様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　６月　８日現在

機関番号：８２６２６
研究種目：若手研究（Ｂ）
研究期間：2009～2011
課題番号：２１７００３２６
研究課題名（和文）　ＭＡＦＦＴアルゴリズムの拡張によるＲＮＡおよびタンパク質の構造多重アラインメント
研究課題名（英文）　Extension of the MAFFT algorithm to RNA and protein structural multiple alignment
研究代表者
加藤　和貴　（KATOH KAZUTAKA）
独立行政法人産業技術総合研究所・生命情報工学研究センター・招聘研究員
研究者番号：７０３７８８６８

研究成果の概要（和文）：多重アラインメントは、基本的な配列解析技術の一つであり、広い応用範囲をもつ。系統樹推定、構造予測などに用いられる。本研究の目的は、研究代表者がこれまでに開発した配列多重アラインメントプログラム MAFFT を拡張して、(1) タンパク質や RNA の構造情報を利用した多重アラインメントを計算可能にすること、および、(2) より使いやすいものにして配列解析に関連する研究に広く貢献すること、である。大阪大学と産総研 CBRC において計算サービスとプログラム配布を開始し､その結果 MAFFT プログラムは広く普及した。また、並列化、既存のアラインメントの拡張といった新規機能を追加した。

研究成果の概要（英文）：Alignment of multiple sequences is a basic technique for sequence analysis, and has various applications, including phylogenetic inference and structure prediction. This research aims (1) to extend the MAFFT multiple sequence alignment program to enable structural alignment of proteins and RNAs, and (2) to enhance its usability to contribute to a wide range of studies that use biological sequences. We started a new web services for alignment calculation and distribution of the program in CBRC, AIST and Osaka University. As a result the MAFFT program has gained worldwide popularity. We also have implemented several new features, such as parallelization, and the addition of new sequences into an existing alignment.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,300,000	390,000	1,690,000
2010 年度	700,000	210,000	910,000
2011 年度	700,000	210,000	910,000
総　計	2,700,000	810,000	3,510,000

研究分野：総合領域
科研費の分科・細目：情報学・生体生命情報学
キーワード：配列アラインメント，構造アライメント，タンパク質，並列計算，バイオインフォマティクス

１．研究開始当初の背景

配列アラインメント手法の研究は古くから行われてきたが、近年特に活発化している。その理由は、ゲノムやその他の大規模シークエンスプロジェクトの進展にともなって従来の方法では処理できないような大量のデータが発生していること、タンパク質や RNA の高次構造情報の利用が必要となってきたことなどである。高次構造の情報を利用することによって、進化的により遠い関係にある配列の比較解析が可能となる。配列レベルの類似性は構造レベルの類似性に比べて保存されやすい傾向があるためである。
本研究の開始前に、研究代表者は、多重配列

アラインメントに関連するいくつかの新規手法を考案し、それらを実装したソ フトウェア MAFFT を開発・公開した。

２．研究の目的

本研究の第一の目的は、MAFFT プログラムに、構造を考慮した多重アラインメント機能を付加することである。第二に、開発した方法が実際に関連分野の研究者に活用されることも重要である。MAFFT プログラムの普及を図るために、ユーザインタフェイスやウェブサービスの改良を行う。

３．研究の方法

(1) タンパク質 決定されている構造データを用いて多重アラインメントの正確さを向上させる方法として、3DCoffee や Promals3D など、いくつかの方法が提案されている。これらは、TCoffee や Promals などの配列アラインメントアルゴリズムと、FAST, DaliLite といった構造アラインメントアルゴリズムの組み合わせによる。これらの既存の方法の問題点は、配列の長さのばらつきが大きい場合、入力配列に対して構造既知の配列の系統的分布が偏っている場合、入力配列、構造の数が多い場合などに対応していないために多くの現実の問題を処理できないことである。

このような入力データに対して、配列レベルのアラインメントにおいて MAFFT はこれらの状況に比較的ロバストであるので、これを用いてタンパク質の構造・配列の統合アラインメントを新たに開発することにメリットがあると考えられる。

組み合わせる構造アラインメントアルゴリズムとして、ASH (Standley et al. 2007)を試した。ただし、ASH は構造を剛体として扱うため、与えられたタンパク質が複数のドメインからなり、ドメインの間の角度が状況によって変化するような構造をうまく処理できないという問題が生じた。そのため、構造をドメインに分割し、ドメイン同士の構造アラインメントを計算し、最後にそれらを連結するという方法をとった。

構造アラインメントは計算量が多いという問題もある。それに対応するために、PDB 上の代表的な構造について、それらをドメインに分割し、その間の構造アラインメントをあらかじめ計算しておくことによって、高速に結果を返すことを可能にするための準備を進めている。

(2) 性能の評価方法　計算したアラインメントの正確さの評価は、簡単な問題ではない。配列アライメントについては、立体構造情報を使ったアラインメントを正解とみなして、それにどの程度近い結果が配列のみの情報から得られるか、という基準で評価することができる。これは、立体構造が配列に比べて進化的に保存されやすいことに基づいている。

しかし、立体構造アラインメントも方法によって結果が異なり、一つの立体構造アラインメントを正解であると簡単に仮定することはできない。その上、本研究では、立体構造アラインメントを使用することによるアラインメントの正確さの向上を図るので、どの立体構造アラインメントが良いか、を知る必要もある。

そこで、広く受け入れられているベンチマークに加えて、以下の新しい基準に基づく評価を行った。すなわち、(1) 遠い進化的関係にある複数の配列と構造のセットのアラインメントを構築し、(2) その中の一つの構造既知の配列について、遠縁な構造情報のみを用いてホモロジーモデリングを行い、正解の構造と比較する。この時、近縁な配列の構造情報は除外する。(3) 良い結果を与えたアラインメントが良いアラインメントであると解釈する。

この方法は、アラインメントの評価方法としては間接的であるが、立体構造アラインメントの計算手法も含めて計算結果の正確さが評価できるというメリットがある。

(3) RNA RNA の構造を考慮したアラインメントについては、MAFFT-X-INS-i と呼ぶ予備的な方法を 2008 年に報告した (Katoh & Toh 2008)。SCRNA (Tabei et al. 2006) と MAFFT を組み合わせたものである。いくつかの点で、この方法の改良を試みた。しかし、正確さや計算時間の面で明確な改善は見られなかった。研究期間中に他の研究グループにより、既存の方法の比較が行われた (Yoon et al. 2010) が、依然 MAFFT-X-INS-i が最も高い正確さを示した。

(4) 並列化 複数のコアを持つ PC の普及が進んだため、マルチスレッド計算への対応を行った。多重アラインメントの計算には並列化の容易な段階と容易でない段階がある。MAFFT プログラムに関しては、a. 全ペアの進化距離を推定する計算の並列化は簡単である。b. 累進法アラインメントの効率的な並列化は容易ではないが、元の計算量が小さいので効果は小さい。c. 反復改善法については、いくつかの異なる並列化を行いパフォーマンスを比較した。

(5) 系統樹推定との関係 研究期間中に、系統樹推定の正確さを用いてアラインメントの正確さを評価する方法が他の研究グループによって報告された (Dessimoz & Gil 2010)。系統樹推定に関する研究は当初の計

画には含まれていなかったが、本研究と強く関連するので、この論文の著者であるDessimoz博士およびGil博士との共同研究として、彼らの方法の拡張を期間の途中から開始した。

(6) 既存のアラインメントの利用 キュレートされたアラインメントのデータベースがいくつか利用可能である。新しい配列が決定されるたびに多重配列アラインメントを新しく計算する代わりに、既存のアラインメントに新しい配列を付け加えられると便利である。このように既存のリソースを利用することは、計算手法としては簡単である一方、アルゴリズムの改良や並列化に匹敵する程度に有効な、大きなデータを処理するための方法である。さらに、既存のアラインメントが機能や構造に基づく付加的な情報を考慮している場合、それらの情報を保持しながら大規模なアラインメントを構築することができる。

(7) その他 塩基配列の方向の自動判定、ユーザが指定する案内木の利用、類似性に基づく多数の配列の簡単な分類、などいくつかの機能を新しく実装し、プログラムの有用性を向上させた。ウェブサーバ上では、近隣結合法による系統樹推定サービスを提供している。アラインメントに含める残基の選択や系統樹に含める配列の選択を対話的に行えるサービスを開始した。

４．研究成果

(1) タンパク質の構造配列アラインメント タンパク質立体構造と配列の両方の情報を用いてアラインメントを計算する方法として、MAFFT と構造アラインメントプログラムASH を組み合わせることを試みた。ASH の開発者であるDaron M. Standley博士らとの共同研究として行った。一般に、与えられた配列のホモログの情報を加えるとアラインメントの正確さは向上すると考えられている。構造配列アラインメントを計算する際もホモログの選択方法は重要であり、適切な選択方法を検討した。これまでに得られた知見に基づいて、計算サービスを大阪大学において開始した。このサービスは、ホモログを収集する SEEKQUENCER、および、構造配列アラインメントMAFFTashの二つの部分からなる。

(2) 並列化 MAFFT プログラムを並列計算に対応させた。メモリを共有する複数コアをもつ PC が近年普及してきているため、その環境で並列的に動作するバージョンを開発した。POSIX Threads ライブラリを用いた。数コア程度の規模の並列計算に適した計算手法を比較検討し、結果を *Bioinformatics* 誌において報告した (Katoh & Toh 2010)。

(3) 既存のアラインメントの利用 ユーザが与えたアラインメントを保持しながら新しい配列を付け加えるために、２種類の異なる方法を実装した。一つは、反復改善法に基づくもので、与えられたアラインメントを保存するように目的関数を変更するものである。上述のタンパク質の立体構造アラインメントを計算するために用いた。もう一つは、累進法に基づく高速で単純な方法である。新しい配列と既存の配列の両方を含む案内木を計算し、それの各ノードにおいてにグループ間アラインメントを行うことによって多重アラインメントを計算するが、既にアライメントが与えられているノードでは、計算をスキップする。以上の二つの方法のサービスを産総研の計算機で開始した。

(4) 系統樹推定との関係 アラインメントの正確さと系統樹推定の正確さの関係に関する予備的な解析を行った。特に、系統樹推定にホモログを加えることの効果に着目した。一般に、ある程度多数の配列を用いた系統樹推定の方が、数本程度のみを用いた系統樹推定に比べて信頼できると考えられている。このことについて、系統樹推定段階とアラインメント計算段階にわけて、定量的に検討した。その結果、ホモログがアラインメントの正確さに与える影響について、従来考えられていたのとは異なる結果が得られた。また、アラインメント計算手法の正確さについても構造に基づく性能評価とは異なる結果が得られた。これまでに得られた中間的な結果を、ISMB 学会、SMBE 学会などで報告した。

(5) 計算サービス 以上の新規機能のうち、配列アライメントに関する計算サービスを、産業技術総合研究所生命情報工学研究センター (CBRC) に設置した計算機で開始した。最終年度 (2011 年度) 一年間の訪問数は、約140,000、ユニークユーザ数は約 35,000 であった。アクセス元の地理的分布は、ヨーロッパ 37%、南北アメリカ 34%、アジア 23%、その他 5%である。

(6) 波及効果 SCOPUS によれば、MAFFT プログラムを記述した過去の論文の被引用回数は、2002 年から 2012 年までの累計で約 3000 回である。このうち、本研究期間の開始年 (2009) 一年間で 400 回、最終年 (2011) 一年間で約 800 回であった。また、本研究期間中に、米国トムソンロイター社は、Fast Breaking Paper (2009 年 10 月) および New Hot Paper (2009 年 11 月) として MAFFT を記述した論文の一つ (Katoh & Toh 2008) を選んだ。

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１件）

① Katoh K, Toh H. (2010) Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* **26**:1899-1900 査読有
DOI: 10.1093/bioinformatics/btq224

〔学会発表〕（計６件）

① Katoh K, Ledergerber C, Dessimoz C, Gil M. Effect of adding homologs in phylogenetic analysis. GCOE バイオインフォマティクス公開セミナー, 日本バイオインフォマティクス学会北海道地域部会セミナー, 2011/8/4, 北海道大学, 札幌
② Katoh K, Ledergerber C, Dessimoz, C, Gil, M. Effect of adding homologs in phylogenetic analysis. SMBE2011, Jul. 27, 2011, Kyoto University, Kyoto, Japan
③ Katoh K, Ledergerber C, Dessimoz, C, Gil, M. Effect of adding homologs in phylogenetic analysis. ISMB/ECCB2011, Jul. 18, 2011, Austria Center Vienna, Vienna, Austria
④ Katoh K. Effect of adding homologs in phylogenetic analysis. Zurich Colloquium for Computational Molecular Evolution, Jul. 14, 2011, ETH Zurich, Zurich, Switzerland
⑤ Katoh K, Standley D M, Dinh H, Nakamura H, Toh H. Two extensions of MAFFT: protein structure-sequence alignment and large-scale sequence alignment. CBRC2010, Jul. 28, 2010, AIST Tokyo Waterfront Center, Tokyo, Japan
⑥ 加藤和貴, Daron M. Standley, 中村春木, 藤 博幸 MAFFTash - a structure-sequence multiple alignment program for proteins 第 32 回日本分子生物学会年会, 2009/12/12, パシフィコ横浜, 横浜

〔図書〕（計１件）

① Katoh K, Asimenos G, Toh H. (2009) Multiple Alignment of DNA Sequences with MAFFT. ‘Bioinformatics for DNA Sequence Analysis’ edited by D. Posada. *Methods in Molecular Biology* **537**:39-64, Springer

〔その他〕
ホームページ等
http://mafft.cbrc.jp/alignment/software/
http://sysimm.ifrec.osaka-u.ac.jp/MAFFTash/
http://sysimm.ifrec.osaka-u.ac.jp/seekquencer/
http://sciencewatch.com/dr/fbp/2009/09octfbp/09octfbpKato/

６．研究組織
(1) 研究代表者

加藤　和貴　(KATOH　KAZUTAKA)
独立行政法人産業技術総合研究所・生命　情報工学研究センター・招聘研究員
研究者番号：７０３７８８６８