平成 23 年度　研究開発実施報告書

ライフサイエンスデータベース統合推進事業「統合化推進プログラム」
平成 23 年度採択　研究代表者

黒川顕
東京工業大学大学院生命理工学研究科・教授

ゲノム・メタゲノム情報を基盤とした微生物 DB の統合

§1. 研究実施体制

(1)「東工大」グループ
① 研究代表者:黒川　顕 （東京工業大学・大学院生命理工学研究科, 教授）
② 研究項目
・メタゲノムデータベースの構築
・メタデータの整備
・スパコンにおける解析システムの開発および実装
・微生物統合 DB「MicrobeDB.jp」の構築

(2)「遺伝研」グループ
① 主たる共同研究者:中村　保一 （国立遺伝学研究所・生命情報・DDBJ 研究センター, 教授）
② 研究項目
・微生物ゲノム基盤情報資源の共用化
・微生物ゲノムアノテーションリファレンスの整備と共用化
・菌株保存情報の整備

(3)「基生研」グループ
① 主たる共同研究者:内山　郁夫 （基礎生物学研究所・ゲノム情報研究室, 助教）
② 研究項目
・比較ゲノム解析に立脚した微生物ゲノム情報の統合化

§2. 研究実施内容

【研究の目的】

ゲノム科学の発展に伴い，微生物のゲノムやメタゲノムなど圧倒的な量のデータが産出されており，これらを横断的にかつ簡便に利用できれば，新たな仮説や研究分野の創出がより容易になると期待される。これを実現するため，本研究開発では，ゲノム情報を核として様々な微生物学上の知識を統合し，幅広い分野での微生物学の発展に資することのできる「微生物エンサイクロペディア：MicrobeDB.jp」の構築を目標とする。

【概要】

多様性を特徴とする微生物においては，蓄積されたデータや知識は膨大かつ多様であり，ゲノムやメタゲノムなどの大規模データも多数産出されていることから，これらを横断的にかつ容易に利用する状態にはない。本研究開発では，ゲノム情報を核として様々な微生物学上の知識を統合し，幅広い分野での微生物学の発展に資することのできる微生物統合 DB「MicrobeDB.jp」を構築している。具体的には，国内外に散在する細菌の各種オミックス情報を広く収集し，遺伝子，ゲノム，環境の3つの軸に沿って遺伝子機能，分類学的情報，菌株保存情報，表現型情報などの知識を整理し，ゲノム情報を核としてセマンティック Web の技術を積極的に取り入れる事で統合する。構築したシステムを微生物学研究者に活用してもらうためのインタフェイスおよびアプリケーションを実装することで，特に微生物学分野のオミックス研究の発展に資することを目標とする。

【進捗状況，研究成果】

1) メタデータの整備

メタゲノムデータのうち入手可能な配列情報およびメタデータを徹底的に集積した上で、すべてのデータに対してメタゲノム解析パイプラインで、統一的なアノテーションを付加する。H23 年度は、これらメタゲノムデータのうちメタデータに焦点を絞り開発を実施した。特に、微生物の環境横断的なメタデータ定義を表現可能な新たなオントロジー「MEO」を新規に開発し、国内外のすべてのメタゲノムデータのメタデータを MEO にマッピングするとともに、TripleStore に実装し SPARQL による推論検索を実現した。MEO は 1,318,245 タームから構成され、ファイルサイズは 1.5GB である。また、MEO はゲノム情報の記述を標準化することを目的とした国際コンソーシアム GSC との協調を視野に入れており、DarwinCore、MIxS、ABCD、WFCC などとの連携を模索した。

2) スパコンにおける解析システムの開発および実装

メタゲノムデータはデータ量が膨大であり、ひとつのメタゲノムプロジェクトで産出されるデータが 1TB にもなる。そこで、メタゲノムデータのアノテーションには、東工大スパコン TSUBAME2.0 や DDBJ スパコンなどの超高速大規模計算機を積極的に利用する必要がある。H23 年度は、マルチ GPU 対応の塩基配列相同性検索ソフトウェア CLAST を開発し、既存のゲノムレファランス配列に対するメタゲノム配列マッピングを高速に実現することができた。CLAST は、BLAST と同様の検索精度を実現しつつ BWA などマッピングツールと同様の速度で検索可能となっている。また、新型シーケンサーによる情報爆発にも対処可能な、TSUBAME2.0 上で動作するメタゲノム解析パイプラインを構築した（研究協力者：東工大・秋山泰教授）。本解析パイプラインには、相同性検

索を高速化するだけでなく、検索結果を高度に統計処理するパイプラインも実装している。

3) 微生物統合 DB「MicrobeDB.jp」の構築

微生物統合 DB「MicrobeDB.jp」のプロトタイプを立上げ、限定されたデータのみでの運用を開始した。

4) 微生物ゲノム基盤情報資源の共用化

a. GTPS2011 版の作成と RDF 化

国際塩基配列データベースから公開されている微生物完全ゲノム配列 3,250 件を対象に、再アノテーションを施したデータベース GTPS を構築し微生物ゲノムの基盤情報として整備した。解析結果は ftp で公開している。さらに、RDF 化は DBCLS などの支援を受けて、GTPS2011 の RDF 化を行い、セマンティック Web 技術による、GTPS と微生物統合データベースにおける他の要素データベースとの連携を可能にした。生成した RDF は 197,070,005 トリプルの規模に至った。

b. 微生物統合データベースへ向けた菌株データベースの試作

製品評価技術基盤機構と理化学研究所のそれぞれの菌株保存施設 NBRC ならびに JCM の協力を得て微生物統合データベースの実現に適合した菌株データベースを構築した。具体的には、菌株のアクセション番号を Subject として、17,367 株のデータを RDF へと変換した。その結果、菌株データと微生物統合データベースにおける他の要素データベースとの連携も可能にした。

c. 日本産菌類集覧の構造化作業

日本菌学会関東支部の協力を得て刊行物「日本産菌類集覧」からの菌類データの機械抽出を試みた。*Abortiporus* 属から *Botryobasidium* 属までについて分類学的情報とともに、種の国内採取記録文献ならびに種の宿主・基物を抽出するルールを構築することができ、今後の微生物データ掘り起しの実行可能性を示すことができた。

5) 微生物ゲノムアノテーションリファレンスの整備と共用化

これまでかずさ DNA 研究所にて開発・維持されてきた、ソーシャルブックマークシステムによるデータ集積と整理を簡便に行う広域ゲノムアノテーション支援システムである「KazusaAnnotation」と、その基盤となる微生物ゲノムデータベースである「MicrobeBase」を国立遺伝学研究所内のサーバに移転するとともに拡張を実施し、本研究開発に於けるゲノム基盤として整備した。その際 KazusaAnnotation は TogoAnnotation へ名称ならびに URL を変更し、種々の最新更新情報を明示するようインタフェイスを改善した。 続いて、本システムを用いて、リファレンスとして重要な菌株あるいは現象について、信頼性の高いマニュアルキュレーションに基づいたアノテーションの高度化を試行した。すでに大腸菌ならびに枯草菌では遺伝子に関する関連文献が集積されたデータベースが複数存在しておりそれらの情報の連結や活用が今後のデータベース統合に必要であるが、目下文献情報の蓄積が不足しており整備が急務であると判断された放線菌 *Streptomyces griseus* IFO13350 株を対象として、遺伝子および遺伝子セットについて言及した文／節を抽出（sentence extraction）、遺伝子の名前をタグとしてさらに抽出 (named entity extraction) を手動で行うアノテーション・キュレーション方法を設計し、これまでに 22 報文から 2166 注釈を登録し公開した。今後は MicrobeBase への入力とデータ管理機能の整備により、GTPS 等から外部のゲノムデータとの連繋・統合をすすめる。

6) 比較ゲノム解析に立脚した微生物ゲノム情報の統合化

微生物ゲノムデータベース MBGD で構築しているオーソログテーブルに基づいて、種々の微生物情報を統合するための基盤を確立する。このため、比較の基盤となる標準オーソログテーブルの高品質化を進めると共に、他グループと連携して、遺伝子の機能や微生物の表現型などの知見をオーソログ解析に基づいて相互に比較し、ゲノムの特徴付けなどに用いる方法などについて検討する。合わせて、微生物ゲノムデータベースのさらなる大規模化に備えたオーソログテーブルの更新体制を確立する。23 年度は基礎となるゲノムデータについて、GTPS を含む複数のデータベースを対応づけて取り込めるようにするとともに、オーソログ分類の際のドメイン分割の改善方法、並びにオーソログの効率的な更新を行うための差分更新手続きについての検討を行った。

a. ゲノムデータの対応付け

GTPS/GenBank/RefSeq の3つのデータベースについて、染色体レベル、および遺伝子レベルでの対応付けを行った。この対応付けに基づいて、本プロジェクトでの連携の中心となる GTPS を標準遺伝子セットとして、不足する情報を RefSeq および GenBank から補う形でデータ更新を行う手続きを作成した。

b. オーソログ分類の精密化

従来の MBGD で作成されている DomClust を用いたオーソログ分類結果を、マルチプルアライメントを用いて改善する。23 年度は DomClust の売りの一つであるドメイン単位のオーソログ対応について、マルチプルアライメント上のスコアを用いて分類結果を評価し、ドメインの融合や、ドメイン境界の修正を行って分類を改善する基本的な手続きについて検討した。その結果、こうした方法で効果的な改善が可能であることが分かったため、現在基本手続きを組み合わせた、全体的な改善手続きの作成を進めている。

c. オーソログの差分更新手続き

大規模なオーソログ解析を効率的に行うため、DomClustno の出力に新たなデータを加えて差分的にオーソログを更新する手続き(MergeTree)を作成した。これを用いて、新規のゲノムやメタゲノムデータのオーソログ解析に基づくアノテーション付けを行う手続きを作成し、ゼロからオーソログ解析を行う場合との違いなどについて検討した。

【今後の見通し】

セマンティック Web 技術によるデータベースの統合化をさらに促進するために，新たなオントロジーの構築をすでに開始している。また，本研究開発に関与するすべてのデータの RDF 化も検討しているが，それらを収納する TripleStore の処理能力を検討しつつ，MicrobeDB.jp プロトタイプのさらなる拡充を図る予定である。

§3. 成果発表等

(3—1)データベースおよびウェブツールの構築と公開

① 公開中のデータベース・ウェブツール等

データベース名： MicrobeDB.jp

概要：本 DB は，微生物に関する多種多様な情報を遺伝子・系統・環境の 3 つの軸に沿って整理統合し，セマンティック Web 技術を利用して単一の検索ウィンドウからそれらの情報を検索可能な統合 DB である。

公開日：H22 年 12 月 12 日

URL：http://microbedb.jp

データベース名： MEO

概要：本 DB は，微生物の生息環境に関するメタデータを記述し整理するためのオントロジーである MEO (Metagenome/Microbes Environmental Ontology)の OWL ファイルと，公共のメタゲノムデータのメタデータを MEO 相手にマッピングしてサンプルごとに整理した結果の RDF ファイルの 2 つから構成された DB である。

公開日：H23 年 1 月 31 日

URL：http://mdb.bio.titech.ac.jp/meo

データベース名：Human Meta BodyMap

概要：世界中で公開されているヒトメタゲノム解析で得られた配列情報を，独自に注意深く再アノテーションしたメタデータに基づき参照できるようにした DB。メタデータ検索機能や配列相同性検索ツール Body-BLAST を用いたメタデータ検索機能も提供している。

公開日：H24 年 3 月 25 日

URL:http://metagenomics.jp/mg/

データベース名：GTPS

概要：本 DB では微生物ゲノムの最新情報を基にした再アノテーション情報を提供している。

公開日：H21年3月（2011年度版はH24年3月）

URL： ftp://gtps.ddbj.nig.ac.jp/2011/ (2011年度版ファイル)

アクセス数：

http版：公開日からH24年2月2日現在　　681,881回

ftp版(2011年版)：公開日からH24年3月31日現在　　32,363回

データベース名：GTPS/RDF

概要：本 DB はセマンティック Web 技術によって，GTPS と微生物統合データベースにおける他の要素データベースとの連携を可能とした GTPS2011 の RDF 化版である。

公開日：H24 年 4 月 26 日

URL： http://gtps.ddbj.nig.ac.jp/rdf/

データベース名：CyanoBase (MicrobeBase)
概要：シアノバクテリアに代表される酸素発生型光合成細菌とその関連バクテリアのゲノム情報を集積したデータベース。
公開日：H7 年 12 月
URL：http://genome.microbedb.jp/cyanobase
アクセス数：
H12 年 1 月から H24 年 4 月 26 日現在　　34,080,424 回

データベース名：RhizoBase (MicrobeBase)
概要：窒素固定植物共生細菌である根粒菌(共生関連領域のみの部分配列を含む) ゲノム情報を集積したデータベース。
公開日：H13 年 1 月
URL：http://genome.microbedb.jp/cyanobase
アクセス数：
H13 年 1 月から H24 年 4 月 26 日現在　　8,212,584 回

データベース名：TogoAnnotation
概要：ソーシャルブックマークによる文献情報集積プラットフォーム
公開日：H19 年 6 月 1 日（H24 年 3 月 6 日 TogoAnnotation として公開）
公開サイト：http://togo.annotation.jp
アクセス数：
H20 年 4 月から H24 年 4 月 26 日現在　　578,544 回（ユニークユーザ数 15,475）

データベース名：MBGD
概要：オーソログ解析に基づいて微生物ゲノムの比較解析を行うためのデータベース。公開されたゲノム全体を含む標準オーソログテーブルに基づいて、各オーソロググループの系統プロファイルの比較などを行えるほか、動的なオーソログ解析機能によって、利用者自身が持つゲノム配列も含めて、興味のある生物種セットに対象を絞った比較を行うこともできる。
公開日：1997 年 7 月 1 日
URL：http://mbgd.genome.ad.jp/
アクセス数：（2011年度のアクセス数をガイドラインに沿った形で集計したもの）
ユニークアドレス数：18,868　ページ数：785,743

（3－2） 原著論文発表

① 発行済論文数（国内（和文） 0 件, 国際（欧文） 0 件）：
② 未発行論文数（“accepted”, “in press”等）（国内（和文） 0 件, 国際 （欧文）0 件）
③ 論文詳細情報