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

令和　４ 年　 ６ 月 ２８ 日現在

機関番号：３４２０４
研究種目：基盤研究(C)（一般）
研究期間：2018～2021
課題番号：１８Ｋ０７１５１

研究課題名（和文）RNAウイルスゲノムのAIによる解析

研究課題名（英文）AI-supported studies of RNA viruses

研究代表者
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

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：新型コロナウイルスは高速に進化するRNAウイルスであり、社会的重要性から大量なゲノム配列が公開されている。我々の開発した連続塩基組成を対象にした教師無しAI：BLSOMはビッグデータ解析に適しており、多様な可視化機能により能率的な知識発見を可能にした。ヒト集団へ侵入後に連続塩基組成を定方向的に変化させており、コウモリで流行しているコロナウィルスの組成から、ヒトで流行している風邪コロナウイルスの組成へと近づく方向であった。20連のような長い連続塩基の集団内組成の時系列解析により、ウイルス集団内で急拡大する突然変異が特定でき、ウイルス増殖に有利な変異とリンクする中立変異の区別が可能な例も示された。

研究成果の学術的意義や社会的意義
人獣共通感染症RNAウイルスはヒト以外の宿主動物から突然にヒトへと侵入し、大半のヒトが効果的な免疫を持たないことから大流行を引き起こすので、人類はこの脅威に常に晒されている。加えてRNAウイルスは高速に進化するので、ワクチンを含む予防や治療薬ならびに診断薬の有効性が失われ易い。この難題の解決には、本研究で開発したゲノム配列の変化方向の特定と予測が重要となる。新型コロナウイルスの場合は基礎研究であっても研究成果の迅速な公開が重要であり、既に7報の査読付き論文として発表し、投稿中の2報もpreprintとして公開した。ウイルスゲノムの配列解析にAIを導入したことも先導的であり学術的意味が高い。

研究成果の概要（英文）：SARS-CoV-2 is a fast-evolving zoonotic RNA virus, and due to its social concern, a large number of genomic sequences have been published. Unsupervised AI for oligonucleotide composition (BLSOM: Batch Learning Self-Organizing Map) developed by us is suitable for big sequence data analysis and enables efficient knowledge discovery with various visualization functions. The oligonucleotide composition of the virus was changed in unidirectional manner (monotonic increase or decrease) after invading the human population; the direction was from the composition of the coronavirus prevalent in bats to that of the cold coronavirus prevalent in humans, as well as from the composition of bat mRNAs to that of human mRNAs. Time-series analysis of the composition of long oligonucleotide such as 20-mers can identify the rapidly expanding mutations in the virus population, and we have developed a method to distinguish advantageous mutations favoring viral growth from neutral mutations.

研究分野：情報解析

キーワード：新型コロナウイルス　COVID-19　人獣共通感染症　AI　RNAウイルス　ウイルス進化　オリゴヌクレオチド　自己組織化マップ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

(1)　インフルエンザウイルス・エボラウイルス・MERS コロナウイルス等の RNA ウイルスが引き起こす人獣共通感染症の脅威に、人類は常に曝されてきた。ゲノムが RNA であることから進化速度が速く、ワクチンや治療薬や診断医薬品の効果が失われ易く対処を困難にしており、従って多数のゲノム配列を継続的に解析する必要がある。蓄積する大量のウイルスゲノム配列から能率的に知識発見を行うには、ビッグデータ処理に適した AI の活用が重要となる。我々が開発してきた連続塩基組成を対象にした「教師無し型で、分離に寄与した原因を説明可能な (unsupervised explainable) AI: BLSOM(一括学習型自己組織化マップ)」は大量塩基配列の解析に適しており、500 万件の配列の大規模解析も可能で、新規性の高い知識発見を可能にしてきた。

(2)　進化は突然変異と呼ばれるランダムな素過程からスタートするので、ゲノム変化の予測は困難と考えられてきた。しかしながら、我々の BLSOM を用いた従来からの解析により、人獣共通感染症 RNA ウイルス類のゲノム配列の変化には、一定の方向性が存在することを見出してきた。多量なゲノム配列が蓄積していたインフルエンザウイルスの解析では、時系列的にゲノムの塩基組成や連続塩基組成が定方向に変化（単調増加や単調減少）しており、興味深いことに、数十年離れて大流行を開始した A 型インフルエンザの亜型間でも明瞭な再現性が見られ、トリで流行している亜型類の組成から遠ざかり、ヒトでのみ流行を繰り返している B 型株の組成へ近づく方向性であった。

(3)　新興感染症の原因となる人獣共通感染症 RNA ウイルスは、突然に非ヒト宿主動物からヒト集団へと侵入するが、大半のヒトが効果的な免疫を持たないことから、大流行を引き起こす可能性がある。このような大流行が起こった際には、社会的な要請とゲノム配列解読技術の目覚ましい発展から、短期間で大量のウイルス株のゲノム配列がデータベースから公開されると予想される。このような新たな RNA ウイルスの大流行を研究開始前から想定しており、そのような大流行が起きた際には、直ちに BLSOM や時系列解析を開始して、変化の方向性を推定することを交付申請書に予め記載していた。

２．研究の目的

(1)　本研究では連続塩基組成を対象にした教師無し型 AI である BLSOM を活用して、人獣共通感染症 RNA ウイルスの高速進化過程の実態、特に非ヒト宿主からヒト集団への侵入後の、方向性のある連続塩基組成の時系列変化の詳細を明らかにし、関与している分子機構を研究し、近未来予測と検証を繰り返す独自性と創造性のある進化学研究を推進する。

(2)　RNA ウイルス類がヒト以外の宿主動物類からヒト集団へと侵入した際に起こる、方向性と再現性のある時系列変化が興味深く、その方向性を生む分子機構の解明が重要となる。非ヒト細胞で増殖していたウイルスがヒトへと感染した際に、ヒト細胞側が最高に良い増殖条件を用意しているとは考えづらい。よりよく増殖するには、ウイルスゲノム側が変化する必要がある。ウイルス側の適応過程と考えられるが、そこに関与する分子機構としては、多様な宿主因子（タンパク質や RNA 等）の差異が関係すると考えられ、ヒトと非ヒト宿主動物の比較ゲノム解析も必要となる。

(3)　得られた成果をもとに、有効性が持続する治療薬や診断用試薬デザインのための有用情報を提供することも研究目的の一つである。

３．研究の方法

(1)　RNA ウイルスゲノムの 2〜3 連のような短い連続塩基の時系列変化には、モノヌクレオチド組成の変化が重要であり、RNA editing 活性を持つ宿主 APOBEC 酵素類の影響が明らかになってきた。４連以上の連続塩基になるとタンパク質や他の RNA 類との相互作用の可能性が想定され、連続塩基の時系列変化を知ることは、ウイルス側の適応過程に関与する分子機構を研究する手掛かりとなる。しかしながら連続塩基長が長くなれば、多数の変数(5 連塩基では 1024 種類)を取り扱う必要が生じるが、我々が先導的に開発してきた BLSOM は正にこの目的に適している。本研究では、人獣共通感染症を引き起こす RNA ウイルス類のなかでも 2019 年に流行を開始し、社会的にも深刻な影響を与えている新型コロナウイルスの大量ゲノム配列を対象に、BLSOM を活用した大規模 AI 解析ならびに時系列解析を行う。連続塩基組成の BLSOM は、配列相同検索には依存しておらずビッグデータ解析に適しており、加えて教師無し型 AI であることから、モデルや仮説なしに敢えて AI に知識発見の主要部分を任せるとの、新規性の高い特徴あるゲノム解析が可能になる。研究の開始時には予想もしていない新規性の高い知識発見を可能にするが、説明可能型 AI であることから、その知識発見に至る過程や理由を教えてくれる。

(2)　20 連のような長い連続塩基は PCR や核酸医薬のターゲットとなるので、長い連続塩基を対象にした技術開発も重要である。20 連のような長い連続塩基は、polyA tail を除いては、新型コロナウイルスゲノム上には 1 コピーしか存在しないので、流行開始後に出現した 20 連塩基に着目することは、ウイルスゲノムに起きた突然変異に着目することになる。言い換えれば、集団内頻度を急速に増大する 20 連配列を解析すれば、急速に集団内頻度を増大する変異を解析することになる。急拡大する変異類には、感染や増殖に有利な変異と、それにリンクする中立変異が混在するはずであり、その区別が興味深い。大量な配列が存在することから、詳細な集団内頻度の時系列解析を行い、その区別を試みる。

４．研究成果

(1)　2019 年から流行を開始した新型コロナウイルスについて、その社会的重要性から大量なゲノム配列が解読された。我々が開発してきた BLSOM は大量なゲノム配列解析に適しており，新型コロナウイルスの流行の全過程について、継続的にゲノムワイドでの変異蓄積の実態解明を試みた。新型コロナウイルスの場合には、様々な変異型が登場し、世界的な大流行を繰り返しており、研究対象の規模や具体的な課題も時系列に変遷する傾向にあった。本研究開発の課題名は「RNA ウイルスゲノムの AI による解析」であるが、新型コロナウイルスの大流行との社会的に重要な課題が登場したことから、時宜を得た研究課題となった。7 報の査読付き論文が発表でき、このウイルスの場合には公開の緊急性が重要であることから、現在投稿中の 2 報も preprint として公開している。最初にウイルスゲノム側での研究成果を主に時系列的に紹介し、最後に宿主ゲノム側の解析の成果を説明する。

(2)　2019 年 12 月より新型コロナウイルスの世界的流行がはじまったが、このウィルスの連続塩基組成に着目した AI ならびに時系列解析を行ったところ、他の人獣共通感染症 RNA ウイルスでの我々の以前からの研究結果と一致して、月単位でも観察できる、連続塩基組成の方向性を持った時系列変化（単調増加や単調減少）が見られた（Wada et al., Gene, 2020）。この方向性を持った変化は、2〜3 連のような短い連続塩基の場合にはクレード（clade）には依存しない変化であり、コウモリで流行しているコロナウィルスの連続塩基組成から、ヒトで流行している風邪コロナウイルスの組成へと近づく方向性を示していた(Iwasaki et al., BMC Microbiol, 2021)。連続塩基

組成の変化の視点からは、近未来予測が可能なことが判明したので、有効性が持続する治療薬や診断用試薬デザインのための有用情報を提供できた。方向性のある変化の分子機構としては、APOBEC3 酵素による RNA editing の効果が我々を含む複数のグループより提唱されている。

15-mer 以上の長い連続塩基になると 新型コロナウイルスゲノム上で大半が 1 コピーしか存在しない。流行開始後に出現した 15-mer に着目することで、新規に入った変異との関係付けが可能になる。集団内頻度を急拡大した変異に関係する 15-mer 類に着目して、BLSOM 解析を行ったところ、教師無しの AI でありながらグレード(clade)別の分離（自己組織化）が起きていた。既知のクレード内、特に GR に関しては明瞭な内部分岐が起きており、少なくとも７種類のサブグループを観察した。説明可能型の AI であり、それらの内部分岐に関与する変異も特定できた（Wada et al., GENE, 2020）。

(3) 新型コロナウイルスゲノム配列の蓄積は膨大であり、GISAID に収録された完全ゲノム配列に絞っても 2020 年内に 40 万件を超えた。そのような大量配列からの能率的な知識発見については、新たな技術上の改良が必要になった。BLSOM 法の強力な可視化機能を拡張し、3D 表示機能を用いることで、世界の流行地域内で流行拡大に強く寄与しているクレード(clade)やそのサブグループの詳細な検出を可能にした。時系列解析を組み合わせて、世界の各地域における危険クレードの特定も行った(Abe et al., Data Science Journal, 2021)。この研究開発は 5 連塩基を主対象にしたが、このような短い連続塩基では、変異との直接的な関連付けは出来ない。より長い連続塩基組成の時系列解析を行い、世界各地で出現した多数の変異の内で、世界規模で急拡大した変異を特定する手法を確立した（Ikemura et al., Genes Genet Syst, 2021）。この手法を用いて、主に 15 連や 20 連塩基に着目し、流行開始の 2019 年の全登録株において非存在で、変異により出現した連続塩基のなかで、各月の集団内頻度が 10%以上は上昇している例を選別し、155 種類の 20 連塩基を特定し、変異との関係付けを行った（Iwasaki et al., BMC Microbiol, 2022）。この手法で見出した、集団内で急拡大している変異部位は機能的重要性が高い可能性があり、有効性の高い核酸医薬の候補になると考えられ，そのデザインの基盤情報が提供できた。

(4) オミクロン株は他のバリアントと系統的に大きくかけ離れており、独自性の高い進化をしてきた。2019 年の新型コロナ流行の開始時には非存在で，オミクロン系統で変異により生じた 20-mer は 10 万種類を超えるが，それらの大半はオミクロン集団内での頻度が極めて低いが、集団内頻度が 50%を超える 20-mer も 1000 種類以上は存在していた。オミクロン系統で起きた代表的変異に起因する 20-mer と考えられる。他のバリアントと共有する変異も見られたが、興味深いことに、変異別に共有するバリアントの種類が異なっていた。オミクロンの特異的な進化的起源を考えると、それらバリアントとオミクロンで共進化的に集団内頻度を拡大した、機能的に有利な変異に起因する可能性が高い。言い換えれば、有利な変異にヒッチハイク的にリンクした中立変異である可能性が低い。これらの大半はアミノ酸を変化させる変異であったが、N 遺伝子直前の遺伝間変異の例も見出された。機能的に有利な変異の候補が遺伝間に存在することは興味深い(Ikemura et al., Preprint bioRxiv, 2022)。

(5) 本研究開始前の研究により、エボラやインフルエンザウイルス等の人獣共通感染症 RNA ウイルスが、非ヒト宿主よりヒト集団へ侵入後に、ゲノムの塩基組成や連続塩基組成を定方向的に変化させることを見出しており、宿主へのウイルスの適応過程と考えていた。この適応過程の分子機構を知るには、宿主動物側での比較ゲノム解析が重要であり、AI を用いたこの視点での研究開発を計画していた。研究開始当初は、まだ新型コロナウイルスは登場しておらず、ヒトゲノムについて AI を用いた連続塩基組成解析を行ったところ、セントメア近傍のヘテロクロマチン領域に、転写因子の結合配列(TFBS)の大規模構造を見出し Mb-level TFBS islands と命名した

（Wada et al., Genes & Genetic Systems, 2020）。この TFBS islands より小規模ではあるが、テロメア近傍に Mb-level CpG islands も見出している。新型コロナウイルスの登場に伴い、その自然宿主であるコウモリに関する関心も高まり、6 種類のコウモリについて高品質のゲノム配列が公開された。それらコウモリとヒトの AI を用いた比較ゲノムを行ったところ、いずれのコウモリについても、テロメア近傍に大規模な Mb-level CpG islands が観察されたが、TFBS islands は小規模であった(Ikemura et al., Preprint Research Square, 2022)。

新型コロナウイルスの宿主への適応過程を解析する目的で、コウモリ由来のコロナウイルス、ヒトの風邪コロナウイルス、ならびに新型コロナウイルスの連続塩基組成の比較解析を行い、併せて宿主側の比較ゲノム解析も行った。ヒト風邪コロナウイルスの連続塩基組成は、コウモリ由来のコロナウイルス組成とは明瞭に異なっており、其々が対応する宿主の mRNA の組成と似通っており、特にヒト風邪コロナウイルス側では、CpG のレベルがコウモリ由来コロナウイルスより顕著に低かった。新型コロナウイルスの連続塩基組成も、時系列的にヒト mRNA の組成に近づく方向性を示したが、興味深いことに CpG のレベルは流行の開始時において既に顕著に低かった。ヒトでの流行を容易にした理由と考えられるので、自然界に存在するコロナウイルスの内で将来に大流行を引き起こす危険株の推定に役立つ知見と考えている(Iwasaki et al., Preprint Research Square, 2022)。

５．主な発表論文等

〔雑誌論文〕　計9件（うち査読付論文　7件／うち国際共著　0件／うちオープンアクセス　9件）

１．著者名 Kennosuke Wada, Yoshiko Wada, Toshimichi Ikemura	４．巻 763S
２．論文標題 Time-series analyses of directional sequence changes in SARS-CoV-2 genomes and an efficient search method for candidates for advantageous mutations for growth in human cells	５．発行年 2020年
３．雑誌名 Gene	６．最初と最後の頁 100038
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.gene.2020.100038	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Yuki Iwasaki, Takashi Abe, Toshimichi Ikemura	４．巻 21
２．論文標題 Human cell-dependent, directional, time-dependent changes in the mono- and oligonucleotide compositions of SARS-CoV-2 genomes	５．発行年 2021年
３．雑誌名 BMC Microbiol.	６．最初と最後の頁 89
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s12866-021-02158-6	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Takashi Abe, Ryuki Furukawa, Yuki Iwasaki, Toshimichi Ikemura	４．巻 20
２．論文標題 Time-series trend of pandemic SARS-CoV-2 variants visualized using batch-learning self-organizing map for oligonucleotide compositions	５．発行年 2021年
３．雑誌名 Data Science Journal	６．最初と最後の頁 29
掲載論文のDOI（デジタルオブジェクト識別子） 10.5334/dsj-2021-029	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Iwasaki Yuki、Abe Takashi、Wada Kennosuke、Wada Yoshiko、Ikemura Toshimichi	４．巻 22
２．論文標題 Unsupervised explainable AI for molecular evolutionary study of forty thousand SARS-CoV-2 genomes	５．発行年 2022年
３．雑誌名 BMC Microbiology	６．最初と最後の頁 73
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s12866-022-02484-3	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Kennosuke Wada, Yoshiko Wada, Toshimichi Ikemura	4．巻 95
2．論文標題 Mb-level CpG and TFBS islands visualized by AI and their roles in the nuclear organization of the human genome	5．発行年 2020年
3．雑誌名 Genes & Genetic Systems	6．最初と最後の頁 29-41
掲載論文のDOI（デジタルオブジェクト識別子） 10.1266/ggs.19-00027	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Toshimichi Ikemura, Yuki Iwasaki, Kennosuke Wada, Yoshiko Wada, Takashi Abe	4．巻 96
2．論文標題 AI for the collective analysis of a massive number of genome sequences: various examples from the small genome of pandemic SARS-CoV-2 to the human genome	5．発行年 2021年
3．雑誌名 Genes Genet Syst	6．最初と最後の頁 165-176
掲載論文のDOI（デジタルオブジェクト識別子） 10.1266/ggs.21-00025	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Yuki Iwasaki, Toshimichi Ikemura, Kennosuke Wada, Yoshiko Wada, Takashi Abe	4．巻 in press
2．論文標題 Comparative genomics of the human genome and six bat genomes using AI: Mb-level CpG and TFBS islands	5．発行年 2022年
3．雑誌名 BMC Genomics	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.21203/rs.3.rs-1323531/v1	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Yuki Iwasaki, Takashi Abe, Toshimichi Ikemura	4．巻 preprint
2．論文標題 Oligonucleotide usage in coronavirus genomes mimics that in exon regions in host genomes	5．発行年 2022年
3．雑誌名 Research Square	6．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.21203/rs.3.rs-1604205/v1	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Toshimichi Ikemura, Yuki Iwasaki, Kennosuke Wada, Yoshiko Wada, Takashi Abe	４．巻 preprint
２．論文標題 AI-based search for convergently expanding, advantageous mutations in SARS-CoV-2 by focusing on oligonucleotide frequencies	５．発行年 2022年
３．雑誌名 bioRxiv	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1101/2022.05.13.491763	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計9件（うち招待講演　3件／うち国際学会　1件）

１．発表者名 Toshimichi Ikemura
２．発表標題 Time-series analysis of directional sequence changes in SARS-CoV-2 genomes and an unsupervised explainable AI for studying corona virus genomes
３．学会等名 International Symposium on Data Science（招待講演）（国際学会）
４．発表年 2020年

１．発表者名 岩﨑　裕貴 、阿部　貴志 、池村　淑道
２．発表標題 宿主細胞特異的な環境に依存した SARS-CoV-2 連続塩基組成の変化
３．学会等名 第 15 回日本ゲノム微生物学会年会
４．発表年 2021年

１．発表者名 和田　健之介、和田　佳子、池村　淑道
２．発表標題 ヒトゲノムに見られるMb規模のCpGとTFBS islandsとそれらの構造と機能
３．学会等名 日本遺伝学会第91回大会
４．発表年 2019年

１．発表者名
和田　健之介、和田　佳子、池村　淑道

２．発表標題
AI に導かれた感染症RNA ウィルスの分子進化研究

３．学会等名
日本遺伝学会　第90回大会（招待講演）

４．発表年
2018年

１．発表者名
和田　健之介、和田　佳子、池村　淑道

２．発表標題
時系列解析とAIを用いた感染症RNAウィルス増殖に関与する宿主 miRNAの推定

３．学会等名
日本進化学会第20回大会

４．発表年
2018年

１．発表者名
池村　淑道、和田　佳子、和田　健之介

２．発表標題
ゲノムのビッグデータからのAIによる想定外の知識発見とその進化学への応用

３．学会等名
第２回木村資生記念進化学セミナー（招待講演）

４．発表年
2018年

１．発表者名
岩﨑　裕貴、阿部　貴志、安藤 大喜、池村　淑道

２．発表標題
コロナウイルスの宿主適応に関連したgenome signatureの抽出

３．学会等名
第68回日本ウイルス学会学術集会

４．発表年
2021年

１．発表者名 岩﨑　裕貴、 安藤　大喜、 阿部　貴志、 池村　淑道
２．発表標題 宿主環境特異的なコロナウイルスゲノムの特徴抽出
３．学会等名 第93回日本遺伝学会大会
４．発表年 2021年

１．発表者名 安藤　大喜、齊川　幸人、岩﨑　裕貴、阿部　貴志、和田　健之介、 和田　佳子、池村　淑道
２．発表標題 新型コロナウイルスの宿主適応に関連したゲノム変化
３．学会等名 第16回日本ゲノム微生物学会年会
４．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	池村　淑道 (Ikemura Toshimichi) (50025475)	長浜バイオ大学・バイオサイエンス学部・客員教授 (34204)	

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	和田　佳子 (Wada Yoshiko)	長浜バイオ大学・バイオサイエンス学部・特任講師 (34204)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関