プレスリリース

大学共同利用機関法人 情報・システム研究機構 国立遺伝学研究所

理化学研究所

本論文は既に公開されています
本情報はすぐにご利用いただけます

2021 年 7 月 12 日

膨大なメタゲノムデータの相同性検索を可能にするシステム「PZLAST」

■ 概要

「環境」中の細菌集団の研究すなわち「マイクロバイオーム研究」が急速に進展したことにより、DDBJ(1)などの公共データベースにはヒト腸内や土壌、河川、海洋など多様な環境に生息する細菌集団のゲノム断片（メタゲノム）データが大量に登録されています。これらのゲノム断片のデータは「遺伝子の宝の山」と言われていますが、その情報量があまりにも莫大であるため、「宝の山」を「発掘」するための解析技術の適用が困難で、「似た配列」を探し出す相同性検索すら難しい状態でした。

情報・システム研究機構　国立遺伝学研究所、理化学研究所、株式会社 PEZY Computing の共同研究グループは、公開中の膨大なゲノム断片から予測したアミノ酸配列データをもとに極めて高速かつ高精度にアミノ酸配列の相同性検索を可能とする Web サービス「PZLAST」を開発しました。「PZLAST」を利用することで「遺伝子の宝の山」に埋もれている新たな遺伝子をその遺伝子が存在する環境の情報などとともに容易に「発掘」することが可能になったのです。

膨大なゲノム断片のデータが検索可能になったことで、薬剤耐性因子、病原因子やウイルスなど特定の遺伝子の環境中での動態や、新たな機能を持つ遺伝子の発見、遺伝子と環境の関係性の解析、創薬など、さまざまな研究の発展に寄与することが期待できます。

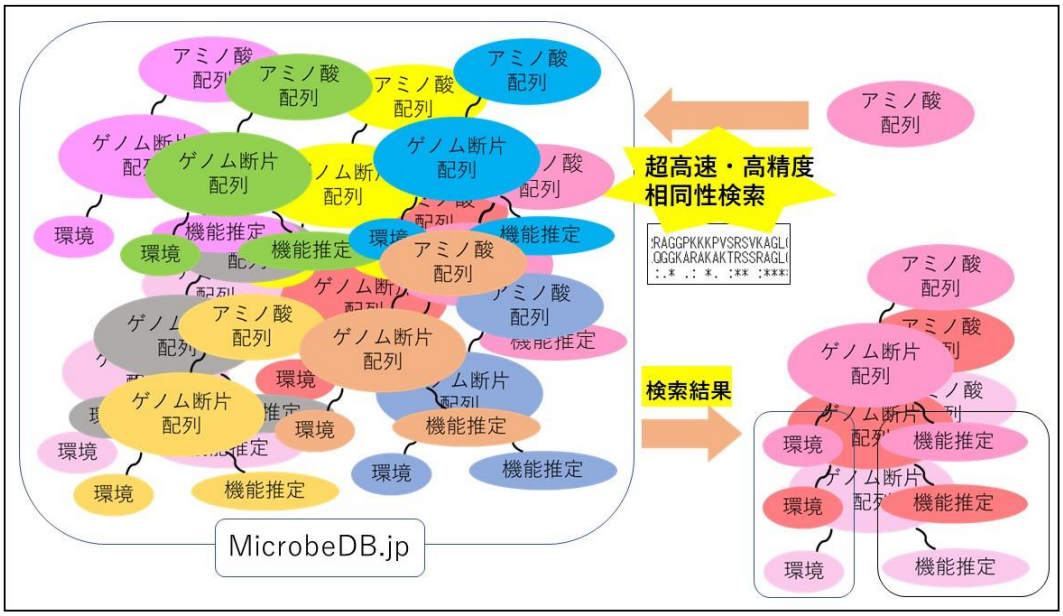

図 1:「PZLAST」の概念図

■ 成果掲載誌

本研究成果は、国際計算生物学会誌「Bioinformatics」に 2021 年 7 月 8 日（日本時間）に掲載されました。

論文タイトル: PZLAST: an ultra-fast amino acid sequence similarity search server against public metagenomes
（メタゲノム遺伝子データに対する超高速相同性検索システム PZLAST）

著者: H. Mori, H. Ishikawa, K. Higashi, Y. Kato, T. Ebisuzaki, K. Kurokawa.
（森宙史、石川仁、東光一、加藤成晃、戎崎俊一、黒川顕）

■ 研究の詳細

● 研究の背景

「環境」中の細菌集団のゲノム研究すなわちマイクロバイオーム研究の急速な進展に伴い、DDBJ などが運営する国際塩基配列データベースにはマイクロバイオームの主体となる微生物集団のゲノム断片配列（メタゲノム配列）データが加速度的に蓄積されつつあります。なかでも「ショットガンメタゲノム配列データ[(2)]」は、未同定、未培養な細菌がもつ未知の遺伝子情報を大量に含むことから「遺伝子の宝の山」とも言われています。この「宝の山」を「発掘」するための解析、すなわちメタゲノム配列データに対する配列相同性検索ができれば、類似遺伝子の環境における動態、機能的に重要な相同遺伝子の発見、さらには機能が未知の遺伝子の機能推定にも活用することができます。しかしながら、公共データベースに蓄積されているショットガンメタゲノム配列データは膨大であり、それらテラバイト(TB)を超える巨大なメタゲノム遺伝子データの配列をもとに、短時間で相同性検索をするのは非現実的でした。

● 本研究の成果

情報・システム研究機構 国立遺伝学研究所、理化学研究所、株式会社 PEZY Computing の共同研究グループは、公開されているメタゲノムデータから予測した遺伝子のアミノ酸配列データ（約 2.5TB）に対して、極めて高速かつ高精度にアミノ酸配列の相同性検索を可能とするウェブサービス「PZLAST」を開発しました。

科学技術計算にも利用される画像処理用プロセッサ GPU は SIMD（Single Instruction Multiple Data）型で、すべてのスレッドが「異なるデータ」に対して「同じ命令」しか実行できません。一方で、本研究グループが開発した PZLAST は MIMD（Multiple Instruction Multiple Data）型メニーコアプロセッサー「PEZY-SC2」上で動作します。MIMD 型プロセッサでは、各スレッドが「異なるデータ」に対して「異なる命令」を実行できます。PZLAST はこの MIMD 型プロセッサの特徴を効率的に活用し、複数の PEZY-SC2 に指令を分散させ、それぞれの PEZY-SC2 において多数のスレッド（15,872 スレッド）を利用、即ち「超並列計算」することで、極めて高速かつ高精度な配列相同性検索を実現したのです。

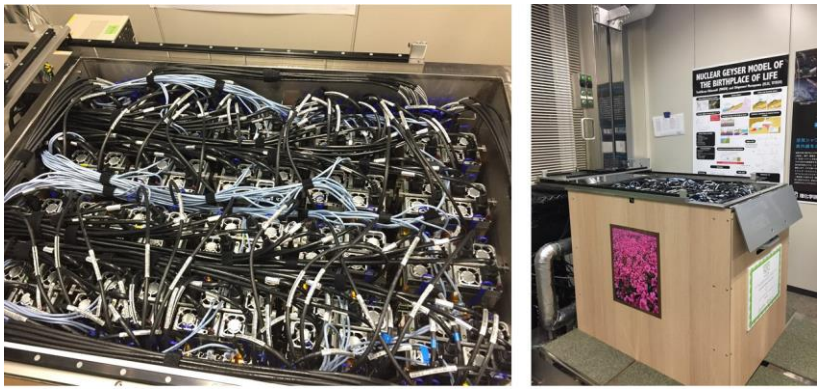

図 2：PZLAST が動作している理化学研究所のスーパーコンピューター「皐月」（黒川顕教授撮影）

PZLAST が相同性検索の際に参照配列として使用するアミノ酸配列データは、同じく国立遺伝学研究所で開発・運用している微生物統合データベース「MicrobeDB.jp[3]」から取得しています。MicrobeDB.jp では、DDBJ から公開されているショットガンメタゲノム配列データから「遺伝子を予測」をするとともにそれら「遺伝子の機能を推定」した上で、サンプルが採取された「環境の情報」と統合してデータベースに格納しています。すなわち、MicrobeDB.jp に格納されているすべての遺伝子データは、それら遺伝子がどのような環境中に存在したのかなど、配列情報と環境情報が紐付けされているのです。

PZLAST では、この MicrobeDB.jp から取得したアミノ酸配列を参照配列として相同性検索をおこないます。参照アミノ酸配列データは、4,339 サンプルのショットガンメタゲノムデータから予測された遺伝子群で、容量は約 2.5 テラバイト（423 億個の予測遺伝子数、アミノ酸約 1.7 兆残基）におよび、「ヒト遺伝子 35 万人分」に相当する巨大なデータとなっているのです（論文発表時）。PZLAST では、MicrobeDB.jp 中の高頻度に更新されるメタゲノムデータをすばやく取り入れた最新の参照配列を対象とした相同性検索を容易に実現する事が可能になったのです。

ユーザーは 1 回の検索で最大 10,000 個の配列を入力することができます（図 3A）。1 回の検索に要する計算時間は約 10 分であり、検索結果は上位 1 万ヒットまでを、Metagenome and Microbes Environmental Ontology （MEO）クラス（図 3B）、Foundational Model of Anatomy オントロジー（FMA）クラス（図 3C）、地理的分布（図 3D）に基づいてまとめ、Web ブラウザ上で可視化します。MEO クラスは相同遺伝子の環境分布、FMA クラスは相同遺伝子のヒトマイクロバイオームにおける分布を示しています。

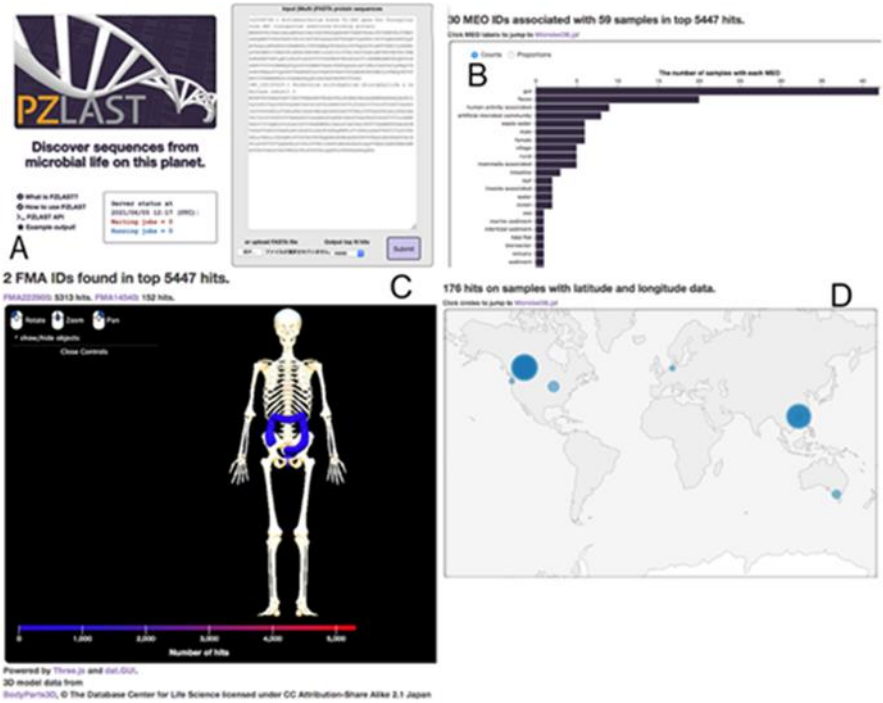

図3：PZLAST ウェブサービスの概要図
A. PZLAST ウェブサービスの検索クエリ入力画面。
B. 相同遺伝子の MEO による環境分布の例
C. 相同遺伝子の FMA によるヒトマイクロバイオームにおける分布の例
D. 相同遺伝子の地理的分布の例

Web ブラウザベースの PZLAST に加えて、単に配列検索の結果を得るための「REST API」サービスも提供しており、GUI による操作なしで検索結果を得ることが可能となっています。

● 今後の期待

PZLAST は、公開されているメタゲノムデータから予測した遺伝子のアミノ酸配列データに対して、データを削減・圧縮することなく、超高速かつ高精度な配列相同性検索を可能とする世界で初めての Web サービスです。膨大なメタゲノム遺伝子データを配列相同性により検索可能になった事で、薬剤耐性因子やウイルスなど

特定の遺伝子の環境中での動態や、新たな機能を持つ遺伝子の発見、遺伝子と環境の関係性の解析、創薬など、多様な課題に対して価値ある情報を提供できることが期待できます。

■　用語解説

(1) DDBJ

DNA Data Bank of Japan。国立遺伝学研究所の生命情報・DDBJ センターが、欧州 EBI および米国 NCBI の 3 局で国際塩基配列データベース連携 INSDC (International Nucleotide Sequence Database Collaboration) を組織し運用する塩基配列データベース。

(2) ショットガンメタゲノム解析

微生物群集のもつ遺伝情報を丸ごと解読する解析手法。

(3) MicrobeDB.jp

JST NBDC「統合化推進プログラム」において国立遺伝学研究所のゲノム進化研究室（黒川教授）が中心となり開発・運用を行なっている微生物の統合データベース（データベース URL: https://microbedb.jp ）。

■　研究体制と支援

本研究は、国立遺伝学研究所の森宙史准教授、東光一特任研究員、黒川顕教授、理化学研究所の加藤成晃協力研究員、戎崎俊一主任研究員、株式会社 PEZY Computing の石川仁氏による共同研究グループによってすすめられました。

本研究は文部科学省高性能汎用計算機高度利用事業「ヘテロジニアス・メニーコア計算機による大規模計算科学（代表：姫野龍太郎）」の支援によりおこなわれました。

■　問い合わせ先

＜研究に関すること＞

- 国立遺伝学研究所　先端ゲノミクス推進センター
  データ解析部門長　森　宙史　（もり　ひろし）

＜報道担当＞

- 国立遺伝学研究所　リサーチ・アドミニストレーター室　広報チーム

- 理化学研究所　広報室

※時節柄、Zoom 会議での取材にも対応できますので、Zoom 会議をご希望の場合には、その旨お知らせください。