

平成 24 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」

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(研究機関別)

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
- ・菌株データの RDF 化
- ・微生物統合 DB「MicrobeDB.jp」の構築

(2) 遺伝研グループ(研究機関別)

① 主たる共同研究者: 中村 保一 (国立遺伝学研究所・生命情報研究センター・大量遺伝情報研究室・教授)

② 研究項目

- ・微生物ゲノム基盤情報資源ならびにアノテーションリファレンスの整備と共用化

(3) 「基生研」グループ(研究機関別)

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ー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 H23 年度に開発済みの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。H24 年度は、微生物の環境横断的なメタデータ定義を表現可能なオントロジー「MEO」を、ドメインオントロジーではなく MicrobeDB.jp によるアプリケーションオントロジーとして再設計を実施し、徹底した軽量化を実現した(ver. 0.3)。また、配列情報入手可能な 24,051 のメタゲノムデータに MEO マッピングを施すとともに、メタゲノム解析パイプラインによりアノテーションを実施した。

2) 菌株データの RDF 化

NBRC(約 16,000 株)および JCM(約 14,000 株)の菌株保存データの RDF 化を実施した。これら2つの菌株データ群はそれぞれが 15 項目のデータから構成されているが、菌株データは配列データとは完全に独立しているため、ゲノムデータなど他のデータとの統合は不可能である。菌株データをゲノムデータ等と統合するために、新たなオントロジー MCCV (Microbial Culture Collection Vocabulary) を設計し、NBRC/JCM の菌株 ID と NCBI Taxonomy ID 間を名寄せした上で菌株 RDF を作成した。この事により、菌株データとゲノムデータ等が完全に統合され、MicrobeDB.jp により横断的な検索が可能となった。

3) 微生物統合 DB「MicrobeDB.jp」の構築(文献 1)

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。本 DB 利用の利便性を向上させるため、上述の MEO, MCCV 以外に以下に挙げる各種

オントロジーを構築した。

- ① FALDO (Feature Annotation Location Description Ontology):ゲノム中の各 feature の位置情報を記述するため (DBCLS)
- ② INSDC Ontology:INSDC エントリの feature と qualifier のターム記述のため
- ③ PDO-SYMP (Pathogenic Disease Ontology with Symptom):細菌が引き起こす感染症の情報および感染症の症状を連結
- ④ GMO (Growth Media Ontology):細菌の培地情報を記述するため (DBCLS)

さらに、MicrobeDB.jp における検索結果を表現するための Stanza の開発に着手した。Stanza は、再利用性、モジュール化、拡張が容易、という点で極めて優れたデータテンプレート(アプリケーション)であるため、統一した仕様となる TogoStanza を、MicrobeDB.jp ならびに Cyanobase をユースケースとして DBCLS と共に策定した。H24 年度中には、約 15 種類の Stanza を開発し、MicrobeDB.jp に実装した。

4) 微生物ゲノム基盤情報資源の共用化

国立遺伝学研究所生命情報研究センターでは、公開微生物ゲノム情報を集約した Genome Information Broker (GIB)-M を構築し提供を続けてきた。平成 24 年度は、前年度までに引き続き、ゲノム配列が決定された菌株情報を公開微生物ゲノム情報を定期的に再アノテーションし全 ORF の品質評価を加えた微生物ゲノムの二次情報資源データベースである Gene Trek in Prokaryote Space (GTPS) を GIB-M をもとに年次更新するとともに、その RDF 化による他の DB との他の要素データベースとの連携を実現した。

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

H24 年度には、ゲノム DB に対してデータ集積と整理を簡便に行うための広域アノテーション支援システム TogoAnnotation (<http://togo.annotation.jp>) の共用化をすすめるための開発を引き続き実施した。具体的には昨年度整備した API を活用し、微生物オミックスデータを格納しまた表現するための拡張開発を行った。具体的には、統合認証クライアントアプリケーションを開発し、Excel ファイルによる TogoAnnotation へのバルク情報インポート機能を追加し文献データ集約によるデータベース高度化を加速するとともに、MicrobeBase の GeneView および GeneSetView ページへのオミックスデータベースエントリの表示機能の実装と MicrobeBase の検索インデックス作成データフローの集約整備および RDF リソースから検索インデックス作成フローを追加することで、近年、微生物に於いても高速大量に集積されつつあるオミックスデータのゲノム情報への統合を支援するための環境整備を実施した。同時に、これまで蓄積してきたシアノバクテリアならびに放線菌の論文情報の収集と更新(それぞれ 418 件、146 件)をすすめるとともに、放線菌については研究者コミュニティと協力してオミックスデータを取り扱うデータ統合整備を実施した。また、それ以外の我が国発のモデル微生物種群リファレンス情報として大腸菌、枯草菌をこの系に追加するための高速な論文収集モデルの開発をすすめた。

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(文献 3)

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。24年度は、前年度に引き続いてゲノムデータソースの統合およびオーソロググループの精密化を進めるとともに、系統群ごとにオーソログ情報をあらかじめ整理しておき、参照できるシステムの開発を行った。合わせて、グループ間の連携に向けて、オーソロググループ情報のRDF化を進めた。

a. ゲノムデータの整備

昨年度作成した、GTPS/GenBank/RefSeqの3つのデータベースを染色体レベルおよび遺伝子レベルで対応付けしてマージする手続きを用いて、MBGDの基となるゲノムデータベースをこの統合データを用いるように改変した。RefSeqに含まれないゲノムも取り込んだことにより、ゲノム数が昨年度末の1532から今年度末は2326と大幅に増加した。

b. 系統群ごとのオーソログ情報の提供

従来作成していた、微生物界全体をカバーする標準オーソログテーブルに加えて、系統群ごとのオーソログテーブルを事前に作成して提供するようにした。その際、科以下の近縁系統群については、当グループで開発したCoreAlignerプログラムを用いて、系統群内でシンテニーが保存された領域のコンセンサスな構造を構築して表示する機能も加えた。これらの結果については、タキソミーツリーに沿って利用者が興味のある系統群を選択して切り替えて表示できるようにした。

c. オーソロググループの精密化

昨年度に引き続いて、DomClustで作成されたドメイン単位のオーソロググループを、マルチプルアライメントを用いて改良する方法の開発を行った。隣接ドメインを含むオーソロググループ対について、マージまたは境界を変更した再分割を行う。その際、昨年度開発したスコアリングシステムを用いて、スコアの和が最大になるような分割を行う。得られたグループについて、系統樹上の重複ノードを切断していく形でさらにグループ分割を行うことによってオーソロググループを得る。この結果をCOGなどのキュレーションされたデータベースを使って評価し、多くの場合に期待通り改善されていることを確認した。今後、より大規模なデータに適用するための効率化を進め、最終的にはMBGD標準オーソログテーブルの精密化に適用する予定。

d. オーソロググループ情報のRDF化

MBGDに含まれるオーソロググループや関連情報のRDF化を行った。SPARQLエンドポイントを立ち上げ、標準オーソロググループの情報を検索できるようにした。また、SPARQLを使って他のデータベースと連携させた検索や、タキソミーを使って保存されたオーソロググループを抽出する検索を行うことなどについて検討した。今後、RDFを用いた他グループのデータベースとの連携を進めると共に、RDFでオーソログ関係を表現するための共通概念を抽出して標準的なオントロジーとして定義することなども試みる予定。

§3. 成果発表等

(3-1) 原著論文発表

- ① 発行済論文数(国内(和文) 0件、国際(欧文) 3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

1. Éamonn Ó Tuama, John Deck, Gabriel Dröge, Markus Döring, Dawn Field, Renzo Kottmann, Juncai Ma, Hiroshi Mori, Norman Morrison, Peter Sterk, Hideaki Sugawara, John Wieczorek, Linhuan Wu, and Pelin Yilmaz, "Meeting Report: Hackathon-Workshop on Darwin Core and MIxS Standards Alignment (February 2012)", *Standards in Genomic Science*, Vol.7, No.1, 2012, (DOI: 10.4056/sigs.3166513)

ゲノムやメタゲノムデータに付属するメタデータ記述の国際標準チェックリストである、MIxS と博物館資料のメタデータの記述のチェックリストである Darwin Core をどのように対応付けるかを話し合った会議の Meeting Report であり、微生物のゲノムやメタゲノムデータを蓄積するデータベースの一つとして、本プロジェクトで開発中の DB である MicrobeDB.jp の名前も挙げられている。

2. Okubo T, Tsukui T, Maita H, Okamoto S, Oshima K, Fujisawa T, Saito A, Futamata H, Hattori R, Shimomura Y, Haruta S, Morimoto S, Wang Y, Sakai Y, Hattori M, Aizawa S, Nagashima KV, Masuda S, Hattori T, Yamashita A, Bao Z, Hayatsu M, Kajiya-Kanegae H, Yoshinaga I, Sakamoto K, Toyota K, Nakao M, Kohara M, Anda M, Niwa R, Jung-Hwan P, Sameshima-Saito R, Tokuda S, Yamamoto S, Yamamoto S, Yokoyama T, Akutsu T, Nakamura Y, Nakahira-Yanaka Y, Takada Hoshino Y, Hirakawa H, Mitsui H, Terasawa K, Itakura M, Sato S, Ikeda-Ohtsubo W, Sakakura N, Kaminuma E, Minamisawa K., "Complete genome sequence of *Bradyrhizobium* sp. S23321: insights into symbiosis evolution in soil oligotrophs", *Microbes Environ*, vol. 27, No. 3, pp 306-315, 2012 (DOI: 10.1264/jsme2.ME11321)
3. Ikuo Uchiyama, Motohiro Mihara, Hiroyo Nishide, Hirokazu Chiba. "MBGD update 2013: the microbial genome database for exploring the diversity of microbial world" *Nucleic Acids Res*, vol 41, pp. D631-D635, 2013. (DOI: 10.1093/nar/gks1006)

(3-2) データベースおよびウェブツール等の構築と公開

公開中のデータベース・ウェブツール等
別紙を参照

(3-3) その他の著作物(総説、書籍など)

1. 森宙史, 山田拓司, 黒川顕, *メタゲノム解析の現状と将来 知識データベースの開発, 情報管*



(3-4) 国際学会および国内学会発表

① 招待講演 (国内 4 件、国際 2 件)

〈国内〉

1. 黒川顕, Development of an integrated analysis system for metagenomics and a global microbial database MicrobeDB.jp, 第 28 回日本微生物生態学会シンポジウム, 豊橋, 9 月 20 日
2. 黒川顕, メタゲノミクスの現状と未来, 情報・システム研究機構シンポジウム 2012「生命科学のビッグデータ革命」, 東京, 11 月 9 日
3. 黒川顕, メタゲノム解析の現状と将来, 社団法人企業研究会 CAMM フォーラム, 東京, 1 月 11 日
4. 内山郁夫, メタゲノム情報と系統分類, 日本進化学会第14回大会, 首都大学東京南大沢キャンパス, 8 月 21-23 日

〈国際〉

1. Ken Kurokawa, Deciphering ancient ecosystems with environmental genomics and “Earth-Life DB”, The 1st ELSI International Symposium, 東京, 3 月 29 日.
2. Yasukazu Nakamura, Integration of databases for microbes and plants from the viewpoint of (meta-)genomics, Joint Conference on Informatics in Biology, Medicine and Pharmacology (生命医薬情報学連合大会), 東京, 10 月 15 日

② 口頭講演 (国内 8 件、国際 0 件)

〈国内〉

1. 森宙史, 内山郁夫, 菅原秀明, 中村保一, 黒川顕, MicrobeDB.jp プロジェクトチーム, 微生物統合データベース MicrobeDB.jp, 第 7 回日本ゲノム微生物学会年会, 長浜バイオ大学, 3 月 8-10 日
2. 山本希, 森宙史, 中村保一, 内山郁夫, 菅原秀明, 黒川顕, MicrobeDB.jp プロジェクトチーム, 微生物統合データベース MicrobeDB.jp の開発, 日本農芸化学会 2013 年度大会, 東北大学川内北キャンパス, 3 月 24-27 日
3. 藤澤貴智, 中村保一, TogoAnnotation を用いたゲノムアノテーションの利用とデータ共有, ラン藻ゲノム交流会 2012, 駒場, 7 月 14 日
4. 藤澤貴智, 照井敬子, 桧原直子, 山田佐知子, 石井英治, 加藤香奈, 手塚武揚, 大西康夫, 神沼英里, 岡本忍, 黒川顕, 中村保一, 放線菌ゲノムアノテーションプラットフォームの構築とリファレンス情報の統合, 日本農芸化学会 2013 年度大会, 仙台, 3 月 27 日
5. 千葉啓和, 内山郁夫, 微生物ゲノムデータベース MBGD におけるオーソログ分類の改良, 日本進化学会第 14 回大会, 首都大学東京南大沢キャンパス, 8 月 21-23 日
6. 内山郁夫, シンテニーの保存性に基づく微生物系統群のコアゲノム構造, 第 35 回日本分子

生物学会年会, 福岡国際会議場, 12月 11-14日

7. 千葉啓和, 内山郁夫, マルチプルアラインメントを用いたドメイン単位のオーソログ分類の改良, 生命情報科学若手の会, 岡崎コンファレンスセンター, 3月 1-3日
8. 内山郁夫, 西出浩世, 三原基広, 千葉啓和, シンテニーに基づくコアゲノムデータベース, 第7回日本ゲノム微生物学会, 3月 8-10日

〈国際〉

なし

③ ポスター発表 (国内 10件、国際 2件)

〈国内〉

1. 森宙史, 岡本忍, 川島秀一, 竹原潤一, 吉野弘二, 山本希, **MicrobeDB.jp** プロジェクトチーム, 微生物の生息環境データの記述および高度な検索のためのオントロジー構築, トーゴーの日シンポジウム 2012, 時事通信ホール, 10月 5日
2. 黒川顕, 中村保一, 内山郁夫, 菅原秀明, **MicrobeDB.jp** プロジェクトチーム, 微生物統合データベース **MicrobeDB.jp**, トーゴーの日シンポジウム 2012, 時事通信ホール, 10月 5日
3. 黒川顕, 中村保一, 内山郁夫, 菅原秀明, **MicrobeDB.jp** プロジェクトチーム, 微生物統合データベース **MicrobeDB.jp**, BioJapan2012, パシフィコ横浜, 10月 10-12日
4. 黒川顕, 中村保一, 内山郁夫, 菅原秀明, **MicrobeDB.jp** プロジェクトチーム, 微生物統合データベース **MicrobeDB.jp**, 第35回日本分子生物学会年会, 福岡国際会議場・マリンメッセ福岡, 12月 11-14日
5. 黒川顕, 中村保一, 内山郁夫, 菅原秀明, **MicrobeDB.jp** プロジェクトチーム, 微生物統合データベース **MicrobeDB.jp**, 第7回日本ゲノム微生物学会年会, 長浜バイオ大学, 3月 8-10日
6. 黒川顕, 中村保一, 内山郁夫, 菅原秀明, **MicrobeDB.jp** プロジェクトチーム, 微生物統合データベース **MicrobeDB.jp**, 日本農芸化学会 2013年度大会, 東北大学川内北キャンパス, 3月 24-27日
7. 藤澤貴智, 岡本忍, 長崎英樹, 神沼英里, 菅原 秀明, 内山郁夫, 黒川顕, 中村保一, **TogoAnnotation: 研究コミュニティのためゲノムアノテーションプラットフォーム構築**, 第35回日本分子生物学会年会, 福岡, 12月 14日
8. 藤澤貴智, 神沼英里, 岡本忍, 菅原秀明, 内山郁夫, 黒川顕, 中村保一, **TogoAnnotation: 研究コミュニティ支援ゲノムアノテーションプラットフォームの整備とセマンティックウェブ技術利用の取り組み**, 第7回日本ゲノム微生物学会, 長浜, 3月 8-10日
9. 千葉啓和, 内山郁夫, 微生物ゲノムデータベース **MBGD** におけるオーソログ分類の改良, 第35回日本分子生物学会年会, 福岡国際会議場, 12月 11-14日
10. 千葉啓和, 内山郁夫, マルチプルアラインメントを用いたドメイン単位のオーソログ分類の改良, 第7回日本ゲノム微生物学会, 長浜, 3月 8-10日

〈国際〉

1. Takatomo Fujisawa, Hideki Nagasaki, Yasukazu Nakamura, Introduction of a



genome annotation platform for the *Marchantia* research community, *Marchantia* Workshop 2012, Aso, Kumamoto, Japan, November 15-17, 2012.

2. Toshiaki Katayama, Shinobu Okamoto, Shuichi Kawashima, Hiroshi Mori, Takamoto Fujisawa, Semantic Integration of the Genome Annotations, Semantic Web Applications and Tools for Life Sciences 2012, Paris, France, November 28-30, 2012.

(3-5) 知財出願

① 平成 24 年度特許出願件数(国内 0 件、海外 0 件)

② 平成 23、24 年度の累積件数(国内 0 件、海外 0 件)

③ その他の知的財産権

他に記載すべき知的財産権があればご記入下さい。(実用新案 意匠 プログラム著作権 等)
なし

(3-6) 受賞・報道等

なし

§4. 研究開発期間中に主催した活動(主催したワークショップ等)

年月日	名称	場所	参加人数	概要
2012年4月19日	チーム内オントロジー開発会議 (非公開)	DBCLS	6人	オントロジー開発のための会議
2012年4月24日	チーム内オントロジー開発会議 (非公開)	DBCLS	6人	オントロジー開発のための会議
2012年4月27日	チーム内DB開発会議(非公開)	東京工業大学 すずかけ台キャンパス	4人	DB開発のための会議
2012年5月10日	チーム内オントロジー開発会議 (非公開)	DBCLS	9人	オントロジー開発のための会議
2012年5月30日	チーム内第5回 全体会議(非公開)	DBCLS	19人	研究進捗報告のための会議
2012年7月5日	チーム内DB開発会議(非公開)	東京工業大学 すずかけ台キャンパス	4人	DB開発のための会議
2012年7月13日	チーム内第6回 全体会議(非公開)	DBCLS	19人	研究進捗報告のための会議
2012年8月2日	チーム内DB開発会議(非公開)	DBCLS	6人	DB開発のための会議
2012年8月16日	チーム内DB開発会議(非公開)	DBCLS	5人	DB開発のための会議
2012年8月30日	チーム内オントロジー開発会議 (非公開)	DBCLS	4人	オントロジー開発のための会議
2012年9月14日	チーム内DB開発会議(非公開)	東京工業大学 すずかけ台キャンパス	4人	DB開発のための会議
2012年10月3日	チーム内DB開発会議(非公開)	DBCLS	6人	オントロジー開発のための会議
2012年10月19日	チーム内第7回 全体会議(非公開)	DBCLS	17人	研究進捗報告のための会議
2012年10月25日	チーム内DB開発会議(非公開)	DBCLS	9人	DB開発のための会議
2012年11月21日	チーム内DB開発会議(非公開)	東京工業大学 すずかけ台キャンパス	4人	DB開発のための会議
2013年1月14日	チーム内DB開発会議(非公開)	東京工業大学 田町キャンパス	6人	DB開発のための会議
2013年3月15日	チーム内DB開発会議(非公開)	DBCLS	8人	DB開発のための会議