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　３ 年　 ５ 月 ２０ 日現在

機関番号：１４４０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2020
課題番号：１６Ｋ０７４６４

研究課題名（和文）MAFFT多重アラインメントプログラムの大量配列データへの対応と機能拡張

研究課題名（英文）Extension of MAFFT multiple sequence alignment program mainly for large data

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）　　3,800,000円

研究成果の概要（和文）：本研究の目的は、研究代表者によって開発され広く使われている多重配列アラインメントプログラムMAFFTの適用範囲をさらに拡大することである。特に、配列決定技術の進歩に伴って必要になってきた巨大アライメントに対応することが第一の目的であった。まず、多重配列アラインメントの案内木の選択について議論があったので、この問題を慎重に検討し、計算量の多い従来の方法が優れているという比較的保守的な結論を得た。そこで、この方法をより大規模な問題に適用可能にするために、技術的な改良を行った。また、より小規模な問題に対して、タンパク質の立体構造を考慮してより正確なアラインメントを求めることも可能とした。

研究成果の学術的意義や社会的意義
本研究の目的は配列解析に役立つ計算プログラムを多くの研究者に提供することであり、直接社会に役立つことは意図していないが、新型コロナウイルスの配列解析において大規模な多重配列アラインメントの計算のためにMAFFTプログラムがよく利用された。このように間接的に役立つことは想定通りであったが、利用頻度は想定を超えた。この計算の高速化の鍵となったアルゴリズムは20年近く前に Katoh et al (2002) で提案したものであり、当時の配列解析のためには過剰性能気味であった。このことは、開発当初は無駄に見える多くの方法の中に、後年役に立つものが少数存在するかもしれない可能性を示している。

研究成果の概要（英文）：The primary purpose is to enable the MAFF program to align large sequence data that is becoming common and necessary as a result of the progress of sequencing technologies. When starging this project, there was an argument about how to select a guide tree for the progressive alignment method for large data. We carefully considered this issue and concluded that a conventional approach works well although resource consuming. Based on this result, we made technical improvements to scale up an existing option of MAFFT. We also improved the accuracy of relatively small scale alignment of protein sequences by incorporating 3D structural information.

This project aims to provide many researchers with useful computer software to help solving real-world problems. As a massive need to analyze SARS-CoV-2 genomes suddenly arose, MAFFT is heavily used, indirectly contributing to solve real-world problems such as the origin of this virus and functional analysis of interaction between virus and host.

研究分野：分子進化

キーワード：多重配列アラインメント　計算プログラム　配列解析　タンパク質　塩基配列

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

配列決定技術の進歩にともなって、アラインメントに含める必要のある配列の数が増加している。研究開始当時、巨大アラインメントを累進法によって構築するためにどのような案内木が適しているかについて、議論があった。すなわち、Boyce et al. (2014) によって、多数の配列からなるアラインメントを構築する時は、従来よく使われてきた案内木に比べてより簡単な、ランダムに生成した鎖状案内木がよい結果を与えることが報告された。これは 30 年来信じられてきた、配列の間の系統関係を考慮することによってより良いアラインメントが得られるはずという前提を疑わせる結果であり、反論もある (Tan et al. 2015)。この問題に対する結論を得ることが最初に必要であった。

２．研究の目的

(1) 上に述べた理由から、Boyce et al. (2014) の再現実験と、彼らとは異なる条件での実験を、研究分担者と共に慎重に行った。ランダムに生成した鎖状案内木が最善の選択肢なのか、計算量が多くより厳密に推定した案内木がより良い結果を与えるのか調べた。その結果、より厳密な方法で案内木を求めた方が良いという、従来の考えを支持する結果が得られた。議論になっていた点だったので、この結果を Yamada et al. (2016) で報告した。ここで試した厳密な方法は、既に MAFFT プログラムに実装されていたものであったが、巨大アラインメントに適用することを想定していなかったので、メモリと計算時間の面で実用的ではなかった。これらの制約の解消が次の段階の目的となった。

(2) 並行して、比較的小規模な問題に対して、タンパク質のアミノ酸配列のアラインメント計算の正確さを向上させることも目指した。タンパク質をコードする遺伝子の進化過程において、多くの場合、立体構造を保持するような中立なアミノ酸置換が蓄積されてきた。比較的機能的制約の弱いタンパク質や分化後長い時間を経たタンパク質ペアの場合、アミノ酸置換の蓄積によって配列上の類似性が低くなってしまった場合も多い。そのような場合でも、立体構造は依然として保存されていて、明確な類似性が見られる場合が多い。立体構造の情報を使って遠い関係にあるタンパク質のアミノ酸配列のアラインメントを正確に行えることが知られている (O’Sullivan et al. 2004)。このことを利用して正確さの改善を目指した。

(3) 2019 年 12 月頃から、新型コロナウイルスのゲノム配列の多重配列アラインメントに対する需要が急増した。比較的長く、類似性の高い配列のアラインメントに対して、Katoh et al. (2002) で提案した方法が有効であったため、本計画で行っている外部向け計算サービスの利用回数が急増し計算資源が逼迫した。それに対応することも目的となった。

(4) ロングリードシークエンサーからのデータのアラインメントが必要になってきている。シーケンサの性能の限界により、挿入欠失エラーレートが高く、特に nanopore シークエンサーの場合置換エラーに強い偏りがある。そのため、この点を考慮した、通常のアラインメントと少し異なる計算を可能にすることも目指した。

３．研究の方法

上で述べた目的 (1) に対して、まず、メモリ使用量を抑制するために、一時データの配置先をメモリからファイルに変更したところ、ディスクアクセスが律速要因になった。そこで、データの読み書きの順番を工夫することによって、メモリ上で行った場合とほぼ同程度の速度が得られた。その結果、計算時間はかかるものの、通常のデスクトップパソコンで巨大アラインメントが計算できるようになった。研究分担者らによって、MPI による並列計算にも対応したため、大規模な並列計算機では高速に計算できるようになった (Nakamura et al. 2018)。

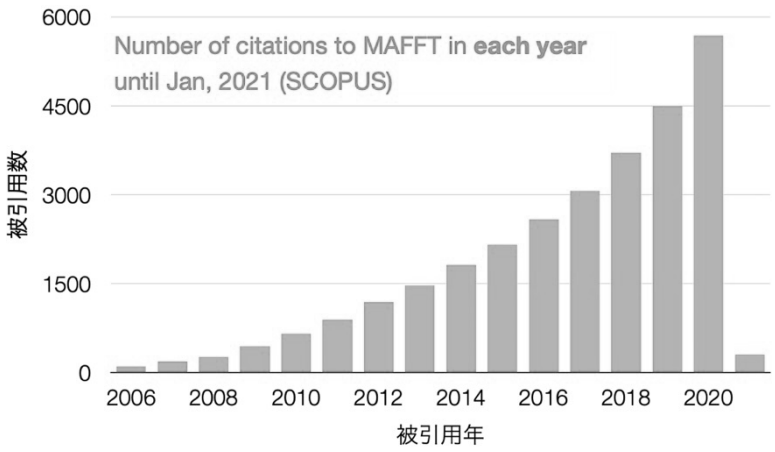
目的 (2) に対して、配列解析に立体構造データを利用するには、いくつかの技術的難点があった。まず、立体構造アラインメントの計算量は配列アラインメントに比べて大きい。また、立体構造データベースの記法があまり統一されていないために、配列上の残基と座標の対応づけを慎重に行う必要がある。これらの点を克服するために、大阪大学 John Rozewicki 研究員らとの共同研究によって、DASH というデータベースを構築した。これは、PDB の全エントリから冗長性を除いたものをドメインに分解し、類似性の見られる全ドメインペアの立体構造アラインメントを計算してサーバに置き、定期的にアップデートするものである。ユーザがローカルな計算機において MAFFT プログラムに DASH オプションをつけて起動すると、REST を通してこのデータベースと通信し、立体構造アラインメントの利用可能なペアを取得し、これらを入力配列に加えて多重配列アラインメントを計算する。DASH サービスおよび MAFFT との組み合わせを Rozewicki et al (2019) で報告した。

目的 (3) に対しては、類似度が高く配列が長い場合の大域的アラインメントは難しい問題ではなく、通常の多重配列アラインメントより簡単なアルゴリズムで充分役に立つので、そのような計算が簡単にできるようにプログラムを応急的に改造し、オンラインサービスと配布版のアップデートを行った。計算資源の拡充も行った。

目的 (4) のロングリードシークエンサーの特性を考慮した多重配列アラインメントを計算するため、LAST-TRAIN (Hamada et al 2017) によるパラメータを用いた多重配列アラインメントに対応した。また、LAST-TRAIN の作者である Martin C. Frith 博士による、部分的にオーバーラップするリードをアセンブルするプログラム lamassemble の開発に協力し、現時点での成果を Frith, Mitsuhashi & Katoh (2020) で報告した。

４．研究成果

以上の拡張を適用した MAFFT プログラムを配布し、前項で述べた個々の論文で詳細を報告した。また、外部向け計算サービス全体を Katoh et al (2019) で解説した。特に目的 (3) に対応した結果、コロナウイルスの配列解析に関する多数の解析で MAFFT プログラムが利用された。例えば、2021 年 1 月から 4 月までに *Nature* 誌に掲載された SARS-CoV-2 に関する 4 報の論文 (Tegally et al 2021; Cele et al 2021; Collier et al 2021; Kemp et al 2021) の解析において MAFFT プログラムが利用された。それ以外の点についても本研究の成果はよく利用された。Web of Science によると、Nakamura et al (2018)、Rozewicki et al (2019)、Katoh et al (2019) の 3 報が、2021 年 5 月の時点において高被引用文献 (被引用数上位 1%) にランクされた。Katoh et al (2019) は、ホットペーパー (被引用数上位 0.1%) にもランクされた。MAFFT プログラムを記述した過去の論文の被引用数を研究開始年 (2016 年) と最終年 (2020 年) のそれぞれ一年間の間で比較すると倍増した。以上のように、本研究およびその前段階の研究の成果は、本研究期間に実現したプログラムの性能改善などによって、科学コミュニティに広く受け入れられた。



５．主な発表論文等

〔雑誌論文〕　計12件（うち査読付論文　8件／うち国際共著　1件／うちオープンアクセス　6件）

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 -
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 -
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 Schritt Dimitri、Li Songling、Rozewicki John、Katoh Kazutaka、Yamashita Kazuo、Volkmuth Wayne、Cavet Guy、Standley Daron M.	４．巻 4
２．論文標題 Repertoire Builder: high-throughput structural modeling of B and T cell receptors	５．発行年 2019年
３．雑誌名 Molecular Systems Design & Engineering	６．最初と最後の頁 761～768
掲載論文のDOI（デジタルオブジェクト識別子） 10.1039/c9me00020h	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 Sone J, (約20名), Katoh K, (約10名), Sobue G	４．巻 51
２．論文標題 Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease	５．発行年 2019年
３．雑誌名 Nature Genetics	６．最初と最後の頁 1215～1221
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41588-019-0459-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 Rozewicki John、Li Songling、Amada Karlou Mar、Standley Daron M、Katoh Kazutaka	４．巻 47
２．論文標題 MAFFT-DASH: integrated protein sequence and structural alignment	５．発行年 2019年
３．雑誌名 Nucleic Acids Research	６．最初と最後の頁 W5～W10
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/nar/gkz342	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 Li Songling、Wilamowski Jan、Teraguchi Shunsuke、van Eerden Floris J.、Rozewicki John、Davila Ana、Xu Zichang、Katoh Kazutaka、Standley Daron M.	4．巻 2048
2．論文標題 Structural Modeling of Lymphocyte Receptors and Their Antigens	5．発行年 2019年
3．雑誌名 Methods in Molecular Biology	6．最初と最後の頁 207～229
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-1-4939-9728-2_17	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 2231
2．論文標題 Analysis of Protein Intermolecular Interactions with MAFFT-DASH	5．発行年 2020年
3．雑誌名 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 Nakamura Tsukasa、Yamada Kazunori D、Tomii Kentaro、Katoh Kazutaka	4．巻 34
2．論文標題 Parallelization of MAFFT for large-scale multiple sequence alignments	5．発行年 2018年
3．雑誌名 Bioinformatics	6．最初と最後の頁 2490～2492
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bioinformatics/bty121	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、Rozewicki John、Yamada Kazunori D	4．巻 20
2．論文標題 MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization	5．発行年 2017年
3．雑誌名 Briefings in Bioinformatics	6．最初と最後の頁 1160～1166
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bib/bbx108	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 Schritt Dimitri、Katoh Kazutaka、Li Songling、Standley Daron M.	4．巻 -
2．論文標題 Modeling Biocatalysts	5．発行年 2017年
3．雑誌名 Future Directions in Biocatalysis (Second Edition), edited by T. Matsuda	6．最初と最後の頁 385～398
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/B978-0-444-63743-7.00019-6	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 Yamada KD, Tomii K, Katoh K	4．巻 32
2．論文標題 Application of the MAFFT sequence alignment program to large data—reexamination of the usefulness of chained guide trees	5．発行年 2016年
3．雑誌名 Bioinformatics	6．最初と最後の頁 3246-3251
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bioinformatics/btw412	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 K, Standley DM	4．巻 32
2．論文標題 A simple method to control over-alignment in the MAFFT multiple sequence alignment program	5．発行年 2016年
3．雑誌名 Bioinformatics	6．最初と最後の頁 1933-1942
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/bioinformatics/btw108	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 lamassemble: Multiple Alignment and Consensus Sequence of Long Reads	5．発行年 2020年
3．雑誌名 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 －

〔学会発表〕　計3件（うち招待講演　3件／うち国際学会　0件）

１．発表者名 加藤和貴
２．発表標題 多重配列アラインメントプログラムMAFFTの 新機能について
３．学会等名 日本進化学会第21回大会（招待講演）
４．発表年 2019年

１．発表者名 加藤和貴
２．発表標題 アラインメント
３．学会等名 木村資生記念進化学セミナー（招待講演）
４．発表年 2017年

１．発表者名 加藤和貴
２．発表標題 多重配列アラインメントの並列計算
３．学会等名 配列解析シンポジウム ~36 years since Smith-Waterman-Gotoh~（招待講演）
４．発表年 2018年

〔図書〕　計1件

１．著者名 Kazutaka Katoh (ed)	４．発行年 2021年
２．出版社 Springer Science+Business Media, LLC	５．総ページ数 321
３．書名 Multiple Sequence Alignment	

〔産業財産権〕

〔その他〕

多重配列アラインメント計算サービス
https://mafft.cbrc.jp/alignment/server/

DASH database
https://sysimm.org/dash/

コロナウイルスの配列解析の紹介
http://www.biken.osaka-u.ac.jp/news_topics/detail/1077

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	山田　和範 (Yamada Kazunori) (20756217)	東北大学・情報科学研究科・准教授 (11301)	
研究分担者	富井　健太郎 (Tomii Kentaro) (40357570)	国立研究開発法人産業技術総合研究所・情報・人間工学領域・研究チーム長 (82626)	

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