

令和 7 年 6 月 19 日現在

機関番号：14401

研究種目：基盤研究(C)（一般）

研究期間：2020～2024

課題番号：20K06767

研究課題名（和文）MAFFT多重配列アラインメントプログラムの機能拡張

研究課題名（英文）Extension of MAFFT multiple sequence alignment program

研究代表者

加藤 和貴（KATOH, Kazutaka）

大阪大学・微生物病研究所・准教授

研究者番号：70378868

交付決定額（研究期間全体）：（直接経費） 3,200,000円

研究成果の概要（和文）：本計画は多重配列アラインメントプログラムMAFFTの機能拡張を目的としていたが、研究期間中に起こったコロナウイルスの感染の拡大に対処するために、このウイルスの配列解析の支援に特に注力した。GISAID EpiCoVデータベース構築に適したMAFFTプログラムの新しいオプションを、担当者の要請に基づき提供した。また、大阪大学で行っている配列アラインメントサービスを拡張してこの解析に適したオプションを利用可能にした。並行して、その他の多様な解析の支援を行った。たとえば、絶滅種の化石に由来する配列と現存種の比較解析、行動分析学的研究のために用いるための改造などである。

研究成果の学術的意義や社会的意義

計画段階では、本研究で開発した多重配列アラインメント手法を用いて種々の分子生物学的解析を支援し、それらを通じて研究成果を社会に還元することを目指していた。しかし、コロナウイルスの解析に開発手法が広く使われた結果、当初の計画以上に直接的かつ短期間の内に成果を実用に応用することができた。例えば、SARS-CoV-2の起源を探る多数の研究（Lu et al. 2020; Zhou et al. 2020など）においてMAFFTによるアラインメントに基づく系統樹推定が頻りに用いられた。さらに、大阪大学で行っているアラインメント計算サービスを通じて、大規模感染に対処する多数の研究者を直接支援した。

研究成果の概要（英文）：The primary objective of this project was to enhance the functionality of the multiple sequence alignment program MAFFT. However, considerable effort was redirected during the research period to support sequence analyses of the coronavirus in response to the global outbreak. New options of the MAFFT program, suitable for constructing the GISAID EpiCoV database, were developed and provided at the request of the relevant personnel. In addition, the sequence alignment service operated at Osaka University was extended to make these options available for such analyses.

In parallel, support was also provided for a variety of other analyses. These included comparative analyses between sequences derived from extinct species and those of extant species, and modifications of the program for applications in behavior analysis.

研究分野：バイオインフォマティクス

キーワード：配列アラインメント 計算サービス コロナウイルス 分子進化

様式 C-19、F-19-1 (共通)

1. 研究開始当初の背景

研究開始当初、多重配列アラインメント手法の改良の方向として以下の点を計画していた。

- 互いに類似度の低い入力配列の間の信頼性の高いアラインメントを得るためには、全ペアのアラインメントの利用や反復改善法など計算コストの高い方法が必要である。このような計算を可能にするためのアルゴリズム上または実装上の工夫が必要である。
- 遠縁のアミノ酸配列の間のアラインメントには、タンパク質の立体構造情報の活用も有効である。
- ロングリードシーケンサーのデータの処理、特に、シーケンサーのエラーに由来する挿入・欠失・置換を、通常の進化過程によるものと同様に扱うのは不適切である。その点を適切に考慮して得られた多重配列アラインメントをコンセンサスの計算に用いれば、信頼性の高い配列が得られるはずである。
- データの性質によって適切なアラインメント計算手法は異なる。適切なオプションを簡単に選択できるように、ウェブサービス、コマンドラインプログラム、GUI のユーザーインターフェイスの向上が必要である。

2. 研究の目的

本研究は多重配列アラインメントプログラム MAFFT の機能拡張を目的とする。当初、上に述べた問題の解決に向けた開発を計画していた。しかし、新型コロナウイルスの感染拡大により方針を変えた。すなわち、コロナウイルスゲノム配列の解析に MAFFT プログラムがよく利用されたため、この問題への対応を目的として、ウェブサービスの改良とユーザーへの対応を行った。

3. 研究の方法

まず、上に挙げた点の中のロングリードシーケンサーのデータの処理の方向への拡張を進めた。しかし、コロナウイルスの感染拡大に伴って、多くの研究者が MAFFT アラインメント計算サービスを使ってウイルスゲノム配列のアラインメントを試みるため、大阪大学に設置している計算サーバーの負荷が増大した。そのため方針を転換した。急増する計算需要に対応するため、以下の技術的対応により、処理能力の増強を図った。

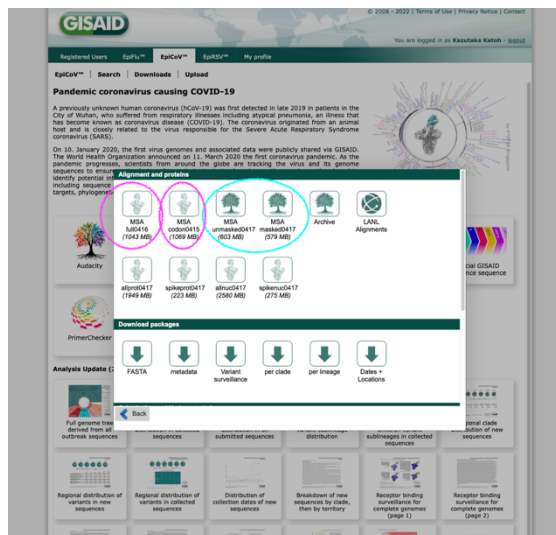
- 共同研究者の Daron Standley 教授の協力を得て、このサービスに割り当てるハードウェアを増強した。
- サービス内部のコードの見直しを行った。
- この問題により適した計算オプションを選択可能にした。
- MAFFT サービスと計算資源を共有している別のウェブサーバーが不正侵入を受けたため、研究室で運用している計算サービスとクラスタ計算機全体のアップデートを進め、セキュリティ対策を強化した。

MAFFT サービスの過負荷状態を受けて、計算資源の効率的な使用のために、計算時間やメモリを多く消費するケースの見直しを行った。多重配列アラインメントの計算量が多くなる典型的な問題は、ドメインシャッフリングを受けたタンパク質や、逆位や転座を含むゲノムデータである。通常の多重配列アラインメントは全配列の相同な座位が同じ順番に現れることを仮定するので、このようなデータのアラインメントは計算できない。このような問題に計算資源を割くことを避けるために、LAST ローカルペアワイスアラインメントプログラムを用いて、対応する座位をあらかじめ切り出し、その部分の多重配列アラインメントを MAFFT によって計算する機能をウェブサーバーに追加した。この機能を実装した動機は計算サービスの負荷の軽減であるが、下に述べる J. T. Fontes 博士らによる研究の支援にこの機能を活用した。

4. 研究成果

本来の計画にしたがって、ロングリードシーケンサーのデータ解析のため東京大学 Martin C. Frith 博士による lamassemble プログラムの開発に協力し、この方法を記述する論文 (Frith, Mitsuhashi, Katoh 2021) を公表した。いくつかの応用 (Nakamura et al 2020; Mitsuhashi et al 2020; Lei et al 2020) にも協力した。

コロナウイルスゲノムの網羅的なデータベースである GISAID EpiCoV の構築に協力した。GISAID の Raphael Tze Chuen Lee 博士の要請に基づき、二つの新しいオプションを提供した。(1) コドンを考慮した塩基配列アラインメントを実装した。現時点では、特定のオプションとの組み合わせたときのみ動作する。すなわち、既存のアラインメントに新しい配列を付け加えるとき、既存のアラインメントにおけるタンパク質コード領域の位置と読み枠をユーザが指定すれば、コドンをなるべく壊さないようにギャップを挿入する。(2) 既存アラインメントを単一配列で代表させて、新しい配列を追加する近似的な計算を利用可能にした。その結果、特に本数の多いアラインメントの構築が容易になった。図に示した GISAID からアラインメントをダウンロードする画面の中の、シアンで囲んだ部分が従来の MAFFT のオプションで計算されたアラインメントへのリンク、マゼンタで囲んだ部分が今回新しく提供したオプションを用いたアラインメントへのリンクである。



SARS-CoV-2 配列のアラインメントを試している過程で、このグループに近縁でコウモリやセンザンコウに感染するグループと、ヒトに感染するようになったグループの間で、アミノ酸置換が起こっている座位が共通していることに気づき報告した (Katoh & Standley 2021)。このように置換しやすい座位の一致は、より遠縁な関係にある SARS-CoV と SARS-CoV-2 の間には見られない。そのため SARS-CoV-2 の系統に共通する感染機構に関わっているいくつかの座位が、正の淘汰を受けて急速に置換されている可能性を指摘した。

MAFFTに限らず、Clustal Omega など多重配列アラインメントに関するいろいろなツールを解説した書籍 Multiple Sequence Alignment (Katoh ed. 2021) を編集した。

ウェブサービスを通じた研究支援に加えて、必要に応じて個別的な支援も行った。

- 京都府立医科大学星野温博士による、逃避変異が出現しにくく、さらにコウモリなどに感染するグループにも効果の見られる、高親和性 ACE2 製剤の研究に協力した (Ikemura et al 2022; Science Translational Medicine)。
- 米国ハーバード大学 Scott Edwards 教授らによる絶滅種モアのゲノム計画のために、モアと現生種シギダチョウの間の染色体全体程度の長さのグローバルアラインメントの計算を支援した (Edwards et al 2024; Science Advances)。
- 一般的な文字列のアラインメントを計算するオプションを実装し、生物学を離れた応用を試みた。ボードゲームを用いた行動分析的解析のためにプログラムを改造し、米国ノーステキサス大学 April Becker 博士らに提供した。
- 遠縁な配列の間の立体構造を考慮したアラインメントの現実の問題への適用として、東京薬科大学山岸明彦教授のグループによる、異なるアミノアシル tRNA 合成酵素 (ARS) の間の進化的関係を推定する研究を支援した。この解析には配列上・立体構造上の類似性の観察できる限界付近にある遠縁の配列のアラインメントが必要である。彼らは、MAFFT プログラムの制約付きアラインメント計算機能を利用して、既知のモチーフに基づくアラインメントを半自動的に求め、複数の祖先的タンパク質の配列を推定し、そこから ARS の基質特異性を推定した。その結果、全生物の最後の共通祖先 (LUCA; コモノート) の段階で既に ARS のアミノ酸特異性が現在と同程度に確立していたと推測された (Furukawa et al 2022; Journal of Molecular Evolution)。さらに前の時期、RNA ワールドにおける翻訳系の進化過程として、彼らは以下のシナリオを示した。タンパク質 ARS が出現する前にリボザイム ARS がまず出現し、リボザイム ARS による翻訳系が成立した。その後、より高い触媒活性や基質特異性を備えたタンパク質性 ARS が自然選択によってリボザイム ARS を置き換えた。さらにその後、細胞内で保持される遺伝情報の増加に伴い、RNA に比べて情報の保存性に優れる DNA が遺伝物質として有利となり、現在の DNA-RNA-タンパク質からなるシステムに移行したというものである。この論文に続いて、東京薬科大学横堀伸一博士らは、複数の祖先的遺伝子を大腸菌で発現させ可溶性タンパク質を得、これらを精製し祖先的 ARS のアミノ酸特異性を検証する実験を進めている。
- 環境 DNA (eDNA) メタバーコーディングによく用いられているミトコンドリアゲノム上の複数

の領域の比較し、どの領域がどのような場合に適しているか明らかにするための、ポルトガル Minho 大学の J. T. Fontes 博士らの解析を支援した。この解析では、NCBI GenBank に登録されている多量の部分的および完全なミトコンドリアゲノムデータから問題の領域を取り出しその部分の多重配列アラインメントを計算するために、方法の項で説明したウェブサービスを用いた。解析の結果を、Fontes et al (2025; Metabarcoding and Metagenomics) で報告した。

- 哺乳類の適応免疫系において、T 細胞受容体遺伝子と IG 遺伝子は、V(D)J 遺伝子再構成によって高度な多様性を実現する。ゲノム上の多数の V, D, J 遺伝子を特定することは、通常の遺伝子発見アルゴリズムには困難であり、手作業が必要である。そのため、この領域が正確にアノテートされた生物種は少ない。このことは、これらの遺伝子のゲノム上の分布に基づく進化学的議論を困難にしている。ヒトまたはマウスの V(D)J 遺伝子とそれ以外の哺乳類のゲノムのローカルアラインメントにもとづいてこれらの遺伝子の位置を推定する時、スコアの閾値が問題となる。既存のアノテーションを教師データとした機械学習によって適切な閾値の決定を試みている。この研究の過程で、大学院生の Zhou Hao 氏がウシゲノム上に新たな TRGJ 遺伝子を発見し報告した (Zhou et al 投稿中)。

5. 主な発表論文等

〔雑誌論文〕 計13件（うち査読付論文 11件／うち国際共著 4件／うちオープンアクセス 8件）

1. 著者名 Ikemura Nariko, Taminishi Shunta, Inaba Tohru, Arimori Takao, Motooka Daisuke, Katoh Kazutaka, Kirita Yuhei, Higuchi Yusuke, Li Songling, Suzuki Tatsuya, Itoh Yumi, Ozaki Yuki, Nakamura Shota, Matoba Satoaki, Standley Daron M., Okamoto Toru, Takagi Junichi, Hoshino Atsushi	4. 巻 14
2. 論文標題 An engineered ACE2 decoy neutralizes the SARS-CoV-2 Omicron variant and confers protection against infection in vivo	5. 発行年 2022年
3. 雑誌名 Science Translational Medicine	6. 最初と最後の頁 eabn7737
掲載論文のDOI (デジタルオブジェクト識別子) 10.1126/scitranslmed.abn7737	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 Standley Daron M., Nakanishi Tokuchiro, Xu Zichang, Haruna Soichiro, Li Songling, Nazlica Sedat Aybars, Katoh Kazutaka	4. 巻 14
2. 論文標題 The evolution of structural genomics	5. 発行年 2022年
3. 雑誌名 Biophysical Reviews	6. 最初と最後の頁 1247 ~ 1253
掲載論文のDOI (デジタルオブジェクト識別子) 10.1007/s12551-022-01031-8	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 Katoh Kazutaka, Standley Daron M.	4. 巻 4
2. 論文標題 Emerging SARS-CoV-2 variants follow a historical pattern recorded in outgroups infecting non-human hosts	5. 発行年 2021年
3. 雑誌名 Communications Biology	6. 最初と最後の頁 1134
掲載論文のDOI (デジタルオブジェクト識別子) 10.1038/s42003-021-02663-4	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 Furukawa Ryutaro, Yokobori Shin-ichi, Sato Riku, Kumagawa Taimu, Nakagawa Mizuho, Katoh Kazutaka, Yamagishi Akihiko	4. 巻 90
2. 論文標題 Amino Acid Specificity of Ancestral Aminoacyl-tRNA Synthetase Prior to the Last Universal Common Ancestor Commonote commonote	5. 発行年 2022年
3. 雑誌名 Journal of Molecular Evolution	6. 最初と最後の頁 73-94
掲載論文のDOI (デジタルオブジェクト識別子) 10.1007/s00239-021-10043-z	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 Nakamura Haruko, Doi Hiroshi, Mitsuhashi Satomi, Miyatake Satoko, Katoh Kazutaka, Frith Martin C., Asano Tetsuya, Kudo Yosuke, Ikeda Takuya, Kubota Shun, Kunii Misako, Kitazawa Yu, Tada Mikiko, Okamoto Mitsuo, Joki Hideto, Takeuchi Hideyuki, Matsumoto Naomichi, Tanaka Fumiaki	4. 巻 65
2. 論文標題 Long-read sequencing identifies the pathogenic nucleotide repeat expansion in RFC1 in a Japanese case of CANVAS	5. 発行年 2020年
3. 雑誌名 Journal of Human Genetics	6. 最初と最後の頁 475 ~ 480
掲載論文のDOI (デジタルオブジェクト識別子) 10.1038/s10038-020-0733-y	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 Mitsuhashi Satomi, Ohori Sachiko, Katoh Kazutaka, Frith Martin C., Matsumoto Naomichi	4. 巻 12
2. 論文標題 A pipeline for complete characterization of complex germline rearrangements from long DNA reads	5. 発行年 2020年
3. 雑誌名 Genome Medicine	6. 最初と最後の頁 67
掲載論文のDOI (デジタルオブジェクト識別子) 10.1186/s13073-020-00762-1	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 Lei Ming, Liang Desheng, Yang Yifeng, Mitsuhashi Satomi, Katoh Kazutaka, Miyake Noriko, Frith Martin C., Wu Lingqian, Matsumoto Naomichi	4. 巻 65
2. 論文標題 Long-read DNA sequencing fully characterized chromothripsis in a patient with Langer-Giedion syndrome and Cornelia de Lange syndrome-4	5. 発行年 2020年
3. 雑誌名 Journal of Human Genetics	6. 最初と最後の頁 667 ~ 674
掲載論文のDOI (デジタルオブジェクト識別子) 10.1038/s10038-020-0754-6	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1. 著者名 Saputri Dianita S., Li Songling, van Eerden Floris J., Rozewicki John, Xu Zichang, Ismanto Hendra S., Davila Ana, Teraguchi Shunsuke, Katoh Kazutaka, Standley Daron M.	4. 巻 11
2. 論文標題 Flexible, Functional, and Familiar: Characteristics of SARS-CoV-2 Spike Protein Evolution	5. 発行年 2020年
3. 雑誌名 Frontiers in Microbiology	6. 最初と最後の頁 2112
掲載論文のDOI (デジタルオブジェクト識別子) 10.3389/fmicb.2020.02112	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 Frith Martin C., Mitsuhashi Satomi, Katoh Kazutaka	4. 巻 2231
2. 論文標題 Iamassemble: Multiple Alignment and Consensus Sequence of Long Reads	5. 発行年 2020年
3. 雑誌名 Multiple Sequence Alignment (Methods in Molecular Biology)	6. 最初と最後の頁 135 ~ 145
掲載論文のDOI (デジタルオブジェクト識別子) 10.1007/978-1-0716-1036-7_9	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 Rozewicki John, Li Songling, Katoh Kazutaka, Standley Daron M.	4. 巻 2232
2. 論文標題 Analysis of Protein Intermolecular Interactions with MAFFT-DASH	5. 発行年 2020年
3. 雑誌名 Multiple Sequence Alignment (Methods in Molecular Biology)	6. 最初と最後の頁 163 ~ 177
掲載論文のDOI (デジタルオブジェクト識別子) 10.1007/978-1-0716-1036-7_11	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 -

1. 著者名 Chang P. K., Chang T.D., Katoh K.	4. 巻 722231
2. 論文標題 Deciphering the origin of <i>Aspergillus flavus</i> NRRL21882, the active biocontrol agent of <i>Afla Guard</i> (R)	5. 発行年 2021年
3. 雑誌名 Letters in Applied Microbiology	6. 最初と最後の頁 509 ~ 516
掲載論文のDOI (デジタルオブジェクト識別子) 10.1111/lam.13433	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1. 著者名 Fontes Joao T., Katoh Kazutaka, Pires Rui, Soares Pedro, Costa Filipe O.	4. 巻 8
2. 論文標題 Benchmarking the discrimination power of commonly used markers and amplicons in marine fish (e)DNA (meta)barcoding	5. 発行年 2024年
3. 雑誌名 Metabarcoding and Metagenomics	6. 最初と最後の頁 e128646
掲載論文のDOI (デジタルオブジェクト識別子) 10.3897/mbmg.8.128646	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 該当する

1. 著者名 Edwards Scott V., Cloutier Alison, Cockburn Glenn, Driver Robert, Grayson Phil, Katoh Kazutaka, Baldwin Maude W., Sackton Timothy B., Baker Allan J.	4. 巻 10
2. 論文標題 A nuclear genome assembly of an extinct flightless bird, the little bush moa	5. 発行年 2024年
3. 雑誌名 Science Advances	6. 最初と最後の頁 eadj6823
掲載論文のDOI (デジタルオブジェクト識別子) 10.1126/sciadv.adj6823	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 該当する

[学会発表] 計5件 (うち招待講演 5件 / うち国際学会 1件)

1. 発表者名 Kazutaka Katoh
2. 発表標題 Off-plan use of technology
3. 学会等名 日本進化学会第 24 回年会 (招待講演)
4. 発表年 2022年

1. 発表者名 Kazutaka Katoh
2. 発表標題 Multiple sequence alignments for predicting antigen-antibody interactions
3. 学会等名 Simons Institute Workshop "Computational Challenges in Very Large-Scale 'Omics'" (招待講演) (国際学会)
4. 発表年 2022年

1. 発表者名 加藤和貴
2. 発表標題 抗原抗体複合体予測における多重配列アラインメントの利用について
3. 学会等名 遺伝研研究会「生命科学を支える分子系統学」(招待講演)
4. 発表年 2022年

1. 発表者名 加藤和貴
2. 発表標題 Annotation of VDJ genes in TCR loci on eutherian genome
3. 学会等名 日本進化学会第25回年会（招待講演）
4. 発表年 2023年

1. 発表者名 加藤和貴
2. 発表標題 非生物学的データのアラインメント
3. 学会等名 遺伝研研究集会 生命科学を支える分子系統学 2024（招待講演）
4. 発表年 2024年

〔図書〕 計1件

1. 著者名 Kato, Kazutaka (Ed.)	4. 発行年 2021年
2. 出版社 Springer	5. 総ページ数 321
3. 書名 Multiple Sequence Alignment (Methods in Molecular Biology)	

〔産業財産権〕

〔その他〕

MAFFT alignment server https://mafft.cbrc.jp/alignment/server/

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
研究協力者	スタンドレー ダロン (Standley Daron) (00448028)	大阪大学・微生物病研究所・教授 (14401)	

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			
米国	University of North Texas			
ポルトガル	University of Minho			
米国	Harvard Univ			