様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２８　年　　６　月　　２　日現在

機関番号：　１４４０１
研究種目：　若手研究(B)
研究期間：　2014～2015
課題番号：　２６８７０３２８

研究課題名（和文）メタゲノミック診断用データベースの構築と高速解析技術の開発

研究課題名（英文）Development of a database for rapid detection of pathogens in infectious diseases

研究代表者

元岡　大祐（Motooka, Daisuke）

大阪大学・微生物病研究所・特任研究員(常勤)

研究者番号：１０６３６８３０

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：臨床検体からの網羅的な病原体探索法として、次世代シークエンサーを用いたメタゲノム解析法が利用されている。本手法は臨床検体から抽出した核酸のシークエンサーによる網羅的解読、解読した配列のGenBank塩基配列データベースに対する相同性検索(BLAST検索)、および系統分類解析から成る。しかし、シーケンサーから得られる配列数は莫大な数であり、相同性検索は大型並列計算機を用いても時間を要することが問題である。そのため本研究では、ヒトおよび病原体と成りうる微生物(ウイルス、細菌、真菌、原虫)で構成される病原体探索用データベースの構築とともに相同性検索法の改良を行い、より迅速な病原体同定法を確立した。

研究成果の概要（英文）：Metagenomic analysis using high-throughput sequencers is now widely applied for comprehensive detection of pathogens in infectious diseases. The pathogen identification approach is based on a similarity search with known proteins or nucleotides along with a taxonomic assignment of massive sequence reads produced by high-throughput sequencers. To adapt this method for clinical application, high-speed and low-cost calculation are essential. We constructed an infection Metagenomics DataBase (iMetDB), a database optimized for metagenomic diagnosis of infectious diseases. Data records of potential human pathogens, including bacteria, viruses, fungi and parasites, were collected from the GenBank database. To test the performance of databases, we carried out metagenomic analysis of influenza virus and norovirus infection cases. The resulting database enabled us perform up to a seven times faster calculation for the same pathogen detection efficiency.

研究分野：　生命情報科学

キーワード：　感染症

１．研究開始当初の背景

感染症は高い死因の 1 つである。特に発展途上国における全死亡の 1/3 が 15 歳未満の子供であり、その大部分は感染症が死亡要因である。それゆえ、感染症の病原体を迅速に検出・同定し、感染症の診断および対策を施すことが重要である。

病原体同定法としては主に、分離・培養、形態観察、PCR 法などが行われているが、それぞれの微生物に適した手法を用いる必要がある。またマイクロアレイ法では、既知の細菌とウイルスの数万種を同時に検査できるキットも発売されているが、全ての細菌やウイルスには対応しておらず、また真菌や原虫などにも対応していないため網羅性に欠ける。一方で、我々が取り組むメタゲノム解析法は、病原体ごとに異なる処理は必要がなく、臨床検体中の遺伝情報を網羅的に探索することで、全ての既知の病原体はもちろんのこと、未知の病原体をも検出・同定し得るものである。

メタゲノム解析法は、臨床検体から抽出した核酸の次世代シークエンサーによる網羅的解読、解読した配列の GenBank 塩基配列データベースに対する相同性検索(BLAST 検索)、および系統分類解析から成る。シークエンサーの著しいスループット向上により、メタゲノム解析法の臨床応用が期待できるようになった。一方で、大量に得られるメタゲノムデータの解析は、数百並列の計算が可能な大型計算機を用いて 1～2 週間、MG-RAST などの Web 解析サーバーを利用した場合では、数ヶ月も要する。また、相同性検索の検索対象である GenBank のデータベースは年間に 2 倍以上のペースで膨大し続けており、臨床応用に向けた妨げとなっている。

２．研究の目的

本研究は、原因不明疾患の患者の臨床検体から網羅的かつ迅速な病原体の検出・同定を行い、感染症の原因を解明することを目的とする。既に我々は、メタゲノム解析法により、呼吸器疾患の患者の咽頭スワブからインフルエンザウイルス、敗血症患者の血液から *Klebsiella variicola* の検出に成功し、病原体同定法を確立しつつある。しかし、情報解析の部分については未発達であり、得られたシーケンスデータを膨大な量のデータベースに対して相同性検索する際に、1 週間から数ヶ月の時間を要してしまう。そのため本研究では、GenBank に代わる病原体探索専用の小さなデータベースの構築と相同性検索・系統分類法の改良、メタゲノムデータ解析法の検討を行い、迅速かつ簡便な病原体同定法を確立することが目的である。

３．研究の方法

本研究ではまず、病原体探索用データベースの構築と系統分類解析法の高速化を行った。また、臨床検体の大部分を占める宿主由来核酸の高速判別法を開発し、これらの結果を基にデータ解析を高速かつ容易に実行できるパイプラインを作成する。次に、臨床検体を用いてメタゲノム解析法を実行し、病原体検出効率や解析法の比較検討行った。

(1) 病原体探索用データベースの構築と系統分類解析法の高速化

GenBank データベースにはマウス、ショウジョウバエなどのモデル生物の登録数が多い。そこで本研究ではまず、ヒトの臨床検体中に存在しえないマウスなどの生物種をデータベースから排除し、病原体探索用データベースには、細菌、ウイルス、真菌、原虫(一部)とヒトのみの情報にした。また近年、土壌細菌のメタゲノム解析などが行われ、”*Uncultured bacterium*”といった名前で登録されているデータを散見される。仮に臨床検体中の核酸配列が、BLAST 検索でこの様な配列にヒットしても有用な情報にはならない。そのため、この様な揺らぎのある情報も病原体探索用データベースから排除した。

BLAST 検索の結果からは相同配列の固有コード、遺伝子名や一致度などの情報は得られるが、どのようなウイルスがどの程度見つかったかを知ることはできない。現在は、固有コードを系統分類データベースと照らし合わせる作業が必要であり、この解析にも時間を要している。そこで本研究では、系統分類解析を高速化するため、作成したデータベースに系統分類情報(界・門・綱・目・科・属・種)を付加した。またこの際、系統分類情報が十分でない配列はデータベースから削除した。

(2) 宿主由来核酸の高速判別法の開発

糞便を除く臨床検体は、大部分を宿主(ヒト)由来核酸が占める。そこで本研究では、メタゲノムデータ中のヒト由来核酸を高速で判別するため、まず全データをヒトゲノムに対してマッピングし、マッピングされなかった配列のみ、つまりヒト由来ではない核酸のみを病原体探索用データベースに対して相同性検索することで高速化を目指した。マッピング法は、非常に相同性が高い配列に限られるが BLAST 検索の 100 倍以上早く、ヒト由来核酸か否かの判定が可能である。臨床検体由来の核酸配列は通常の RNA-Seq などと比較すると短く、また増幅反応による PCR エラーなども含みやすい。そのため、種々のマッピングソフトを用いて、マッピング効率を比較した。今回は bowtie2、bwa、CLC、ELAND、MAQ、SOAP、Stampy、RazerS3 と言ったソフトウェアを用い、リファレンス配列としては hg19 を用いた。

(3)病原体探索用データベースと高速解析パイプラインの臨床応用

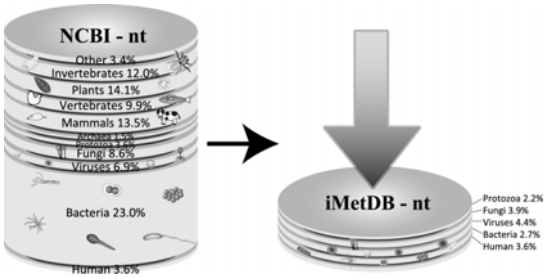
上記手法により作成した病原体探索用データベースと宿主由来核酸の高速判別法を

含むパイプラインを作成し、臨床検体を用いて性能を評価した。臨床検体としては、ノロウイルス陽性患者の糞便、インフルエンザウイルス陽性患者の鼻腔ぬぐいを用いた。イルミナ社MiSeqを用いてメタゲノム解析を行い、得られたデータについて解析した。

４．研究成果

(1) 病原体探索用データベースの構築と系統分類解析法の高速化

病原体探索用データベースの構築と高速な系統分類解析法の開発を目指した。「研究の方法」に記載した手法により作成した病原体探索用のデータベース(iMetDB : infection Metagenomics DataBase)は、核酸配列データベース登録数が全データベースの 4 分の 1 に、アミノ酸配列データベース登録数が、3 分の 2 に減少した(図 1)。本データベースの作成により、核酸配列を対象とした相同性検索では従来の 4 倍早く計算ができるようになった。これまで 2 週間要した計算であれば、3〜4 日で済むようになり、より現実的な時間内で解析できるようになった。



2,074 万配列　　497 万配列

図 1. NCBI-nt と病原体探索用データベース (iMetDB-nt)の構成生物比とデータサイズ

また、従来は BLAST 検索後の系統分類解析は時間を要していた。例えば、SQL データベースを利用した場合に 1 時間を要したデータ量の系統分類解析は今回作成したデータベースを使用した場合には、1 分以内に終わらせることが出来た。また、この作業はある程度のプログラミング処理が必要であり、敷居が高い作業であったが、今回構築したデータベースを使用することで、たった 1 つのコマンドで系統分類解析が完了するようになった。

(2) 宿主由来核酸の高速判別法の開発

ノロウイルス陽性患者の糞便、インフルエンザウイルス陽性患者の鼻腔ぬぐいサンプルについてメタゲノム解析を実施し、NCBI-nt に対する BLAST 検索を行った。そして宿主(ヒト)由来と判定された核酸配列のみを抜き出し、それらが様々なマッピング手法によりどの程度、ヒトと判定できるかを確認した。種々のマッピングソフトウェアを使用して比較した結果、Stampy や CLC を用いた場合、ほぼ全ての配列をヒトと判定することが出来た(図 2)。

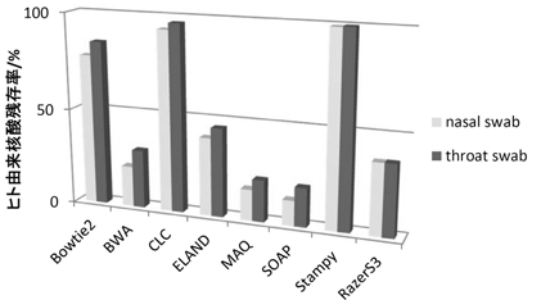


図 2. 種々のマッピングソフトによるヒト由来核酸の非マッピング率

(3)病原体探索用データベースと高速解析パイプラインの臨床応用

上記(1)および(2)にて作成したデータベース(iMetDB)と宿主由来核酸の高速判定法を組み合わせた解析パイプラインを使用して、ノロウイルス、インフルエンザウイルス感染検体についてメタゲノム解析を行った。その結果、ノロウイルス陽性検体については、得られた全配列について NCBI-nt に対する BLAST 検索を行った場合に 12 時間要した計算が、本研究にて開発した解析法を用いることで、1.7 時間と 7 倍の高速化に成功した。インフルエンザ陽性検体については、1 時間が 20 分と 3 倍の高速化に成功した。また、データベースサイズは小さくなったが、病原体の検出効率が同じであることが確認できた(図 3)。

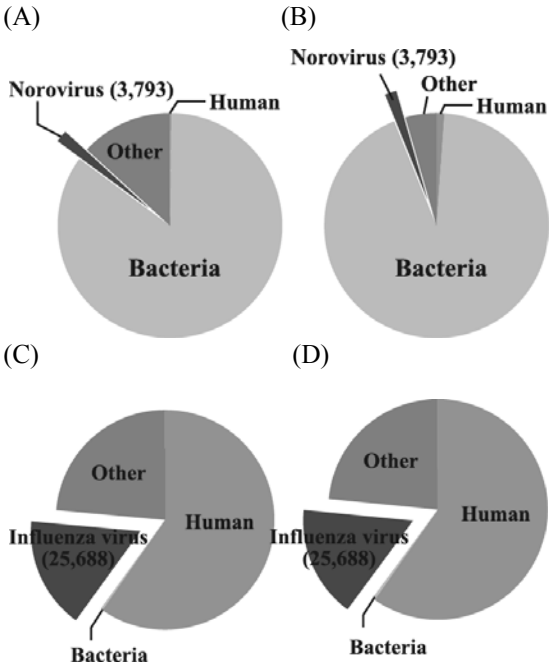


図 3. メタゲノムデータの生物種割合 (A, B) ノロウイルス陽性糞便、(C, D) インフルエンザ陽性鼻腔拭い。A と C は解析に NCBI-nt を、B と D は iMetDB-nt を使用。

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 3 件）

①Perlejewski K, Popiel M, Laskus T, Nakamura S, Motooka D, Stokowy T, Lipowski D, Pollak A, Lechowicz U, Caraballo Cortés K, Stępień A, Radkowski M, Bukowska-Ośko I, Next-generation sequencing (NGS) in the identification of encephalitis-causing viruses: Unexpected detection of human herpesvirus 1 while searching for RNA pathogens, *J Virol Methods*, 2015, 査読有, doi: 10.1016/j.jviromet.2015.09.010

②Motooka D, Nakamura S, Hagiwara K, Nakaya T, Viral detection by high-throughput sequencing, *Methods Mol Biol*, 1236, 125-34, 2015, **査読有**, doi: 10.1007/978-1-4939-1743-3_11

③Miyamoto M, Motooka D, Gotoh K, Imai T, Yoshitake K, Goto N, Iida T, Yasunaga T, Horii T, Arakawa K, Kasahara M, Nakamura S, Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes, *BMC Genomics*, 15, 699, 2014, **査読有**, doi: 10.1186/1471-2164-15-699

〔学会発表〕（計 1 件）

①元岡大祐、メタゲノミック診断用データベース(iMetDB)の構築と高速解析パイプライン、NGS 現場の会第四回研究会、2015 年 7 月 1 日～7 月 3 日、つくば国際会議場

〔その他〕

ホームページ等

http://imet.gen-info.osaka-u.ac.jp/en/imetdb.html

６．研究組織

(1)研究代表者

元岡　大祐（MOTOOKA, Daisuke）

大阪大学・微生物病研究所・特任研究員(常勤)

研究者番号：10636830