様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　６月　８日現在

機関番号：１２６０８
研究種目：若手研究(B)
研究期間：2010～2011
課題番号：２２７１０２０３
研究課題名（和文）　真菌を含む微生物メタゲノムからの遺伝子予測および種分類

研究課題名（英文）　Gene prediction and classification on the microbial metagenomes including fungal species

研究代表者
野口　英樹（NOGUCHI HIDEKI）
東京工業大学・大学院生命理工学研究科・特任准教授
研究者番号：50333349

研究成果の概要（和文）：本研究では、複数の生物種のゲノム断片が混じったメタゲノムデータから遺伝子領域を予測するための情報科学的手法の開発を行った。遺伝子領域では使用される塩基に偏りが見られるが、その偏りは生物種ごとに異なる。本研究では、ゲノム断片の GC 含量からその生物種における塩基出現頻度を推定することで、原核生物より複雑な遺伝子構造を持つ真核生物のメタゲノムデータからでも高い精度での遺伝子予測が可能であることを示した。

研究成果の概要（英文）：In this study, I've constructed a gene-finding tool for metagenomic sequence data comprising various eukaryotic genomes. Nucleotide compositions in the protein-coding regions are different among species, although they are key information on protein coding. I indicated in this study that species-specific nucleotide compositions could be estimated by GC contents of the genome fragments and are effective enough to predict genes with high accuracies even in the eukaryotic genomes.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	1,900,000	570,000	2,470,000
2011 年度	1,200,000	360,000	1,560,000
年度			
年度			
年度			
総　計	3,100,000	930,000	4,030,000

研究分野：複合新領域
科研費の分科・細目：ゲノム科学・応用ゲノム科学 B
キーワード：メタゲノム、生物情報科学、遺伝子予測

１．研究開始当初の背景

特定の自然環境中に棲む全ての生物種のゲノム（遺伝子）情報を調べる方法としてメタゲノム解析が注目されている。メタゲノム解析では、未知の生物種や単離・培養が困難な生物種を含む全生物種のゲノム配列デー

タを得ることができる。しかし一方で、得られるデータ中にはこれらの生物種のゲノムが未分類のまま混在している（つまり個々のゲノム配列の由来生物種が不明)。そのうえ、個々の配列は短い断片であることが多く、そこから有用な統計情報を得ることは困難なため、遺伝子領域同定などの配列決定後の解析を行うことも容易ではなかった。

これまでの研究で、原核生物については、ゲノムの GC 含量から推定した遺伝子統計量（コドン使用頻度や遺伝子長の分布など）を用いることで、未知生物種の遺伝子を高精度で予測可能であることを示した。しかし、原核生物と違って遺伝子密度が低く、より複雑な遺伝子構造を持つ真核生物において同様のアプローチで遺伝子領域を同定可能かどうかは分かっていなかった。

２．研究の目的

真核生物、特に真菌を対象に、これらの生物種が不明な場合でも、遺伝子予測に必要なパラメータを学習することなく高精度な予測が行える遺伝子予測手法の開発を目指す。本手法が実現されれば、真菌などの真核生物が重要な役割を担っている環境サンプルのメタゲノム解析が加速されるものと考えられる。

３．研究の方法

（１）遺伝子統計量のモデル化

全ゲノムが決定された真菌・原生生物のゲノム配列と遺伝子領域のアノテーションを用いて、コドン使用頻度（モノコドン・ダイコドン)、遺伝子・エキソン長の分布、イントロン長の分布、スプライス部位（ドナー・アクセプター）の配列パターンなどの統計量を算出して特徴を精査し、適宜ロジスティック回帰分析やマルコフモデルを用いてモデル化を行う。

（２）遺伝子予測ツールの構築

構築した各種統計量のモデルを統合して、最終的な遺伝子構造を予測するためのプログラムを開発する。ここでは膨大な量の組み合わせ問題を効率よく解く必要があるため、動的計画法を用いて最も尤もらしい遺伝子・エキソン候補の組み合わせを求める。

４．研究成果

（１）遺伝子統計量のモデル化

①コドン使用頻度とゲノム GC 含量の関係

公共データベースに登録された３１種の真菌ゲノム配列を用いて、ゲノム GC 含量とコドン使用頻度の関係を調べた（図１)。原核生物の場合と同様に両者の間には強い相関関係が見られ、シグモイド関数により近似することで、配列の GC 含量のみからコドン使用頻度を高精度に推定可能であることが示された。また、細菌−古細菌で GC 含量とコドン使用頻度の関係に有為な差が見られたように、真菌にも細菌・古細菌とは異なる傾向が見られたことから、コドン使用頻度（＝コドンの出現確率）に基づいて個々の遺伝子の尤度を計算し比較することで、その遺伝子が真菌・細菌・古細菌のいずれ由来のものかを判別することも可能であると考えられる。

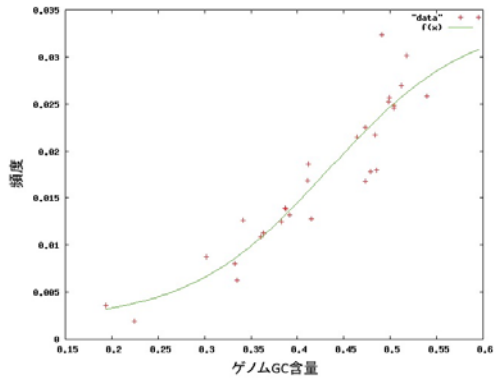

図１　ゲノム GC 含量とコドン CAG の使用頻度

②スプライス部位のコンセンサス配列

スプライス部位に関しては、調べた約 20 万のイントロン中 98.8%が GT-AG ルールに従っており、次に GC-AG が 0.8%、その他が 0.4%であった。高等真核生物でみられる AT-AC などのスプライスサイトはごくわずかしか存在しなかった。次に、GT-AG ルールに従うものについて、生物種ごと（100 以上のイントロンを持つもの）により詳細な配列パターンを調べ、クラスタリングを行った。アクセプター部位に関しては明確な差は見られなかったが、ドナー部位は大きく３つのグループに分類可能であることが分かった（図２）。

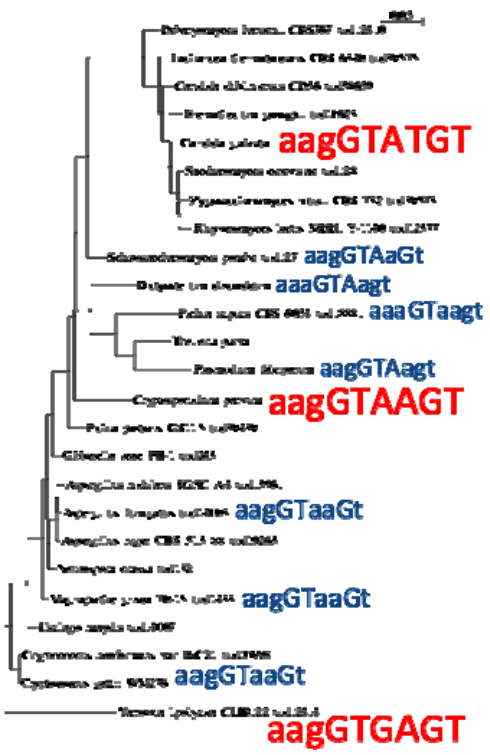

図２　ドナーサイトの配列パターン

Saccharomycetales に属する種の多く（一番上のクラスター）は、GTATGT のコンセンサス配列を持ち、*Yarrowia*（一番下）のみが異なるパターンを示した。これらの生物種の遺伝子はいずれもイントロンをほとんど持たない一方で、ドナー部位の保存度は非常に高かった。同様にイントロンをほとんど持たない原生生物である *Cryptosporidium parvum* は GTAAGT の強いコンセンサスを持っていた。それ以外のイントロンの多い生物種は、これらのコンセンサスを併せ持つ中間的なパターンを示した。イントロン数の大小は、ドナー部位の配列パターンだけでなく、遺伝子密度やエキソン長の分布といった他の統計量に与える影響も大きい。本研究では、３つの強いコンセンサスを持つ代わりにイントロンを持つことに対するペナルティーを大きくしたモデルと、弱いコンセンサスだがイントロンを入り易くしたモデルを用意し、いずれ

のモデルを用いるかで遺伝子密度、エキソン長のパラメータも変えて遺伝子スコアを計算したうえでスコアを比較して、より尤もらしいモデルを選択することとした。

（2）遺伝子予測ツールの構築

構築した、モノコドン・ダイコドン使用頻度のゲノム GC 含量による回帰モデルやドナー部位の重み行列、アクセプター部位の2次のマルコフモデル、またエキソン・イントロン長の頻度分布や遺伝子／非遺伝子領域における GC 含量の差といった統計量を統合して、ゲノム配列中の複数の遺伝子の構造を予測するプログラムを構築した。上記の各モデルは入力配列に応じて確率に基づいたスコア（対数オッズスコア）を出力する。そこで、本プログラムでは、それらのスコアの和が最大となる遺伝子・エキソン-イントロンの組み合わせを、動的計画法を用いて計算することで遺伝子予測を実現している。

構築した遺伝子予測ツールによる予測精度を表1に示す。ここでは、イントロンが多く遺伝子密度の低い生物種の例として *Aspergillus niger* の予測結果を、また逆にほとんどイントロンを持たない生物種の例として *Saccharomyces cerevisiae* の予測結果を示してある。予測精度はエキソン単位で、感度（既知の全エキソンの何パーセントが予測できたか）および特異性（全予測エキソンの何パーセントが正解か）を評価した。

表1　エキソンごとの感度と特異性

	感度	特異性
A. niger	81.4%	83.7%
S. cerevisiae	83.5%	90.0%

いずれの生物種でも 80%を超える高い精度で遺伝子領域を予測できていることが分かる。特に、イントロンの少ない *S. cerevisiae* では特異性も 90%に達しており、原核生物における遺伝子予測精度に迫る高い性能を達成できた。これらの予測精度は、入力ゲノム配列の生物種が不明、すなわち個々の生物種に特化したパラメータ学習を行っていない状態で達成したものであり、メタゲノム解析を行う上で十分実用的な高い数値であると言える。

一方でこれらの精度の高さは、イントロン（とそれに付随するパラメータ）のモデルの切り替えがうまくいってはじめて達成できるものでもある。現状では、長いゲノム配列が与えられたときにはこの切り替えはうまく働くが、配列が短くなるほど選択するモデルを間違えるケースも多くなっている。配列が短い場合にも正しいモデルを選択して予測するためには、あらかじめ別の統計量などを使って生物種を分類しておくなどの工夫が必要になってくるであろう。また、今回遺伝子統計量のモデル化に用いた生物種には、60%を超える高い GC 含量のゲノム配列が含まれていない（データベースに登録されていない）。そのため、今回のモデルを高 GC 含量のゲノムに適用した場合の予測精度は未知数である点にも注意が必要である。

このようにいくつかの問題点は残っているものの、未知の生物種に対してパラメータの学習というステップを踏まずに高い遺伝子予測精度を達成できた本研究の意義は大きい。今後は本手法を用いることで、原核生物だけでなくあらゆる生物種が関わる複雑な環境を対象にしたメタゲノム解析が進展できるものと期待できる。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計０件）

〔学会発表〕（計０件）

〔図書〕（計０件）

６．研究組織

(1)研究代表者

野口　英樹（NOGUCHI HIDEKI）

東京工業大学・大学院生命理工学研究科・特任准教授

研究者番号：50333349