helicobacter pylori J99*、*Bacillus subtilis subsp. subtilis str. 168*、*Escherichia coli K12*、*Chlorobium tepidum TLS*、*Corynebacterium jeikeium K411*、*Burkholderia pseudomallei K96243 chr.1*、*Burkholderia pseudomallei K96243 chr.2*、*Methanocaldococcus jannaschii DSM 2661*、*Archaeoglobus fulgidus DSM 4304*、*Natronomonas pharaonis DSM 2160* を表す。

２．４．２　断片ゲノムに対する予測結果

今回の目的であるメタゲノム配列からの遺伝子予測の性能を評価するために、テストデータからランダムに生成した 700 bp の断片ゲノムに対して本手法の適用を行った。まず今回のテストデータの特徴として、１断片あたり遺伝子は平均 1.4 個含まれており、データの全遺伝子中約 92%が断片化した遺伝子であった。またデータ中にはスタートコドンだけのような極端に短い遺伝子断片も含まれていたので、このような無意味な遺伝子を取り除く意味とまた生物学的な価値を考慮して今回 60 bp 未満の断片遺伝子は予測の評価から取り除いた。

表 2.2 に断片ゲノムに対する予測結果を示す。表が示すように、本手法は断片配列に対しても高い予測精度を達成することができた。しかし完全長ゲノムに対する予測結果（表 2.1）と比べて Sn はほぼ同程度の精度を達成できたが、Sp には低下が見られた。この理由は ORF が断片化されたために短い ORF がたくさん存在し、それを偽 ORF とうまく分類できず擬陽性が増えた事が原因に挙げられる。一方 Exact match は完全長ゲノムに比べて向上が見られたが、この理由は ORF が断片化されたためにスタートコドンが欠落した ORF がたくさん発生したためである。

ドメイン分類に関しては平均 88.4%と断片ゲノムに対して高い分類精度を達成できたが、完全長ゲノムのときと比較すると分類精度に低下が見られた。また分類精度は生物種ごとに大きなばらつきがあったので、分類精度が低かった個々の生物種についてそれぞれ考察を行った。まず *W.endosymbiont* についてだが、これは完全長ゲノムのときにも予測がうまくいかなかった生物種であり理由は前のセクションで述べた通りである。次に *C.tepidum* についてだが、これは前のセクションでも述べたようにこのゲノムは Archaea 由来の遺伝子を多く含んでいるためにドメイン分類の精度は低い（[24]によれば約 12%）。しかし遺伝子予測精度を見ると完全長ゲノムのときよりも向上が見られた。これは今回ゲノムが断片化したためにそれらが Archaea モデルに正しく分類され、結果としてドメイン分類の精度は低くなるが、その分 Archaea 由来の遺伝子を正しく予測できたためである。このようにドメイン分類に精度の低下は必ずしも分類に失敗した事を意味するものではない事が言える。また *M.jannaschii*、*N.pharaonis* の 2 つの Archaea は 20～30%の割合でドメイン分類に失敗しているが、これらの遺伝子予測は高い精度を誇っていた。この理由としてこれらのゲノム中に Bacteria に類似した遺伝子が存在していることが考えられる。実際に *M.jannaschii* はエネルギー生産と細胞分裂、そして代謝に関連した遺伝子を Bacteria と共有していることが知られている [25]。

表 2.2　700 bp の断片ゲノムに対する予測結果

テストデータ	GC%	Exact match	Sn	Sp	ドメイン分類の精度
Bacteria					
B. aphidicola	26.3	90.9%	98.2%	92.7%	98.6%
P. marinus	31.2	87.6%	95.5%	92.7%	90.9%
W. endosymbiont	34.2	80.8%	93.1%	76.0%	72.8%
H. pylori	39.2	77.7%	92.6%	92.7%	95.1%
B. subtilis	43.5	73.5%	92.3%	92.5%	92.9%
E. coli	50.8	81.2%	95.3%	93.2%	97.9%
C. tepidum	56.5	73.2%	88.1%	89.6%	78.4%
C. jeikeium	61.4	78.5%	94.0%	91.4%	85.8%
B. pseudomallei chr.1	67.9	81.2%	96.8%	87.9%	93.3%
B. pseudomallei chr.2	68.5	80.5%	96.6%	85.7%	93.0%
Archaea					
M. jannaschii	31.3	82.4%	97.8%	94.1%	70.2%
A. fulgidus	48.6	81.5%	95.8%	93.7%	99.3%
N. pharaonis	63.4	86.2%	97.1%	93.0%	80.8%
平均		81.2%	94.9%	90.4%	88.4%

今まで 700 bp の断片配列に関する予測性能を示してきたが、実際のメタゲノムデータには様々な長さの断片が存在する。よって 100 bp から 1000 bp までの断片を同様にテストして、本手法の予測モデルがどの程度の長さの断片にまで適応できるのかを検証した（図 2.6)。図では断片が短くになるにつれて緩やかに予測精度は落ちていくが、急激に予測精度が落ちてくるのは 100～200 bp のごく短い断片のときだけである。よって本手法は実際のメタゲノムデータに対しても十分に予測を行えることがわかる。

また実際の断片に対する別の問題としてシーケンスエラーがある。シーケンサーから読まれる元々の one-pass 配列の長さは 900 bp 程度だが、配列の末端には

技術的な問題で多くのシーケンスエラーやクローニングした際のベクター配列が含まれる。よって末端の配列はトリミングされて実際の one-pass 配列は 700 bp 程度の長さのものとなり [1,3]、そしてほとんどエラーが含まれない信頼性の高い配列になる。この信頼性は Phred スコアというもので評価されていて、トリミング後の Phred スコアは平均 40 でありこれは 99.99%の精度を意味する。よって今回は実際の Phred スコアから得られたエラー分布を利用して断片ゲノムにランダムに塩基置換を起こし、そのときの予測精度がどう変化していくのかを検証した（図 2.7)。図から読み取れるようにエラー率が高くなるにつれて精度は下がっていくが、その低下は緩やかなものであり実際のエラー率の付近でも信頼性の高い精度が保証される事がわかる。よって実際の断片ゲノムに対しても今回のテストデータと同等の予測精度が保証される事がわかる。

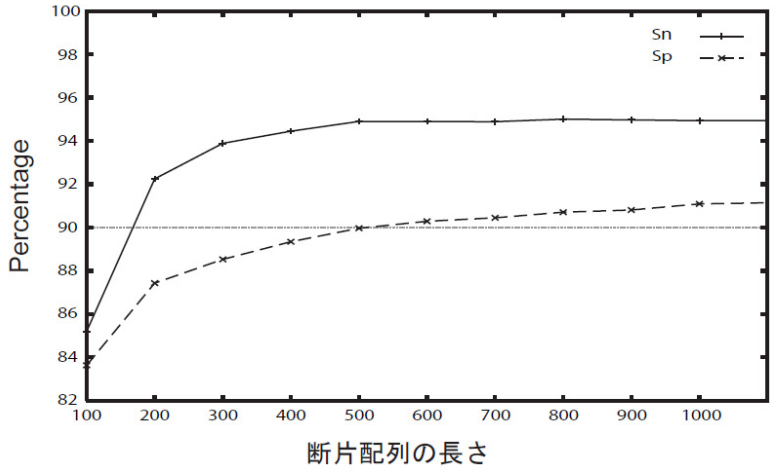


図 2.6 様々な長さの断片配列に対する予測精度。各長さの断片はテストデータの生物種から作られた。感度と特異度はテストデータの平均値を表す。

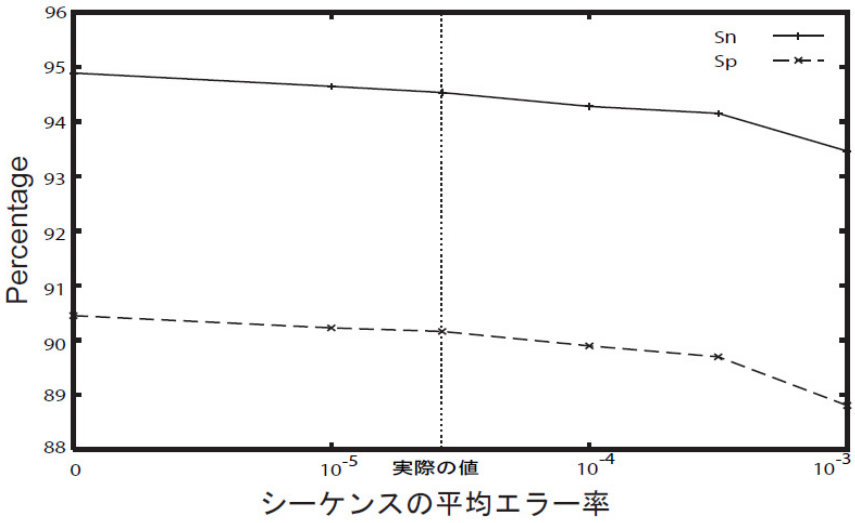


図 2.7 シーケンスエラーに対する予測性能の変化。シーケンスエラーが増えるにつれて予測精度がどのように変化していくのかを調べた。元のテストデータの状態をエラー率 0 として、実際の one-pass 配列のエラー分布をもとにテストデータに塩基置換を挿入していった。横軸はそのときの平均エラー率を表す。感度(Sn)と特異性(Sp)は 700 bp の断片テストデータに対する平均を表す。

２．４．３　本手法で用いた指標の有効性の検証

このセクションでは、本手法で採用した指標が実際の予測にどの程度効果をもたらしたのかを検証した。具体的にはダイコドン使用頻度の利用、Bacteria と Archaea の２種類のコドンモデル、その他の３つのモデル（ORF 長のモデル、最も上流にあるスタートコドンからの距離のモデル、ORF 間の距離のモデル）がどの程度予測に効いたのかを検証した。

一般にダイコドン使用頻度はコドンの出現する条件付確率を見ているので、モ

ノコドン使用頻度を用いるよりも高い予測精度を達成できる。しかし今回はGC%からコドン使用頻度を回帰しているので回帰の当てはまりの良さもまた重要な要因となってくるが、実際に決定係数を求めたところモノコドンのほうが良い結果が得られた。よってモノコドンとダイコドン使用頻度のどちらを使用した方が良いかは一概には言えない事になる。そこで両方の指標でテストを行い予測精度の比較を行ったところ、結果としてダイコドン使用頻度を用いた方が良い結果が得られた（図 2.3)。よって本手法で用いたダイコドン使用頻度はより良い選択であったことが言える。またその他の３つのモデルを追加したときには、する前と比べて Exact match、Sp に大きな向上が見られた。よって今回用いたその他の３つのモデルに関しても有効な指標であった事が言える。

今回我々は Bacteria と Archaea の２種類のコドンモデルを用意したが、これが生物種をまとめて回帰した単独のコドンモデルを用いた予測と比べてどの程度の効果が得られたのかを検証した（図 2.8)。図から読み取れるように２種類のモデルの方がより高い精度を達成できた事がわかり、特に Archaea のテストデータに関して大きな予測精度の向上が見られた。これは Bacteria と Archaea の間に有意なコドン使用頻度の差があるにも関わらず単独のモデルではこれらをまとめて回帰していたため、データ数の多い Bacteria のコドン使用頻度に大きく影響されていたためだと考えられる。また、２種類のモデルを用いたときの Bacteria データの予測精度に関しては Archaea モデルの影響があるにも関わらずほとんど低下は見られなかった。よって本手法で構築した２種類のモデルはより有効な手段であったと言える。

表 2.3 本手法で構築した予測モデルの有効性の検証

予測モデル	Exact match	Sn	Sp
ダイコドンモデルとその他3つのモデル（本手法）	81.2%	94.9%	90.4%
ダイコドンモデル	70.0%	94.0%	86.7%
モノコドンモデル	67.7%	93.8%	84.3%

その他の３つのモデルは ORF 長のモデル、最も上流にあるスタートコドンからの距離のモデル、ORF 間の距離のモデルを指す。

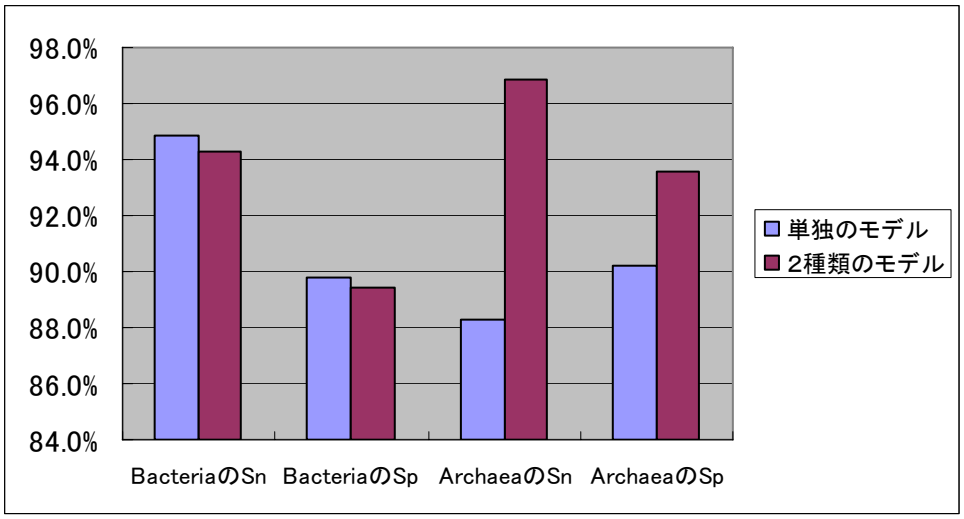


図 2.8 単独のモデルと今回用いた２種類モデルでの予測性能比較。単独のモデルは生物種をまとめて回帰する事によりモデルを構築した。一方２種類のモデルは生物種を Bacteria と Archaea に分けて回帰し、モデルを構築した。テストデータに対して今回提案した２種類のモデルの方がより高い精度を達成できた事がわかる。

２．４．４　Sargasso Sea データセットへの適用

今までの各セクションにおいて本手法の予測性能を様々な角度から検証してきた。そしてこの信頼性を元に実際のメタゲノムデータである Sargasso Sea データ[21]に対して本手法を適用し、論文の著者たちが予測した遺伝子セットとの比較を行った。このデータ中には約 80 万個のコンティグ配列があり、その平均は約 1000bp、そしてトータルの長さは約 820M bp である（図 2.9)。

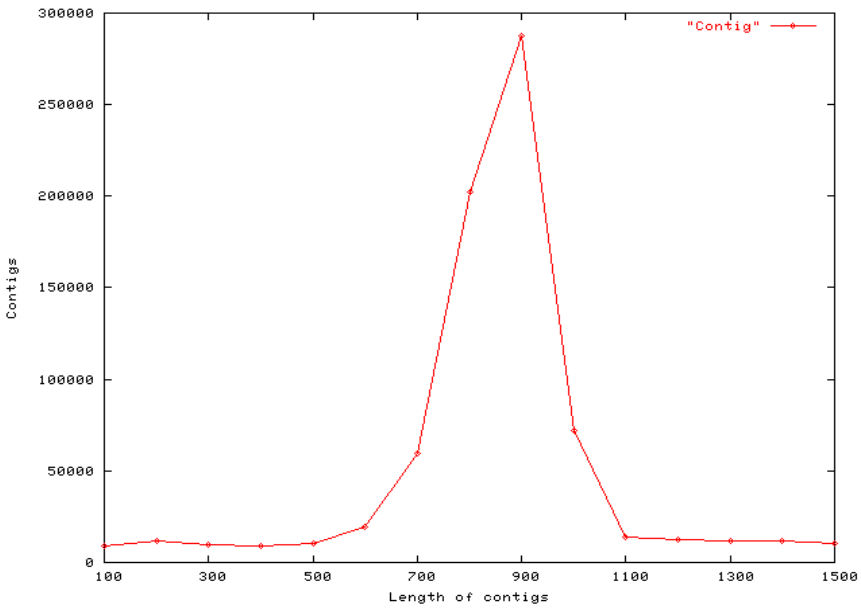


図 2.9 Saragasso Sea データ中のコンティグ長の分布。横軸はコンティグ長、縦軸はコンティグ数を表す。約 900 bp にピークがあり、全体の平均長は 1006 bp である。

論文[21]の著者たちはこのデータから約 100 万個の遺伝子を予測したが、この予測方法は２通りの相同性検索に基づいていた。１つはデータベース中にある既存の遺伝子に対して相同性検索を行ったもので、もう１つは今回の Sargasso デ

ータの中で相同性検索を行ったものである。２つ目の方法から得られる配列は必ずしも遺伝子とは言えないが、それでもゲノム中の保存されている領域を特定しているのでそれなりに信頼性の高い予測方法だと考えられる。一方で本手法をこのデータに適用したところ約 140 万個の遺伝子を予測することができた。そしてその内訳を論文の著者たちが予測した遺伝子セットと比較すると、１つ目の方法で予測された遺伝子セットの約 96%を、また２つ目の方法で予測された遺伝子セットの約 92%をカバーしていた。このカバー率を元に考えると、今回新たに予測できた 40 万個の遺伝子は信頼性の高い新規遺伝子であると考えられる。

また本手法の優位性をさらに検証するために Sargasso データのコンティグ長の分布（図 2.9）と既存の原核生物ゲノムの遺伝子密度（生物種によらずほぼ一定）を用いて理論的に推定される各長さにおける遺伝子数の分布を計算し、本手法で予測された遺伝子セットと論文でアノテートされた遺伝子セットの分布との比較を行った（図 2.10）。理論値から推定された遺伝子を正解とすると、本手法はより高い精度で遺伝子を予測できた事になる。また論文でアノテートされた遺伝子セットと比較すると、特に 1000 bp 以下の遺伝子に対して高い感度で予測を達成できている事がわかる。よってメタゲノムデータに対しては、特に短い遺伝子に関して相同性検索による遺伝子発見には限界がある事がわかり、本研究の重要性が改めて確認されたと言える。

しかしながら図 2.10 に見られるように、本手法により得られた分布は理論的に推定された分布に対していくつかの違いもある。まず 200 bp あたりの短い遺伝子で多くの擬陽性が発生しているが、これは断片ゲノムのテストのときと同様に ORF が断片化して偽 ORF との分類がうまくできなかったためである。一方で 900 bp の付近では逆に見落としが生じているがそれには以下の理由が考えられる。ま

ず図 2.9 によればコンティグ長のピークは 900bp のあたりにきていて、また図 2.10 の理論値では 900bp のところにピークが来ている。よってこれら約 900bp の断片は、すべてが ORF 領域であるものの割合が高い。一方でこれらの断片は one-pass 配列の平均長（700bp）より長いので、あまりトリミングされていない可能性が高い。よってこれらの断片の末端には多くのシーケンスエラーが含まれている可能性が高く、その結果偽ストップコドンの出現や読み枠のずれが発生し ORF が複数に分断されたと考えられる。以上の理由で 900bp 付近の ORF の多くが本手法では分断された状態で予測されたために、今回のような分布の違いが生じたと考えられる。

また今回のドメイン分類によると全データ中の約 90%が Bacteria に分類されたが、これは論文中で推定された割合とほぼ等しかった。本手法で提案したドメイン分類は遺伝子を分類するもので、必ずしもゲノムの生物種を推定するものではない。しかし原核生物は遺伝子密度が高いので断片中にはほとんど遺伝子が含まれていて、その結果断片ゲノムの生物種を推定することと同等の役割を果たしたと考えられる。よって本手法はメタゲノムデータから、遺伝子を予測するだけでなく、また環境がどのような生物種から構成さているのかを大まかに知るための有効な手法であると考えられる。

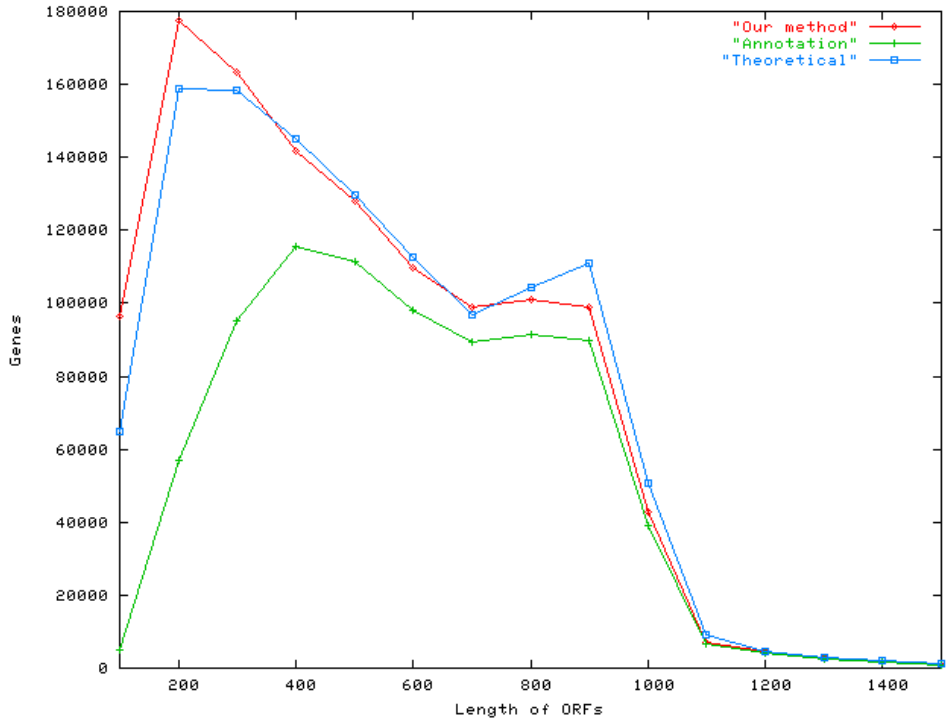


図 2.10　3 種類のアプローチにより予測された遺伝子数の分布の比較。それぞれの分布は本手法（赤）、論文の著者らが行った相同性検索に基づく方法（緑）、理論値（青）から得られたものである。理論値はデータセット中のコンティグ長の分布（図 2.9）と既存の原核生物ゲノムの遺伝子密度（生物種によらずほぼ一定）から推定したものである。横軸は遺伝子長、縦軸は遺伝子数を表す。本手法は相同性検索による方法に比べて、より理論値に近い遺伝子数を予測できていることがわかる。

２．５　真菌へのターゲットの拡張

２．５．１　背景

微生物をターゲットとした解析から得られるメタゲノムデータセットの中には、わずかながら真核生物特有のマーカー遺伝子である 18S rRNA の存在が確認されている [2,3]。このことはゲノムを採取する際のフィルタリングの条件を考えると、データセット中に単細胞真核生物のゲノムが含まれていることを意味する。現在知られている単細胞真核生物には真菌（Fungi）やアルベオラータ（alveorata）、ケルコゾア（cercozoa）などがあり、実際にこれらと相同性を示す配列がデータ中から確認されている。よって前述した原核生物をターゲットとした遺伝子予測の手法を単細胞真核生物にも拡張しようと試みたが、残念ながら現在完全長ゲノムが読まれている単細胞真核生物はわずかしかない。しかしその中でも Fungi は比較的多くのゲノムが読まれていて解析を行えるゲノムデータがそろっている。よって本章では Fungi に注目し原核生物に適用した遺伝子予測とドメイン分類の手法の拡張を試みた。

２．５．２　利用したデータと手法

表 2.3 は現在 NBCI の ftp サイトから利用できる完全長 Fungi ゲノムの一覧で

ある。一般に、真核生物のゲノムにはイントロンが含まれているが、Fungi の多くはイントロンを含んでいない。よって本研究ではイントロンの構造を含む遺伝子を予測のターゲットから除き、原核生物のときと同じ方法で ORF 構造を抽出した。そしてイントロンを大量に含んでいる *A.fumigatus*、*C.neoformans*、*S.pombe* の３つのゲノムデータは今回対象から外した。また *C.glabrata* に関しては遺伝子のアノテーションリストにおかしな点がいくつか見受けられたので、これも今回の対象から外した。結果として７種類のゲノムデータを利用して Fungi モデルの構築を目指した。

表 2.3 現在利用可能な Fungi の完全長ゲノムデータとその特徴

生物種	染色体の数	GC%	遺伝子密度	イントロンを含む遺伝子の割合	平均 ORF 長
Aspergillus_fumigatus	8	48.7%	24.8%	77.7%	1443
Candida_albicans	1	33.5%	30.9%	2.7%	1458
Candida_glabrata_CBS138	13	38.7%	31.8%	1.6%	1510
Cryptococcus_neoformans_var_JEC21	14	48.6%	27.1%	96.9%	1611
Debaryomyces_hansenii_CBS767	7	36.3%	37.7%	5.0%	1340
Encephalitozoon_cuniculi	11	47.3%	43.1%	0.7%	1080
Eremothecium_gossypii	7	52.0%	39.8%	4.6%	1472
Kluyveromyces_lactis_NRRL_Y-1140	6	38.8%	35.2%	2.5%	1406
Saccharomyces_cerevisiae	16	38.4%	35.6%	5.2%	1485
Schizosaccharomyces_pombe	3	36.1%	27.9%	44.7%	1395
Yarrowia_lipolytica_CLIB99	6	49.0%	22.8%	10.1%	1441

遺伝子密度はゲノム中のエキソン領域の割合を表す。

まず、コドンモデルを構築する際には回帰を用いるので、ゲノムデータの GC% にある程度の幅が必要である。今回扱うデータは 33.5%～52.0%と原核生物に比

べてあまり幅は無かったが回帰を行うには十分であった。また一般に真核生物は遺伝子密度が低いこともあり、遺伝子領域と非遺伝子領域では GC%に違いがある。実際に計測したところ、Fungi でも遺伝子領域の GC%が若干高い傾向が見られた。しかしこの差を今回コドンモデルに組み込んでも予測結果に変化は見られなかった。よってモデル構築の際には原核生物同様ゲノム全体の GC%から回帰を行った。ただし Fungi のスタートコドンはほぼ 100%が ATG なので、スタートコドンに関しては今回モデルを作らなかった。

次に、Fungi の平均 ORF 長は原核生物の平均 950bp と比べて長い（表 2.3)。この事は Fungi の遺伝子を予測する上で、また原核生物と Fungi のドメイン分類を行う上で有効な指標となりうる。よって今回 Fungi に対する ORF 長のモデルを新たに構築した。また Fungi の最も上流に観測されるスタートコドンから真のスタートコドンまでの距離を観測したところ、全体で 95%以上の遺伝子が最も上流のスタートコドンを採用していた。これは真核生物の翻訳開始地点の決定は難しく、まだしっかりとアノテーションされていないためだと思われる。しかしながらバックグランドとの比を取ると GC%によってスコアに差が出てくるので、この指標に関しても Fungi で新たにモデルを構築した。

次に隣接する ORF 間の向きを考慮した距離を観測したところ、原核生物との大きな違いが観測された（図 2.9)。まず Fungi は遺伝子密度が低いので、原核生物と比べて全体的に ORF 間の距離は広く分布していた。またオペロン構造をもたないので Tandem な向きをとる ORF 間の距離は原核生物に比べて長く、顕著な違いが見られた。次に Tail-to-Tail の向きの ORF 間はプロモータ領域を伴わないので、他の２つと比べて分布の距離は短い事が観測された。また Tandem と head-to-head の構造は ORF 間に存在すると考えられるプロモータの数が違うに

も関わらずほとんど同じ分布を描いた。これは興味深い事実である。そしてこれらの情報を予測に活用するために、Fungi に対する ORF 間の距離のモデルを新たに構築した。

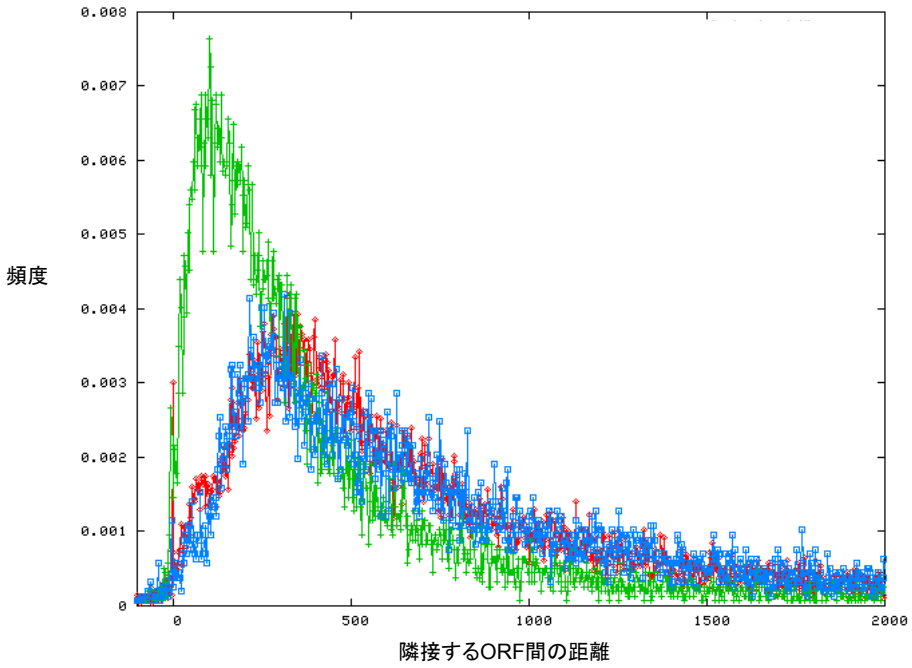


図 2.9 Fungi に対する隣接する ORF 間の向きを考慮した距離の分布。ORF 間の距離の分布を Tandem（赤）、Head-to-head（青）、Tail-to-Tail（緑）の向きに分けてそれぞれ描いた。原核生物の分布（図 2.3.e）と比較すると Tandem の距離に顕著な違いが見られるが、これは Fungi がオペロン構造を持たないためである。また Tail-to-Tail はプロモータ領域を間に伴わないので、他の分布に比べて距離が短い。また Tandem と Head-to-head では遺伝子間に存在するプロモータの数が異なるにも関わらず、ほとんど等しい分布を描いた。

２．５．３　結果と考察

Bacteria と Archaea のモデルに加えて新たに構築した Fungi のモデルを統合し、３つのドメインからなる遺伝子予測モデルを今回作成した。そしてその性能を評価するために、原核生物のときと同様の方法で作成した Fungi の断片テスト配列をテストした。しかし今回利用できたゲノムデータは７生物種と少なかったので、データ中から１つをテストデータとして選び残りの６つデータでモデルを構築する方法をすべてのペアで繰り返す方法で評価を行った（表 2.4)。結果を見ると各生物種のデータに対して実用的な予測精度が得られている事がわかり、特に感度（Sn）に関しては平均 93.4%と高い精度を達成した。この事は Fungi もまたコドン使用頻度と GC%の間に強い相関があることを示している。

しかし、特異性（Sp）に関しては原核生物と比べて大きく劣る結果となった。その原因としてまず微量ながらイントロンの存在が挙げられる。今回の手法ではイントロンの構造を持つ遺伝子でも内部コドンの塩基組成から遺伝子領域として認識はするが、ドナーサイトとアクセプターサイトを認識しないために結果としてアノテーションした際に不正解となってしまう。実際今回テストした生物種のなかでも、１番多くイントロンを含んでいる *Y.lipolytica* は１番特異性が低かった。よって断片配列からイントロンの構造を認識するモデルの構築は今後の課題である。また、次に考えられる原因として遺伝子密度が低い事が挙げられるが、これに対処するためには遺伝子間の距離のモデルなどをより精密に作っていくなどの検討が必要であろう。

表 2.4 Fungi の 700bp のテスト断片配列に対する予測結果

生物種	GC%	Sn	Sp	ドメイン分類の精度
Candida_albicans	33.5%	94.6%	70.7%	61.1%
Debaryomyces_hansenii_CBS767	36.3%	93.4%	81.9%	84.6%
Saccharomyces_cerevisiae	38.4%	94.0%	78.5%	85.5%
Kluyveromyces_lactis_NRRL_Y-1140	38.8%	94.5%	78.2%	88.0%
Encephalitozoon_cuniculi	47.3%	93.2%	88.1%	64.5%
Yarrowia_lipolytica_CLIB99	49.0%	91.5%	53.6%	76.3%
Eremothecium_gossypii	52.0%	94.0%	83.6%	79.3%
平均		93.4%	76.4%	78.7%

また今回の結果において、ドメイン分類の精度は平均 78.7%と原核生物の分類と比べて劣る結果となった。よってドメイン分類の精度が遺伝子の予測精度に響いていると考え Fungi のモデルのみを用いてテストを行ったが、３つのモデルを用いたときと比べて予測精度に変化はみられなかった。この事は Fungi のゲノム内に存在するいくつかの遺伝子は、Bacteria と Archaea のモデルを用いても予測が可能であることを示唆していた。よって今回構築したこれら３つのドメインモデルが Bacteria と Archaea、Fungi のゲノムデータに対して、それぞれどの程度の予測性能を示すのかを検証した（図 2.10)。結果を見たところ Bacteria モデルだけで予測を行った場合でも、Bacteria や Archaea の遺伝子だけでなく Fungi の遺伝子に関しても半数以上を予測する事ができた。そして Fungi のモデルを追加したときには、残りの遺伝子をさらに予測する事ができた。また注目すべきことは、モデルを統合したことによる元のモデルからの予測性能の低下（トレードオフ）がほとんど見られなかった事である。よってこれら３つのモデルが効果的に働いている事が検証された。

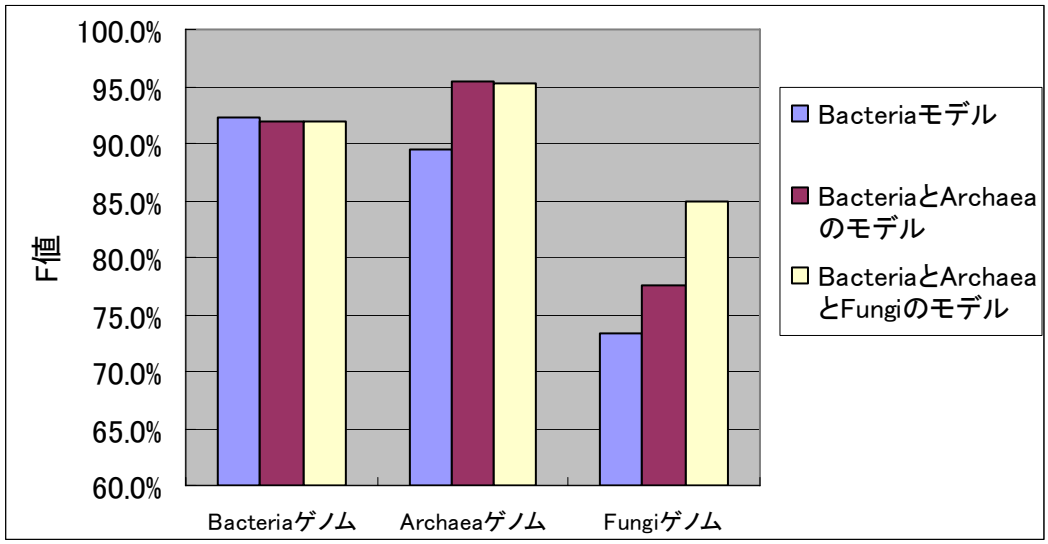


図 2.10　今回構築した３つのモデルの効果の検証。 Bacteria、Archaea、Fungi の 700bp 断片配列に対して、３つのモデルをいろいろな組み合わせで適用しそれぞれの精度の F 値（Sn と Sp の調和平均）を求めた。Bacteria モデルだけで予測を行っても、大部分の Archaea と多くの Fungi 遺伝子を予測することができた。残りの予測できなかった遺伝子は、それぞれ対応したドメインのモデルを用いたときにさらに予測する事ができた。またモデルを追加したことによる元のモデルからの予測性能の低下も見られなかった。

次に３つのドメインの間で、コドン使用頻度がどの程度異なるかを生物種全体のゲノムを用いて比較を行った。図 2.11 は３つのドメインに属する各完全長ゲノムに対して構築した３つのモデルを適用し、得られた３種類の相対スコアの位置をプロットしたものである。モデルのフィッティングの度合いから、ほとんどのゲノムがドメインレベルできれいに分類できた事がわかる。よって完全長ゲノムではほぼ 100%の割合でドメイン分類が可能という事になる。しかし例外的にいくつかのゲノムはうまく分類できなかったものがあり、例えば２つの超高熱菌の Bacteria（*Aquifex aeolicus*、*Thermotoga martima*）は Archaea モデルに高いフィッティングを示した（最も右に位置する緑の２点）。この２つのゲノムは系統

木の中でも Bacteria と Archaea の分岐が起こった直後に Bacteria 内で分岐した位置にいる。そのような事から考えても、今回の結果はある意味妥当であったとも考えられる。また他の Bacteria の 1 種である *Anaplasma marginale* もまた Archaea のモデルにフィッティングを示した（右から 3 番目の緑の点）。このゲノムは系統的にはαプロテオバクテリアの 1 種で Archaea との関連性は見られないが、牛の共生細菌であり特殊な環境に生息しているので他のαプロテオバクテリアとは異なる傾向を示したのかもしれない。また図 2.10 の右端に外れて位置している Fungi の *Encephalitozoon cuniculi*（染色体ごとにプロット）は、これもまた Archaea のモデルにフィッティングを示した。この生物種もまた共生細菌の 1 種であり他の Fungi に比べて ORF の平均長が短く（原核生物の平均長にほぼ等しい）、イントロンもほぼ皆無である（表 2.3）。よってこの生物種は Fungi の中でも特異的なゲノム構造を持っているので興味深い。以上の結果を考察すると、コドン使用頻度は進化の過程だけではなく生息している環境の要因にもまた大きく影響するという事がいえるだろう。

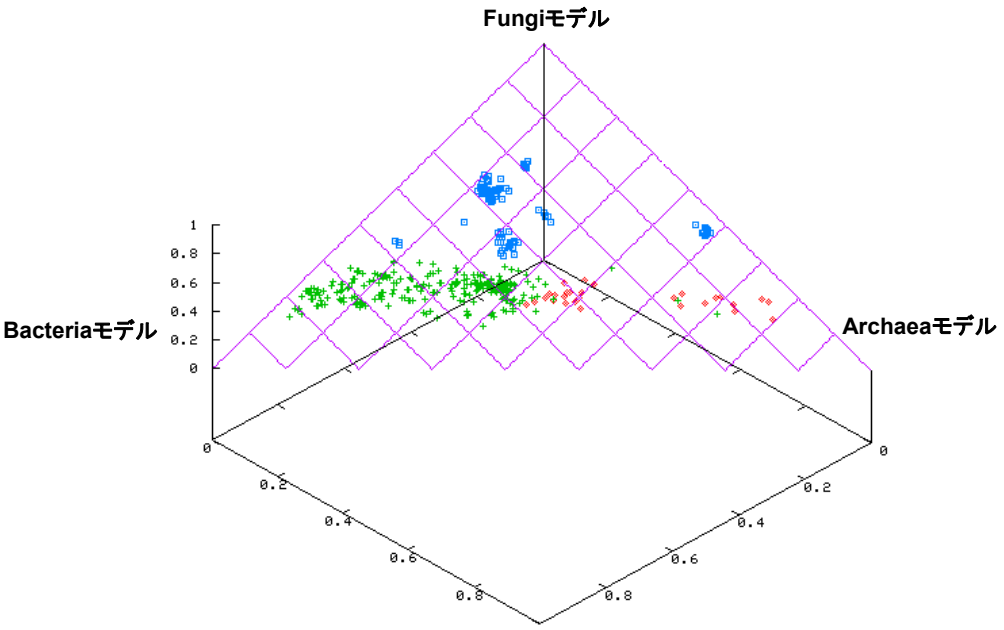


図 2.11 各完全長ゲノムに対する、3 種類のコドンモデルから得た相対スコアのプロット。もしスコアの比がすべて等しい場合には、点は中心にくる。Bacteria（緑）、Archaea（赤）、Fungi（青）のゲノムは、それぞれ対応したモデルにフィティングを示し、結果としてきれいな分類ができている。ただしいくつかのゲノムはうまく分類できず、特に Bacteria の 2 つの超高熱菌（最も右にある緑の 2 点）に関しては Archaea モデルに強いフィッティングを示した。

２．６　まとめ

メタゲノム配列からの遺伝子予測は与えられる配列の生物種が未知であるという問題から、既存の遺伝子予測手法は適用ができない。そこで本章ではコドン使用頻度と GC%の相関を利用して、与えられた配列の GC%からダイコドン使用頻度を推定するアプローチをとった。さらにこのときドメイン間のコドン使用頻度の違いに注目して Bacteria、Archaea、Fungi の 3 つのコドンモデルを構築した。そして予測精度のさらなる向上のため ORF 長の分布、最も上流にあるスタートコドンからの距離の分布、隣接する ORF 間の向きを考慮した距離の分布を指標として取り入れた。これらの指標は原核生物と Fungi の 2 つのモデルを構築した。そしてこれらのモデルを組み合わせて全体の予測モデルを構築した。

結果として本手法は 700 bp の断片ゲノムに対して高い予測精度を達成できた。原核生物（Bacteria、Archaea）では感度と特異性それぞれ 95%と 90%、そして Fungi では 93%と 76%を達成できた。真菌では微量のイントロンの存在と低い遺伝子密度のために、擬陽性が多く発生して特異性の低下が見られた。この特異性の改善は今後の課題である。

またドメイン分類では 700 bp の断片ゲノムに対して原核生物の間では平均約 90%の分類精度を達成できた。この精度は外来性遺伝子の問題もあり一概に正しいものとは言えないが、それでもメタゲノムデータ中のドメインの構成を知るには有効な手段であると考えられる。

３．メタゲノム配列からの系統分類

３．１　背景

微生物をターゲットとしたメタゲノムデータからの系統分類は、たくさんの生物種の断片配列を含んでいるので困難な問題である。このような状況の中で現在最もよく使われているアプローチの１つが 16S rRNA を使った方法である。16S rRNA は高度に保存されている遺伝子であり、この領域を含むゲノム配列を高い精度で正しい系統に分類することができる [27,28]。しかしメタゲノムデータ中に占める 16S rRNAs の割合は Saragasso Sea[2]サンプルでは全コンティグ中の 0.06%、Minnesota soil[3]のサンプル中では 0.017%とわずかにしか存在せず、多くの断片配列の生物種は未知のままである。よってこの方法では真の系統分類を実現したとは言えず、残った多くの断片配列を分類するためには違ったアプローチが必要になってくる。そこで本研究ではこれら多くの断片配列を系統分類する方法を検討した。

まず断片配列の系統分類を行うにあたって、本研究はゲノム中の塩基組成に注目をした。塩基組成はゲノム内で高度に保存されており、オリゴヌクレオチドの頻度を見ると生物種ごとに偏った傾向が見られる[29]。そしてこの塩基組成を用

いていくつかの生物種の完全長ゲノムから系統分類を試みた研究が過去に報告されている[30]。この研究によると完全長ゲノムの塩基組成を用いた分類は 16S rRNA を用いた分類と一定のレベルの一致が見られ、さらにそれは遺伝子領域内の塩基組成において顕著に見られたと述べている。そう考えた場合今回のメタゲノム断片に対しても、ゲノム全体の塩基組成の頻度を見るより遺伝子領域の頻度を見る方がより良い系統分類ができると考えられる。また原核生物のゲノムは遺伝子密度が非常に高いので(80~90%)、ほとんどすべての断片に遺伝子は含まれていると考えてよい。よって本研究では２章で構築した遺伝子予測手法を用いてまずメタゲノム断片から遺伝子領域を特定し、次に特定された遺伝子領域のダイコドン使用頻度から系統分類を行う事を試みた。

実際に行った今回の手順としてはまずダイコドン使用頻度を用いるという理由から、断片配列を検証する前に遺伝子を用いてどの程度系統分類が可能であるかを検証した（遺伝子の平均長は約 1000 bp なのでメタゲノムデータから出てくる断片と仮定しても矛盾は無い)。そしてメタゲノムデータ中に含まれる難培養性の新規生物種を想定して、既知の生物種と未知の生物種の両方のケースを想定して遺伝子の系統分類を試みた。

ところで近年になって前述したようなゲノムの塩基組成を用いて、メタゲノム配列の系統分類を行った研究がいくつか報告されている [31,32]。しかしこれらはゲノム全体のオリゴヌクレオチドの頻度を使っていて、現状では 1000 bp 以下の短い断片に対して必ずしも良い精度を達成できているとは言えない。よって今回のように遺伝子領域のダイコドン頻度に注目して系統分類を行うことはそれで重要な研究であると考えられる。

３．２ 利用したデータ

本研究では遺伝子とゲノム断片からの生物種分類を行うために、NCBI ftp サイト [20] にある真正細菌（Bacteria）と古細菌（Archaea）の完全長ゲノムデータと遺伝子のアノテーションリスト、そして各生物種に付けられた系統名リストを利用した。同じ種ではゲノム構造にほとんど差が無いので各生物種から 1 つの株のみを選び結果、Bacteria205 種と Archaea23 種のゲノムデータを利用して系統分類を行った。

次にトレーニングデータとテストデータの作成方法について述べる。まず、水平伝播遺伝子は他の生物種から来た外来遺伝子なのでゲノム内の他の遺伝子と異なったコドン使用頻度を使っている。よって水平伝播遺伝子は今回の対象から取り除く事にした。また 16S rRNA などのマーカー遺伝子は変異が起こりにくく高度に保存されているので、これもまたゲノム内の他の遺伝子と異なるコドン使用頻度を持っている。よってマーカー遺伝子も今回の対象から取り除くことにした。実際にこれらの遺伝子を取り除くにあたっては、これらの遺伝子リストをピックアップしているデータベースの情報を利用した [33,34]。

次に、残った遺伝子セットをランダムに選び 10:1 の比で分け前者をトレーニングデータとし後者をテストデータとした。そしてテスト用の断片配列は評価用の遺伝子データのみを含むような形で作成し断片の長さはすべて 1000bp とした。また実際に遺伝子と断片配列のテストをするときには各生物種のテストデータからランダムに 50 個ずつ抽出して分類のテストを行った。

３．３ 手法

３．３．１　原核生物の系統関係について

今回原核生物の遺伝子と断片配列からの系統分類を行うにあたって、現在これらがどのような形で系統分類されているのかを最初に説明する。

原核生物の系統は一般に領域(Domain)、門(Phylum)、網(Class)、目(Order)、科（Family)、属（Genus)、種（Species）のような階層的な分類がなされている。例えばNCBI-Taxonomomy[35]の中では大腸菌のK12株(*Escherichia coli K12*)は各階層で *Bacteria*、*Proteobacteria*、*Gammaproteobacteria*、*Enterobacteriales*、*Enterobacteriaceae*、*Escherichia*、*Escherichia coli* の系統名が付けられている。これらは絶対的な定義ではなく他の専門書などを見比べると種によっては稀に違う系統名がつけられたりしているが、それでもおおよそのコンセンサスは取れているので今回はNCBI-Taxonomy の分類名を正解として扱った。

今回利用した生物種のゲノムデータは表 3.1 のような系統数でそれぞれ構成されている。ここで今回便宜を図るために各階層の系統名をグループ、そしてグループ内に所属する実際の系統名をメンバーと呼ぶことにした。表 3.1 は子を複数持つ木の構造を形成していると思っていただきたい。そして一番最下層のSpecies のメンバーの数が今回利用したゲノムデータの生物種の数である。よって今回の系統分類の評価方法は Species の各メンバーが持っている遺伝子と断片配列を与えたときに、それらがどのレベルのグループまで正しく分類できたのかを評価す

るものとした。つまり Order まで正しく分類できたときには Domain、Phylum、Class でも正しく分類できた事になるし、Species まで正しく分類できたものはすべてのグループで正しく分類できた事になる。

表 3.1 今回利用した生物種が構成する系統グループとメンバー

グループ	メンバーの数
領域（Domain）	2
門（Phylum）	17
網（Class）	36
目（Order）	73
科（Family）	106
属（Genus）	145
種（Species）	228

本研究では各階層の系統名をグループ、グループ内の各系統名をメンバーとした。

３．３．２　分類方法

本研究は各生物種の遺伝子と断片配列を入力（以降この２つをまとめて入力配列と呼ぶ）として系統分類を行い、各グループごとの分類精度を評価するようにした。そしてその分類を行う際の指標としてダイコドン使用頻度を用いた。ところがここで問題となるのが一般に遺伝子の平均長は短いので入力配列から直接頻度を計算すると、その頻度は生物種が本来使っているダイコドン使用頻度に対し

て大きな誤差を含むものとなってしまう。よって本手法は各生物種のトレーニングデータにある十分な量の遺伝子からあらかじめダイコドン使用頻度を学習しておき、その頻度を用いて入力配列を評価することにした。つまり今回用いた生物種は228種なので、228個のダイコドンモデルをトレーニング用の遺伝子セットを学習してあらかじめ構築しておく。これを$M=\{M_1,M_2,\cdots,M_{228}\}$とする。そして入力配列に対して228個のモデルで評価を行い、最もスコアの高いものを正解とした。いかに具体的な計算方法と評価方法を示す。

入力配列gの中にある遺伝子領域を$g=\{g_1,g_2,\cdots,g_N\}$とする。ここで入力配列が遺伝子の場合$N=1$であり、断片配列の場合は$N\geq 1$である（ここで$N=0$は考慮しない)。次に任意の$g_i(1\leq i\leq N)$に対して内部コドンをc_j、そしてスタートコドンとストップコドンを除いた連続した内部コドンの長さをlとすると$g_i=(c_1,c_2,c_3,\cdots,c_l)$となる。このとき任意の生物種のダイコドンモデル$M_m(1\leq m\leq 228)$から得られる$c_{j-1}c_j$の頻度確率を$p(c_j\mid c_{j-1})$とすると、$M_m$から得られる$g_i$に対するスコア$S_m(g_i)$は、

$$S_m(g_i)=\log p(c_1)+\sum_{k=2}^{l}\log p(c_k\mid c_{k-1}),\qquad (11)$$

となり、このとき入力配列gに対するスコアは$S_m(g)$は、

$$S_m(g)=\sum_{i=1}^{N}S_m(g_i),\qquad (12)$$

となる。そしてこれを228種類のダイコドンモデルで評価したとき、gは

$$S(g) = (S_1(g), S_2(g), \cdots, S_{228}(g)), \qquad (13)$$

で示す 228 種類ののスコアを持つことになる。そしてこの中から最大スコアを示したものを正解のモデルとして予測する。ここで g の正解の生物種を $a(1 \le a \le 228)$、モデルから予測された生物種を $b(1 \le b \le 228)$ とする。このとき既知の生物種に対する分類では正解を

$$S_b(g) = \max_{1 \le k \le 228} S_k(g), \qquad (14)$$

の条件で選ぶものとし、未知の生物種に対する分類では正解を

$$S_b(g) = \max_{1 \le k \le 228} S_k(g), \qquad (\text{ただし } k \ne a) \qquad (15)$$

の条件で選ぶものとした。

そして a の Species でのメンバー名を Species(a)などとすると、選ばれたモデルから分類評価は次のように決まる。

if　Species(a) = Species(b)　then　Species 以上のグループで分類成功

else if　Genus(a) = Genus(b)　then　Genus 以上のグループで分類成功

else if　Family(a) = Family(b)　then　Family 以上のグループで分類成功

else if　Order(a) = Order(b)　then　Order 以上のグループで分類成功

else if　Class(a) = Class(b)　then　Class 以上のグループで分類成功

else if　Phylum(a) = Phylum(b)　then　Phylum 以上のグループで分類成功

else if　Domain(a) = Domain(b)　then　Domain のグループで分類成功

else　分類失敗

３．４ 結果と考察

３．３．３　遺伝子の系統分類

今回の研究の目的は断片配列に対しての系統分類であるが、まずダイコドン使用頻度でどの程度分類が可能であるかを検証するために遺伝子単位での分類を行った。評価は既知の生物種の遺伝子についてまずどの程度分類できるかを行い、そして未知の生物種の遺伝子が存在するときにどの程度分類できるのかをシミュレートした。

３．３．３．１　既知生物種の遺伝子に対する系統分類

各生物種が使っているコドン使用頻度は種特異的で、実際にゲノム内に存在する遺伝子のコドン組成は非常に類似している。またコドン使用頻度は系統関係が近いほど類似傾向がある [31]。生物種が持っている１遺伝子から得られる頻度情報は、多くの遺伝子から得られる頻度情報に比べてあいまいにはなるが、それでも近い頻度情報を示すはずである。よって今回扱った 228 種の生物種においてトレーニング用の遺伝子セットでダイコドンモデルを構築し、テスト用の遺伝子セットがどの程度分類できるのかを検証した。分類評価はグループごとで行い、分類精度の評価の指標としてグループ内の平均感度（avgSn）と平均特異度（avgSp）

$$\mathrm{avgSn} = \left(\sum_{k=1}^{N} \frac{tp_k}{t_k} \right) \cdot \frac{1}{N}, \qquad (16)$$

$$\mathrm{avgSp} = \left(\sum_{k=1}^{N} \frac{tp_k}{p_k} \right) \cdot \frac{1}{N}, \qquad (17)$$

を用いた。ここで N はグループ内のメンバー数であり、tp_k は k 番目のメンバーに正しく分類された数、t_k はメンバー内の数、p_k はメンバー k と分類された数である。

図 3.1 は分類の結果を示したもので、すべてのグループに対して高い精度での分類が達成できた。しかしその中でも Species グループでの分類精度はあまり良くなかったが、これは Species 間ではゲノムに相同性が高いもの同士が多く存在

するためにそれらを分類することができなかったためである。それでもグループの階層が上がるにつれて分類精度は徐々に向上していき、残りのグループでは Sn、Sp ともにほぼ 90%以上を達成した。次に、さらに細かい考察を行うために各メンバーの分類精度を調べた。例として Phylum グループのメンバー結果を表 3.2 に示す。表での Sn, Sp は個別のメンバーの感度 $\frac{tp_k}{t_k}$ と特異性 $\frac{tp_k}{p_k}$ をそれぞれ表す。結果を見るとすべてのメンバーにおいて高い分類精度を達成でき、メンバーによる精度の偏りはほとんど見られなかった。またこの精度はメンバー内に存在する種の数にも依存していなかった。よってこの先新たな生物種のゲノムが読まれメンバー内の種の数が増えても、またグループ内で新しいメンバーが増えても今回と同様の分類精度が保てるということが言える。このように既知遺伝子に関しては高い信頼性での分類が可能であった。

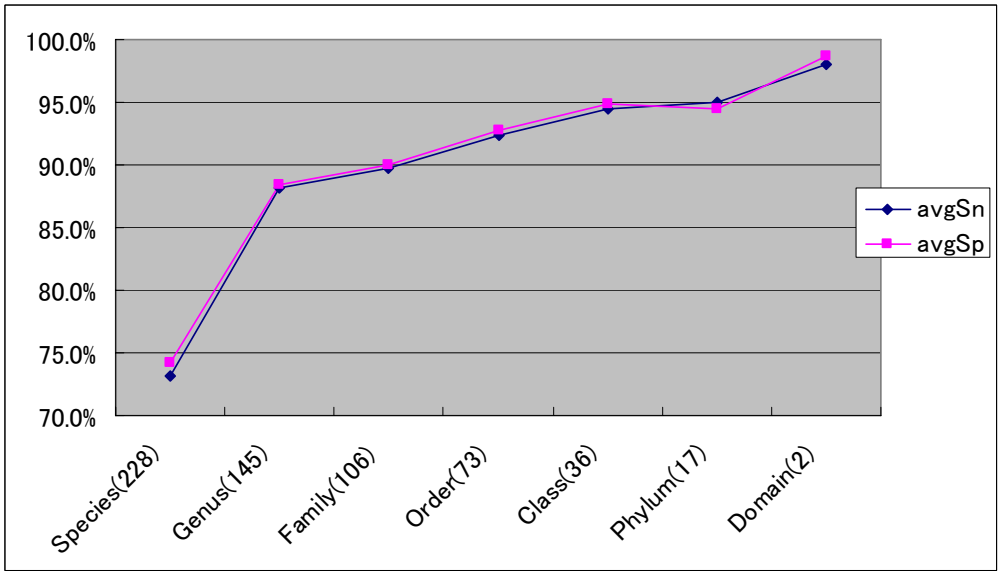


図 3.1 既知生物種の遺伝子に対する分類結果。グループごとの平均感度(Sn)と平均特異度(Sp)をプロットした。横軸はグループ（メンバー数）を表す。Species グループではゲノムの類似性が高いものが多く存在するのでその分、分類精度は落ちる。

表 3.2 Phylum 内メンバーの既知遺伝子に対する分類精度

Phylum（生物種の数）	Sn	Sp			
Actinobacteria (18)	98.3%	95.4%	Euryarchaeota (19)	95.4%	98.2%
Aquificae (1)	100.0%	98.0%	Firmicutes (46)	93.7%	95.6%
Bacteroidetes (4)	90.5%	96.3%	Fusobacteria (1)	90.0%	88.2%
Chlamydiae (6)	97.7%	88.8%	Nanoarchaeota (1)	94.0%	97.9%
Chlorobi (3)	87.3%	94.2%	Planctomycetes (1)	96.0%	85.7%
Chloroflexi (2)	94.0%	100.0%	Proteobacteria (104)	97.1%	96.8%
Crenarchaeota (5)	95.2%	93.0%	Spirochaetes (5)	94.4%	87.7%
Cyanobacteria (9)	95.3%	96.0%	Thermotogae (1)	98.0%	96.1%
Deinococcus-Thermus (2)	98.0%	98.0%	平均	95.2%	95.5%

３．３．３．２　未知生物種の遺伝子に対する系統分類

直前の検証により既知遺伝子は高い精度で分類できる事がわかった。しかしメタゲノムデータ中には難培養性の未知生物種も存在する。よって今回はこれら未知生物種の遺伝子が、どの程度分類できるのかを検証するためのシミュレーションを行った。具体的には、与えられた遺伝子に対してその生物種由来のダイコドンモデルをマスクした合計 227 個のモデルの中から１番スコアの高いものを正解として、今回の未知生物種の遺伝子に対する分類を行った。

しかし未知生物種の分類を既知のときと同様にグループごとに評価しようとしたとき、メンバー内に生物種が自分自身しかない場合は正解のモデルがなくなっ

てしまう。よって今回は各グループの評価において正解のモデルがある場合、つまりメンバー内に２種以上の生物種が存在するものだけを評価の対象とした。そして既知生物種の遺伝子に対しても評価の対象を揃えて両者の分類精度の比較を行った。評価方法は同様 avgSn と avgSp を用いた。

表 3.2 に分類結果を示す。結果によると未知生物種の遺伝子に対してはグループ全体で分類精度が低下する形となった。そしてグループにわたっての分類精度の変化の様子は、既知生物種の分類精度の様子とは異なっていた。これは Genus 内では比較的ゲノムに相同性が高くダイコドン使用頻度にも高い類似性が見られるので、既知生物種では分類精度が悪くなり逆に未知生物種に対しては分類精度が良くなるためである。そうするとグループの階層が上がるにつれて既知生物種の分類精度は上昇するので未知生物種の分類精度は下降していくと考えられるが、Order の分類精度を底として上昇する様子が見られた。これはグループ内のメンバーの数が減少していったためだと思われる。しかしそれでも Domain の分類精度で極端な上昇が見られたのは、やはりドメイン間ではダイコドン使用頻度に大きな違いがあるためだと思われる。よって Domain レベルの分類は未知生物種の遺伝子に対しても十分効果を発揮できる結果となった。

次に精度の詳細を知るために、グループ内の各メンバーの分類精度を調べた。既知生物種のときと同様に例として Phylum のメンバーの分類精度を示す（表 3.3)。結果を見ると既知生物種のときと違いメンバー間でかなりばらつきが生じていた。これはメンバーによっては、その中にいる生物種のダイコドン使用頻度の類似度にかなりばらつきがあるということになる。その中でも Chloroflexi は未知生物種に対しても高い精度で分類ができたが、これは他のメンバーと比べてダイコドン使用頻度に違いが見られ、またメンバー内に属する２つの生物種

（*Dehalococcoides_ethenogenes*、*Dehalococcoides_sp.*）が同じ Genus でダイコドン使用頻度が非常に類似していたためである。一方で Deinococcus-Thermus は既知遺伝子では高い分類精度を示していたのだが、未知遺伝子では全く分類ができなかった。これはこのメンバーに属する２つの生物種のうち *Deinococcus_radiodurans* は常温菌で *Thermus_thermophilus* は高熱菌だったために２種の間でダイコドン使用頻度が大きく異なっていたためである[36]。また２章でも述べたようにコドン使用頻度はゲノム GC%に強く依存するので、これもまた分類精度を低下させる原因となる。例えば Proteobacteria や Firmicutes のようにメンバー内に多くの生物種がいて GC%の幅があれば、未知生物種の遺伝子が与えられてもメンバー内の近い GC%の生物種にフィッティングを示す。しかし生物種が少ないにも関わらず GC%に幅を持つメンバーは、未知生物種の遺伝子が与えられたときに GC%の近い他のメンバーの生物種にフィッティングを示すために分類に失敗する。よって Chrorobi や Spirochaetes では分類精度が特に低かった。以上の事からダイコドン使用頻度は系統ごとの類似性を持っている事は確かだが、環境に対する選択圧や GC%への依存を考慮すると、メンバーによっては未知生物種の遺伝子に対して極端に分類が悪くなってしまう。よって現状ではこのようなメンバーに対しては、ダイコドン使用頻度を用いるだけでは限界があると言える。

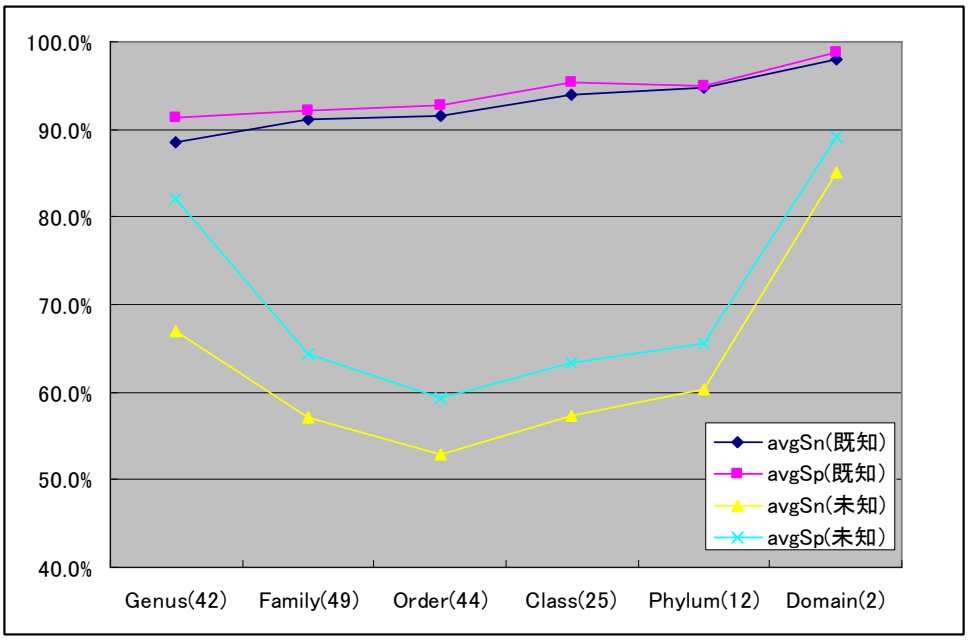


図 3.2 未知、既知生物種の遺伝子に対する分類結果の比較。グループごとの分類結果に対する平均感度(avgSn)と平均特異度(avgSp)をプロットした。横軸はグループ名（メンバー数）を表す。未知生物種でも Genus レベルではお互いのダイコドン使用頻度は比較的似ているので、より高い精度での分類ができた。Domain レベルでは未知遺伝子対しても高い分類性能が得られた。

表 3.3 Phylum の各メンバーの遺伝子に対する分類精度の比較

Phylum（生物種の数）	未知生物種		既知生物種	
	Sn	Sp	Sn	Sp
Actinobacteria (18)	84.9%	68.3%	98.3%	95.4%
Bacteroidetes (4)	53.0%	69.7%	90.5%	96.3%
Chlamydiae (6)	86.7%	67.7%	97.7%	88.8%
Chlorobi (3)	18.7%	45.9%	87.3%	94.2%
Chloroflexi (2)	92.0%	98.9%	94.0%	100.0%
Crenarchaeota (5)	60.4%	69.6%	95.2%	93.0%
Cyanobacteria (9)	52.9%	71.9%	95.3%	96.0%
Deinococcus-Thermus (2)	0.0%	0.0%	98.0%	98.0%
Euryarchaeota (19)	68.4%	83.8%	95.4%	98.2%
Firmicutes (46)	77.9%	84.7%	93.7%	95.6%
Proteobacteria (104)	89.7%	83.1%	97.1%	96.8%
Spirochaetes (5)	40.8%	41.8%	94.4%	87.7%
平均	60.4%	65.5%	94.7%	95.0%

メンバーは未知生物種の遺伝子の分類に対して評価可能なもの、つまり生物種数が２種以上あるものを選んだ

３．３．４　断片配列の系統分類

遺伝子に対する分類のときと違い、断片配列からはまず遺伝子領域を特定する必要がある。そこで本手法では２章で構築した遺伝子予測手法を用いて与えられた断片配列の遺伝子領域を最初に決めた。ここで断片配列には遺伝子領域が複数

存在しうるので、本手法は各遺伝子領域のスコアの合計値を１つのダイコドンモデルから得られたスコアとした。そして既知生物種に対しての分類のときには 228 個のダイコドンモデルから最もスコアの高かったモデルを正解とし、未知生物種に対する評価のときには与えられた配列の正解の生物種をマスクして 227 個のダイコドンモデルから最もスコアの高かったモデルを正解とした。

今回評価用の断片配列の長さは 1000bp とし、各生物種に対してトレーニングデータに使用した遺伝子を含まない領域から 50 個ずつ用意した。そして断片配列に対してアノテーションの遺伝子領域を用いて分類を行ったものと、本手法で予測された遺伝子領域を用いて分類を行ったものの２種類を評価した。評価の指標として今回は F 値の平均（avgSn と avgSp の調和平均）を用いた。図 3.3 に既知の生物種の配列に対する結果を、図 3.4 に未知の生物種の配列に対する分類結果を示す。２つの図から見て取れるように 1000 bp の断片配列と遺伝子との分類精度にほとんど差は見られなかった。この理由は今回用意した 1000bp の断片配列には平均 1.7 個の遺伝子が含まれていて、また遺伝子領域をもたない断片の割合は全体のわずか 2%としかなかったためである。また 1000 bp の断片に関して本手法を用いて遺伝子領域を予測したものとアノテーションを用いたものでは、予測誤差があるために本手法の方が弱冠精度に低下が見られたがその差はほとんど変わらなかった。このようにほとんどの短い断片配列には遺伝子領域が蜜に存在しまたそれらを正確に予測することも可能なので、遺伝子が分類可能であれば短い断片配列も同様に分類が可能である事がわかる。

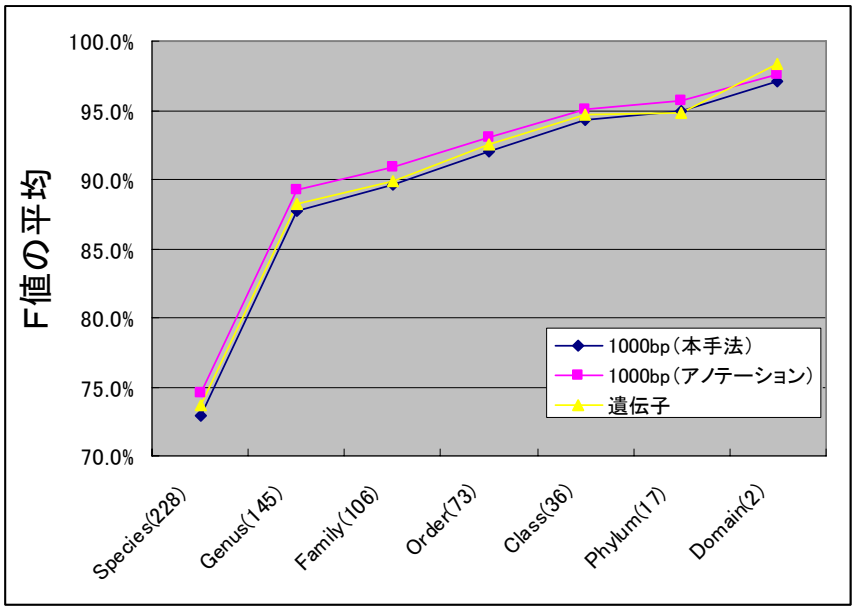


図 3.3 既知の生物種に対する 1000 bp 断片と遺伝子の分類結果の比較。1000 bp の断片に関してはアノテーションと本手法で予測した遺伝子領域の２通りで評価を行った。３つの間にほとんど差は見られない。

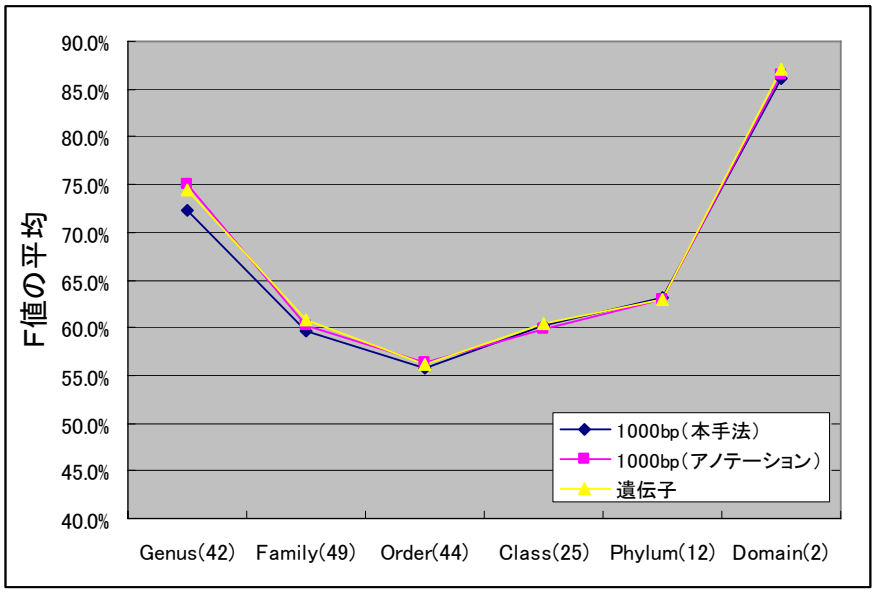


図 3.4 未知の生物種に対する 1000 bp 断片と遺伝子の分類結果の比較。1000 bp の断片に関してはアノテーションと本手法で予測した遺伝子領域の２通りで評価をいった。既知遺伝子のときと同様３つの間にほとんど差は見られなかった。

３．３．５　関連研究の紹介

今回我々は生物種ごとのダイコドンモデルを構築して、断片配列に対して最もスコアの高かったものを正解として系統分類を行った。しかし異なるアプローチとして、グループ内の各メンバーを１つのクラスとしてみなし、グループごとの多クラス問題を解く方法が考えられる(例えば Domain では Bacteria と Archaea の２クラス問題を解く)。実はこの方法を用いて断片配列の系統分類を行った研究が、去年 12 月の Nature Methods 誌において発表された [32]。

この研究は各クラス（メンバー）に属する断片配列の分類にゲノム全体のオリゴヌクレオチドの頻度を用いて、その多クラス問題を SVM(Support vector machine)という機械学習の手法を用いて解くことで系統分類を行ったものである。扱ったデータが違うので本手法との直接の比較はできないが、以下に論文で公開されている 1000 bp の断片配列に対する分類精度を参考までに載せる（表 3.4)。

表 3.4　1000 bp の断片配列（本研究とデータセット異なる）に対する関連研究の分類精度。

	既知の生物種		未知の生物種	
グループ	avgSn	avgSp	avgSn	avgSp
Genus	7.1%	96.7%	4.4%	87.9%
Order	25.1%	95.0%	6.4%	96.1%
Class	30.8%	93.2%	30.7%	83.0%
Phylum	50.3%	92.1%	40.6%	79.5%
Domain	57.7%	88.7%	57.7%	81.3%

関連研究では 10k bp 以上の断片ゲノムに対しては比較的良い分類精度を達成していたが、表にある結果のように 1000 bp の断片に対しては良い分類精度を達成できていなかった。本手法の結果（図 3.3 と図 3.4）と比較しても、特に既知の生物種に関しては本手法の方が良い分類が達成できている事がうかがえる。この理由としてオリゴヌクレオチドやダイコドン使用頻度も含めたゲノムの塩基組成は、種特異的ではあるがメンバー内では共通の特異性があまり見られない。よって関連研究のようにメンバーごとの塩基組成の情報を学習するよりも、本手法のように生物種ごとの塩基組成を学習した方がうまく分類ができたと考えられる。また同様の理由で関連研究の手法は、未知の生物種の分類に関しても学習データでうまく特徴空間が作れず良い分類精度が達成できなかったものだと思われる。

また関連研究の結果を見ると Genus レベルでの分類精度が既知、未知の生物種に関わらず極端に低下している事が見受けられる。一方で本手法は Genus レベルでも既知、未知に関わらず比較的高い精度で分類が達成できた。この理由として関連研究で用いたゲノム全体のオリゴヌクレオチドの頻度よりも、本手法で用いたダイコドン使用頻度がより種特異的な偏りがあるために、それが Genus レベルでも良い分類を可能にした事が考えられる。

いずれにせよ未知の生物種に関しては未だ良い分類ができていないことになる。よって、メタゲノム断片から未知の生物種の系統分類は今後の大きなテーマであると言えよう。

３．５　まとめ

本章ではメタゲノムデータから出てくる様々な生物種が混在した断片配列の系統分類を行うために、ゲノムの塩基組成に注目した。ゲノム全体の塩基組成には生物種固有の偏りがあり、またそれは遺伝子領域において顕著に見られる。よって本章では２章で構築した遺伝子予測手法を用いて与えられた断片配列の遺伝子領域を特定し、その領域内のダイコドン使用頻度から系統分類を行うアプローチをとった。

結果として既知の生物種に関しては 1000 bp の断片ゲノムに対して Genus 以上の系統グループで感度、特異性共に約 90%以上での高い分類精度を達成した。一方で未知の生物種に関しては Genus と Domain の系統グループでは一定の分類精度を達成できたが、それ以外の系統グループではあまり良い精度は得られなかった。よってこれらの分類に関しては今回用いたダイコドン使用頻度だけでは限界があり、今後新たな指標を取り入れた分類方法が必要である。

本手法は未知の生物種に対してはあまり良い分類精度は得られなかったが、実際のメタゲノムデータの中には既知と未知の両方の生物種が存在している。よって既知の生物種に対する高い分類精度を考えた場合、本手法は実際のメタゲノムデータに対しても十分に実用性はあると考えられる。

４．結論

近年行われ始めたメタゲノム解析からは多くの生物種が混在した断片ゲノムが生成された。このようなゲノムデータに対して有用な遺伝子予測手法は確立されておらず、また生物種を推定する手法も確立はされていない。そういった理由から本研究はこれら２つの問題に取り組んだ。

メタゲノム配列からの遺伝子予測は与えられる配列の生物種が未知であるという問題から、既存の遺伝子予測手法は適用ができない。そこで本研究ではコドン使用頻度と GC%の相関を利用して、与えられた配列の GC%からダイコドン使用頻度を推定するアプローチをとった。さらにこのときドメイン間のコドン使用頻度の違いに注目して Bacteria、Archaea、Fungi の 3 つのコドンモデルを構築した。そして予測精度のさらなる向上のため ORF 長の分布、最も上流にあるスタートコドンからの距離の分布、隣接する ORF 間の向きを考慮した距離の分布を指標として取り入れた。これらの指標は原核生物と Fungi の２つのモデルを構築した。そしてこれらのモデルを組み合わせて全体の予測モデルを構築した。

結果として完全長ゲノムに対しては GC%からの近似によるコドン使用頻度の推定にも関わらず高い予測精度を達成でき、その精度は既存のツールと比較してもほとんど差は無かった。既存のツールは生物種個々の遺伝子を学習してパラメータを推定しているので、与えられる配列の生物種情報が既知である必要があり、生物種ごとに異なる予測モデルを用意する必要がある。一方で本手法は与えられた配列の GC%のみからパラメータの推定を行っているので、与えられる配列の生物種に関係なく１つのモデルで予測を行える。そのような利点を考えると本手

法は完全長ゲノムに対しても十分な実用性があると考えられる。

断片ゲノムに対しても完全長ゲノムと変わらない精度での予測制度を達成できた。原核生物（Bacteria、Archaea）では感度と特異性それぞれ 95%と 90%、そして Fungi では 93%と 76%を達成できた。真菌では微量のイントロンの存在と低い遺伝子密度のために、擬陽性が多く発生して特異性の低下が見られた。この特異性の改善は今後の課題である。

また本手法は遺伝子予測を行う過程で、３つのドメイン間のコドン使用頻度の違いを用いたドメイン分類を提案した。この手法により 700 bp の断片ゲノムに対して原核生物の間では平均約 90%の分類精度を達成でき、また Fungi の断片ゲノムに対しても約 80%の精度での分類を達成できた。また正しく分類できなかった生物種の中には外来性遺伝子を含むものもいくつか存在していた。よって今回不正解となった断片でも外来性遺伝子を含むものとして別のドメインに分類されたとすれば、それは正解とも考えられる。

様々な生物種が混在したメタゲノムデータ中の断片ゲノムを正しい系統に分類することは重要なテーマである。本研究では断片ゲノム上にある遺伝子領域のダイコドン使用頻度に注目して系統分類を行うアプローチをとった。今回は原核生物 228 種の遺伝子と 1000 bp の両方をテストデータとして分類の評価を行った。また評価の際には既知生物種と未知生物種の両方のケースを想定して分類性能を評価した。

結果として既知生物種の遺伝子と 1000 bp の断片ゲノムに対しては Genus 以上の系統グループで F 値 90%以上での高い分類精度を達成できた。これはコドン使用頻度が種特異的である事は以前から知られていたが、それが遺伝子や 1000

bp の断片を分類できる程の強い偏りを持つという事は今回始めて確認された。よってこれから新たな生物種のゲノムが読まれていけば、この手法はメタゲノムデータに対してより強力な分類性能を達成できるだろう。

未知遺伝子に対する系統分類は既知遺伝子と比べると良い分類精度を達成できなかった。これはコドン使用頻度が種特異的ではあるが、Order や Class などの高次の系統グループでは使用頻度のばらつきが大きかったためである。それでも Genus と Domain のグループでは一定の精度を達成できた。これは Genus 内のメンバーは GC%に幅が無いためにお互いのコドン使用頻度が類似していて、未知の生物種に対しても正しい分類が行われたためである。また Domain に関しては Bacteria と Archaea の間で有意なコドン使用頻度の差が見られるので、未知の生物種に対してもその差は認識するには十分だったためである。

このように未知の生物種に対する分類では今回あまり良い精度は達成できず、今後の課題となった。しかし実際のメタゲノムデータには既知の生物種も多く存在している。よって本手法の既知と未知の両方の生物種に対する平均の分類精度を考えた場合、実際のデータに対しても適用は十分に可能である。

５．参考文献

[1] Tyson,G.W., Chapman,J., Hugenholtz,P. et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428, 37-43.

[2] Venter,J.C., Remington,K., Heidelberg,J.F. et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304, 66-74.

[3] Tringle,S.G., Mering,C.V., Kobayashi,A. et al. (2005) Comparative metagenomics of microbial communities. *Science*, 308, 554-557.

[4] Hugenholtz,P. (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biol.*, 3, reviews0003.1-0003.8.

[5] Rappe,M. and Giovannoni,S. (2003) The uncultured microbial majority. *Annu. Rev. Microbiol.*, 57, 369-394.

[6] Chen,K and Pachter,L. (2005) Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput. Biol.*, 1, 106-112.

[7] Fickett,J.W. (1981) Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res.*, 10, 5303-5318.

[8] Staden,R. (1984) Measurements of the effects of that coding for a protein has on a DNA sequence and their use for finding genes. *Nucleic Acids Res.*, 12, 551-567.

[9] Borodovsky,M.Y and McIninch,J.D. (1993) GeneMark: parallel gene recognition for both DNA strands. *Comput. Chem.*, 17, 123-153.

[10] Lukashin,A.V. and Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.*, 26, 1107-1115.

[11] Salzberg,S.L., Delcher,A.L., Kasif,S. at al. (1998) Microbial gene identification using interpolated Markov model. *Nucleic Acids Res.*, 26, 544-548.

[12] Delcher,A.L., Harmon,D., Kasif,S. et al. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, 27, 4636-4641.

[13] Frishman,D., Mironov,A., Mewes,H.-W. et al. (1998) Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res.*, 26, 2941-2947.

[14] Yada,T., Nakao,M., Totoki,Y. et al. (2001) A novel bacterial gene-finding system with improved accuracy in locating start codons. *DNA Res.*, 8, 97-106.

[15] Hayes,W.S and Borodovsky,M. (1998) How to interpret an anonymous bacterial genome: machine learning approach to gene identification. *Genome Res.*, 8, 1154-1171.

[16] Audic,S. and Claverie,J.M. (1998) Self-identification of protein-coding regions in microbial genomes. *Proc. Natl Acad. Sci. USA*, 95, 10026-10031.

[17] Chen,L.L. and Zhang,C.T. (2003) Svevn GC-rich microbial genomes adopt similar codon usage patterns regardless of their phylogenetic lineages. *Biochem. and Biophys. Res. Comm.*, 306,310-317.

[18] Besemer,J. and Borodovsky,M. (1999) Heuristic approach to deriving models for gene finding. *Nucleic Acids Res.*, 27, 3911-3920.

[19] Cambell,A., Mrazek,J. and Karlin,S. (1999) Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl. Acad. Sci.*, 96, 9184-9189.

[20] http://www.ncbi.nlm.nih.gov/Ftp/

[21] http://www.venterinstitute.org/sargasso/

[22] Zhu,H.Q, Hu,G.Q., Ouyang,Z.Q. et al. (2004) Accuracy improvement for identifying translation initiation sites in microbial genomes. *Bioinformatics*, 20, 3308-3317.

[23] Nielsen,P. and Krough,A. (2005) Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics*, 21, 4322-4329.

[24] Eisen,J.A., Nelson,K.E., Paulsen,I.T. et al. (2002) The complete genome sequence of *Chlorobium tepdium* TLS, a photosynthetic, anaerobic, green-sulfur bacterium. *Proc. Natl Acad. Sci*. USA, 99, 9509-9514.

[25] Bult,C.J., White,O., Olsen,G.J. et al. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, 273, 1058-1073.

[26] Sridhar,S.H., William,S.H., Artemis,G.H. et al. (1999) Bacterial start site prediction. *Nucleic Acids Res.*, 27, 3577-3582.

[27] Woese,C.R., and Fox,G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci. USA.*, 74, 5088-5090.

[28] Woese,C.R. (1987) Bacterial evolution. *Microbial. Rev.*, 51, 221-271.

[29] Karlin,S. and Burge,C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet.*, 11, 283-290.

[30] Pride,D.T., Meinersmann,R.J., Wassenaar,T.M. (2003) Evolutionaly implications of microbial genome tetranucleotide frequency biases. *Genome Res.*, 13, 145-158.

[31] Abe,T., Sugawara, H., Kinouchi, M. et al. (2005) Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. *DNA Res.*, 12, 281-290.

[32] McHardy,A.C., Martin,H.G., Tsirigos,A. et al. (2007) Accurate phylogenetic

classification of variable-length DNA fragments. *Nat. Methods*, 4, 63-72.

[33] Aristotellis,T., and Isidore,R. (2005) A Sensitive, Support-Vector-Machine Method for the Detection of Horizontal Gene Transfers in Viral, Archaeal and Bacterial Genomes. *Nucleic Acids Res.*, 33, 3699-3707.

[34] http://cbcsrv.watson.ibm.com/HGT_SVM/

[35] http://www.ncbi.nlm.nih.gov/Taxonomy/

[36] Lynn,D.J., Singer,G.A. and Hickey,D.A. Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res.*, 30, 4272-4277.

謝辞

本研究を進めるに当たりまして、的確なご指導と恵まれた研究環境を提供して下さった高木利久教授に心より感謝いたします。そして本研究を完成させるまでに終始暖かいご指導を頂いた野口英樹先生に心より感謝いたします。また文章校正や口頭発表の指導においてお世話になりました道菅伸介さんに心より感謝いたします。またその他にも研究生活をしていく中でいろいろとご協力をいただいた、小池麻子助教授、Steven Kraines 助教授、星山大輔先生、牧野貴樹先生、水谷治央先生、小野尚孝さん、山本泰智さん、江島恭子さん、斉藤美紀さん、深川浩志さん、石井奈都さん、岩崎渉君、巻島健志君、山口大輔さん、赤田庸平君、酒田理人君、富岡直子さん、武藤祐子さん、岡田信之さんに対しても心よりお礼を申し上げます。

皆様の暖かいご協力がなければ本研究を進めることはできませんでした。
この場を借りて改めて感謝の意を表します。