Vol.2020-BIO-61 No.12
2020/3/13

Genome Identifier: 微生物ゲノム配列用系統解析ツール

堀池徳祐 [1]　下山祐紀 [2]

概要：昨今ではハイスループットシーケンサーや情報解析技術の急速な進歩と普及により，環境微生物のゲノムも配列決定されるようになり，それらの数は将来増加すると予想されている．環境微生物などの系統的な位置が不明な生物に対する系統解析はゲノム配列決定後に実行されることが多い．16S rRNA はこのような系統解析にしばしば用いられているが，遺伝子重複や水平遺伝子伝播が報告されており，常に信頼できる系統関係が推定できるとは限らない．この問題を克服するには，複数の遺伝子（またはタンパク質）を使用する必要があるが，系統推定するために必要なオーソログを手作業で収集するのは困難である．そこで，我々はこの問題を克服するために「Genome Identifier」を開発した．このパイプラインは新たに配列決定された環境微生物のゲノムデータから遺伝子領域を予測し，その配列情報を基に近縁種からオーソログを収集し，配列を連結し，種の系統樹を推定できる．

キーワード：系統解析，ゲノム

1. はじめに

近年，ハイスループットシークエンサーの普及により，多くの非モデル生物および環境微生物などの系統関係が未知な生物についても配列決定されており，全ゲノム配列データの量は将来も増加すると予想される．一般的に，系統関係が未知な生物の全ゲノム配列を決定した後には，オープンリーディングフレーム（ORF）を推定し，次に配列データを使用して機能および分類の解析が行われる．たとえば, xBASE2 [1], RAST [2], および Prokka [3]は機能解析に，AMPHORA2 [4]および PhyloPhlAn [5]は分類学的解析に使用される．

細菌の系統解析においては，16SrRNA およびいくつかの保存されたタンパク質がよく用いられる．しかし，16S rRNA においても遺伝子水平伝播，遺伝子重複，遺伝子の欠失が報告されており[6][7], その系統樹は常に信頼できるとは限らない[8][9][10][11][12][13][14]．オーソログ配列連結系統樹はこの矛盾を回避するために開発され，[15][16][17]，種の系統関係を推定するための一般的な方法になった[17][18][19]．したがって，サンプル生物種を含む「種の系統関係」を推定するには，個々のオーソログ系統樹に基づく推定よりもオーソログ配列連結系統樹に基づく推定の方が適している．しかし，2 つの困難な点により，全自動でゲノム配列データから ORF を推定して系統解析までを行うプログラムは開発されてこなかった．1 つ目は，系統解析に用いる配列（遺伝子あるいはタンパク質）を適切に選ばなくてはならないということである．種の数が多い場合，利用可能なすべての配列を用いると計算時間が長くなりすぎてしまう．したがって，適切なマーカー配列（系統解析に用いる，保存性の高いタンパク質）を選択する必要がある．2 つ目は，系統解析のためにサンプルの近縁種を選択し，そのホモログ配列データを収集することである．一般に，各 ORF のホモログ配列データは BLAST+ [20]などの相同性検索プログラムによって検出される．しかし，各ホモログの類似性ランキングリスト内の生物種の順序はクエリー配列によって異なることがしばしば観察される．したがって，研究者は各類似度ランキングリストで生物種の順序を比較して，系統解析に用いる生物種を手作業で選択する必要がある．筆者らは上記の 2 点を克服し，ゲノム配列からマーカータンパク質の配列連結系統樹推定までを全自動で行う「Genome Identifier」を開発した．このパイプラインでは適切な近縁種のグループを自動的に選択し，ほぼ全ての細菌が共有する 31 個のマーカータンパク質またはほぼ全ての古細菌が共有する 104 個のマーカータンパク質のホモログを AMPHORA2 [4]により検出し，それを基に連結系統樹を推定する．

2. 方法

Genome Identifier には BLAST + ver.2.50[20]，HMMER ver.3.1b2[21], Prodigal ver.2.6.3[22], AMPHORA 2 [4], MAFFT ver.7.307[23]，trimAL ver.1.4 [24]，FastTree 2 ver.2.1.9 [25]がバンドルされているため，これらはインストール不要であるが，Bioperl [26]はあらかじめインストールしておく必要がある．Genome Identifier は，Linux オペレーティングシステムで動作する（動作テストは Ubuntu 16.06 LTS および 18.04 LTS で行った）．プログラムとユーザーガイドは https://sites.google.com/view/GenomeIdentifier で入手できる．入力ファイルとしてサンプル生物種ゲノムのヌクレオチド配列が記載された FASTA 形式のファイルが必要である．Genome Identifier は，以下のプロセスで構成される．（1）Prodigal を用いて，サンプル生物種のゲノム配列からタンパク質コーディング領域を予測し，タンパク質配列に翻訳する．（2）マーカータンパク質（系統内で高度に保存されたタンパク質であり，シングルコピー遺伝子によりコード

1 静岡大学農学部
Faculty of Agriculture, Shizuoka University
2 静岡大学大学院総合科学技術研究科
Graduate School of Integrated Science and Technology, Shizuoka University

ⓒ2020 Information Processing Society of Japan

される）を，AMPHORA2 を用いて検出する．Genome Identifier は 31 個の細菌マーカータンパク質または 104 個の古細菌マーカータンパク質を使用する．（３）マーカータンパク質をクエリー配列として使用し，BLAST +を用いてホモログを検出する．BLAST +のデータベースは事前に構築され，Genome Identifier に組み込まれている．相同性検索の E 値の閾値は 10^{-10} である．配列一致率と配列カバレージのデフォルトの閾値は，それぞれ 25%と 50%である．（４）各生物種のマーカータンパク質に対する平均配列一致率を計算する．平均配列一致率のランキングに基づいて，系統解析に使用される種を自動的に選択する．デフォルトでは，平均配列一致率（%）の高い順から 25 種を選択する．（５）MAFFT を使用して，各タンパク質のマルチプルアライメントを実行する．各タンパク質のアラインメントデータを連結し，一つの大きなアラインメントデータにする．次に，非保存領域を trimAl を用いて削除する．（６）連結系統樹を FastTree 2 を用いて推定する．FastTree2 の全パラメーターはデフォルト値である．

Genome Identifier の目的は，新たにゲノム配列が決定された，系統関係が不明な細菌とその近縁種の系統樹を自動で推定することである．理想的には，推定された系統樹と正しい種系統樹を比較して，プログラム（または方法）を検証する必要がある．しかし，正しい種の系統樹を入手することは事実上不可能である．したがって，系統樹比較の代わりに，推定された系統樹上のサンプル種の系統情報を確認し，NCBI 分類データベース内の分類情報と比較することとした．サンプルデータには MBGD (Microbiol Genome Database)[27]で「門ごとに 1 つのゲノムを選択」を使用して自動的に選択された 36 種の原核生物を用いた．この 36 種をそれぞれ新しく配列決定された系統未知の細菌のもの（サンプル生物種）であると仮定し，Genome Identifier を使用して種の系統的位置を正しく推定できるかどうかを以下の手順でテストした．まず入力データとなるサンプル生物種のゲノム配列データを NCBI データベースから FASTA 形式で取得した．Genome Identifier を使用してサンプル生物種とその近縁種の連結系統樹を推定した．プログラムの全てのパラメーターはデフォルト値とした．サンプル生物種がある系統のクレード（例：目，科，属）に位置する場合，その種はその系統に属すと推定した．この作業は手作業で行った．ちなみにこのテスト用のサンプル生物種の全タンパク質データは，すでにデータベースに登録されている．したがって系統樹を推定する際，常にサンプル種と最近縁な種（同一種）としてサンプル生物種の既登録配列が検出された．したがって，最近縁種を推定する際に，この登録データを無視した．最後に系統樹から推定したサンプル種の系統を NCBI 分類学データベースに基づいた分類群と比較した．

Genome Identifier はウイルスゲノムにも対応している．しかし，ウイルスゲノムには広く共有されている遺伝子がないため，マーカータンパク質を利用できない．したがって，Genome Identifier は Prodigal を使用してウイルスゲノムから全ての ORF を推定し，アミノ酸配列に変換した後，BLAST +（閾値 E 値は 10^{-10}）を使用した相同性検索により近縁ウイルスのホモログを検出する．このホモログを使用して，その系統樹を推定する．この系統樹推定の結果から，サンプルのウイルスとその近縁種との系統関係を推定できる．

3. 結果と考察

36 種の原核生物のうち 35 種の系統樹を推定できたが，*Nanoarchaeum equitans* Kin4-M の系統樹は推定できなかった．*Nanoarchaeum equitans* Kin4-M は，最小ゲノムを持つ古細菌として知られ，マーカータンパク質を 11 個も欠いているため，Genome Identifier では系統推定できなかったと考えられる． 35 種の原核生物のうち 29 種の系統が系統樹から推定された． 29 種の原核生物のうち，7 種が種レベルで推定できた．さらに属レベルでの 13 種，科レベルで 3 種，目レベルで 3 種，綱レベルで 2 種，門レベル 1 種，推定できた．一方，他の 6 つの原核生物の系統は近縁種を持たないため，推定できなかった．Genome Identifier により得られた系統樹から推定された全ての種の系統は NCBI 分類データベースの情報と一致しており，各原核生物の系統樹が Genome Identifier により正しく推定されたことを示唆した．

4. おわりに

Genome Identifier は環境微生物など，系統関係が不明なバクテリアのゲノム配列が得られた後，その系統解析を全自動で行うパイプラインである．本研究では原核生物を網羅する 36 門から選んだ 36 種を対象にテストし，29 種で系統樹を推定できた．これらの系統樹は分類情報と矛盾がなく，信頼できるものと考えられる．したがって，Genome Identifier は，原核生物の様々な種において，その近縁種の系統関係を推測するために広く使用できることが示唆された．

謝辞　静岡大学の平田久笑氏と柏原美紗子氏に様々な助言をいただいた．謹んで感謝の意を表する．

参考文献

[1] Chaudhuri, R. R., Loman, N. J., Snyder, L. A., Bailey, C. M., Stekel, D. J., and Pallen, M. J. (2008) xBASE2: a comprehensive resource for comparative bacterial genomics. Nucleic Acids Res. 36, D543–6.
[2] Aziz, R. K., Bartels, D., Best, A. A., Dejongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., et al. (2008) The RAST server: rapid annotations using subsystems technology. BMC Genomics. 9, 75.
[3] Seemann, T. (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–9.

ⓒ2020 Information Processing Society of Japan

[4] Wu, M., and Scott, A. J. (2012) Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics 28, 1033–4.
[5] Segata, N., Börnigen, D., Morgan, X. C., Huttenhower, C. (2014) PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat Commun. 4, 2304.
[6] Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V., and Polz, M. F. (2004) Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. J Bacteriol. 186, 2629–35.
[7] Kitahara, K., and Miyazaki, K. (2013) Revisiting bacterial phylogeny: Natural and experimental evidence for horizontal gene transfer of 16S rRNA. Mob. Genet. Elements 3, e24210.
[8] Saruhashi, S., Hamada, K., Horiike, T., and Shinozawa, T. (2007) Determination of whole prokaryotic phylogeny by the development of a random extraction method. Gene 392, 157–63.
[9] Tian, R. M., Cai, L., Zhang, W. P., Cao, H. L., and Qian, P. Y. (2015) Rare events of intragenus and intraspecies horizontal transfer of the 16S rRNA gene. Genome Biol. Evol. 7, 2310–20.
[10] Fitch, W. M. (2000) Homology: a personal view on some of the problems. Trends Genet. 16, 227–31.
[11] Delsuc, F., Brinkmann, H., and Philippe, H. (2005) Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6, 361–75.
[12] Koonin, E. V. (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 39, 309–38.
[13] Snel, B., Huynen, M., and Dutilh, B. E. (2005) Genome trees and the nature of genome evolution. Annu Rev Microbiol. 59, 191–209.
[14] Horiike, T., Miyata, D., Hamada, K., Saruhashi, S., Shinozawa, T., Kumar, S., Chakraborty, R., Komiyama, T., and Tateno, Y. (2009) Phylogenetic construction of 17 bacterial phyla by new method and carefully selected orthologs. Gene 429, 59–64.
[15] Baldauf, S., L., Roger, A., J., Wenk-Siefert, I., and Doolittle, W., F. (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. Science. 290, 972–7.
[16] Brown, J., R., Douady, C., J., Italia, M., J., Marshall, W., E., Stanhope, M., J. (2000) Universal trees based on large combined protein sequence data sets. Nat Genet. 28, 28–5.
[17] Horiike, T., Minai, R., Miyata, D., Nakamura, Y., and Tateno, Y. (2016) Ortholog-Finder: A tool for constructing an ortholog data set. Genome Biol Evol. 8,446–57.
[18] Gadagkar, S., R., Rosenberg, M., S., and Kumar, S. (2005) Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. J Exp Zool B Mol Dev Evol. 304(1):64–74.
[19] Thiergart, T., Landan, G., and Martin, W.F. (2014) Concatenated alignments and the case of the disappearing tree. BMC Evol Biol. 14:266.
[20] Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden TL. (2009) BLAST+: architecture and applications. BMC Bioinformatics 10, 421.
[21] Finn, R. D., Clements, J., and Eddy, S. R. (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39, 29–37.
[22] Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119.
[23] Katoh, K., and Standley, D. M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–80.
[24] Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–73.
[25] Price, M. N., Dehal, P. S., and Arkin, A. P. (2010) FastTree 2 – approximately maximum-likelihood trees for large alignments. PLoS One 5, e9490.
[26] Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 12, 1611–8.
[27] Uchiyama, I., Mihara, M., Nishide, H., and Chiba, H. MBGD update 2015: (2015) microbial genome database for flexible ortholog analysis utilizing a diverse set of genomic data. Nucleic Acids Res. 43, D270–6.

ⒸⒸ2020 Information Processing Society of Japan