研究開発課題別中間評価結果

➢ 課題情報

研究開発課題名	「マイクロバイオーム研究を先導するハブを目指した微生物統合データベースの特化型開発」
研究代表名	森 宙史

➢ 中間評価結果：

これまで開発してきた微生物統合データベース MicrobeDB.jp に、微生物の表現型やメタゲノム由来のゲノムなどの新しいデータを統合化するとともに、マイクロバイオーム関連のデータ検索・解析機能を大幅に強化し、自由度の高い検索・データ取得を実現することで、国際的に隆盛を極めるマイクロバイオーム研究に特化した国際的なデータハブ Microbiome Datahub として再構築する研究開発課題である。

218,653 の MAG（Metagenome-Assembled Genome）データを Microbiome Datahub に格納した。これまで 16S rRNA 遺伝子配列を用いていたマイクロバイオームの系統組成解析パイプラインを k-mer 組成に変更し、速度を 8 倍以上に向上させた。マイクロバイオームの遺伝子組成解析パイプラインについては、深層学習ベースの DeepGO による遺伝子組成推定により 10 倍以上の速度向上を達成したが、遺伝子機能のアノテーション精度に問題があったことから、MAG データに対して高精度な KOfamScan を実行し、MAG 単位で遺伝子機能組成を解析することで対応した。MAG の機能オーソロググループアノテーションでは、MBGD（Microbial Genome Database for Comparative Analysis）のデフォルトオーソログクラスタに対して PZLAST を適用することで、オーソログ ID を高速にアサインできるようにした。理研バイオリソースセンター微生物材料開発室(JCM)と製品評価機構バイオロジカルリソースセンター(NBRC)がそれぞれ所有する菌株データに微生物 Phenotype オントロジーをアノテーション付けし、Phenotype 情報での検索基盤を整えた。LLM を使って論文から抽出し整理したメタデータと BioSample との対応関係を整理し、BioSample をクラスタリングすることでマイクロバイオームのメタ解析に使用できるサンプルメタデータ俯瞰ツール EMBERS を開発した。EMBERS を開発したことで、MEO や HMADO 等のオントロジーを用いたメタデータアノテーションに関し、アノテーションとチェックにかかる時間を 70%以上低減することができた。キラーデータセット・アプリケーションとしては、Microbiome Datahub のデータをユーザが自由に取得できる API を開発し、データ ID を指定した URL にアクセスするだけで、BioProject のメタデータや系統組成データ、MAG のメタデータ、MAG の配列データを取得できるようにした。Microbiome Datahub のメタゲノムアミノ酸配列データを活用し、ユーザが入力したメタゲノムアミノ酸配列データから微生物群集が生息する環境温度を推定するツール Metagenomic Thermometer（メタゲノム温度計）を開発し、論文発表と web アプリケーションの公開を行った。また、約 21 万 MAG 由来の遺伝子のアミノ酸配列データに対する配列類似性検索を行える機能（PZLAST-MAG）を開発した（2024 年度中に公開予定）。現在、Microbiome Datahub はテスト版を公開しているが、2024 年中には正式版を公開する予定。セキュリティ上の問題から、前身の MicrobeDB.jp は 2024 年度初めに閉鎖し、Microbiome Datahub のテストサイトに誘導するようにしている。

上記のように、データベース構築に必要となる個々の要素技術の開発においては一定の進捗が見られる。特に、LLM を活用した文献やデータベースからのメタデータ抽出は、統合化推進プログラムの他の研究開発課題においても有益であることから、EMBERS の横展開を期待する。一方で、Microbiome Datahub の正式版の公開が当初計画よりも遅れているなど、研究開発の進捗状況に若干の懸念が認められることから、研究開発期間内に十分な成

果を得るためには、研究計画の一部見直しを行う必要がある。

以上