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
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
氏名（学籍番号）　朴　重鎬（５６９０９）
指導教員名　　高木　利久 教授

キーワード：メタゲノム解析、遺伝子予測、系統分類、ダイコドン使用頻度

１．背景と目的

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必要なパラメータを個々の生物種の遺伝子情報を元に学習しているが、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個々の配列の生物種情報を得る事ができないためである。またそれ以外にも、未知のものも含む様々な生物種が混在したメタ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。今回遺伝子予測では原核生物と真菌（Fungi）を対象として扱い、系統分類では原核生物のみを対象として扱った。

２．メタゲノム配列からの遺伝子予測

２．１　手法

本研究は与えられた断片配列の GC%からダイコドン使用頻度を推定して遺伝子予測を行うアプローチをとった（図 1）。ダイコドンを用いたのは１つ前のコドンとの条件付確率を見る事により予測精度が向上する事が知られているためである。またドメイン(Bacteria、Archaea、Fungi)の間ではコドン使用頻度に有意な差がある。よって実際には３つのドメインを分けてそれぞれ回帰を行い、３種類のダイコドンモデルを構築した。さらに予測精度の向上のため、その他に３つの指標（ORF 長の分布、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、隣接する ORF 間の向きを考慮した距離の分布）を取り入れた。

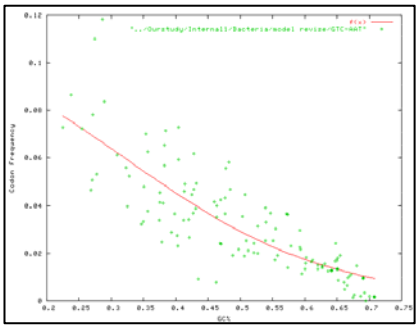

図 1 GC%によるダイコドン使用頻度の推定

予測の手順は次の通りである。1)与えられた断片配列中のすべての ORF 候補を抽出する。2)各 ORF 候補に対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真のスタートコドンまでの距離の分布によるスコア付けを行う。3)ORF 候補全体に対して、ORF 間の向きを考慮した距離の分布によるスコアと 2)で得られたスコアを用いてゲノ

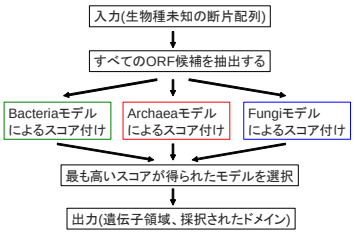

図 2 予測の手順

ム上での最適な ORF の組み合わせを計算する。2)と 3)の手順を Bacteria、Archaea、Fungi の３つのモデルで行い、最もスコアの高かったモデルを正解として、その予測遺伝子とドメインを出力する（図 2)。

２．２．結果と考察

Bacteria9 種、Archaea3 種、Fungi7 種の完全長ゲノムからランダムに抽出した 700 bp の断片配列をテストデータとして、本手法の性能評価を行った。予測精度は遺伝子予測に関しては感度（Sn）と特異度(Sp)で、またドメイン分類に関しては正しいドメインに分類された断片ゲノムの割合でそれぞれ評価を行った。表 1 にテストデータについてのドメインごとの予測精度の平均を示す。原核生物においては Sn、Sp 共に高い予測精度を達成でき、また Fungi では Sn に関して高い精度を達成できた。この結果から本手法は短い断片に対しても十分に予測を行える事がわかる。一方で Fungi において Sp の低下が見られたが、これは微量のイントロンの存在と遺伝子密度が低いために、擬陽性が増えたことが原因に挙げられる。これに関しては断片配列からイントロンを認識できるモデルを構築するなど今後の改善策が必要である。またドメイン分類に関しては平均 80%と遺伝子予測の精度に比べると劣る結果となった。しかし分類に失敗したものの中には Bacteria ゲノム中の Archaea 由来の遺伝子を含む断片ゲノムが、Archaea のドメインに分類されるなどの結果が見られた。よってドメイン分類の精度の低下は、必ずしも分類に失敗している事を意味しているわけではないと言える。

表 1　700 bp の断片配列に対する予測結果

ドメイン	Sn	Sp	ドメイン分類の精度
Bacteria	94.7%	89.4%	85.3%
Archaea	96.7%	93.9%	81.5%
Fungi	93.4%	78.4%	78.7%

３．メタゲノム配列からの系統分類

３．１　手法

本研究は断片ゲノムの分類を行うにあたってゲノム全体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。よって本手法では生物種未知の断片配列から２で提案した手法を用いて遺伝子領域を特定し、その領域内のダイコドンの組成から系統分類を行うアプローチをとった。

具体的には現在ゲノムが読まれている Bacteria と Arhcaea の全 228 種からダイコドン使用頻度のモデルを構築し、与えられた断片ゲノムに対して 228 個のモデルでスコア評価し最もスコアの高いモデルを正解とした。そしてお互いの系統関係の一致度を Genus から Domain までの５段階で評価した。またメタゲノムデータ中に存在する難培養性の新規生物種を仮定して、与えられた配列に対して自分自身のモデルをマスクし 227 個のモデルから最適なモデルを選ぶ方法により未知の生物種に対する評価も行った。

３．２　結果と考察

1000 bp の断片ゲノムに対して、既知の生物種と未知の生物種の両方のケースでの分類精度を評価した(図 3)。評価には F-measure（感度と特異性の調和平均）を用いた。1000 bp 中には平均 1.7 個の遺伝子が存在し、また遺伝子が１つも無い断片の割合はわずか 2%である。よってほぼすべての断片に対して本手法は適用可能である。既知の生物種では高い精度での分類が可能であったが、これはコドン使用頻度が 1000 bp という短い断片に対しても十分に分類可能な指標である事を示している。一方未知の生物種に関しては Genus と Domain のレベルでは比較的良い精度が得られたが、他の系統では大きな分類精度の低下が見られた。これはコドン使用頻度が種レベルでは強い偏りを持つが、系統グループ内では共通の偏りがあまり見られないためである。本研究では未知の生物種に関しては課題を残したが実際のメタゲノムデータには既知と未知の生物種が混在しているので、両方の精度を考慮すれば本手法の実用性は十分にあると考えられる。

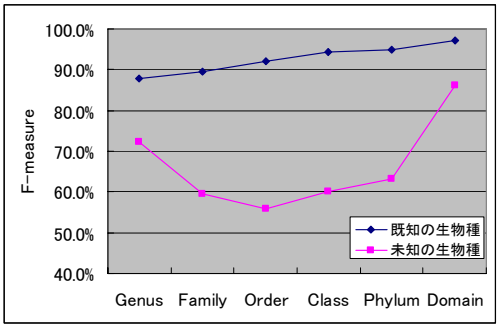

図 3 1000 bp の断片配列に対する系統分類