様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１４　日現在

機関番号：　１３１０１
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０３２７

研究課題名（和文）ゲノムビックデータ解析のための高速データマイニングシステムの開発

研究課題名（英文）Development of a novel bioinformatics method to analyze big genome sequence data for efficient knowledge discovery

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）　　3,700,000円

研究成果の概要（和文）：ゲノムビックデータからの効率的な知識発見に向け、より高速化した解析手法の開発が求められている。一括学習型自己組織化マップ（BLSOM）の可視化や分離能などの特長は損なわず、爆発的な増加に対応できる新手法として、自己圧縮型BLSOMを開発した。本手法を中心に、メタゲノム配列に対する生物系統推定法、ならびに、機能未知タンパク質アミノ酸配列に対する機能推定システムに適応し、超大量データからの効率的なデータマイニングシステムの確立を行った。

研究成果の概要（英文）：As the result of extensive decoding of genome sequences, novel tools are needed for comprehensive analyses of available big sequence data. We previously developed a BLSOM, which can cluster genomic fragment sequences according to phylotype solely dependent on oligonucleotide composition. However, a large-scale BLSOM needs a large computational resource.
We have developed Self-Compressing BLSOM (SC-BLSOM) for reduction of computation time, which allows us comprehensive analysis of big sequence data. The strategy of SC-BLSOM is to hierarchically construct BLSOMs according to data class such as phylotype. SC-BLSOM could be constructed faster than BLSOM and cluster the sequences according to phylotype with high accuracy. We have also developed a new method to predict protein function on the basis of similarity in oligonucleotide composition. The proteins could be related to function-known proteins. These methods are useful to analyze big sequence data for efficient knowledge discovery.

研究分野：　生命・健康・医療情報学

キーワード：　一括学習型自己組織化マップ　自己圧縮BLSOM　連続塩基組成　連続アミノ酸組成　メタゲノム　系統推定

１．研究開始当初の背景

我々は、広範な生物種に由来する超大量ゲノム配列を対象に、ゲノム配列の 3 連続や 4 連続塩基の頻度に着目することで、生物種固有の特徴を俯瞰的に把握可能とする一括学習型自己組織化マップ（Batch-Learning Self=Organizing Map, BLSOM）を開発した。BLSOM は生物種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のみで、生物種ごとに高精度に分離（自己組織化）できる強力なクラスタリング能を持ち、その結果を容易に可視化できる。さらに、並列計算に適したアルゴリズムになっており、地球シミュレータなどの高性能計算機を用いた超大規模解析もいち早く可能としている。現在公開されている全既知生物のゲノム情報を対象にした BLSOM 解析の結果、原核生物や真核生物ばかりではなく、ウイルス類についても、連続塩基頻度の類似度のみで、配列断片が高精度に分離することが明らかになった。我々は、この知見を基に、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。国内外の実験研究者との共同研究を通じて、メタゲノム配列データを対象に BLSOM 解析を行い、論文発表を行ってきた。

２．研究の目的

これまで開発してきた BLSOM を用いたメタゲノム配列に対する系統推定には、地球上に生息する全既知生物が持つゲノムの特徴を網羅的、かつ、俯瞰的に把握する必要がある。BLSOM は地球シミュレータなどの高性能計算機上での高度な並列化を行い、大規模計算を実現させてきたが、全既知生物のゲノム配列を対象にした場合、1 ヶ月以上もの計算時間が必要である。ゲノムビックデータのための超高速化した解析手法の開発が求められている。

本研究では、これまで研究開発を行ってきた BLSOM の可視化や分離能などの特長は損なわず、爆発的なゲノム配列データの増加に対応できる新規解析手法として、自己圧縮型 BLSOM を開発した。さらに、開発した解析手法を、これまで、BLSOM で開発を行ってきたメタゲノム配列に対する系統推定法、ならびに、機能未知タンパク質アミノ酸配列に対する機能推定システムへの適用を行った。

３．研究の方法

（１）自己圧縮 BLSOM（Self-Compress BLSOM, SC-BLSOM）

BLSOM の計算時間増加率は、おおよそ(データの増加量×2)^{2}と近似でき、データ件数が 2 倍増加すれば、計算時間は約 16 倍増加する。ゲノムビックデータ時代に、最新データへの更新は容易ではない。超大量データに対応できる新規解析手法として、自己圧縮 BLSOM の開発を行った（図１）。

BLSOM は、対象となる入力データの特徴を 2 次元格子点のマップ上に配置した入力データと同じ形式を持つリファレンスベクトル（代表ベクトル）に反映させてクラスタリングを行う。入力データの特徴をリファレンスベクトルに要約、もしくは、圧縮していると言える。この特長を活かし、はじめに入力データを分割し、分割したデータごとに BLSOM を行ない（1 階層目）、1 階層目で得られた入力データの特徴が反映されたリファレンスベクトルを入力データの代わりに利用した BLSOM を行う（2 階層目）。階層構造を持つ複数の BLSOM を行うことで、元のデータの特徴を保持したまま、少ないデータ数での BLSOM が可能になる。

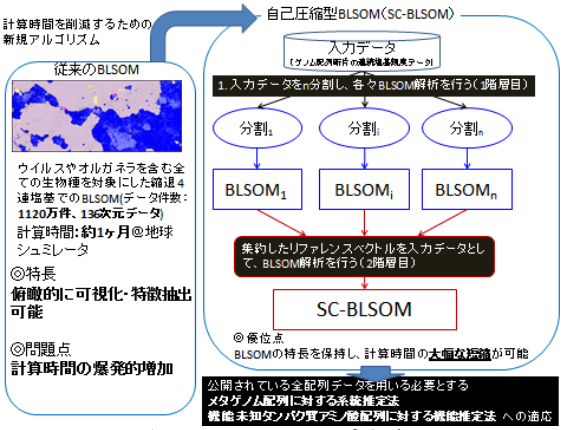

図１　SC-BLSOM の概要

（２）オリゴペプチド組成の距離関係に着目したタンパク質機能推定法の開発

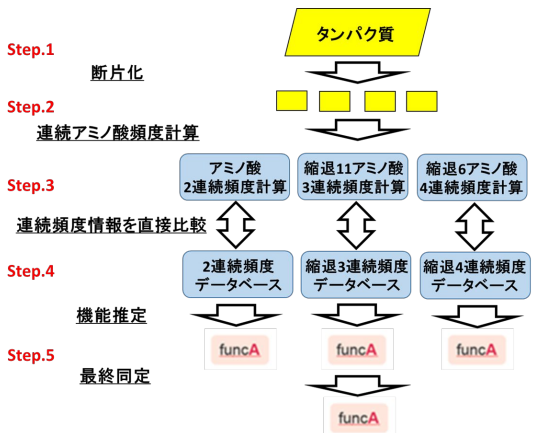

図２ タンパク質機能推定法の概要

開発したタンパク質アミノ酸配列のオリゴペプチド頻度の距離関係に基づく類似度からタンパク質の機能推定を行う手法の概要を図２に示す。参照するデータベースとして、NCBI より提供されている COG (Clusters of Orthologous Groups of Proteins)のタンパク質アミノ酸配列を使用した。COG が提供

しているタンパク質機能カテゴリは、殆どの微生物ゲノムプロジェクトがアノテーションを行う際に使用されるほど有用なものである。

まずはじめに、COG に登録されている 100 アミノ酸（aa）以上のアミノ酸配列を対象に、各アミノ酸配列を断片化サイズ 200 aa として、断片化を行った。ここで、以前の BLSOM 解析において、断片化を行ったほうがタンパク質の機能推定精度が良かったため、断片化した配列を用いることにした。断片化した配列に対し、2 連続アミノ酸、3 連続アミノ酸、4 連続アミノ酸の使用頻度を計算する。ここで、3 連続アミノ酸は、20 のアミノ酸を 11 のカテゴリーに縮退（11^3 = 1331 次元のベクトルデータ）し、4 連続アミノ酸は、20 のアミノ酸を 6 のカテゴリに縮退した使用度数（6^4 = 1296 次元のベクトルデータ）を計算した。この各オリゴペプチド使用頻度データをリファレンスデータベースとして用いる。機能予測を行う場合は、まずはじめに、クエリー配列に対し、断片化サイズ 200aa、ステップサイズ 10aa で断片化を行う（図 2 の Step 1）。断片化した配列ごとに、2 連続アミノ酸、3 連続アミノ酸、4 連続アミノ酸の使用頻度を計算する（図 2 の Step 2）。次に、各連続アミノ酸使用頻度を用いて、対応する連続アミノ酸使用頻度のリファレンスデータベース中の全断片配列とのユークリッド距離を計算し、そのユークリッド距離が最小となる断片配列を探索し、その連続アミノ酸使用頻度データベースでの推定候補を選出する（図 2 の Step 3）。この Step 3 を、断片化した配列分だけ繰り返し、実行する。次に、クエリー配列の全断片配列を対象に、リファレンスデータベースごとで選出された推定候補において、選出された機能（ここでは、COG で付与されている ID としている。）のうち、クエリー配列に対する断片配列数の 60%を超えていた場合に、そのクエリー配列に対する各リファレンスベースからの最終候補とする（図 2 の Step 4）。最後に、各リファレンスベースからの最終候補として選出された機能が 3 つのリファレンスデータベースで完全に一致していた場合をそのクエリー配列に対する最終推定結果とする。一致しなかった場合は、そのクエリー配列に対して、機能推定できなかったとする（図 2 の Step 5）。

４．研究成果

（１）自己圧縮 BLSOM（Self-Compressing BLSOM, SC-BLSOM）の開発

SC-BLSOM の基本的な性能をテストするために、計算時間とクラスタリング能を測り、従来の BLSOM と比較した。使用した入力データは、原核生物完全長ゲノム 817 種よりランダムに 10kb ごと抜き出し元の配列長の 1/10 の長さになるまで併合した塩基配列データとし、断片化サイズ 5kb、縮退 4 連続塩基頻度の条件で、従来の BLSOM と SC-BLSOM を行った。従来の BLSOM では、weight vectors 数を入力データ数の 50%となるようにした。SC-BLSOM は、2 階層とし、各階層での BLSOM の weight vectors 数は入力データ数の 50%となるようにした。1 階層目での入力データの分割数は、生物系統の Phylum（門）を利用し、対象ゲノムが属する Phylum 数である 20 とした。入力データ数は 90,998 配列断片である。本解析で使用した計算機環境は、CPU は Intel(R) Xeon(R) CPU X5472@3.00GHz, メモリ 32Gbyte, OS は CentOS 5.11 である。従来の BLSOM と SC-BLSOM の 2 階層目の分類結果を図 3 に示す。従来の BLSOM と SC-BLSOM の計算時間と分解能（本来の Phylum 由来の配列断）を表１に示す。ここで、クラスタリング能力とは、得られた BLSOM マップ上で単一の Phylum のみが分類されていた格子点数の割合とした。

SC-BLSOM は従来の BLSOM と比べ、計算時間が最大で約 1/40 にまで減少し、SC-BLSOM が非常に高速であると言える。

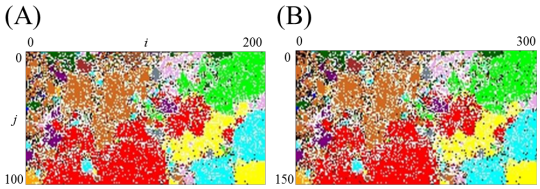

図 3 SC-BLSOM（A）と BLSOM（B）との解析結果。ここで、各色は、各 Phylum を示す。

表１. SC-BLSOM と BLSOM の実行時間と分解能

	BLSOM	SC-BLSOM
計算時間 (分)	831	21
分解能 (%)	93.5	94.1

（２）SC-BLSOM を用いたメタゲノム配列に対する系統推定法の開発

自己圧縮型 BLSOM（SC-BLSOM）を用いて、連続塩基組成に基づくメタゲノム配列に対する生物系統推定法への適応のための条件検討を行った。

従来の BLSOM では、断片化サイズ 5kb に断片化したゲノム配列を対象にしていたが、近年産出されるメタゲノム配列は、5kb 以下の配列も多く、より短く断片化した配列を基にした生物系統推定法が必要であったが、分離能と計算時間の問題もあり、実現が難しかった。しかし、SC-BLSOM は計算時間の大幅な短縮が可能であり、BLSOM では困難であった問題も解消しつつ、より大規模な生物系統推定システムの開発が可能である。検証した生物系統推定のワークフローを図 4 に示す。

はじめに、参照用のマップの検証として、断片化サイズを含めた SC-BLSOM に適した解析条件の検証を行い、小さい断片化サイズ

でも BLSOM とほぼ同等以上の精度を得ることができた。また、メタゲノム配列に対する推定アルゴリズムの検証として、SC-BLSOM は各階層でのリファレンスベクトルを次の階層での入力データとしているため、各階層間で関係性が保持されており、最上位層でのマッピングのみで最下層までの分類結果を一括して取得する方法を検証し、従来法よりもより高速化できた。

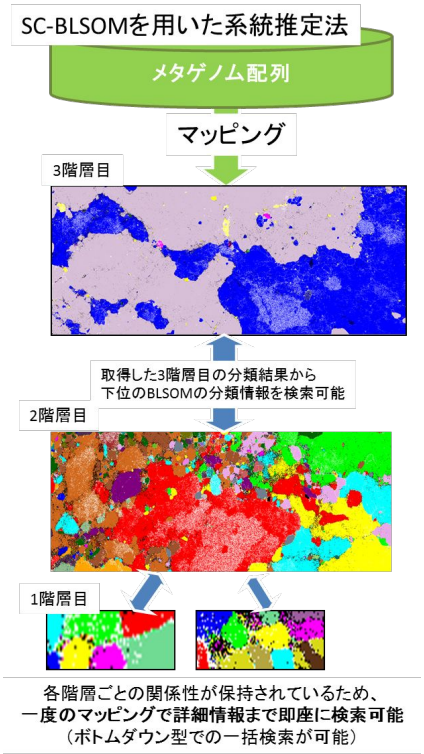

図 4 SC-BLSOM を用いた系統推定ワークフローの概要

（3）オリゴペプチド組成の距離関係に着目したタンパク質機能推定法の開発

Sargasso 海由来のメタゲノム配列データを用いて、開発手法との性能比較を行った。このうち、従来の相同性検索において、機能推定できた（すなわち、COGID を付与することができた）タンパク質 4,240 件を評価用データとした。

COG データベースに登録されているアミノ酸配列を断片化サイズ 200 aa で断片化を行い、2 連続アミノ酸、3 連続アミノ酸、4 連続アミノ酸の使用頻度計算したものをデータベースとし、評価用データについても同様に断片化サイズ 200aa で断片化した配列をクエリーとして、機能推定を行った。その結果を表 2 に示す。なお、表 2 における推定結果の割合の値は、（COGID が一致した件数）/（機能既知なタンパク質の件数）とした。

表 2 評価用データでの推定結果

データベースの種類	評価用データセットの推定結果 (%)
2 連続アミノ酸	97.38%
縮退 11 アミノ酸 3 連続	97.55%
縮退 6 アミノ酸 4 連続	97.24%
3 種で一致していた場合	97.05%

本手法において、相同性検索結果とほぼ同等の推定結果を得ることができた。また、相同性が低く機能推定が困難なタンパク質アミノ酸配列を対象に、機能推定を行ったところ、機能推定可能な配列を多く検出できた。

さらに、データベース更新に時間がかかることもなく、増加を続けるタンパク質にも容易に対応でき、相同性検索や機能モチーフ検索を補完する適用範囲の広いタンパク質の機能推定法を確立することができた。

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　9 件）

1. Ryo Nakao†, Takashi Abe†, Chihiro Sugimoto（†共同筆頭著者，他 1 人，2 番目）. Horizontally transferred genetic elements in the tsetse fly genome: an alignment-free clustering approach using batch learning self-organising map (BLSOM). *BioMed Research International*, Volume 2016, Article ID 3164624, 2016.(http://dx.doi.org/10.1155/2016/3164624)（査読有）.
2. Kazuki Fujinawa, Takashi Abe, Kazuya Watanabe（他 3 人，5 番目）. Genomic features of uncultured methylotrophs in activated-sludge microbiomes grown under different enrichment procedures. *Scientific Reports*, 6, 26650, 2016（査読有）.
3. Akihito Kikuchi, Toshimichi Ikemura, Takashi Abe. Development of Self-Compressing BLSOM for Comprehensive Analysis of Big Sequence Data. *BioMed Research International*, Volume 2015, Article ID 506052, 2015. (http://dx.doi.org/10.1155/2015/506052)（査読有）.
4. Shizuka Eshima, Takashi Abe, Ikuo Saiki（他 2 人, 3 番目）. Multi-pathway cellular analysis on crude natural medicines from Japanese Kampo prescriptions. *PLoS One*, 10, e0128872, 2015（査読有）.
5. Yoshiko Wada, Takashi Abe, Toshimichi Ikemura（他 3 人, 3 番目）. CG-containing oligonucleotides and transcription factor

binding motifs evidently enriched in human pericentric regions. *Genes & Genetic Systems*, 90, 43-53, 2015（**査読有**）.

6. Sho Ninomiya†, Takashi Abe†, Masayuki Nashimoto（**†共同筆頭著者，他 5 人, 3 番目**）. Potential Small Guide RNAs for tRNase ZL from Human Plasma, Peripheral Blood Mononuclear Cells, and Cultured Cell Lines. *PLoS One*, 10, e0118631, 2015. (https://doi.org/10.1371/journal.pone.0118631)（**査読有**）.
7. Ayaka Yamamuro, Atsushi Kouzuma, Takashi Abe, Kazuya Watanabe. Metagenomic analyses reveal the involvement of syntrophic consortia in methanol/electricity conversion in microbial fuel cells. *PLoS ONE*, 9, e98425, 2014. (https://doi.org/10.1371/journal.pone.0098425)（**査読有**）.
8. Yuki Iwasaki, Takashi Abe, Norihiro Okada, Toshimichi Ikemura（**他 2 人, 2 番目**）. Evolutionary Changes in Vertebrate Genome Signatures with Special Focus on Coelacanth. *DNA Research*, 21, 459-467, 2014（**査読有**）.
9. 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, 2014（**査読有**）.

〔学会発表〕（計　45 件）
（招待講演と国際会議での発表のみを記載）

1. **阿部貴志. 連続塩基組成に基づくウイルスゲノムの多様性の解明. 第 1 回内在性ウイルス様エレメント研究会**, 2016 **年** 12 **月** 16 **日（京都大学（京都府・京都市））（招待講演）**.
2. Ryo Nakao, Takashi Abe, Yongjin Qiu, May Thu, and Chihiro Sugimoto. First isolation of Chlamdiae from an ixodid tick collected in Japan. ISME2016, 21-26 Aug. 2016 (Montreal, Canada).
3. Takashi Abe, Shigehiko Kanaya, and Toshimichi Ikemura. Phylogenetic estimation and classification of metagenomic sequences on the basis of batch-learning self-organizing map. ISME2016, 21-26 Aug. 2016 (Montreal, Canada).
4. **阿部貴志.** A bioinformatics analysis for efficient knowledge discovery from big sequence data with BLSOM**. 第** 89 **回日本細菌学会総会,** 2016 **年** 3 **月** 23 **日～**25 **日（大阪国際交流センター（大阪府・大阪市））（招待講演）**.
5. Akihito Kikuchi, Shigehiko Kanaya, Toshimichi Ikemura, Takashi Abe. Development of Self-Compress BLSOM for comprehending big sequence data. GIW2014, 15-17 Dec. 2014 (Tokyo, Japan).
6. Takashi Abe, Hachiro Inokuchi, Yuko Yamada, Akira Muto and Toshimichi Ikemura. tRNADB-CE: tRNA gene database curated manually by experts. GIW2014, 15-17 Dec. 2014 (Tokyo, Japan).
7. **阿部貴志. メタゲノム解析を活用した新規微生物群の効率的な探索. 第** 157 **回日本獣医学会学術集会**, 2014 **年** 9 **月** 11 **日（北海道大学（北海道・札幌市））（招待講演）**.
8. Yongjin Qiu, Ryo Nakao, Takashi Abe, and Chihiro Sugimoto, Tick virome analysis using a high-throughput sequencing technology. TTP8, 24-29 Aug. 2014 (Cape Town, South Africa).

〔図書〕（計　2 件）

1. **阿部貴志（第** 3 **章 配列解析，分担執筆）バイオインフォマティクス入門，日本バイオインフォマティクス学会編集，慶応義塾大学出版会,** 90-113, 2015 **年** 08 **月**.
2. 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* (Ed. Karen E. Nelson), 666-670, Springer, 2015（**査読無**）.

〔その他〕
ホームページ等
1. 研究紹介
http://bioinfo.ie.niigata-u.ac.jp/

アウトリート活動
1. 阿部貴志，講習会講師. 平成 28 年度 新潟大学高度技術研修 「統計ソフトウェア「R」を活用したデータ分析コース」，2016 年 11 月（新潟県新潟市）
2. 阿部貴志，高校での出前講義講師，福島県立白河高等学校，2016 年 9 月（福島県白河市）

6．研究組織
(1)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