科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　２５　日現在

機関番号：１３１０１
研究種目：若手研究(B)
研究期間：2011～2013
課題番号：２３７１０２４２

研究課題名（和文）メタゲノム資源からの新規微生物探索とその代謝経路予測による環境浄化システムの解明

研究課題名（英文）A bioinformatics strategy for revealing microbial community structures and metabolic pathways by reconstructing multiple genomes from metagenomic sequences.

研究代表者
阿部　貴志（Abe, Takashi）

新潟大学・自然科学系・准教授

研究者番号：３０３９０６２８

交付決定額（研究期間全体）：（直接経費）　　3,400,000円、（間接経費）　　1,020,000円

研究成果の概要（和文）：環境メタゲノム資源から、連続塩基・アミノ酸組成に基づく一括学習型自己組織化マップ（BLSOM）を基に、環境中の微生物叢のための生物系統推定ソフトウェア、地球環境改善に役立つ新規微生物ゲノムの検出手法、ならびに、それらが持つ環境浄化システムに関与する有用遺伝子候補の探索手法を開発した。新規性の高いゲノム断片配列を微生物ゲノム別に再構成するための手法が開発でき、環境が保有する環境浄化システムを構成する微生物や代謝遺伝子セットの全体像が把握でき、様々な微生物が持つ環境浄化システムの全体像把握に向けた情報学的スクリーニング法としての活用も期待できる。

研究成果の概要（英文）：Metagenomics studies of uncultivable microorganisms in clinical and environmental samples should allow extensive surveys of genes useful in medical and industrial applications. BLSOM is the most suitable for phylogenetic assignment of metagenomic sequences because fragmental sequences can be clustered according to phylotypes, solely depending on oligonucleotide composition.
We constructed oligonucleotide-BLSOMs for all available sequences from species-known genomes, and by mapping metagenomic sequences on this large-scale BLSOMs, we could predict phylotypes of individual metagenomic sequences, revealing a microbial community structure of uncultured microorganisms including viruses. This software will be freely available at our website. We also developed another BLSOM strategy for clustering metagenomic sequences according to genome, and for function prediction of poorly-characterized protein genes obtained from metagenome analysis.

研究分野：応用ゲノム科学

科研費の分科・細目：メタゲノム

キーワード：メタゲノム　一括学習型自己組織化マップ　連続塩基　連続アミノ酸　生物系統推定　タンパク質機能推定

１．研究開始当初の背景

我々は、ゲノム配列の 3 連続や 4 連続塩基などの連続塩基頻度に着目し、超大量ゲノム配列から生物種固有の配列特徴を俯瞰的に理解可能とする一括学習型自己組織化マップ(BLSOM) を開発した。BLSOM は生物種名の情報を計算の途中で一切与えずに、連続塩基の出現頻度の類似性だけで、ゲノム配列断片を生物種ごとに高精度に分離(自己組織化)させる強力なクラスタリング能力を持つ。さらに、並列計算に適したアルゴリズムとなっており、地球シミュレータのような高性能計算機を用いた超大規模解析も可能である。2008 年以降、次世代シークエンサーの登場によって、大量のゲノム配列の解読が一度に可能となった。大規模メタゲノム解析も精力的に実施され、微生物が持つ環境浄化システムの全体像把握が進められている。環境メタゲノム配列は、混合ゲノム試料をクローン化した際に短く断片化（泣き別れ）した大量のゲノム配列断片である。もとのゲノムへの再構成が難しいため、環境メタゲノム配列にどれだけの生物種がどのような割合で混在しているのか、代謝系遺伝子セットを単独の微生物種が保有しているのか等を推定することが困難である。相同性検索とは全く異なった原理に基づく超大規模データ解析技術の確立が求められている。
我々は、連続塩基頻度のみに着目した BLSOM で、ゲノム配列断片の大半を生物種により高精度に分類できる知見を基に、メタゲノム配列に対する系統推定が可能なことを世界に先駆けて見出した。日本国内の実験研究者との共同研究を通じて、実験グループが解読したメタゲノム配列データを対象に BLSOM 解析を行い、論文発表を行っている。

２．研究の目的

環境メタゲノム資源から、地球環境改善に役立つ新規微生物ゲノムや、それらが持つ環境浄化システムに関与する有用遺伝子候補の探索手法を開発する。新規性の高いゲノム配列断片を微生物ゲノム別に再構成するためのアノテーション手法が開発できれば、環境が保有する環境浄化システムを構成する微生物や代謝遺伝子セットの全体像が把握できる。既知微生物による浄化に関与する酵素群をカタログ化することによって、様々な微生物が持つ環境浄化システムの全体像把握に向けた情報学的スクリーニング法としての活用も期待できる。

３．研究の方法

我々が開発してきた連続塩基組成を基にした一括学習型自己組織化マップ(BLSOM)を用いて、環境メタゲノム配列に対する系統推定法の開発を行ってきたが、本研究では、BLSOM を活用して、大量メタゲノム配列からの更なる知識発見を可能とするため、以下の 4 点についての研究開発を行う。

① 生物系統ソフトウェアの開発と公開
② 短いメタゲノム配列断片に対する系統推定精度の向上に向けた改良
③ メタゲノム配列群からの新規微生物ゲノム検出手法の開発
④ 有用遺伝子探索のためのタンパク質機能推定システムの開発

新規性の高いゲノム配列断片を微生物ゲノム別に再構成するための検出手法を開発し、環境が保有する環境浄化システムの構成要素の全体像の把握を試みる。

４．研究成果

（１）生物系統推定ソフトウェアの公開

メタゲノム解析により、全地球レベルでの生物生態系の把握を目標にした大規模解析も可能になってきた。メタゲノム配列データは新規性の高い微生物種が優占種となる場合も多く、既知微生物配列との配列相同性検索では、新規微生物種の存在を検出することが困難である。BLSOM を用いて、連続塩基頻度のみに着目することでゲノム配列断片の大半を生物種ごとに高精度に分類できる知見を基に、既知微生物を対象にした BLSOM 解析結果にメタゲノム配列を照合（マッピング）することで、生物系統が推定できる手法を開発してきた。本手法を組み込んだソフトウェアとして、PEMS（Phylogenetic Estimation of Metagenomic sequence using BLSOM）を公開した
（http://bioinfo.ie.niigata-u.ac.jp/?PEMS_Soft）。

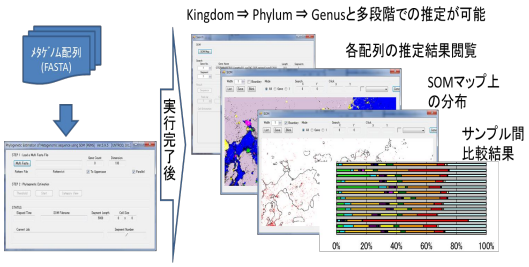

図1．生物系統推定ソフトウェア PEMS の概要

本ソフトウェアの概要を図１に示す。入力画面にて、マルチ FASTA 形式のメタゲノム配列を投入し、実行する。その際、STEP2 の【Threshold】にて、近傍中に含まれる生物種の割合（デフォルト 40）を変更できる。なお、プログラムでの最小長は、300 塩基であり、300 塩基未満は自動的に除去される。本プログラムの実行時間は利用する PC の CPU とメモリ量で大きく変わるため、大量のメタゲノム配列データを投入する場合には、注意

が必要である。実行完了後、【Category View】をクリックすることで、各配列の推定結果一覧、各レベルでの BLSOM マップ上でのマッピング結果、各系統レベルでの推定結果の集計データの取得などができる。また、マッピングを行う BLSOM マップを換えることで、真核生物やウイルス等の他の生物系統についても推定可能である。国内外の実験グループとの共同研究により、本ソフトウェアを活用して、論文発表も行った（4, 9)。

（２）短いメタゲノム配列断片に対する系統推定精度の向上に向けた改良

PEMS は、300 塩基以上のメタゲノム配列を想定した手法であり、次世代シーケンサから産出される大量の短いメタゲノム配列（100～300 塩基程度）に対しては、照合の際の正規化に伴うエラーが生じやすく、生物系統推定の精度が低下する。
短いメタゲノム配列に対する高精度な生物系統推定法の確立を目指し、ゲノム配列断片の 4 連続塩基頻度計算時に 1 塩基や 2 連続塩基組成を考慮した連続塩基配列組成計算法を開発し、既知生物ゲノムを対象にした BLSOM マップを作成し、検証を行った。原核生物完全長ゲノム 808 種を対象に、ゲノム配列断片の 4 連続塩基の実測値とゲノム配列中の 1 塩基組成から算出された期待値の比を用いて BLSOM 解析を行うことで、断片化サイズ 2.5kb でも断片化サイズ 5kb と同等以上の分解能が得られた。照合するための既知生物全ゲノムの BLSOM マップに用いる断片化サイズを短くすることができ、短いメタゲノム配列に対する生物系統推定の精度向上が可能となった。
さらに、短いメタゲノム配列断片に対する系統推定を行うための分子マーカーとして、tRNA 遺伝子に着目した。tRNA 遺伝子は、配列長が 76～120 塩基程度であり、次世代シークエンサーより産出される 100 塩基程度の短いゲノム断片配列にも tRNA 遺伝子がコードされており、既存の方法では困難であった短い断片配列を対象にした系統推定に有効な手段の一つとなりえる。tRNA 遺伝子間の系統保存性を検証したところ、保存性が非常に高く、系統分子マーカーとして利用可能であった（2, 19)。我々が構築している tRNA 遺伝子データベース tRNADB-CE に、メタゲノム配列から探索された tRNA 遺伝子も格納し、既知生物種由来の tRNA 遺伝子との相同性によって、生物系統推定可能な機能を開発し、公開している。

（３）メタゲノム配列群からの新規微生物ゲノム検出法の開発

次世代シークエンサーから産出される大量の短いメタゲノム配列データに対して新規微生物ゲノムごとの検出が非常に困難である。取得されたメタゲノム配列情報のみを使用し、環境特異的な微生物群集の構造を検出する手法として、着目するメタゲノム配列と各配列に対して 1 塩基組成や 2 連続塩基組成を保持して作成したランダム配列を混合させた BLSOM 解析を開発した。次世代シークエンサー由来の短いメタゲノム配列データにも対応できるかを検証するために、既知微生物 3 種を対象に、BSLOM での解析条件（反映させる塩基組成や BLSOM 実行時に使用する連続塩基）の検討を行い、断片化サイズ 300bp、２連続塩基を反映させたランダム配列を加えた縮退４連続塩基にて、各既知微生物の 80%程度を、生物種情報を用いることなく、クラスタとして分離することができた。現在、メタゲノム解析で主に使用される次世代シーケンサのロシュ社 GS FLX Titanium の平均長（350bp）よりも短い配列を対象にでき、実際のメタゲノムデータを用いた解析では、新規性の高い微生物を含むゲノムごとのクラスタを検出することができた。本手法によって、次世代シーケンサ由来メタゲノム配列データに対して、効率的な新規微生物ゲノムの検出が可能である。

（４）有用遺伝子探索のためのタンパク質機能推定システムの開発

配列相同性検索で機能が推定出来ないタンパク質の配列が、利用価値が低いままに大量にデータベースに集積している。配列相同性検索に依存しない情報学的手法の確立が重要である。オリゴペプチド頻度を対象にした BLSOM を用いることで、タンパク質類が機能を反映した分離を起こすことを見出した。この成果を用いて、植物由来の２次代謝関連酵素タンパク質の特徴抽出を目的に、公開されている全植物と原核生物由来のタンパク質アミノ酸配列（721,266 配列）を対象に、200 アミノ酸ごとに断片化した配列（配列断片数：1,752,300）の２連続アミノ酸組成（400 次元）に基づく BLSOM 解析を行った。植物由来の２次代謝関連酵素タンパク質としてテルペン、アルカノイド、フラボノイド、イソフラボノイドに着目し、作成した BLSOM マップへ各々のアミノ酸配列をマップしたところ、酵素タンパク質ごとにクラスタが形成されており、機能に特化したアミノ酸配列組成を持つことが明らかとなった（8)。また、BLSOM 上でマップされた格子点のばらつきを見ることで、２次代謝関連酵素の機能としての多様度を示す指標となることが明らかとなった。また、機能ごとにクラスタが形成されていた領域には、既知未知な植物由来タンパク質も多く含まれており、それらが２次代謝関連酵素としての機能を持つ可能性が高いと考えられる。
本手法を用いることで、大量に蓄積されるタンパク質アミノ酸配列群に対し、機能既知タンパク質類の特徴抽出、ならびに、機能未知

タンパク質への機能推定をシステマティックに行うことが可能であり、産業的・医学的に有用な機能を持つタンパク質類への応用が可能である。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 19 件）

1. Takashi Abe, Yuta Hamao, Toshimichi Ikemura. Visualization of genome signatures of eukaryote genomes by Batch-Learning Self-Organizing Map (BLSOM) with a special emphasis on Drosophila genomes. BioMed Research International, 2014, Article ID 985706, doi:10.1155/2014/985706, 2014（**査読有**）.
2. Takashi Abe, Hachiro Inokuchi, Yuko Yamada, Akira Muto, Yuki Iwasaki, Toshimichi Ikemura. tRNADB-CE: tRNA gene database well-timed in the era of big sequence data. Frontiers in GENETICS, 5, 114. doi:10.3389/fgene.2014.00114, 2014（**査読有**）.
3. Yuki Iwasaki, Takashi Abe, Toshimichi Ikemura（**他** 2 **人、**2 **番目**）. A Novel Bioinformatics Strategy to Analyze Microbial Big Sequence Data for Efficient Knowledge Discovery: Batch-Learning Self-Organizing Map (BLSOM). Microorganisms. 1, 137-157, doi:10.3390/microorganisms1010137, 2013（**査読有**）.
4. Atsushi Kouzuma, Takashi Abe, Kazuya Watanabe（**他** 3 **人、**5 **番目**）. Comparative metagenomics of anode- associated microbiomes developed in rice paddy-field microbial fuel cells. *PLoS ONE*, 8, e77443, doi:10.1371/journal.pone.0077443, 2013（**査読有**）.
5. Yuki Iwasaki, Takashi Abe, Toshimichi Ikemura（**他** 3 **人、**2 **番目**）. Nobel bioinformatics strategies for prediction of directional sequence changes in influenza virus genomes and for surveillance of potentially hazardous strains. *BMC Infectious Diseases*, 13:386, doi:10.1186/1471-2334-13-386, 2013（**査読有**）.
6. Yuki Iwasaki, Takashi Abe, Toshimichi Ikemura（**他** 3 **人、**4 **番目**）. Notable clustering of transcription-factor-binding motifs in human pericentric regions and its biological significance. *Chromosome Research*, 21, 461-474, 2013（**査読有**）.
7. Kyoko Hayashida, Takashi Abe, Chihiro Sugimoto（**他** 7 **人、**2 **番目**）. Whole-genome sequencing of Theileria parva strains provides insight into parasite migration and diversification in the African continent. *DNA Research*, 20, 209-220, 2013（**査読有**）.
8. Shun Ikeda*, Takashi Abe*, Shigehiko Kanaya （*equal contribution, **他** 8 **人、**2 **番目**）. Systematization of diversity of protein sequences in enzymes related to secondary metabolic pathways in plants in the context of big data biology inspired by KNApSAcK Motorcycle database. *Plant and Cell Physiology*, 54, 711-727, 2013（**査読有**）.
9. Ryo Nakao*, Takashi Abe*, Chihiro Sugimoto （*equal contribution, **他** 4 **人、**2 **番目**）. A novel approach, based on BLSOMs (Batch Learning Self-Organizing Maps), to the microbiome analysis of ticks. *ISME Journal*, 7, 1003-1015, 2013（**査読有**）.
10. Hiroaki Sakai, Takashi Abe, Takeshi Itoh（**他** 14 **名、**10 **番目**）. Rice Annotation Project Database (RAP-DB): An integrative and interactive database for rice genomics. *Plant and Cell Physiology*, 54, e6(1-11), 2013（**査読有**）.
11. Kyoko Hayashida*, Yuichiro Hara*, Takashi Abe* （*equal contribution, **他** 27 **人、**3 **番目**）. Comparative genome analysis of three eukaryotic parasites with differing abilities of leukocyte transformation reveals key mediators of Theileria-induced leukocyte-transformation. *mBio*, 3, e00204-12, doi:10.1128/mBio.00204-12, 2012（**査読有**）.
12. Ryosuke Nakai, Takashi Abe, Takeshi Naganuma（**他８人、２番目**）. Diverse RuBisCO genes responsible for CO2 fixation in an Antarctic moss pillar. *Polar Biology*, 35, 1641-1650, 2012（**査読有**）.
13. Ryosuke Nakai, Takashi Abe, Takeshi Naganuma（**他８人、２番目**）. Eukaryotic phylotypes in aquatic moss pillars inhabiting a freshwater lake in East Antarctica, based on 18S rRNA gene analysis. *Polar Biology*, 35, 1495-1504, 2012（**査読有**）.
14. Ryosuke Nakai, Takashi Abe, Takeshi Naganuma（**他８人、２番目**）. Microflorae of aquatic moss pillars in a freshwater lake, East Antarctica, based on fatty acid and 16S rRNA gene analyses. *Polar Biology*. 35, 425-433, 2012（**査読有**）.
15. Yuki Iwasaki, Toshimichi Ikemura, Takashi Abe（**他** 2 **名、**5 **番目**）. A Novel Bioinformatics Strategy to Predict Direcitional Changes of Influenza A Virus Genome Sequences. *WSOM 2011*, LNCS 6731, 198-206, 2011（**査読有**）.
16. Yuki Iwasaki*, Takashi Abe*, Toshimichi Ikemura （*equal contribution、**他** 2 **人、**2

番目）. Prediction of directional changes of influenza A virus genome sequences with emphasis on pandemic H1N1/09 as a model case. *DNA Research*, 18, 125-136, 2011（査読有）.

17. Hiroshi Uehara, Toshimichi Ikemura, Takashi Abe（他 2 名、5 番目）. A novel bioinformatics strategy for searching industrially useful genome resources from metagenomic sequence libraries. *Genes & Genetic Systems*, 86, 53-66, 2011（査読有）.
18. Ryosuke Nakai, Takashi Abe, Haruko Takeyama and Takeshi Naganuma. Metagenomic analysis of 0.2-μm-passable microorganisms in deep-sea hydrothermal fluid. *Marine Biotechnology*, 13, 900-908, 2011（査読有）.
19. Takashi Abe, Toshimichi Ikemura, Hachiro Inokuchi（他 8 名、1 番目）. tRNADB-CE 2011: tRNA gene database curated manually by experts. *Nucleic Acids Research*, 39, D210-D213, 2011（査読有）.

〔学会発表〕（計　54 件）
（国際会議での口頭発表のみを記載）

1. Takashi Abe, Ryo Nakao, Toshimichi Ikemura, Frans Jongejan, and Chihiro Sugimoto. Metagenomic analysis for unveiling of microbial diversities within tick guts. International Union of Microbiological Societies 2011 Congress, 8 Sept., 2011 (Sapporo, Japan)
2. Chihiro Sugimoto, Ryo Nakao, Toshimichi Ikemura, Frans Jongejan, and Takashi Abe. Metagenomic approach to identify tick-borne pathogens by using ultra high throughput DNA sequencing and data analyzing technologies. 7th International Conference on Ticks and Tick-borne Pathogens (TTP-7), 28 August, 2011 (Zaragoza, Spain)
3. Takashi Abe, Yuki Iwasaki, Hiroshi Uehara, Yuuta Hamano, Kennosuke Wada and Toshimichi Ikemura. Visualization of Genome Signatures with BLSOM and its Application to Eukaryotic and Viral genomes. Society for Molecular Biology and Evolution 2011, 26 July, 2011 (Kyoto, Japan)
4. Yuki Iwasaki, Kennosuke Wada, Masae Itoh, Toshimichi Ikemura, and Takashi Abe, A Novel Bioinformatics Strategy to Predict Direcitional Changes of Influenza A Virus Genome Sequences. Workshop on Self-Organizing Map 2011, 15 June 2011 (Espoo, Finland)

他、国際学会 9 件、国内学会 41 件の発表を行った。

〔図書〕（計　7 件）

1. 阿部貴志，金谷重彦，池村淑道．一括学習型自己組織化マップ（BLSOM）を用いた大量メタゲノム解析．生命のビックデータ利用の最前線（植田充美　監修），シーエムシー出版，104-112，2014．
2. 池村淑道，阿部貴志．16 章　ゲノミクス．ベーシックマスター分子生物学（東中川徹，大山隆，清水光弘共編），オーム社，2013．
3. Yuki Iwasaki, Toshimichi Ikemura, Kennosuke Wada, Yoshiko Wada, Takashi Abe. Novel bioinformatics method to analyze more than 10,000 influenza virus strains　easily at once: Batch-Learning Self Organizing Map (BLSOM). Advances in Viral Genomes Research (Eds: J. Borrelli and Y. Giannini), Nova Science Publishers, Inc. 95-112, 2013.
4. 岩崎裕貴，和田健之介，阿部貴志，池村淑道．インフルエンザウイルスゲノム配列の変化方向の予測，アドバンスシミュレーション，12，56-57，2012．
5. 阿部貴志，中尾亮，杉本千尋．メタゲノム解析による微生物群集構造の解明への一括学習型自己組織化マップ（BLSOM）の活用，生物工学会誌，90(12)，765-768，2012．
6. 岩崎裕貴，阿部貴志，和田健之介，池村淑道．新規情報学的手法を用いたインフルエンザウイルスのゲノム塩基配列の変化の方向性の予測，生物工学会誌，90(12)，769-772，2012．
7. Takashi Abe, Hachiro Inokuchi, Yuko Yamada, Akira Muto, Yuki Iwasaki, Toshimichi Ikemura. tRNADB-CE and use of tRNAs as phylogenetic markers for metagenomic sequences, Encyclopedia of Metagenomics, Springer, in press.

〔その他〕
ホームページ等

1. 研究紹介：
   http://bioinfo.ie.niigata-u.ac.jp/
2. ソフトウェア PEMS：
   http://bioinfo.ie.niigata-u.ac.jp/?PEMS_Soft
3. tRNADB-CE:
   http://trna.ie.niigata-u.ac.jp/

６．研究組織
(1)研究代表者
阿部　貴志（Takashi Abe）
新潟大学・自然科学系・准教授
研究者番号：30390628