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　３０　年　　９　月　　７　日現在

機関番号：　３４２０４
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2017
課題番号：　２６３３０３３４

研究課題名（和文）生命科学分野の多様なビッグデータからの能率的知識発見手法の開発

研究課題名（英文）Efficient knowledge discovery from life-science big data

研究代表者
池村　淑道（Ikemura, Toshimichi）

長浜バイオ大学・バイオサイエンス学部・客員教授

研究者番号：５００２５４７５

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：オリゴヌクレオチド組成に着目したBLSOM(一括学習型SOM)はゲノム配列のビッグデータ解析に適しており、教師なし機械学習であることから想定外の新知識発見を可能にする。感染症RNAウイルスゲノムを対象にしたAIを用いた解析で、時系列的に方向性や再現性のあるオリゴヌクレオチド組成の変化を見出した。この成果を応用して効果が持続すると期待できる核酸医薬をデザインする手法を開発した。高等動物のBLSOM解析により、セントロメアの近傍領域でエピゲノミクスの代表的マーカであるCpGを含む特定の5連塩基や、多様な転写因子の結合配列が高密度に集中する領域を見出し核内配置における役割に関するモデルを提唱した。

研究成果の概要（英文）：We have developed a BLSOM (Batch-learning SOM) for oligonucleotide compositions, which is suitable for big data analysis of genome sequences. It is an unsupervised machine-learning and thus enables unexpected knowledge discoveries. By using the AI-method for genomes of RNA viruses such as influenza virus, ebolavirus and MERS coronavirus, we found time-series and reproducible changes in oligonucleotide composition and, for these highly mutable viruses, we have developed a method for designing oligonucleotide drugs with long-term efficacy. By BLSOM analyses on higher animals such as human and Coelacanth, we found regions enriched by specific 5-mers that include CpG, a representative epigenetic marker, or by binding sequences of various transcription factors. These specific regions cluster densely in the vicinity of centromeres, which are known to form chromocenters. Therefore, we proposed a model on their roles in nuclear organization.

研究分野：　ゲノム進化の情報解析

キーワード：　ビッグデータ　人工知能　時系列解析　メタゲノム解析　核酸医薬　自己組織化マップ　感染症ウイルス　オリゴヌクレイチド

１．研究開始当初の背景
生命科学分野には多様なビッグデータが蓄積しており、そこからの能率的な知識発見をする AI 等の情報学的手法の確立が急務といえる。我々のグループはゲノム配列のビッグデータを解析するのに適した教師無し機械学習「BLSOM」法を世界に先駆けて確立し、特許化してきた。

２．研究の目的
本研究開発では、生命科学分野の多様なビッグデータについて、オリゴヌクレオチド等の連文字頻度組成に着目し、以下の個別課題の解決を実践しながら、応用範囲の広い技術として発展させる。
（１）高等生物のゲノム機能解析への BLSOM の応用。
（2）分子進化速度の著しく速い病原性 RNA ウイルス類のゲノム配列解析への BLSOM の応用と実用的課題への適用。
（３）大量なメタゲノム配列ならびに tRNA 遺伝子配列に関する情報解析。
（４）ゲノム配列の更なるビッグデータ化を想定した BLSOM の改良。

３．研究の方法
我々の開発した BLSOM は、教師無しの学習アルゴリズムによる AI であり、特定のモデルや予備知識なしに、新規性の高い知識発見を可能にする。加えて、優れた画像表示能を備えており、大量情報からの能率的な知識発見に強力な手段を提供している。オリゴヌクレオチド組成に着目した BLSOM では、ゲノム断片配列が生物種や生物系統により分離（自己組織化）するので、大量ゲノム配列の生物系統の推定を可能にする。加えて、分子進化速度の著しく速い病原性 RNA ウイルスゲノムについては、オリゴヌクレオチド組成の時系列解析が有用な解析手法を提供する。

４．研究成果
（１）高等動物のゲノム機能解析への BLSOM の応用。ヒトゲノムについて、100kb の配列中の 5 連塩基組成の BLSOM 解析を行った所、セントロメアの近傍領域で、エピゲノミクスの代表的マーカと関係する CpG を含む特定の 5 連塩基や、多様な転写因子の結合配列が高密度に集中する領域を見出し、ヘテロクロマチン化と核内配置における役割についてのモデルを提唱した（論文 5）。ショウジョウバエとシーラカンスゲノムを含む広範な生物種に関する連続塩基組成の特徴（genome signature）の特定と、その生物学的意義については、論文 6, 8, 9 として発表した。
（２）分子進化速度の著しく速い病原性 RNA ウイルス類のゲノム配列解析への BLSOM の応用と実用課題への適用。2009 年から流行を開始した新型インフルエンザや西アフリカで流行していたエボラのように、人類は常にウイルスの引き起こす感染症の危険に曝されているが、その社会的な重要性から、大量な病原性ウイルスのゲノム配列が蓄積している。これらの大量ゲノム配列のオリゴヌクレオチド組成解析を行い、分子進化速度の速い RNA ウイルスについて、明瞭に方向性のある分子進化過程を明らかにできた(論文３)。自然宿主の細胞からヒト細胞へと増殖の場を変えた際に起こる、方向性のあるゲノム配列の変化であり、進化速度の速い RNA ウイルスに対して、持続性のある診断法や治療法（特に核酸医薬）の開発に役立つ基盤ゲノム情報を得ることができた(論文１)。
（３）大量なメタゲノム配列ならびに tRNA 遺伝子配列に関する情報解析。次世代シーケンサーが産出するメタゲノム配列を含む大量ゲノム配列は、比較的に短い配列が大半を占める。100 塩基程度の短い配列であっても、tRNA については完全長の遺伝子が得られる。我々のグループはこの tRNA 遺伝子のビッグデータを DB として公開しているが(論文７)、そのビッグデータより、AI を用いた知識発見を行った(論文２)。この研究成果は、ビッグデータ化した tRNA 遺伝子の DB の AI による高品質化にも役立つ。
（４）ゲノム配列の更なるビッグデータ化を想定した BLSOM 法の改良。環境試料由来のメタゲノム配列の場合、原核ならびに真核生物だけでなく、ウイルスの配列も混在する可能性が高い。メタゲノム配列の生物系統推定を BLSOM で行う場合、予め生物種既知の全ゲノム配列で大規模 BLSOM を作成しておく必要がある。ES のような我が国を代表するような HPC を用いても、この大規模 BLSOM の作成が困難になってきた。この課題を克服する目的で、既知生物種の生物系統別に BLSOM を作成し、それらを順次結合する BLSOM 法を開発した（論文４）。

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）全て査読あり。

① Wada K, Wada Y, Iwasaki Y, Ikemura T. Time-series oligonucleotide count to assign antiviral siRNAs with long utility fit in the big data era. *Gene Therapy* 24, 668-673. doi: 10.1038/gt.2017.76, 2017.

② Iwasaki Y, Abe T, Wada K, Wada Y, Ikemura T. An artificial intelligence approach fit for tRNA gene studies in the era of big sequence data. *Genes Genet Syst.* 2017 Sep 12;92(1):43-54. doi: 10.1266/ggs.16-00068, 2017.

③ Wada Y, Wada K, Iwasaki Y, Kanaya S, Ikemura T. Directional and reoccurring sequence change in zoonotic RNA virus genomes visualized by time-series word count. *Sci Rep.* 2016 Nov 3;6:36197. doi: 10.1038/srep36197, 2016.

④ Kikuchi A, Ikemura T, Abe T. Development of Self-Compressing BLSOM for Comprehensive Analysis of Big Sequence Data. *BioMed Res Int*, 2015: 506052, 2015. doi: 10.1151/2015/506052.

⑤ Wada Y, Iwasaki Y, Abe T, Wada K, Tooyama I, Ikemura T. CG-containing oligonucleotides and transcription factor-binding motifs are enriched in human pericentric regions. *Genes Genet Syst*, 90(1), 43-53, 2015. doi: 10.1266/ggs.90.43.

⑥ Iwasaki Y, Abe T, Okada N, Wada K, Wada Y, Ikemura T. Evolutionary changes in vertebrate genome signatures with special focus on coelacanth. *DNA Res.*, 21, 459-467, 2014. doi: 10.1093/dnares/dsu012.

⑦ Abe T, Inokuchi H, Yamada Y, Muto A, Iwasaki Y, Ikemura T. tRNADB-CE: tRNA gene database well-timed in the era of big sequence data. *Front Genet*, 5, 114. doi: 10.3389/fgene.2014.00114, 2014.

⑧ Abe T, Hamano Y, Ikemura T. Visualization of genome signatures of eukaryote genomes by Batch-Learning Self-Organizing Map (BLSOM) with a special emphasis on Drosophila genomes. *Biomed Res Int*, 2014: 985706, 2014. doi: 10.1155/2014/985706.

⑨ Yu Bai, Yuki Iwasaki, Shigehiko Kanaya, Yue Zhao, and Toshimichi Ikemura, A Novel Bioinformatics Method for Efficient Knowledge Discovery by BLSOM from Big Genomic Sequence Data, *Biomed Res Int*, 2014: 765648, 2014. doi: 10.1155/2014/765648.

〔学会発表〕（計 22 件）

① 池村淑道 、和田健之介 、和田佳子、岩崎裕貴。AI が明らかにした病原性 RNA ウイルスゲノムの方向性と再現性のある変化と薬効性の失われ難い核酸医薬設計。第 12 回日本ゲノム微生物学会年会。 2018 年 3 月（京都）。

② 和田健之介、和田佳子、岩崎裕貴、池村淑道。人工知能に導かれた病原性RNAウイルスの弱みを知る。日本遺伝学会　第89回大会。2017年9月（岡山）。

③ 和田健之介、和田佳子、岩崎裕貴、池村淑道。AIに導かれた病原性RNAウイルスの分子進化研究。日本進化学会　第19回大会。2017年8月（京都）。

④池村淑道、和田佳子、岩崎裕貴、和田健之介。人類に脅威を与える RNA ウイルス類の弱みを AI で探る：薬効が失われ難い siRNA のデザイン。第 11 回日本ゲノム微生物学会年会 2017 年 3 月（藤沢）。

⑤ Takashi Abe, Shigehiko Kanaya, and Toshimichi Ikemura. Phylogenetic estimation and classification of metagenomic sequences on the basis of batch-learning self-organizing map. ISME2016 (Montreal, Canada) 2016 年 10 月.

⑥ 和田佳子、和田健之介、岩崎裕貴、金谷重彦、池村淑道。ビッグデータを活用した予言<->検証の速やかなサイクルを可能にする分子進化学。日本遺伝学会第 88 回大会。2016 年 9 月（三島）。

⑦ 和田佳子、和田健之介、岩崎裕貴、金谷重彦、池村淑道。大量 RNA ウイルスゲノム配列のビッグデータ解析が可能にする新規性の高い分子進化学。日本進化学会第 18 回年会。2016 年 8 月（東京）。

⑧ 和田佳子、 和田健之介、 岩崎 裕貴、 金

谷 重彦、 池村淑道。エボラ・インフルエンザウイルスの方向性のあるゲノム配列変化。第 10 回日本ゲノム微生物学会年会。2016 年 3 月（東京）。

⑨ Toshimichi Ikemura. Big data analyses on genome sequences with special focus on oligonucleotide word count. ISM High Performance Computing Conference. Tokyo, Japan. 2015年10月.

⑩ 池村淑道、 和田佳子、 岩崎裕貴、 和田健之介、 金谷重彦。オリゴヌクレオチドのビッグデータ解析が検出する人獣共通感染症 RNA ウイルスゲノムの方向性のある変化。第 87 回日本遺伝学会。2015 年 9 月（仙台）。

⑪ Yoshiko Wada, Kennosuke Wada, Daisuke Isoda, Yuki Iwasaki, Shigehiko Kanaya and Toshimichi Ikemura. Big data bioinformatics for designing therapeutic oligonucleotides for ebolavirus disease. Taskforce Infectious Disease: Ebola Paris 2015. Institut Pasteur - Paris, France. 2015 年 5 月.

⑫ 和田佳子、岩崎裕貴 、磯田大典 、阿部貴志、和田健之介、池村淑道。ビッグデータ解析によるエボラ、インフルエンザ、エイズウイルス用の核酸医薬のデザイン。第 9 回日本ゲノム微生物学会年会。2015 年 3 月 (神戸)。

⑬ Akihito Kikuchi, Shigehiko Kanaya, Toshimichi Ikemura and Takashi Abe. Development of Self-Compress BLSOM for comprehending big sequence data. GIW2014 (Tokyo, Japan), 2014 年 12 月.

⑭ Takashi Abe, Hachiro Inokuchi, Yuko Yamada, Akira Muto and Toshimichi Ikemura. tRNADB-CE: tRNA gene database curated manually by experts. GIW2014, (Tokyo, Japan) 2014年12月.

⑮ 岩崎裕貴, 阿部貴志, 和田佳子, 和田健之介, 池村淑道。ビッグデータ時代の集団進化遺伝学・ゲノム研究のための教師なし学習解析。日本遺伝学会第 86 回大会。2014 年 9 月（長浜市）。

他 7 件

〔図書〕（計　1 件）
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），104-112，シーエムシー出版，2014 年.シーエムシー出版

〔産業財産権〕

○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
池村　淑道（IKEMURA, Toshimichi）
長浜バイオ大学・バイオサイエンス学部・客員教授
研究者番号：50025475

(2)研究分担者
和田　健之介（WADA, Kennosuke）
長浜バイオ大学・バイオサイエンス学部・教授
研究者番号：90231026

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
（　　）