様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月　 ６ 日現在

機関番号：３２６６５
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ１２０６８

研究課題名（和文）短鎖連続塩基配列に基づく細菌ゲノム・菌叢解析と機械学習による口臭予測への応用

研究課題名（英文）Bacterial Genome and Microbiota Analysis Based on Short Nucleotide Sequences and Its Application to Halitosis Prediction by Machine Learning

研究代表者
中野　善夫（NAKANO, Yoshio）

日本大学・歯学部・教授

研究者番号：８０２５３４５９

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：　5塩基連続配列の出現頻度に基づき系統樹解析を行えるプログラムを構築・公開することができた。この方法でIpomoea属植物の葉緑体ゲノムを用いて系統解析を行ったところ、得られた系統樹は既知の系統樹と一致したことから、細菌のみならず広い生物種を対象にした系統解析に貢献できることが明らかとなった。PLOS Oneに論文が掲載された。
　一方、これまでに得られた唾液サンプル中の口腔内細菌メタゲノム解析したところ、DNAの由来となる菌のほとんどが口腔内細菌としてよく知られている細菌由来であり、５塩基配列出現頻度に基づく系統樹解析により、塩基配列の相同性に依存しない方法で口腔内細菌叢の類似度を比較できた。

研究成果の学術的意義や社会的意義
　近縁種の系統解析はもっとも簡単で広く利用されている方法として16S rRNA遺伝子の塩基配列の相同性に基づく手法があるが、近縁種ではその配列の相同性が高すぎて系統解析ができない場合が少なくない。そのようなときに用いられる手法にmulti-locus sequence analysisがあるが、標的遺伝子を決めるのに多大な労力と時間を要求するものとなっている。今回報告した５塩基配列の出現頻度による系統解析は、そのような遺伝子の相同性に依存しない手法であり、遺伝子情報がまったくなくても何の準備もなく解析が実行できるものであり、それをアプリ化したことにより、広くその解析法を提供できる成果といえる。

研究成果の概要（英文）：We have developed and released a program for phylogenetic analysis based on the frequency of penta-nucleotide sequences. The results were published in PLOS One.
　On the other hand, metagenomic analysis of oral bacteria in saliva samples obtained so far showed that most of the bacteria from which DNA was derived were from well-known oral bacteria, and phylogenetic tree analysis based on the frequency of penta-nucleotide sequences allowed comparison of similarities in oral microbiota in a manner independent of nucleotide sequence homology. The phylogenetic tree analysis based on the frequency of penta-nucleotide sequences allowed us to compare the similarity of oral bacterial flora in a sequence-independent manner.

研究分野：分子生物学

キーワード：口腔細菌　系統樹　メタゲノム　細菌叢

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

1. 研究開始当初の背景

腸内細菌をはじめとするヒトと共棲する細菌叢とヒトの健康との関係が注目されている。ヒトの体に生態系を確立して生涯ともに生存する細菌叢は、その細胞数はヒト細胞の数をも上回り、常に宿主と相互作用を行い、両者が「超個体」を形成しているという考え方も提唱された。これまでの菌叢解析には 16S rRNA 配列が使われてきたが、いくつかの問題や不便があると指摘されている。(1) 一つの菌が複数の 16S rRNA 遺伝子を保有していて種によってその数が異なること、(2) PCR による増幅が必要であるために増幅効率に偏りが生ずること、(3) 極めて相同性の高い配列の増幅を繰り返すため PCR 増幅時にキメラ分子が生成すること、(4) 精度の高い分析には 500 塩基以上の正確な塩基配列が必要であること、(5) 数は少ないが 16S rRNA 遺伝子にも水平伝播の例が報告されていること等である。このようなときに用いられるのがmulti-locus sequence analysis/typing (MLSA/MLST) であるが、これは標的とする遺伝子を複数選び、その相同性の変動がどれくらいで、その組み合わせが系統解析にふさわしいかどうかをあらかじめ検証する必要がある。すでに手法が確立されている菌種間でない場合、かなりの手間と労力を要求されることになる。

生物種の塩基配列の並び方の組み合わせには、種固有の偏りがある。これまで、連続配列の出現頻度を利用した解析では、自己組織化マップと組み合わせたメタゲノム解析などが報告されてきたが、自己組織化マップ以外のパターン認識や機械学習等の手法と組み合わせた報告がほとんどない。本研究では、連続塩基配列の出現頻度とサポートベクターマシン（SVM）や深層学習を組合わせ、ゲノム配列中の特長ある領域の抽出、ゲノムやゲノム領域の比較、細菌叢の特徴の抽出と細菌叢同士の比較、さらにヒトの口腔内細菌叢や腸内細菌叢から宿主であるヒトの健康状態を判定する手法の開発を目指した。

2. 研究の目的

16S rRNA 配列に依存しない細菌叢構成比を解析する方法はこれまで、手間のかかるMLSA法しかなかった。そもそも塩基配列の相同性に依存しない系統解析が可能になれば、さらにそれが誰でも簡便に導入できるようになれば、原核生物のゲノム解析に新たな解析方法をもたらすと期待できる。また、 塩基配列解析時のエラーがランダムに発生するものであるならば、本方法はそのようなエラーの影響も少なく、nanopore のような携帯型シーケンサーで得られた結果でもその場ですぐ に解析に利用できる。さらにその解析方法をヒトの健康状態の予測にも活用できる。例えば口腔内細菌が原因である口臭そのものの測定は条件の再現性が難しいが、口腔内細菌叢の測定ならば、口臭の原因となる口腔内の状況を再現性よく把握できる。本研究ではモデルとして口臭の予測を行うが、同様の手法はさまざまな対象に応用できる。

本研究では、識別困難菌種の系統解析と種判別器の確立を第一の目的とした。細菌ゲノム中のN個連続塩基配列の出現頻度の分布を調べ種の系統解析を行う。大腸菌と赤痢菌のように 16S rRNA に基づく解析では系統を分離識別できないような組合わせの例を用いて、その有効性を検証した。他にYersinia属細菌やその近縁種、あるいは朝顔とその近縁種についても解析して細菌以外の解析対象に対する有用性を検証する。次に、この短鎖出現頻度に基づく菌叢解析を次の目標とした。上記の原理に基づき口腔内細菌叢のN個連続塩基配列解析を行い、各サンプルの細菌叢における出現頻度から、口臭の有無を深層学習あるいは SVM によって予測する。さらに、口腔内細菌叢に対して、上記のゲ

ノム解析の場合と同様に平均的なサンプルとは異なるものを OneClass-SVM で選び出し、口腔内の健康状態に何らかの異常のある菌叢を見つけ出し、それが例えば口臭の指標として有効かどうかを検証するというものである。

3. 研究の方法

まず、Yersinia属細菌ゲノムの塩基配列をNCBIのデータベースより得て３，４，５，６塩基配列の出現頻度を数え上げた。Yersinia属細菌は16S rRNA遺伝子の塩基配列で系統分析ができないほどその配列が近いことで知られている。この結果と比較するために従来の16S rRNA配列相同性に基づく系統樹と、比較対象を adk (adenylate kinase)、argA (amino acid acetyltransferase)、aroA (3-phosphoshikimate 1-carboxyvinyltransferase)、glnA (glutamine synthetase)、thrA (aspartokinase-homoserine dehydrogenase I)、tmk (thymidilate kinase) および、 trpE (anthranilate synthase component I) としたMLST法に基づく系統樹と比較した。さらに、大腸菌（Escherichia coki）と赤痢菌（Shigella）、その近縁種E. albertiiを含めて解析した。E. coliとShigellaについては、Rのe1071パッケージを使ったSVMで、 leave-one-out cross-validation method、すなわち１サンプルを抜き出し、それ以外のデータで学習してその１サンプルを判定することを全サンプルで繰り返す方法により、菌種を判定できるかどうかを検証した。

この５塩基配列出現頻度に基づく系統解析を簡便に実施するプログラムをR/Shinyのシステムで構築し、一般公開した。また、コマンドラインでの実施のためにPythonと組み合わせて使いやすいプログラムを作成した。

口臭患者由来の唾液サンプルよりDNAを抽出し、メタゲノム解析を行った。菌叢全体を相同性に依存しない方法、すなわち上記の５塩基配列出現頻度に基づく相互の距離によってグループ化して、口臭の有無との関連を考察する。サンプル採取方法、DNA抽出精製方法は従来実施してきた方法を踏襲している。

4. 研究成果

まず、短鎖塩基配列の出現頻度に基づく系統解析であるが、使用する連続塩基数は５が最適だと考え、以降の実験を５塩基連続配列の出現頻度で解析した。３〜４塩基では不十分、５，６塩基で良好な分離がみられた。５塩基で十分であれば、不必要に計算処理を増やすことはないので、５塩基で実施するのが現実的であった。これをプログラム言語RやPythonになじみのない研究者も活用できるようにR/Shinyシステムによってオンラインで公開した（ https://phy5.shinyapps.io/Phy5R/ ）。また、これをコマンドラインで利用できるようにするのが有用であるとの論文査読者の示唆により、Pythonの中にRスクリプトを組み込むことにより、研究者が自分のPC上で実行する解析に活用できるようにした。Gitを利用してこれも広く公開することにした（ https://github.com/YoshioNakano2021/phy5 ）。５塩基配列出現頻度集計表、系統距離表と系統樹ファイル（newick format）も生成するようになっている。このアプリケーションについては、PLOS ONE誌に論文が掲載された（https://doi.org/10.1371/journal. pone.0268847）。

110のYersinia属細菌に関して本方法を適応してみると、16S rRNA遺伝子配列の相同性による解析では分離できなかったものが、はっきりと異なる系統樹上のグループとして示された。MLST法でもそれに近い分離はできたが、解析の手間がまったく違う。E. coli、Shigella、E. albertiiに属する1186の細菌種に由来するゲノム配列を解析し、種による分離のみならず血清型によっても系統樹上でグル

ープを形成することが確認でき、細菌の系統解析に極めて有用であることが示された。

以上の結果に基づいて大腸菌／赤痢菌の判別をSVMによって行ったところ、一つのShigellaがE. coliと判定された例を除いて、すべて本来の種として判定できた。大腸菌と赤痢菌をこの方法で分類する場面はなかなかないかも知れないが、まだ特徴が完全に明らかになっていない菌種の場合には有用であろう。

この方法を、朝顔を含むIpomoea属の葉緑体ゲノムに対して活用してみたところ、これまでに報告されている形態に基づく系統樹と極めて類似した系統樹を描くことができ、まったくなじみのない分野の解析対象であっても、誰でもただちに活用できることが示唆された。

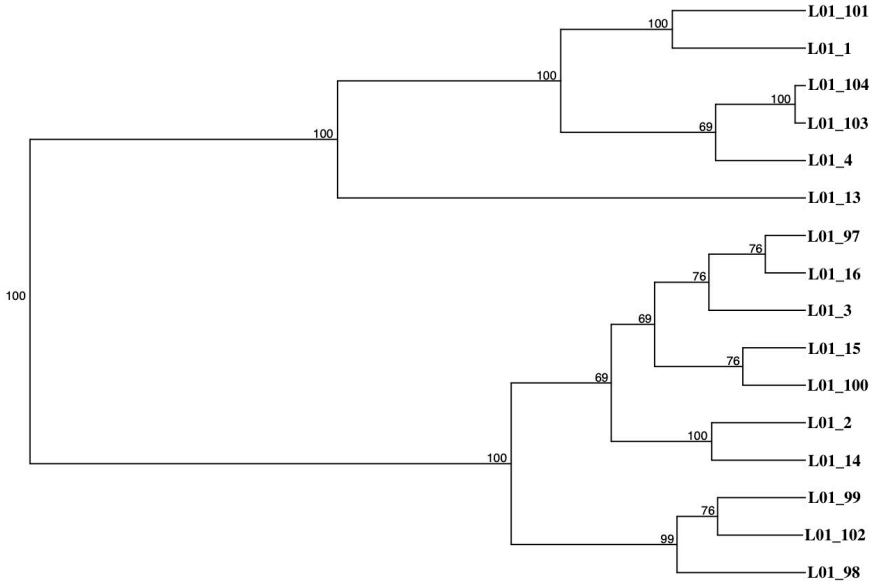

図１：メタゲノムサンプルの５塩基配列出現頻度に基づく系統樹

この第一の目的に関しては論文の発表まで到達したが、第二の口腔内細菌叢を５塩基配列出現頻度によって分類して口臭の予測を試みる方法の確立については、まずコロナ禍によって福岡歯科大学の口臭外来への来院者が激減するという状況に見舞われサンプル採取が予定通りできなかったことと、予算の３割減額が予想以上に厳しく、100サンプルほどは採取した唾液試料より口腔内細菌のゲノムDNA生成まではできたものの、このショットガンシーケンスまでは資金不足でできなかった。実施できた16サンプルで解析してみたところ、５塩基配列出現頻度でサンプルの類似性に基づく系統樹を描くことはできた（図１）が、それを使った口臭予測には統計的に意味のある数に及ばなかった。ヒトゲノム配列データベースを用いて口腔内細菌ゲノムDNAサンプル中にどれくらいヒト由来のDNA断片が混入しているかを調べたところ、25-80％のヒト由来塩基配列が見出された（平均49％）。まず、fastpでDNA断片の5’と3’のクオリティの低い領域のトリミングと残存しているかも知れないアダプターのトリミングと行なった後、KneadDataを使ってこの宿主ゲノム（ヒトゲノム）の除去を実施した。除去後の配列をsendsketchコマンドを使って確認してみると、そのほとんどすべてがNeissseria属細菌、Prevotella属細菌のように口腔内細菌としてよく知られている細菌由来であることが確認できた。ここで得られた配列を用いて、Megahitを使ったアセンブルを行なった。現在はこれをコンティグの分類、配列が由来するゲノムのアサイン等を行なっているところである。

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Suzuki Nao、Nakano Yoshio、Yoneda Masahiro、Hirofuji Takao、Hanioka Takashi	４．巻 8
２．論文標題 The effects of cigarette smoking on the salivary and tongue microbiome	５．発行年 2021年
３．雑誌名 Clinical and Experimental Dental Research	６．最初と最後の頁 449～456
掲載論文のDOI（デジタルオブジェクト識別子） 10.1002/cre2.489	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Watanabe Atsuko、Kawada-Matsuo Miki、Le Mi Nguyen-Tra、Hisatsune Junzo、Oogai Yuichi、Nakano Yoshio、Nakata Masanobu、Miyawaki Shouichi、Sugai Motoyuki、Komatsuzawa Hitoshi	４．巻 11
２．論文標題 Comprehensive analysis of bacteriocins in Streptococcus mutans	５．発行年 2021年
３．雑誌名 Scientific Reports	６．最初と最後の頁 1～13
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41598-021-92370-1	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Nakano Yoshio、Domon Yusaku、Yamagishi Kenji	４．巻 18
２．論文標題 Phylogenetic trees of closely related bacterial species and subspecies based on frequencies of short nucleotide sequences	５．発行年 2023年
３．雑誌名 PLOS ONE	６．最初と最後の頁 e0268847～
掲載論文のDOI（デジタルオブジェクト識別子） 10.1371/journal.pone.0268847	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計0件

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	谷口　奈央 (Taniguchi Nao) (60372885)	福岡歯科大学・口腔歯学部・教授 (37114)	

６．研究組織（つづき）

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	渡辺　孝康 (Watanabe Takayasu) (70725514)	日本大学・歯学部・講師 (32665)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関