社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－MPS－67 (45)
2007－BIO－11 (45)
2007／12／21

# 統計的予測法を用いた原核生物ゲノムにおける遺伝子発見法の開発

柴田有輝、花田篤志、横瀬拓也、高橋紘一
近畿大学大学院 総合理工学研究科

近年、様々なゲノムの解析が進む中、情報論的な技術を用いる遺伝子発見法の開発が重要となってきた。そこで我々は、マルコフモデルを用いて原核生物ゲノム中の塩基配列に潜む統計的な偏りを自動的に検出し、遺伝子領域をORFとして探索することで、約60%の遺伝子同定に成功した。さらに隠れマルコフモデルを用いて、開始コドンの上流にあるShine-Dalgarno (SD) 配列に基づいてORFを整形し、より正確な遺伝子領域の同定を試みた。その結果、開始コドン上流にSD配列の存在する生物では、遺伝子同定率が5～20% 向上した。

# Development of gene finding method for prokaryote genomes by statistical prediction approach

Y.Shibata, A.Hanada, T.Yokose, K.Takahashi
Interdisciplinary Graduate School of Science and Engineering, Kinki University

Recently, the automatic gene finding has become more important. We found prokaryote gene coding regions using Markov model which learned the statistical biases of base sequences automatically. About 60% genes are found by reforming coding regions as ORFs. Furthermore, we rearranged the position of start codons based on Shine-Dalgarno (SD) sequence by hidden Markov model. We could find additional 5~20% genes for prokaryote genomes which have SD sequences.

## 1. 序論

近年、ゲノムの解読が急速に進んでいる。ゲノムは4種の塩基 (A, T, C, G) の配列であり、その中に遺伝子領域が存在する。遺伝子がどこにあり、どのような順番で読み出され、それらが細胞の中でどのように働くのかを明らかにすることは、生物の仕組みを知るうえで非常に重要である。

遺伝子探索の方法は、大きく分けて2種類ある。第1の方法は、マルチプルアラインメント法である。第2の方法はマルコフモデル(MM)法で、塩基配列の特徴に注目する統計的な探索法である。従来のマルコフモデルは、既知の遺伝子情報を用いて類似の生物の遺伝子を探索している。しかし、この方法では新種の生物の遺伝子を探索できない。そこで、マルコフモデルにゲノム自身が持っている統計的特徴を学習させ、自動的に遺伝子を探索する方法が考えられる。これにより、未知の生物の遺伝子を探索することができる。

1998年、AudicとClaverieは、類似の生物の遺伝子情報という知識を用いずに、進化の過程で形成された遺伝子の特徴に基づく塩基の統計的偏りを自動的に学習するマルコフモデル

の可能性を示唆した[1]。我々は、この方法に基づく原核生物の遺伝子探索方法を開発した。

## 2. 遺伝子

### 2-1. 遺伝子領域

ゲノムには、遺伝子としてアミノ酸を指定するコード領域 (coding region) と、それ以外の非コード領域 (non-coding region) がある。さらにコード領域には、上流の 5' 端から下流の 3' 端方向に書かれている順方向遺伝子領域 (direct coding region) と、逆に 3' 端から 5' 端方向に書かれている逆方向遺伝子領域 (reverse coding region または shadow region) がある。逆方向遺伝子領域では、2 本鎖で構成されるゲノムの相補鎖上の遺伝子が翻訳される。

### 2-2. コドンバイアス

遺伝子領域では、3 塩基からなる $4^3 = 64$ 種のコドンが使われている。そのうち 3 種 (TAA、TGA、TAG) は終止コドンで、残りの 61 種のコドンは 20 種のアミノ酸をコーディングしている。多くの種のゲノム配列には統計的偏り（コドンバイアス）があり、塩基はランダムに出現しているように見えても、実際には G の次に A が現れることが多いなどの傾向を見出すことができる。特に、遺伝子領域や Shine-Dalgarno 配列などのシグナル配列においては、コドンや塩基の出現の偏りが著しい。このような偏りを表現できる方法の一つが、次節で説明するマルコフモデルである。

## 3. 統計的予測法

### 3-1. マルコフモデル

マルコフモデルは図 1 のように、状態を丸印で、状態遷移を矢印で表され、各状態には塩基が一意に割り当てられている。このモデルでは、

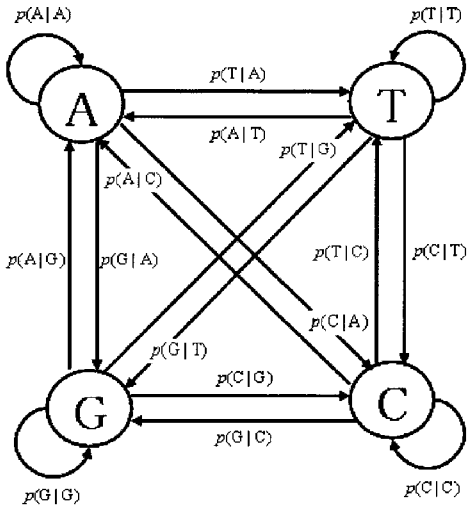


図 1. 1 次マルコフモデル

ある状態に至ったときに、その状態に対応する塩基が出力される。したがって、ある塩基配列の出現確率は、その状態に至るまでの遷移確率の積で求めることができる。例えば、塩基配列 AGGT が出力される確率は、P(AGGT) = P(T|G)P(G|G)P(G|A)P(A)となる。

図 1 のマルコフモデルでは、4 つの各塩基が出力される確率が直前の塩基にのみ依存して出力されるモデルであり、1 次マルコフモデルと呼ばれる。実際のゲノム解析に用いられるマルコフモデルでは、数個前からの塩基の並びに依存する 2〜5 次マルコフモデルが多用されている。本研究では 5 次と 2 次のマルコフモデルを使用した。

### 3-2. 6 クラス統計的予測法

1 つのアミノ酸を特定するコドンは数種類存在し（同義コドン）、多くの場合コドンの第 1、第 2 塩基は共通で、第 3 塩基のみが異なる。同じアミノ酸を指定するコドンの使用頻度は、各生物種によって偏りが見られるため、遺伝子

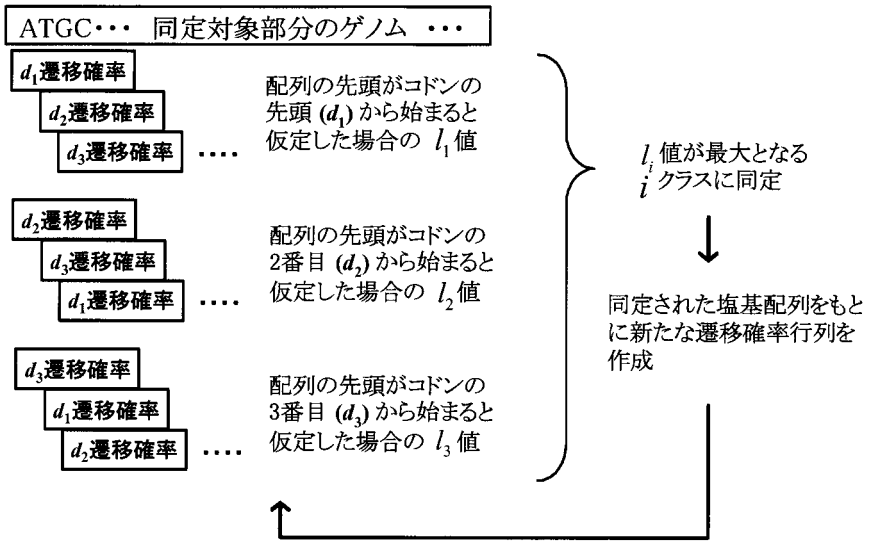


図 2. 6 クラス統計的予測法

逆方向遺伝子領域でも同様の手続きを行い、合計 6 つのクラスに分ける。

発見の精度を上げるために、そのような特徴を捉えやすいマルコフモデルを構築する必要がある。また、塩基配列上のコドンは、1 塩基ずれるだけで別のコドンと解釈されてしまうため、塩基配列から正しいコドンを見つけるには、3 通りの読み枠を考慮しなければならない。

そこで、これらの条件を満たすマルコフモデルとして、6 クラス統計的予測法を用いる必要がある [2] 。ただし、この 6 クラス統計的予測法は、後述するように、順方向遺伝子領域と逆方向遺伝子領域ごとに 3 通りの読み枠を考えるため、合計 6 つの遷移確率行列が必要になるが、まったくの予備知識無しに 6 つの遷移確率行列を作成するのは困難であり、誤った推定を招きかねない。

そこで、前段階の処理として、まず 3 クラス統計的予測法を用いて、後に述べる 6 クラス統計的予測法と同様の方法で、ゲノムから順方向遺伝子領域、逆方向遺伝子領域、非遺伝子領域のそれぞれの領域の大まかな特徴を抽出した。原核生物は遺伝子領域に比べ、非遺伝子領域が著しく少ないという特徴を持つことから、3 つの領域のうち最も塩基数が少ないものを非遺伝子領域と同定した。

### 3-3. 読み枠の学習

3 クラス統計的予測法で分けられた 3 つの領域から、遺伝子領域のみを 3 つの読み枠を考慮して、合計 6 つのクラスの遷移確率行列を作成した。これを初期値として、図 2 のように 6 クラス統計的予測法を行った。初期の時点で作成された遷移確率行列は、コドンの読み枠を考慮しない 3 クラス統計的予測法によって求められたコード領域を用いて作成されているため、3 種の読み枠をほぼ均等に含んだものとなる。従って、これら 6 つのクラスはほぼクラス間で特長のない遷移確率行列を持つことになる。

続いて、上記のように作成した遷移確率行列を用いて、遺伝子領域をゲノムの先頭から走査し直し、確率が最も高くなるクラスに振り分ける。振り分けられた配列を元に、それぞれのク

ラスで新たな遷移確率行列を作成し直す。この手順を繰り返し行うことで、次第にそれぞれのクラスが読み枠ごとの特徴を持った遷移確率行列に変化していく。これを、遷移確率行列が変化しなくなるまで繰り返した。

実験の結果、大腸菌 K12 では、全遺伝子領域中の塩基数のうち、98.9% (同定できた塩基数 対 正確な遺伝子中の塩基数) の塩基を発見することができた。

なお、本研究の対象である原核生物のゲノムデータは塩基配列データベース Gene Bank (NCBI) [3] より取得した。

## 4. Open Reading Frame (ORF)

開始コドンで始まり終止コドンで終わる塩基配列を Open Reading Frame (ORF) と呼ぶ。マルコフモデルによって求められたコード領域は統計的に求められたものなので、正確に読み枠を特定するまでには至らない。また、統計的予測法では、非遺伝子領域以外の 2 つのクラスが順方向遺伝子領域、逆方向遺伝子領域のどちらに属するかを判断できない。

そこで、まず始めに、遺伝子領域とみなされた 2 つのクラスのうち、どちらが順方向遺伝子領域でどちらが逆方向遺伝子領域なのかを決定しなければならない。そのために、それぞれのクラスに属する配列のうち始めの数百塩基を調べ、開始コドンとなりうる ATG が出現している割合を比べ、高いほうが順方向遺伝子領域と同定した。

次に、統計的予測法によって求められたコード領域を、読み枠を考慮して終止コドン候補 (TAA, TAG, TGA) から開始コドン候補 (ATG, GTG, TTG) までの塩基数がある一定数以上 (本研究では 210) なら、遺伝子領域とみなすことにより、ORF を整形した。1 つの終止コドン候補に対して複数の開始コドン候補が考えられる場合は、そのうち最も長いものを ORF とした。

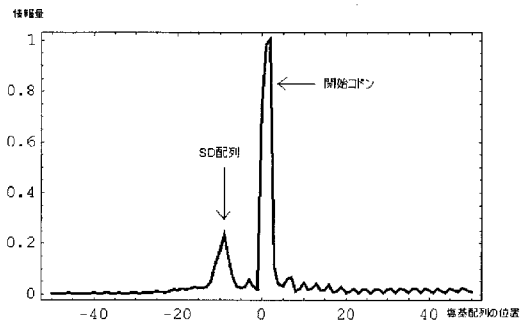


図 3. 大腸菌 K12 の開始コドン周辺の情報量

## 5. Shine-Dalgarno (SD) 配列を利用した開始コドンの発見

6 クラス統計的予測法によって求められた大腸菌 K12 の ORF と、正確な遺伝子との一致の割合は 61.3 %である。しかし、終止コドンのみ一致している割合は、93.0%と非常に高い。これは、開始コドンを決める際、単純に最長の開始コドン候補を採用しているためである。したがって、6 クラス統計的予測法によって求められた ORF の開始コドンを正確に縮小することができれば、遺伝子との一致の割合が最高で 93.0 %にまで達する可能性がある。

正確な開始コドンを発見するために、遺伝子の開始コドン上流約 -5 ~ -15 の領域にある翻訳開始のシグナル配列である Shine-Dalgarno (SD) 配列を利用する方法が考えられる。図 3 に大腸菌 K12 の開始コドン周辺の平均情報量を示す。SD 配列は確かに高い情報量を持つことが分かる。そこで、次節に述べる隠れマルコフモデルを用いて SD 配列に相当するシグナル部分配列を抽出することで、より正確な開始コドンを見つけることができる。

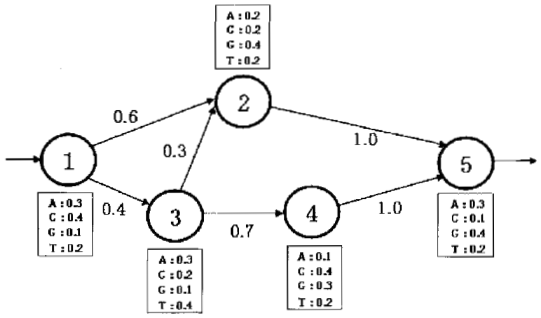


図4. 隠れマルコフモデル

## 6. 隠れマルコフモデル

### 6-1. 隠れマルコフモデル

隠れマルコフモデル (HMM) は塩基配列のパターンを探索することができるモデルである。通常のマルコフモデルでは、各状態は必ず1つの塩基を表すが、隠れマルコフモデルでは、図4に示すように、各状態において観測される塩基は出現確率により確率的に決められる。

隠れマルコフモデルにおいて、ある経路を通り塩基パターンが出力される確率は、通過した各状態における各塩基の出力確率と、経路上の遷移確率との積となる。例えば、図4で経路1 → 3 → 4 → 5を通り、T, A, C, Gという配列が出力される確率は、$0.2\times0.4\times0.3\times0.7\times0.4\times1.0\times0.4=0.0027$ となる。このような隠れマルコフモデルにおいて、塩基配列の出力される確率またはその対数が最大となる経路を求めるアルゴリズムを Viterbi アルゴリズムという。

### 6-2. 隠れマルコフモデルのSD配列同定問題への応用

SD配列は揺らぎを持っており、開始コドンの上流約 -15 ～ -5 の範囲に存在するとされている。またSD配列は、その部位が連続して存在しているわけではなく、不連続に存在している場合もある。よって、シグナル配列を表現する際は、揺らぎを考慮したモデルが必要となる。隠れマルコフモデルでは揺らぎを考慮したモデルの構築が可能であり、例えばシグナル配列がAGGAになる可能性は50 %, GAAGになる可能性は25 %のように各々のパターンに対して、確率を割り当てることができる。このように、様々な配列パターンを隠れマルコフモデルに学習させることができる。

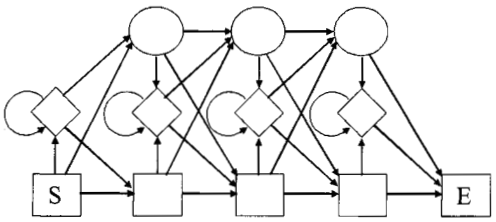


図5. プロファイル隠れマルコフモデル

隠れマルコフモデルを構築し、与えられた配列が出力される確率を求めるには、以下に示す手順が必要である。まず、隠れマルコフモデルの構造を決定する。シグナル配列の特性をうまく表現できるように、状態をどれだけ用意し、どの状態とどの状態を接続するかを決める。次に、与えられた既知のシグナル配列をもとに、モデル中の状態遷移確率および塩基出現確率を推定（学習）する。最後に、与えられた配列がモデルに基づいてどれくらいの確率で出力されるかを計算する。

ここで問題となるのは、最適な構造をどう構築するかである。モデルの推定ができない場合、すべての可能性に対応できるプロファイル隠れマルコフモデルがよく用いられる。

### 6-3. プロファイル隠れマルコフモデル

プロファイル隠れマルコフモデルをうまく構築すれば、高い確率でシグナル配列を出力するモデルを作ることができる。図5にその例を

示す。四角形の状態は一致状態、ひし形の状態は挿入状態、円形の状態は欠損状態を表す。

このモデルを用いて、用意された数多くの配列からシグナル配列を抽出する。まず、初期の塩基の出現確率や状態間の遷移確率をランダムに割り振る。次に、Viterbi アルゴリズムを用いて、それぞれのシグナル配列候補の最適な経路を決定し、それを元に塩基の出現確率と状態間の遷移確率を新たに作り直す。そして再度、Viterbi アルゴリズムを用いて最適な経路を決定する。この操作を繰り返すことで、プロファイル隠れマルコフモデルは、多数回観測されるシグナル配列が一致状態に来るように学習される。

SD 配列を出力しやすい最適なプロファイル隠れモデルが構築できた場合、一致状態で A や G などの SD 配列に多く見られる塩基(プリン塩基) 配列が出力されるようになる。

### 6-4. 実装

6 クラス統計的予測法によって求まった ORF の上流に存在する SD 配列に該当する部分配列を用いてプロファイル隠れマルコフモデルの学習を行い、SD 配列に特化したモデルを構築した。SD 配列の長さが平均 4.8 ということから、本研究では一致状態を 5 つとした。6 クラス統計的予測法によって求まった ORF は約 60 %の遺伝子を同定しているので、プロファイル隠れマルコフモデルに SD配列の特徴をある程度学習させることができると考えられる。学習の後、それぞれの ORF に対応する SD 配列の情報量を計算した。

次に、ORF の開始コドンを順方向遺伝子領域なら下流へ、逆方向遺伝子領域なら上流へ 3 塩基分ずつ縮小する。その過程で開始コドン候補(ATG, TTG, GTG) がみつかる度に SD 配列の情報量を計算し、終止コドンに到達するまでに

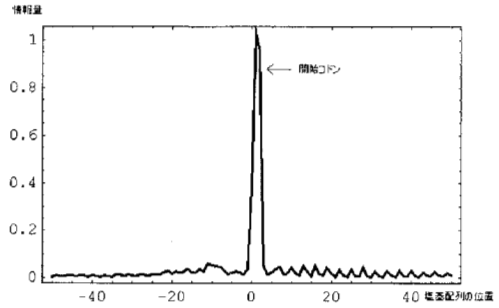


図 6. コレラ菌の開始コドン周辺の情報量

最も大きい情報量を出した開始コドン候補を新しい開始コドンとした。しかし、このようにして求めた新しい ORF の遺伝子同定率は、予想に反して縮小前と比べ下がっていた。これは、塩基が A,T,C,G の 4 種類しかないことから、偶然に SD 配列と類似した部分が現れ、思わぬ部分でSD配列部分の情報量が高くなっていることが原因と考えられる。そこで、6 クラス統計的予測法での結果を最大限に尊重し、そこから離れれば離れるほどスコアを低くするような重み付けを行った。これにより、各生物の遺伝子同定率が大きく向上した。

### 6-5. 繰り返しによる精度向上

新しく求まった ORF は、縮小を行う前よりもより正確な ORF と考えられる。そこで、新しい ORF の SD 配列に該当する部分配列を用いて再度プロファイル隠れマルコフモデルを構築することで、より正確な SD 配列の特徴を持った塩基の出現確率行列と、状態間の遷移確率行列が得られる。次いで再び、新しく構築されたプロファイル隠れマルコフモデルで開始コドン候補を縮小する。これを、出現確率、遷移確率が一定になるまで繰り返す。この方法を用いることにより、ほとんどの原核生物でより多くの遺伝子を同定することができた。

### 6-6. 問題点

図 6 に示すコレラ菌やラン藻のように、原核生物には SD 配列が存在しないものもある。この場合、SD 配列を元に開始コドンを縮小するとより悪い結果となった。そこで、SD 配列はプリン塩基 (A,G) に富んでいるという情報を用い、隠れマルコフモデルの構築時に、一致状態としてプリン塩基以外の塩基に収束した場合には、SD 配列がない、もしくは ORF で求めた遺伝子領域の精度が悪すぎると判断して、SD 配列を利用した開始コドンの縮小を行わなかった。実際、これらの生物の遺伝子同定率を計算してみたが、やはり下がっていたので妥当な判断だったといえる。

また、SD 配列がないにもかかわらず、ORF が不完全なために一致状態がプリン塩基に収束してしまう生物もあった。今後は、これらの生物をどう見分け、さらに遺伝子同定率を上げていくかが重要となる。

## 7. 結論と考察

次ページの表 1 に、本研究で実験した原核生物の遺伝子同定率を示す。

原核生物のゲノムでは 6 クラス統計的予測法を用いることで 60%程度の正確な遺伝子を発見することができた。また、終止コドンのみが探索された ORF も含めると約 90%の遺伝子を自動的に発見することができた。

さらに、SD 配列が存在する原核生物では、隠れマルコフモデルを用いて ORF を整形することで新たに 5～20%の遺伝子を同定できた。しかし、コレラ菌などのように SD 配列が顕著に出現しない原核生物の場合、プロファイル隠れマルコフモデルに SD 配列の特徴を学習させることが困難となり、誤った ORF に整形してしまう。このような生物の遺伝子発見率を上げるには、他の方法を考える必要がある。

また、遺伝子中にはオペロンと呼ばれる遺伝子間隔 (スペーサー) が短い、もしくは遺伝子同士がわずかに重なっている遺伝子群が存在する。オペロンは、オペロン群の先頭の遺伝子以外 SD 配列を持たないため、プロファイル隠れマルコフモデルを構築するときに SD 配列でない部分のノイズを与えることになる。また、ORF を整形する際に、オペロン中の遺伝子領域の上流には SD 配列が存在しないため、間違った ORF に整形されてしまう可能性が非常に高くなる。

以上のように、これらのオペロンの問題は、本研究の大きな問題点と考えられる。オペロンをどう見つけ、処理していくかが今後の課題である。

## 8. 謝辞

本論文を推敲するにあたり、多大な助言をいただきました北湯口佳恭さん、森口理絵さんに深く感謝の意を表します。

## 9. 参考文献

[1] S.Audic, J.M.Claverie, Proc.Natl. Acad. Sci. USA, 95, 10026 - 10031 ( 1998 )

[2] M.Borodovsky, J.McIninch, Comput. Chem., 17, 123 - 133 ( 1993 )

[3] NCBI, http://www.ncbi.nlm.nih.gov/

[4] 富田勝、斎藤輪太郎、 “バイオインフォマティクスの基礎 ゲノム解析プログラミングを中心に”、サイエンス社

[5] 中原大悟、”マルコフモデルによる原核生物遺伝子発見法の開発” 平成 18 年度修士論文 ( 2007 )

[6] 中島啓之、”マルコフモデルを用いた原核生物遺伝子発見” 平成 17 年度修士論文 ( 2006 )

表 1.　各原核生物の遺伝子同定率

各一致率は、同定できた遺伝子数 対 正確な遺伝子数 の比で求めた。

HMM での斜線部分は一致状態がプリン塩基以外に収束したため打ち切ったことを示す。

| 生物名 | ドメイン | 門 | 正確な遺伝子数 | ORF 配列数 | ORF 完全一致率 | HMM 完全一致率 | 終止コドン一致率 |
|---|---|---|---|---|---|---|---|
| Bifidobacterium_longum<br>乳酸菌 | 真正細菌 | アクチノバクテリア門 | 1727 | 1740 | 52.8% | 65.0% | 89.2% |
| Mycobacterium_bovis<br>牛型結核菌 | 真正細菌 | アクチノバクテリア門 | 3920 | 3698 | 48.3% | | 90.1% |
| Synechocystis<br>ラン藻 | 真正細菌 | シアノバクテリア門 | 3171 | 3090 | 71.6% | | 94.7% |
| Bacillus_Subtilis<br>枯草菌 | 真正細菌 | ファーミキューテス門 | 4105 | 3896 | 55.8% | 79.9% | 92.7% |
| Brucella_abortus<br>牛流産菌 | 真正細菌 | プロテオバクテリア門 | 2030 | 1897 | 45.7% | 61.3% | 82.2% |
| Escherichia_coli<br>大腸菌 | 真正細菌 | プロテオバクテリア門 | 4133 | 4054 | 61.3% | 81.1% | 93.0% |
| Haemophilus_influenzae<br>インフルエンザ菌 | 真正細菌 | プロテオバクテリア門 | 1657 | 1687 | 75.7% | 81.2% | 96.2% |
| Helicobacter_pylori<br>ピロリ菌 | 真正細菌 | プロテオバクテリア門 | 1576 | 1503 | 74.9% | 71.9% | 93.0% |
| Vibrio_cholerae<br>コレラ菌 | 真正細菌 | プロテオバクテリア門 | 2742 | 2515 | 82.6% | 61.9% | 88.9% |