ライフサイエンスデータベース統合推進事業

統合データ解析トライアル

研究開発課題

「MicrobeDB.jp データを用いたメタゲノム解析

Web アプリケーションの開発」

研究開発終了報告書

研究開発期間： 平成25年9月～平成26年1月

研究代表者 :森宙史

（東京工業大学大学院生命理工学研究科、

助教）

§1 研究開発のねらい

細菌群集は地球上のいたるところに生息しており、地球上の物質循環の基盤を担うとともに、人間活動に密接に関与している。それらの群集を構成する細菌の多くは未だ培養困難であるため、培養法に立脚した従来の細菌学的手法では細菌群集の詳細を解明することは困難であった。細菌群集から DNA をまるごと抽出し、細菌群集の系統組成や遺伝子機能組成を明らかにするメタゲノム解析が、新型シーケンサーの普及に伴い、細菌群集に対する主な解析手法として広く用いられている。特に、ここ 1-2 年で小型の新型シーケンサーが普及し、2-4 日ほどで数百万から数千万配列のメタゲノムシーケンシングが行えるようになってきている。しかしながら、新型シーケンサーから得られたメタゲノム解析データを、実験系の研究者でも容易に解析できるようにする解析ツールは未だ存在せず、大きな問題となっている。

使い勝手の良いメタゲノム解析ツールがほとんど存在しない理由の一つとして、メタゲノム解析はゲノム解析同様、他のメタゲノム解析データとの比較解析を行うことで、初めてその細菌群集の特徴を明らかにできるため、関連する既存のメタゲノム解析データも同一の解析手法で解析し直す必要があることが挙げられる。そのため、研究者がメタゲノム解析データを用いてバイオインフォマティクスの解析を行う際には、自分がシーケンスしたメタゲノム解析データに加えて、関連する既存のメタゲノム解析データについても、公共の塩基配列データベース（DB）から塩基配列データをダウンロードして自分で解析を行う必要があり、バイオインフォマティクスの解析手法に詳しくない実験系の研究者には、非常に荷が重い。したがって、自分のメタゲノム解析データを容易に解析できる解析ツールと、その解析ツールを用いて既存のメタゲノム解析データを解析した結果が整理された DB の両方を開発することが、実験系の研究者が容易にメタゲノム解析データを解析できるようになるために必須であると言える。

本研究開発では、この現状を打開するために、ユーザが持つメタ 16S・メタゲノム解析データを容易に解析可能な解析ツールを開発することを目的とした。具体的には、ユーザがアップロードしたメタ 16S・メタゲノム解析データの fastq 配列データを、独自の解析パイプラインを用いて解析し、メタ 16S 解析データであれば、系統組成を、メタゲノム解析データであれば、系統組成と遺伝子機能組成を推定する。その後、JST NBDC の統合化推進プログラムの採択課題のうち、微生物関連の統合 DB として開発されている MicrobeDB.jp で公開されている、公共の塩基配列 DB から取得した既存のメタ 16S・メタゲノム解析データを、今回用いる解析パイプラインで解析した結果のデータを、ユーザのメタ 16S・メタゲノム解析データとの比較解析に利用する。両者は同じ解析パイプラインで解析されているため、問題無く比較が可能となる。既存のメタ 16S・メタゲノム解析データについてそれらの解析結果を利用することで、大量の計算を個々の研究者が行う手間を大幅に削減し、自分のサンプルは系統組成および遺伝子機能組成がどの程度特異的であるか、また、どのような環境由来のサンプルと類似しているか等の情報を提供することで、ユーザが持つサンプルが由来した細菌群集についての理解を促進することが期待できる。

§2 研究成果

(1) MicrobeDB.jp の解析手法を用いたメタゲノム解析 Web アプリケーション MeGAP の開発

MicrobeDB.jp で用いられているメタ 16S・メタゲノム解析手法を、ユーザが持つメタ 16S・メタゲノム解析データに対応できるように改良して、ユーザから fastq ファイルまたは圧縮された fastq.gz ファイルを受け取って、メタ 16S 解析データであれば系統組成、メタゲノム解析データであれば系統組成と遺伝子機能組成を推定して結果をメールでユーザに返す Web アプリケーション MeGAP (MetaGenome Annotation Pipeline)(http://fs2.bio.titech.ac.jp/megap)を開発した。

具体的な解析ステップとしては、メタ 16S 解析データであれば、(i)塩基配列のクオリティコントロール、(ii)配列相同性による配列のクラスタリング、(iii)PCR で生じたキメラ配列のフィルタリング、(iv)各配列中の連続塩基頻度を計算し、連続塩基頻度が類似した既知の系統(Ribosomal Database Project 由来)に配列をアサイン、(v)全配列の系統アサインメント結果を、それぞれの系統分類階層(Genus, Family, Order, Class, Phylum) で集計、という流れになっている。

また、メタゲノム解析データであれば、(i)塩基配列のクオリティコントロール、(ii)各メタゲノム配列を完全ゲノムが解読済みの Archaea と Bacteria の全てのゲノム配列相手に配列相同性検索、(iii)タンパク質コーディング遺伝子の領域と一定以上の配列相同性があったメタゲノム配列のみに結果を絞り込む、(iv) 完全ゲノムが解読済みのゲノム中の各タンパク質のオーソログ情報を基に、各メタゲノム配列に機能アノテーション情報を付加、(v)そのメタゲノム解析データの系統組成、遺伝子機能組成を集計、という流れになっている。メタ 16S・メタゲノム解析の両方で、上記の方法および利用する配列 DB は MicrobeDB.jp で既存のメタ 16S・メタゲノム解析データからの系統組成と遺伝子機能組成を推定する際に用いているものと同一である。なお、計算時間としては、新型シーケンサー由来の数十万から数百万本の配列データであれば、現状、メタ 16S 解析であれば数分から数十分、メタゲノム解析であれば、数時間から数十時間となっている。

上記の MeGAP を利用することにより、新型シーケンサー由来の大量の配列データから、ユーザは fastq ファイルをアップロードしてメールが返ってくるのを待つだけで、細菌群集を特徴付ける基本的な要素である、その群集の系統組成と遺伝子機能組成の情報を容易に得ることが可能になった。

(2) 既存のメタゲノム解析データとの比較解析 Web アプリケーション MeGAP-MicrobeDB.jp の実装

あるメタ 16S・メタゲノム解析データを MeGAP で解析した結果の系統組成および遺伝子機能組成を、MicrobeDB.jp の既存のメタ 16S・メタゲノム解析データと比較解析する Web アプリケーション、MeGAP-MicrobeDB.jp (http://fs2.bio.titech.ac.jp/megaptomicrobedb.jp)を開発した。

MeGAP-MicrobeDB.jp では、ユークリッド距離や 1-Pearson 相関係数等のいくつかの距離尺度で、ユーザがアップロードした系統組成および遺伝子機能組成データと、

MicrobeDB.jp 中の既存のメタ 16S･メタゲノム解析データの系統組成および遺伝子機能組成データとの間で距離を計算し、類似した既存のサンプルを発見した後、系統組成や遺伝子機能組成の棒グラフや、それらのサンプルの詳細が記述された MicrobeDB.jp ページへのリンク等をメールでユーザに返信する。

MeGAP-MicrobeDB.jp を利用することにより、ユーザは MeGAP から得られた系統組成および遺伝子機能組成データをアップロードしてメールが返ってくるのを待つだけで、そのメタ 16S･メタゲノム解析データと類似した系統組成および遺伝子機能組成を持つ既存のメタ 16S･メタゲノム解析データがどれかがわかるようになった。MicrobeDB.jp でそれらの既存のメタ 16S･メタゲノム解析データが由来した環境の詳細な情報を参照することによって、どのような環境パラメータが共通しているため、細菌群集の系統組成や遺伝子機能組成が類似する結果になったのかを推測することが出来るようになった。

§3 研究開発計画および計画に対する達成状況

（1）達成状況

当初の研究開発計画では、(1)MicrobeDB.jp の解析手法を用いたメタゲノム解析ツールの開発、(2)既存のメタゲノム解析データとの比較解析機能の実装、の 2 つを行うと計画していた。これらは両方とも実現できた。ただし、(2)既存のメタゲノム解析データとの比較解析機能の実装の中で挙げていた、Metagenome/Microbes Environmental Ontology (MEO)、および Metagenome Sample Vocabulary (MSV) 等のオントロジーの語彙を選択することで、MicrobeDB.jp の既存のメタ 16S・メタゲノム解析データの中から比較解析対象とするデータを絞り込む機能については、今回の開発期間中には実装することが出来なかった。環境という抽象的なものを扱うため、それぞれのユーザが考える、細菌群集が生息する環境のオントロジーが、MicrobeDB.jp で定義している細菌群集の生息環境のオントロジーと完全に一致することは稀であり、比較解析機能を実装する過程でこの絞り込みがどの程度有効であるか疑問が生じたためである。

（2）ツールの将来性への展望

新型シーケンサーの小型化およびハイスループット化は今後も続くと予想され、既に USB 型シーケンサーも実用化しつつある。細菌は地球上のほとんどの場所に生息しており、その系統組成および遺伝子機能組成は環境条件が変わると変動する。つまり、細菌群集の組成は、環境の高精度なモニタリング指標として利用可能である。これらの、新型シーケンサーの小型化およびハイスループット化と、環境と細菌群集の関連性の 2 つの要素をあわせて考えると、今後はメタ 16S・メタゲノム解析のデータは、(i) 細菌群集自体の詳細を解明する従来と同様の目的、(ii) 細菌群集の組成を単なる環境のモニタリング指標としてのみ利用する目的、の 2 通りの目的で生産されると予想できる。

(i)の目的のためには、例えば細菌群集に様々な環境刺激を加えた場合にどの

ように群集が変動するかをそれぞれのタイムポイントでメタゲノム解析し、MeGAP で解析して得られた系統組成および遺伝子機能組成データを基に、詳細に統計比較をする等の緻密な実験デザインの元に研究を行う必要がある。したがって、ユーザが持つ複数のメタ 16S・メタゲノム解析データを比較解析可能なように、MeGAP-MicrobeDB.jp を改良する必要がある。また、MicrobeDB.jp に存在する、各系統の細菌の詳細や各遺伝子の詳細な情報を容易に参照可能なように、ユーザに返す結果をより充実させる必要がある。

(ii)の目的のためには、細菌群集の系統や遺伝子機能の詳細について、ほとんど知識が無い研究者にとっても使いやすいツールにする必要がある。MeGAP および MeGAP-MicrobeDB.jp は、ユーザ側で必要な処理を出来るだけ少なくすることを一つの開発方針として、研究開発を行って来た。その結果として、MeGAP と MeGAP-MicrobeDB.jp は、どちらもデータをアップロードしてユーザにメールで結果を返す形の Web アプリケーションになり、ユーザはアップロードするファイルを選択して解析実行ボタンを押すだけで良くなった。しかしながら、MeGAP-MicrobeDB.jp がユーザに返す結果はあくまで既存のどのサンプルとユーザがアップロードしたデータが類似しているか、であり、既存のいくつかのサンプルとの類似性が何を意味するのかについては、それらの既存のサンプルが由来した環境パラメータの共通性や、共通して優占する系統や遺伝子の詳細について十分に調べて考察する必要がある。類似したサンプル間で何が共通しており、何が異なるのかを容易に抽出することが出来るように、多次元尺度構成法等に代表される統計解析の手法の有効性を検討しつつ、MeGAP-MicrobeDB.jp により良い比較解析手法を実装していく必要がある。

細菌群集を解析する上記の 2 つの目的の両方で、知識発見に至るまでの道筋はどちらも容易ではないが、上述のような様々な改良を今後も MeGAP と MeGAP-MicrobeDB.jp に継続的に行うことによって、それらのツールを微生物学に限らない多くの研究者に利用してもらえるようになり、比較解析を行う上での基盤となっている、統合化推進プログラムで開発されている MicrobeDB.jp の利用者も増えると確信している。

§4 研究参加者

氏名	所属	役職	研究開発項目	参加時期
○森宙史	東京工業大学大学院生命理工学研究科	助教	研究代表者	H25.9-H26.1

§5 成果発表等

(1)原著論文発表　(国内(和文)誌　0 件、国際（欧文）誌　0 件)

(2)その他の著作物(総説、書籍など)

(3)国際学会発表及び主要な国内学会発表

① 招待講演　(国内会議　0 件、国際会議　0 件)

② 口頭発表　(国内会議　0 件、国際会議　0 件)

③ ポスター発表　(国内会議　0 件、国際会議　0 件)

(4)知財出願

①国内出願（0 件）

②海外出願（0 件）

③その他の知的財産権

なし

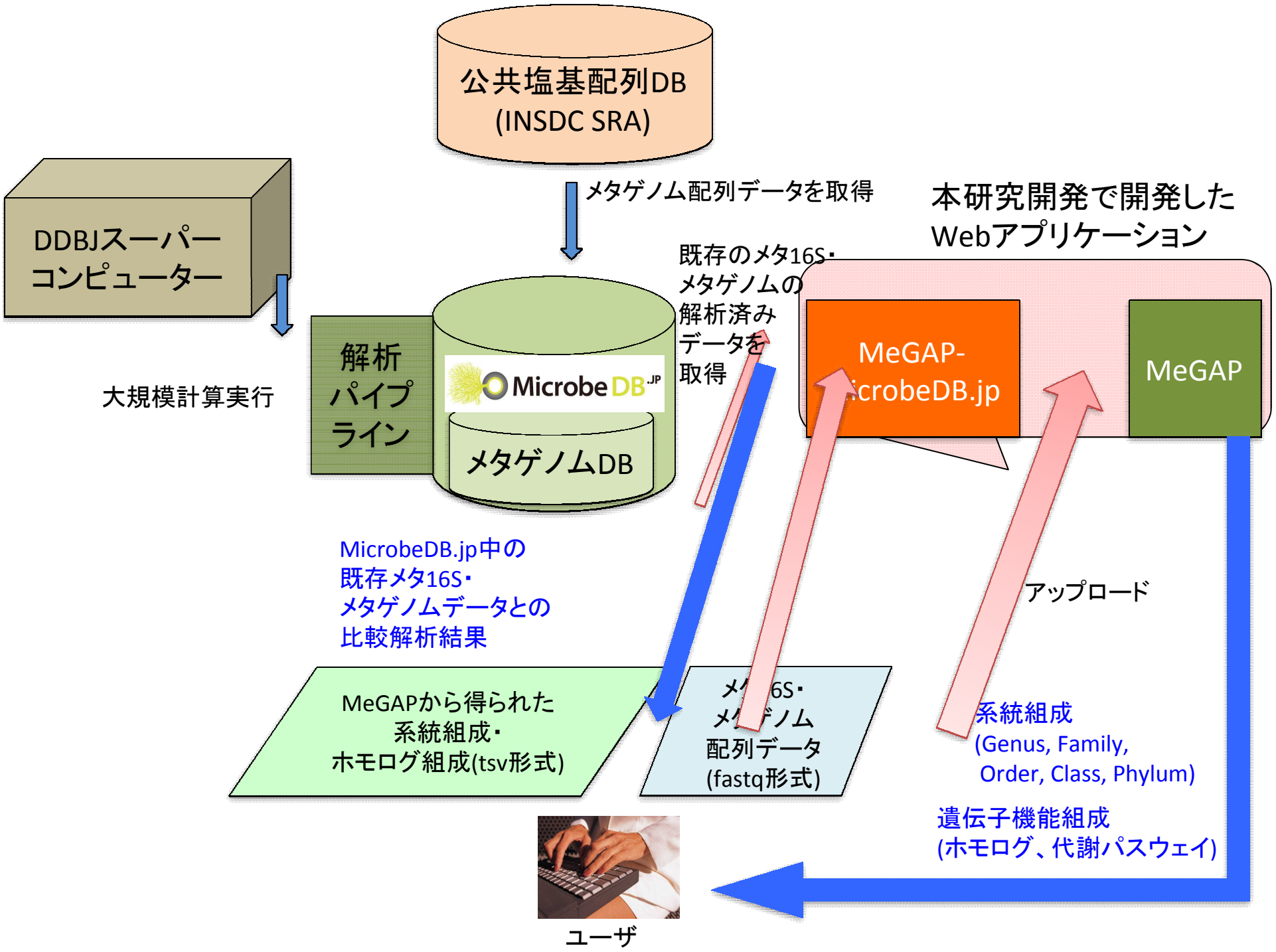
(5)受賞・報道等

なし

§6 自己評価

本研究開発で開発した MeGAP と MeGAP-MicrobeDB.jp は、バイオインフォマティクスの知識をほとんど持たない研究者でも、fastq ファイルをアップロードしてボタンを押すだけで利用可能である。今後も継続的に両ツールの改良を行って行き、潜在的なユーザが多いと考えられる微生物学系の学会で発表すると共に早急に論文化して宣伝し、多くの研究者に使ってもらうことによって、細菌群集の研究を行う上で標準的な手法となりつつあるメタ 16S・メタゲノム解析においてボトルネックになっていた、バイオインフォマティクスの解析の部分が、大きく改善できると期待している。しかしながら、MeGAP と MeGAP-MicrobeDB.jp はまだまだプロトタイプであり、例えば入力を fasta 形式にも対応できるようにすることや、特にメタゲノム解析について、アルゴリズムを改良してより高速にする、巨大な配列データにどのように対応するのか等、より多くのユーザに使ってもらうツールにするためには今後の継続的な改良が必須である。

以上


公共塩基配列DB
(INSDC SRA)
メタゲノム配列データを取得
本研究開発で開発した
Webアプリケーション
DDBJスーパー
コンピューター
既存のメタ16S・
メタゲノムの
解析済み
データを
取得
MeGAP-
MicrobeDB.jp
MeGAP
解析
パイプ
ライン
MicrobeDB.JP
メタゲノムDB
大規模計算実行
MicrobeDB.jp中の
既存メタ16S・
メタゲノムデータとの
比較解析結果
アップロード
MeGAPから得られた
系統組成・
ホモログ組成(tsv形式)
メタ16S・
メタゲノム
配列データ
(fastq形式)
系統組成
(Genus, Family,
Order, Class, Phylum)
遺伝子機能組成
(ホモログ、代謝パスウェイ)
ユーザ