メタゲノムオーソログ遺伝子統合解析システムの開発

東京工業大学大学院生命理工学研究科
黒川　顕

Development of an integrated analysis system for metagenomics

Ken Kurokawa
Graduate School of Bioscience and Biotechnology, Tokyo Institute of Technology

Microbes are essential for every part of life on Earth. Numerous microbes inhabit the biosphere, many of which are uncharacterized or uncultivable. They form a complex microbial community that deeply affects against surrounding environments. Metagenome analysis provides a radically new way of examining such complex microbial community without isolation or cultivation of individual bacterial community members. However, metagenome analysis is more complex than common genome analysis, because an analysis target is composed of enormous bacterial strains instead of a single strain. Here, to untangle the complexity of metagenome analysis procedure, we developed an integrated analysis system for metagenomics with new gene prediction tool, visualizing tools for massive metagenomic data, and also developed an analysis pipeline for metagenomics by integrating with all the developed tools.

1. はじめに

環境中の細菌は、数百から数万種の集団（細菌叢）を形成し、細菌間相互作用だけでなく、環境と密接に関連しながら複雑なシステムを構成している。これら細菌叢をまるごとゲノム解析する「メタゲノム解析」により、環境の根幹を形成する細菌叢の生命システムを明らかにすることが可能となりつつある。メタゲノム解析では極めて大量の配列データが公開されるが、これら大量のメタゲノム遺伝子を比較し、環境の違いによる生命システムの相違を見出すためには、まず未知の遺伝子を含むこれら大量の遺伝子を配列相同性にてクラスタリングし、かつその冗長度を見極めることが重要となる。本研究では、メタゲノム配列から遺伝子配列を予測、クラスタリングし、それらを既存のオーソログ遺伝子データベースと組み合わせ機能別にデータベース化することを目標とした。将来的には、環境中の様々な要素をインデックス化し、遺伝子クラスターと組み合わせて解析することで、生態系の根幹を形成する細菌叢と環境との因果関係に焦点を絞った解析、「エンバイロメントーム解析」が実現できると考えており、本研究はその解析基盤を形成する重要な研究になる。

また、本研究に着手しはじめた頃、大量の配列を一度にシークエンス可能な新型シークエンサーが登場し、あらゆる環境におけるメタゲノム解析が本格化した[1]。これら大量のメタゲノムデータから有用な知見を効率よく発見するためには、上述した通りメタゲノムデータからの遺伝子配列の予測およびクラスタリングや、それらを機能別にデータベース化、可視化する統合解析システムの

開発ならびに解析パイプラインの整備が必須となる。そこで我々は、メタゲノム配列断片からの遺伝子予測ソフトMetaGene（MG）およびその発展であるMetaGeneAnnotator（MGA）を開発し、入手可能なメタゲノムデータに応用して得られた大量の遺伝子群のデータベース化などを実施した。さらに、メタゲノムデータの各種可視化技術の開発、シークエンスにより得られた配列を既存のゲノム配列にマッピングし表示するシステム、KEGGに代表される代謝パスウェイマップへの分類群付加マッピング技術などを開発するとともに、本システムをヒト腸内メタゲノムデータに対して応用し世界に先駆けてヒトメタゲノム研究の成果として発表した[2, 3]。

2. 研究開発の成果

2.1 メタゲノム配列断片からの遺伝子予測法の開発

メタゲノム解析においても通常のゲノム解析同様、塩基配列から遺伝子コード領域を予測する事は、生物学的な知見を得る上で最も重要な解析のひとつである。細菌の遺伝子予測は隠れマルコフモデルなどの確率的手法をもちいるのが一般的である。これは全遺伝子配列が得られる近縁種が存在した場合、それら既知の全遺伝子配列情報を教師データとして学習することで非常に高い確度で新規遺伝子領域を予測することが可能となる。しかし、メタゲノム解析においては、多種多様な細菌叢由来の配列を対象としているため、教師データを一意に決定することができないため、予測確度が著しく低下する。

本研究が開始される直前に、共同研究者らによりメタゲノム断片配列から遺伝子を予測するソフトウェアMetaGene（MG）が開発された[4]。細菌の遺伝子配列においてはGC含量とdi-codonの使用頻度との強い相関が存在する。この傾向を利用してMGでは、入力された配列断片の由来が未知であっても、特異的な遺伝子群による学習をすることなく遺伝子領域を予測することが可能である。しかしMGは、典型的な細菌の遺伝子予測は可能であるが、プロファージなどの外来遺伝子は予測する事が困難であった。そこで我々はこのMGをさらに改善し、プロファージ遺伝子群の確立モデルを作成し、さらにRibosome binding site（RBS）の特性を加えることで、プロファージなどの外来遺伝子も含めより短い配列断片から遺伝子領域を正確に予測可能なMetaGeneAnnotator (MGA)を開発した[5]。

本ソフトウェアの予測確度は、MGよりも若干劣るものの非常に高く、ゲノム配列からの遺伝子予測において最も信頼性の高いソフトウェアの一つであるGeneMarkSやGlimmer3とほとんど同様の結果もしくはより高い予測精度を示した。また、700bpの断片配列からの遺伝子予測においても極めて高い精度で予測が可能であることが示せた。

2.2 メタゲノム解析法および可視化技術開発

2.2a ゲノムマッピング

ゲノム配列が既知である種の近縁種が細菌叢に存在する場合、メタゲノム解析により得られた細菌叢由来の配列断片を、すでに明らかになっているゲノム配列に対してマッピングすることで、細菌叢に含まれる当該種のゲノム多様性を論じる事が可能となる。例えば、ヒト腸内細菌叢のメタゲノム解析で得られた配列断片を、大腸菌K-12株のゲノム配列にマッピングすることで、標準株であるK-12株とは異なるヒト腸内に存在する大腸菌の特異性を特定することができる。メタゲノム解析では大量の配列断片が得られるが、新型シークエンサー由来の短い配列断片だけでなく、従来のSanger型シークエンサー由来の比較的長い配列断片においてもゲノム配列上の相同領域をより高速

に検索可能な相同性検索ソフトウェアを別途開発し[6]、ヒト腸内細菌叢メタゲノムデータを対象にゲノムマッピングによるゲノム多様性解析を実施した。また、マッピング結果を可視化するGUI付きソフトウェアも併せて開発した（**図１**）。

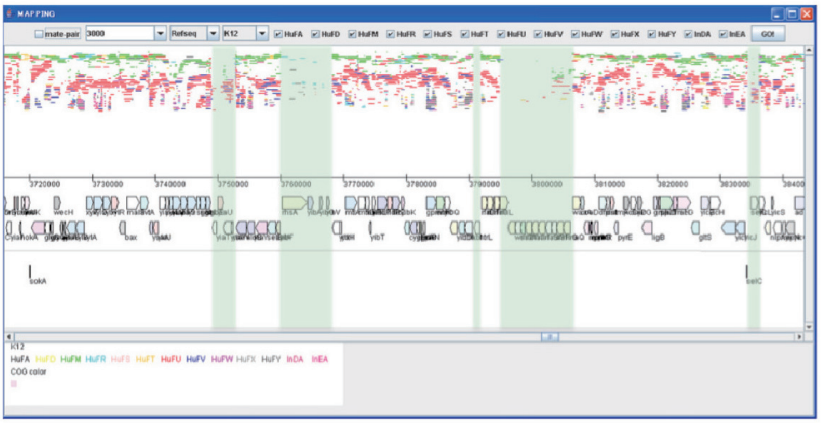

図１

2.2b 代謝パスウェイマッピング

メタゲノム解析は、群集構造を形成している細菌叢のすべての遺伝子を明らかにする事を可能とする解析手法である。個別菌のゲノム解析においては、明らかとなった遺伝子をKEGGなどの代謝パスウェイデータベースに対して検索する事により、どのような代謝経路を有しているかを解析することが可能となる。本解析手法をメタゲノムに応用し、細菌叢全体としてどのような代謝経路が存在するのか、さらには、集団内における細菌種間の代謝相互作用を解明するためには、メタゲノム解析により得られた遺伝子を、遺伝子由来の分類群情報を付加した上で代謝パスウェイマップ上にマッピングする必要がある。メタゲノム遺伝子の代謝パスウェイマップ上へのマッピングは、データベースに対する配列相同性検索により実施可能であるが、遺伝子由来の分類群を同定するために、①データベース配列検索におけるトップヒット、②データベース配列検索パターンのクラスタリング、③配列そのものの特徴から分類予測、の３つの手法を導入することで代謝パスウェイマップ上に分類群情報を付加したマッピングを実現した（**図２**）。

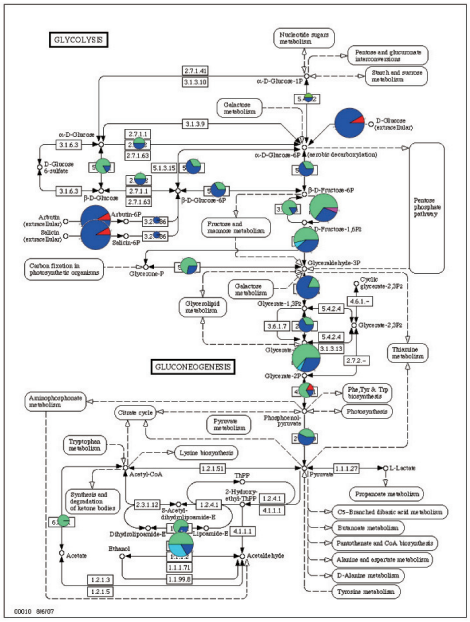

図２

2.2c 新規16S rRNA系統解析技術の開発

細菌叢のメタゲノム解析において、群集にどのような種が存在するかを明らかにする事が重要な課題のひとつとなっている。群集構造を種レベルで明らかにするためには、遺伝子配列から種を分類する際に良いマーカーとなる16S rRNA遺伝子を特異的にシークエンスし系統解析を実施する。しかしながら、新型シークエンサーの登場により膨大な配列が得られるようになり、これらを通常

の手法で系統解析する事が困難となりつつある。また、サンプル間での系統比較を実施する場合は、さらに情報量が増大するため、解析結果から有意義な知見を見出す事が困難となる。そこで本研究では、より容易にサンプル間の相違を表現する事を可能とする新たな系統解析手法および可視化技術を開発した。本手法は、配列の相違度を系統樹により表現するのではなく、配列間の進化距離を完全に反映しつつサークル状にプロットさせることで、細菌叢に存在する細菌種の全体像を得られるだけでなく、サンプル間の比較を容易にし、サンプル特異的な分類群も容易に見出すことを可能とした（**図３**）。

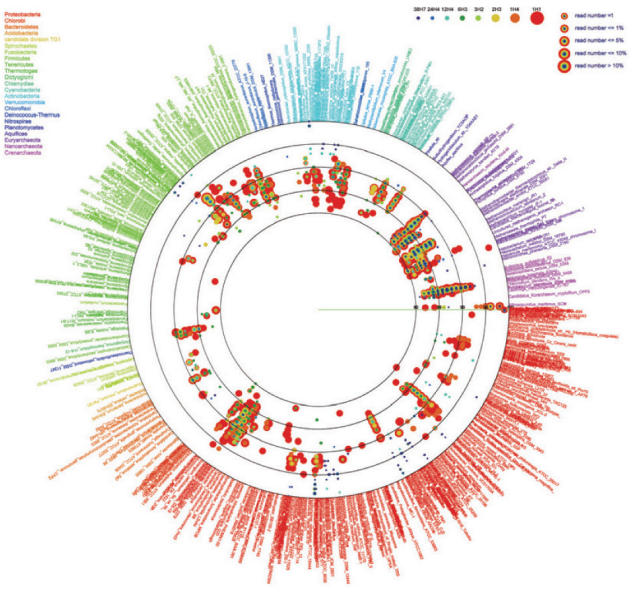

図３

2.3 ヒトメタゲノムボディマップ

新型シークエンサーによる大量16S rRNAシークエンスに対応するために、特にヒトメタゲノム研究に特化する形で、ヒトメタ16S rRNA BodyMapを併せて開発した。本システムは、現在入手可能なヒト由来バクテリア16S rRNAデータ53,460本を収集し、大腸菌16S rRNA遺伝子に対応する位置、さらに分類情報を相同性に基づき注意深く再アノテーションし、これらすべてのメタデータを格納したデータベースとなっている。また、統合データベースセンターで開発されているアナトモグラフィーAPIを利用することで、細菌のヒト各部位における分布を視覚的に表現し、データベースに対する検索システムも搭載している（**図４**）。ヒトメタゲノム国際コンソーシアムを中心として、腸内、口腔内、皮膚や膣内などヒトの各部位におけるメタゲノムデータが急速に蓄積されており、16S rRNAのみならずこれらすべてのメタゲノムデータを本システムに取り込むことで、ヒトの口腔、腸、膣、皮膚など各部位における細菌叢を俯瞰できるだけでなく、各部位間での比較や他の環境との比較により、ヒト細菌叢の共生の進化を論じる事が可能となる。

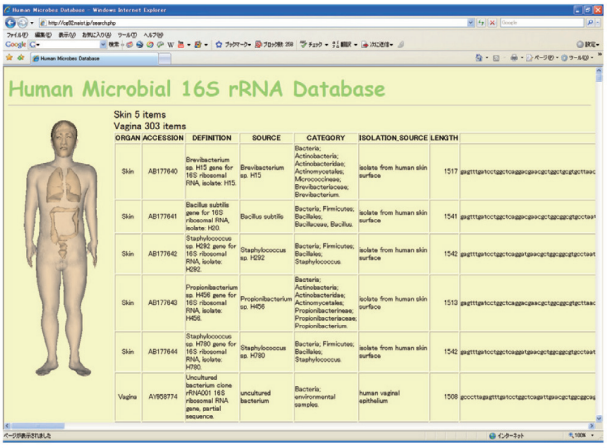

図４

2.4 メタゲノム解析パイプライン整備

本研究で開発した上述の技術をすべて結集し、メタゲノム研究における解析パイプラインを構築した（**図５**）。さらに、現在入手可能なメタゲノムデータ、特にヒトメタゲノムプロジェクトを中心に、得られたすべての配列を対象にMGAによる遺伝子予測をおこない、予測遺伝子群データベースを構築するとともに、メタゲノムコンティグビューア、Taxonomy逆引きビューア、クラスタリング解析結果ビューア、種分類情報可視化ビューアによる大量データの可視化など、すべての解析結果を可視化した上でウェブを通して公開した[7]。

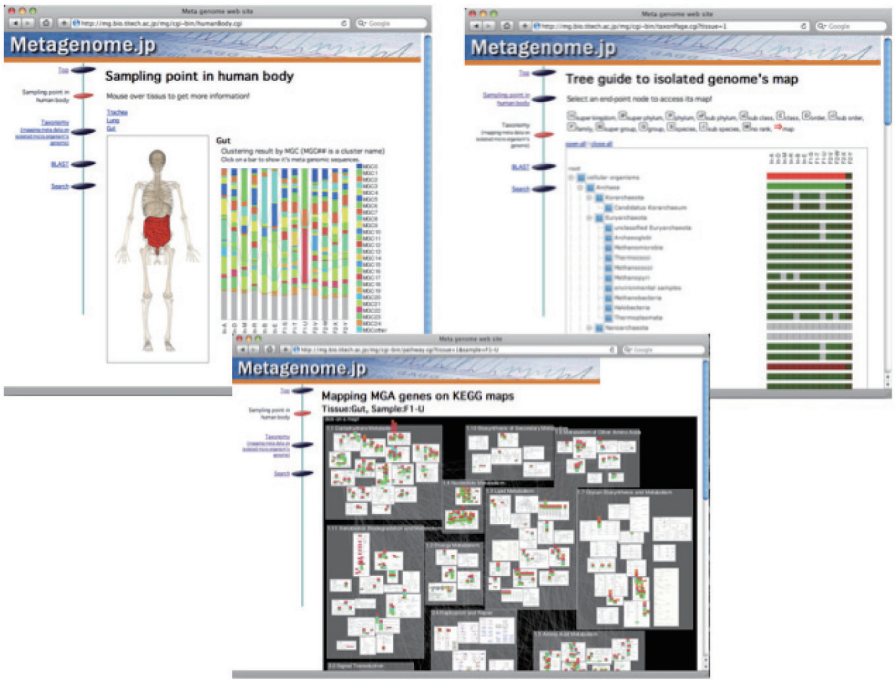

図５

3. まとめ

本研究では、環境中の培養困難な細菌叢をまるごと解析可能なメタゲノム解析における技術開発ならびに新技術を応用した解析パイプラインの整備を実施した。これら技術はメタゲノム解析のみならず、個別種を対象としたゲノム解析研究や宿主と共生している原虫および細菌叢の生物間相互作用の研究から、例えば食品加工における衛生管理、品質管理など広範な分野において応用可能である。

また、最初に述べたように、新型シークエンサーの登場により、これまで以上に猛烈な勢いでメタゲノム解析が実施されている現状では、新規技術に立脚した基盤システムが極めて重要となる。我々はすでに新型シークエンサーでのメタゲノム解析の可能性の検討に着手しており[8]、本研究で開発したシステムを新型シークエンサー由来のメタゲノムデータに応用することを目指し研究を継続している。今後さらなるデータ爆発に備えて、より高速に解析可能なアルゴリズムの開発ならびに本システムの拡張など含めて、本研究をさらに発展させていきたいと考えている。

4. 研究開発実施体制

代表研究者　黒川　顕（東京工業大学大学院生命理工学研究科）

研究開発題目　メタゲノムオーソログ遺伝子統合解析システムの開発

(1) データベース・クラスタリング開発グループ

　グループリーダー　黒川　顕（東京工業大学大学院生命理工学研究科）

(2) アノテーション手法開発グループ

　グループリーダー　平川　英樹（かずさDNA研究所）

(3) 比較メタゲノムグループ

　グループリーダー　服部　正平（東京大学大学院新領域創成科学研究科）

5. 参考文献

[1] 黒川 顕, 服部正平 メタゲノムデータベース, 実験医学4月増刊号「バイオインフォマティクスツールの開発と生命研究への応用の最前線」, 37-42, 羊土社, 東京, 2008.

[2] Kurokawa, K., Itoh, T., Kuwahara, T., Oshima, K., Toh, H., Toyoda, A., Takami, H., Morita, H., Sharma, V. K., Srivastava, T. P., Taylor, T. D., Noguchi, H., Mori, H., Ogura, Y., Ehrich, D. S., Itoh, K., Takagi, T., Sakaki, Y., Hayashi, T., and Hattori, M. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Res., 14:169-181, 2007.

[3] Hattori, M., and Taylor, T.D. The human intestinal microbiome: a new frontier of human biology. DNA Res., 16:1-12, 2009.

[4] Noguchi, H., Park, J., and Takagi, T. MetaGene: prokaryotic gene ˋnding from environmental genome shotgun sequences. Nucleic Acids Res., 34:5623-5630, 2006.

[5] Noguchi, H., Taniguchi, T., and Itoh, T. MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res., 15:387-396, 2008.

[6] Goto, N., Kurokawa, K., and Yasunaga, T. Analysis of invariant sequences in 266 complete genomes. Gene, 401:172-180, 2007.

[7] http://metagenome.jp/

[8] 森宙史，林哲也，黒川顕メタゲノム研究の最前線，蛋白質核酸酵素，54:1264-1270, 2009.