

研究開発課題別事後評価結果

1. 研究開発課題名

メタゲノム統合解析システムの開発

2. 代表研究者名

黒川 顕（東京工業大学大学院生命理工学研究科 教授）

3. 代表研究者による成果開発概要

3-1. 研究目的

自然環境中の様々な細菌叢を、培養することなくまるごとゲノム解析する「メタゲノム解析」が実現化し、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ってきた。細菌群集を構成する種や遺伝子のリストを記述することに主眼をおいたメタゲノム研究は、新型シーケンサーの普及によりゲノム科学分野および環境微生物学分野などにおいて一般化すると言っても過言ではない。このような状況下において、申請者らがこれまで開発してきたシステムの集大成とも言えるメタゲノム解析ワークフローの整備は、国内外のメタゲノム研究にとって研究遂行上必須なものとなる。

メタゲノム研究を実施する際、細菌叢を取り巻く環境因子の各要素を記述したメタデータは欠く事のできない重要な情報となる。多岐に渡る高精度なメタデータとメタゲノムデータを統合し、環境と遺伝子プールとの関係性を解析する事が今後の重要な課題となる。本研究課題では、メタゲノム解析のワークフロー整備を主な目的とし、メタゲノムデータベースシステム整備およびワークフロー作成、免疫研究に資するヒトメタゲノム情報解析技術の開発、メタデータと遺伝子データの統合解析技術の開発、スパコンによるメタゲノム解析、の4項目を中心にシステム開発することを目標とした。

3-2. 研究開発概要

本課題ではメタゲノム解析の中でも特にヒトメタゲノム解析に着目し、大量のメタゲノムデータから有用な知識を効率よく発見するためのインフラ構築を目的に、ヒトメタゲノムデータベース Human Meta BodyMap の整備、メタ 16S rRNA 解析用ソフト VITCOMIC の開発、ヒトメタゲノム検索システム Body-BLAST の開発を実施した。ヒトメタゲノムデータに特化したデータベースシステム「Human Meta BodyMap」は、これまで開発してきたメタゲノムデータベースの基盤技術ならびに解析パイプライン、メタゲノムデータの各種可視化技術を統合するとともに、世界中で公開されている新型シーケンサーによる 69,876 サンプルにおよぶヒトメタゲノムデータも収録している。また、これら配列は、85 カテゴリーからなるメタデータにより注釈付けされているが、オントロジーを整備した上でメタデータを XML 化し本データベースに併せて収録し、メタゲノムデータおよびメタデータによる横断的検索を実現した。

新型シーケンサーによるメタ 16S rRNA 解析は、これまでの解析手法では実施困難となりつつある。そこで、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。また、複数のサンプルを同時に解析し、視覚的かつ統計的に比較できるよう機能を拡張し、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。

Human Meta BodyMap に収録しているすべてのデータに対して、配列相同性検索が可能な Body-BLAST を開発した。本システムは、データベースに対して BLAST による相同性検索を実施するが、通常の BLAST 検索のようにアラインメント等の配列相同性検索結果を出力するのではなく、配列相同性検索を介し、高い相同性が得られた配列のメタデータを結果として出力する。

これまでに開発してきたメタゲノム解析パイプラインを、東工大スパコン TSUBAME2.0 に実装し、メタゲノム解析における大量相同性検索を高速に解析可能なシステムを開発した。さらに、GPGPU による相同性検索ソフトウェアを併せて開発し、より低電力かつ高速なメタゲノム解析、すなわちリアルタイムメタゲノム解析を実現した。

3 - 3 . 研究成果概要

新型シーケンサーによるメタ 16S rRNA 解析において、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 VITCOMIC を開発した。本手法は、メタ 16S rRNA データを既存の 16S rRNA 配列に対して blastn によるマッピングを実施するが、得られた結果を順序正しく並べる事でグローバルアラインメントの代替とする技術を開発し、blastn を利用しながら可能な限り正確にアラインメントする手法を構築した。発表論文は、highly accessed paper に評価されるとともに、国内外からすでに 1,000 件近い解析実績を挙げている。

これまで開発してきたメタゲノム解析パイプラインは、リード長が長い配列によるメタゲノム解析が対象となっていた。しかし、Illumina に代表されるリード長が短い配列によるメタゲノムデータが産出されつつあり、解析パイプラインならびに本システムを、それらに対応させる必要が生じた。まず、短い配列において生物学的情報を正確に得る事ができるかを検討した結果、60bp 以上の配列長があれば、概ね 80%以上の確度で系統や遺伝子機能をアサインできる事を確認した。この結果を受け、各種フィルタリング、統計的解析手法を新たに開発し、新型シーケンサーに対応した解析パイプラインを構築した。EU ヒトメタゲノムプロジェクト MetaHIT との共同研究においてもこれら基盤の統計解析が議論され、得られた情報を基に多変量解析技術を駆使することでヒト腸内細菌叢の分類である「Enterotype」という新たな概念を提唱し、腸内細菌叢は人種や性別を問わず大きく 3 つの Enterotype に分類される事を見出した。

また、Human Meta BodyMap には、これまで開発してきたすべての要素を集合するとともに、世界中に散在しているヒトメタゲノムデータを可能な限り集積し、メタデータを整理した上でシステムに格納した。これらデータへのアクセスは、プロジェクトやサンプルからの検索に加え、メタデータからの検索も可能としている。さらに、これらデータをデータベースとして BLAST による配列相同性検索を可能とする Body-BLAST を実装している。Body-BLAST を利用すれば、配列相同性を介したメタデータ検索が可能であり、その結果は BodyParts3D により視覚的に表現される。

4 . 事後評価結果

4 - 1 . 当初計画の達成度

当初の研究開発計画から見た進捗状況や達成度等

当初の研究開発計画のうち、免疫研究に資するヒトメタゲノム情報解析技術の開発をコスト面から困難であると断念したが、その代わりに、高速シーケンサーからの大量のヒトメタゲノム配列データに対応できる可視化手法や解析パイプラインなど当初想定外の

研究開発の実現を達成できた。

当初計画では想定されていなかった新たな展開

新型シーケンサーによるデータ産出量が想定を大幅に上回っていたが、スパコンとGPGPUを活用することにより対応可能となり、その結果、「リアルタイムメタゲノム解析」を世界に先駆けて実現したことは非常に望ましい展開になったと評価できる。

4 - 2 . 知的財産権、外部発表（論文等）等研究開発成果の状況

外部発表（論文等）

海外との共同研究も含め多くの論文が出ている。また、開発したソフトウェア、解析手法、解析結果いずれの論文も発表されている。

データベースの公開状況

本課題で開発されたソフトウェアは、管理上の手違いから現在は公開できない状態となっている。今後の公開が待たれる。

知的財産権

特許は出していないが、非常に有用なデータベース・ソフトウェアツール群が開発できた。

4 - 3 . 研究開発成果の公開による波及効果

データベース、プログラムや論文、学会発表など、これまでの成果公開によるバイオインフォマティクス領域およびライフサイエンス分野全体への影響や効果について

論文の引用回数、ソフトウェアの解析実績ともに、本課題による成果の影響の大きさを示している。ただし、データベースの公開が遅れているため、それらの利用を通じての技術普及が遅れると思われる。

国内外のデータベース高度化・標準化の動向と比較して本課題の意義

メタゲノム解析ソフトは海外を中心にいくつか開発されていて、日本オリジナルのものはない。本研究ではヒトゲノム解析に特化していることと、配列データではなくメタデータを出力することなどの特徴がある。また、本研究で開発したメタゲノム断片配列からの遺伝子予測ソフトの意義は非常に高いと考え、今後さらに発展させて国際標準となることが期待される。

4 - 4 . 成果の実用化の可能性及び成果から予想される波及効果

成果を今後発展展開することによりバイオインフォマティクス領域およびライフサイエンス分野全体に影響や効果を及ぼすと期待できるか。

高速シーケンサーを使ったメタゲノム解析という研究領域では、そのデータを比較的容易に解析できるツールを提供することは、ライフサイエンス分野全体へ大きな影響、貢献がある。日本初の解析システムとしてアピールできる可能性がある。

今後、成果の生命科学、研究基盤、産業への貢献度

本研究では新型シーケンサーを用いたメタゲノム情報解析の基盤技術を確立しており、今後の生命科学研究に大きく貢献することが期待できる。また、微生物工学においてゲノ

ムレベルでの理解が可能となったため産業への応用も期待できる。

4 - 5 . 総合評価

研究開発計画以上の成果が見られ、バイオインフォマティクス研究の進展に大きく貢献した。

日本におけるメタゲノム解析の代表的な研究としての位置付けを確立している。メタゲノム情報解析のための有用かつ独自性のあるソフトウェアツール・データベースが構築され、幅広く利用されている。また、それらを利用して実際のデータを解析した結果、有用な生物学的知見も得られた。しかし、これらのデータからどれだけ細菌叢の実態について予見ができるかについては、さらなる実験的検証が必要である。ここで行われたメタゲノム解析は国際的に最先端の研究であり、今後の展開が期待される。

5 . 主な論文発表等

- *1. Arumugam M, Raes J, Pelletier E, Paslier D L, Yamada T, Mende D R, Fernandes G R, Tap J, Bruls T, Batto JM, Bertalan M, Borrue I N, Casillas F, Fernandez L, Gautier L, Hansen T, Hattori M, Hayashi T, Kleerebezem M, Kurokawa K, Leclerc M, Levenez F, Manichanh C, Nielsen H B, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal E G, Wang J, Guarner F, Pedersen O, de Vos W M, Brunak S, Doré J, MetaHIT Consortium, Weissenbach J, Ehrlich S D, Bork P (2011) Enterotypes of the human gut microbiome. *Nature*, 473:174-180.
2. Nozawa T, Furukawa N, Aikawa C, Watanabe T, Haobam B, Kurokawa K, Maruyama F, and Nakagawa I (2011) CRISPR inhibition of prophage acquisition in *Streptococcus pyogenes*. *PLoS one*, 6:e19543.
3. Watanabe T, Maruyama F, Nozawa T, Aoki A, Okano S, Shibata Y, Oshima K, Kurokawa K, Hattori M, Nakagawa I, and Abiko Y (2011) Complete genome sequence of the periodontogenic bacterium *Porphyromonas gingivalis* TDC60. *J Bacteriol.* Epub Jun 24.
4. Tajima N, Sato S, Maruyama F, Kaneko T, Sasaki NV, Kurokawa K, Ohta H, Kanasaki Y, Yoshikawa H, Tabata S, Ikeuchi M, Sato N (2011) Genomic Structure of the Cyanobacterium *Synechocystis* sp. PCC 6803 Strain GT-S. *DNA Res.*, 18:393-399.
- *5. Kuwahara T, Ogura Y, Oshima K, Kurokawa K, Ooka T, Hirakawa H, Itoh T, Nakayama-Imahiji H, Ichimura M, Itoh K, Ishifune C, Maekawa Y, Yasutomo K, Hattori M, Hayashi T (2011) The Lifestyle of the Segmented Filamentous Bacterium: A Non-Culturable Gut-Associated Immunostimulating Microbe Inferred by Whole-Genome Sequencing. *DNA Res.*, 18:291-303
- *6. Mori, H., Maruyama, F., and Kurokawa, K. VITCOMIC: visualization tool for taxonomic compositions of microbial communities based on 16S rRNA gene sequences., *BMC Bioinfo.*, 11:332, 2010.
7. Nakayama, K., Kurokawa, K., Fukuhara, M., Urakami, H., Yamamoto, S., Yamazaki, K., Ogura, Y., Ooka, T., and Hayashi, T. Genome comparison and phylogenetic analysis of *Orientia tsutsugamushi* strains., *DNA Res.*, 17:281-291., 2010.

8. Ogura, Y., Ooka, T., Iguchi, A., Toh, H., Asadulghani, M., Oshima, K., Kodama, T., Abe, H., Nakayama, K., Kurokawa, K., Tobe, T., Hattori, M., and Hayashi, T. Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*. *Proc. Natl. Acad. Sci. USA*, 106:17939-17944, 2009.
9. Masahira, H., Oshima, K., Kim, S., Kurokawa, K., Toh, H., and Taylor, T. Metagenomics and genomics decoding human gut microbiome, *Genes Genet. Syst.*, 84:479-479, 2009.
10. Toh, H., Oshima, K., Toyoda, A., Ogura, Y., Ooka, T., Sasamoto, H., Park, S.H., Iyoda, S., Kurokawa, K., Morita, H., Itoh, K., Taylor, T.D., Hayashi, T., and Hattori, M. Complete genome sequence of the wild-type commensal *Escherichia coli* strain SE15 belonging to phylogenetic group B2., *J. Bacteriol.*, 192:1165-1166, 2009.
11. Hirakawa H., Akita H., Fujiwara T., Sugai M., Kuhara S. Structural insight into the binding mode between the targeting domain of ALE-1 (92AA) and pentaglycine of peptidoglycan. *Protein Eng. Des. Sel.*, 22, 385-391, 2009.
- *12. 森宙史, 丸山史人, 黒川 顕メタゲノムインフォマティクス, 難培養微生物研究の最新技術 II, pp82-91, 2010, シーエムシー出版, 東京
- *13. 森宙史, 丸山史人, 黒川 顕(2010) メタゲノムデータベース, メタゲノム解析技術の最前線, シーエムシー出版, pp42-53

* は当課題の成果として主要な論文

参考

論文発表	国内 2 件、海外 11 件
招待、基調講演	国内 8 件、海外 0 件
口頭発表	国内 7 件、海外 0 件
ポスター発表	国内 1 件、海外 2 件
特許	国内 0 件、国際 0 件

成果プログラムおよびデータベース等

Human Meta BodyMap (VITCOMIC、Body-BLAST 含む)

ヒトメタゲノム配列データに対する遺伝子構造予測結果の可視化機能、そのクラスタリング結果を人体模式図や代謝パスウェイにマッピング・可視化する機能、ゲノム配列既知の個別菌との配列類似関係を可視化する機能、立体構造解析支援機能を提供する。これらのビューア上でサンプル個体間の相違を比較解析することも可能となっている。また公開されているヒトメタゲノムプロジェクトのメタデータ検索、配列相同性検索Body-BLASTを実装している。

(H24 年に公開予定)