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　２９　年　　５　月　１７　日現在

機関番号：　１２６０１
研究種目：　基盤研究(B)（一般）
研究期間：　2013～2016
課題番号：　２５２８０００２

研究課題名（和文）ＳＭＡＤによるビッグデータ類似検索超高速化とその応用

研究課題名（英文）Fast Similarity Search on Big Data based on SMAD and its applications

研究代表者
渋谷　哲朗（Shibuya, Tetsuo）

東京大学・医科学研究所・准教授

研究者番号：６０３９６８９３

交付決定額（研究期間全体）：（直接経費）　　8,000,000円

研究成果の概要（和文）：統計的モデルにもとづく全く新しいアルゴリズム設計パラダイムであるＳＭＡＤに基づき、タンパク質立体構造データベースを中心とした様々なデータベース上での高速なビッグデータ検索技術の研究開発の実現、およびそれらを活用した応用アルゴリズムの実現をめざして研究を行った。その結果、タンパク質機能予測などを精度を落とさずに高速化することに成功したほか、より幅広いタンパク質立体構造検索の高速検索も実現した。また、次世代シークエンサーデータの解析アルゴリズムなどでも新たな解析手法などを開発することに成功した。

研究成果の概要（英文）：We aimed to develop very fast indexing and searching algorithms for big-data databases, especially the protein 3-D structure databases, and also aimed to develop application algorithms utilizing them. We succeeded in developing dramatically faster protein function prediction algorithms without any loss of accuracy. We also succeeded in developing faster algorithms for protein 3-D structure searching for wider applications. We also developed several analysis algorithms for next-generation sequencer data.

研究分野：　バイオインフォマティクス

キーワード：　アルゴリズム　ビッグデータ　検索　タンパク質立体構造　次世代シークエンサー

１．研究開始当初の背景

2009 年に、統計的モデルにもとづく全く新しいアルゴリズム設計パラダイムＳＭＡＤ(Statistical Model-based Algorithm Design）が本研究代表者の渋谷によって編み出され、それにもとづき構造生物学の最重要データベースであるタンパク質立体構造データベースの基本類似検索の速度が劇的に向上した。これまでの情報科学におけるアルゴリズム設計のほとんどは、最悪性能あるいは単純なランダム入力に対する平均性能を解析し、それらの性能が高いアルゴリズムの設計を目指したものである一方、ＳＭＡＤは、対象とするデータベースの精緻な統計学的モデルに基づいてアルゴリズムを設計・性能解析を行い、それにより従来のアルゴリズムの性能評価やそれに基づくアルゴリズム設計では達成できなかった性能を、実用上も理論上も新たに達成できる可能性のある新しい情報科学理論であった。そこで、本研究では、この理論に基づき、新たなデータベース上での検索高速化や、それを活用した新たな応用を開拓することをめざした。

２．研究の目的

ＳＭＡＤは、データベース中のデータを精緻に統計モデル化することによってアルゴリズムの高速化をはかる一般的枠組であり、対象の統計モデルをうまく設計することができれば、当然のことながらタンパク質立体構造データベースに限らず、多くの複雑な巨大データについても同様の検索の高速化が可能である。さらに、それらの検索の高速化に成功すれば、たとえば生体分子のデータベースならば、これまで検索が遅く実現不可能あるいは困難であった網羅的な大規模データベース検索にもとづいた分子機能解析が可能となると考えられた。

３．研究の方法

(1) ＳＭＡＤの登場によって、タンパク質立体構造の基本検索の速度は大幅に向上したが、検索速度の向上および検索時間に制限がある際の検索精度の向上、あるいは問題の制限が異なる場合の高速化などについて、より一層の高度化を図ることをめざした。さらに、それらの技術が他の問題に対して適用可能かどうかの研究を進めた。

(2) また、そのような大規模高速検索ができるようになることで、それを活用したより高度な機能解析技術を実現できる可能性があり、そのような高度な生体分子機能解析技術の開発を狙って研究を進めた。

４．研究成果

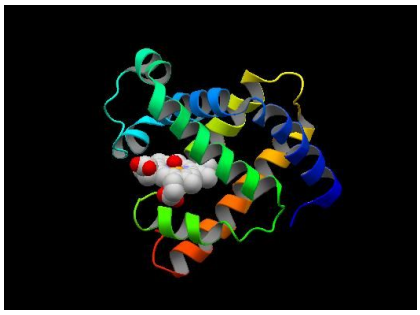
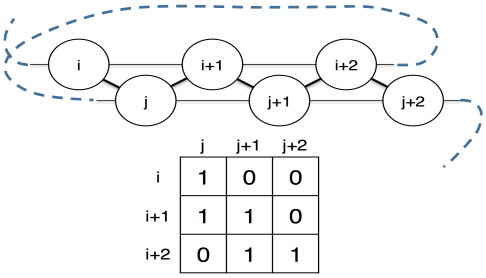


図１　タンパク質立体構造



	j	j+1	j+2
i	1	0	0
i+1	1	1	0
i+2	0	1	1

図２　コンタクトマップ表現

(1) 大規模行列索引を活用したタンパク質立体構造からの超高速高精度機能予測の実現

本研究の目標の一つは、タンパク質立体構造大規模索引技術を活用した大規模データベースへスケーラブル可能な機能予測アルゴリズムの構築であった。本研究では、これに対して、I-suffix tree とよばれる大規模行列索引アルゴリズムを活用して、超高速なタンパク質立体構造のコンタクトマップ表現（図１、図２）に対する部分構造索引技術を新たに開発した。これによって、従来、計算量が非常に大きく実用性の面で非常に困難のある立体構造アラインメントという技術を通してのみしか高精度な機能予測ができなかった問題に対して、この大規模索引を用いることで、精度をまったく落とさないまま構造からの超高速機能予測を実現することに成功した。

(2) 大規模ギャップ付き文字列索引を活用した超高速高精度タンパク質機能予測の実現

本研究では、タンパク質立体構造にとどまらない様々なデータに対し、その高速な処理を行うことを目標としていた。そこで、タンパク質の立体構造ではなく、塩基配列の持つ文字構造にも着目し、それを利用して機能予測を行う研究も行った。ここにおいても大規模索引技術を活用することによって高精度・高速な生体分子機能解析技術の実現を行った。大規模文字列データに対する索引構造としては、従来より接尾辞配列、接尾辞木、

FM-index などが知られている。これらは、データベース中の部分文字列を効率よく索引化し、高速な検索を実現する。しかしながらこれらを用いてタンパク質の機能予測などに用いるには、索引が連続した部分文字列の厳密一致のみしか考慮しておらず、高精度な機能予測への応用が困難であった。本研究では、これに対応するため、ギャップを含むような文字列パタンの索引が可能な b-suffix array という接尾辞配列のさらなる拡張を用いて、柔軟なタンパク質パタンを網羅的に索引化することに成功し、それによって、従来知られている最高精度の機能予測とくらべても精度をほぼ落とさずに、理論計算量も実際の計算量もより高速な新たなカーネルに基づく学習機械を構築することに成功した（図３）。

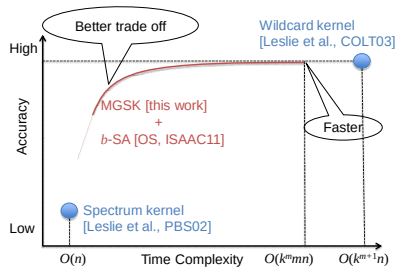


図３　b-suffix array に基づくカーネルによる高速化

(3) 塩基順序を限定しないタンパク質立体構造の超高速検索の実現

SMAD 技術がこれまで得意としていたタンパク質立体構造検索に関しても新たな成果を得ることに成功した。従来 SMAD によって高速化に成功していたのは、タンパク質立体構造中の連続部分構造の類似検索であった。一方、タンパク質立体構造の機能部位の中には、連続していない別々の部分が合わさって、重要な機能を構成していることがあることが知られており、従来の SMAD をベースとした検索方法ではそのような構造の検索には対応できていなかった。本研究では、これに対し、タンパク質立体構造の塩基順序を考慮しない部分構造を考え、塩基順序を保存せずに構造が類似するような部分構造の検索に関しても新たなアルゴリズムを開発した。そもそもそのような部分構造の組み合わせは、分子長に関して指数的に存在するため、そのような検索は実際には極めて困難であり、従来手法では、極めて小さなデータセットに対してしか計算することができなかった。しかし、本研究では、SMAD 技術を応用することで、理論的にも実際てきにも高速な検索技術の開発に成功した。これによって、従来のアルゴリズムに比べ、10 倍～100 倍の高速化に成功した。

(4) 統合化ディープラーニングによる高精度タンパク質機能部位予測の実現

SMAD においては、対象データベースのモデルを考え、それを活用してアルゴリズムの高速化することを目指してきた。一方、機械学習などの各種アルゴリズムも同様のモデルを考えて設計されるため、本研究では、そのような考察が、それらの機械学習アルゴリズムの進化につながることも考えながら研究を行い、実際に、本研究では、新たなタンパク質立体構造の機能部位予測アルゴリズムの開発に成功している。機能部位予測においては、タンパク質立体構造は部位によって異なる機能を持つことがあり、タンパク質のどの部位がどのような構造を持つかを高精度に予測することが求められる。一方、それらの部位は部位によって、その持つ特徴が著しく異なることが知られている。そこで、本研究では、それらの異なる部位予測を異なる特徴抽出技術を用いて予測するディープラーニングに基づくアルゴリズムを開発し、それらを統合して、高精度なアルゴリズムを実現することに成功した。

(5) ニューラルネットワーク解析

また、ディープラーニングに関連して、ニューラルネットワークを実際に解析するにあたり、ニューラルネットワークのどの部位が重要な役割を果たしているのかを解析する手法の開発に成功した。

(6) 次世代シークエンサーデータ解析技術の開発

また、別の応用先として、次世代シークエンサーによる超大規模リードの解析への応用についても成果を得ることができた。次世代シークエンサー技術の発達により、従来とは比較にならない大量の DNA データが算出されるようになり、それらに対する高速な解析技術が要求されるようになっている。本研究では、そのような次世代シークエンサーから読まれた大量のリードを succinct de Bruijn graph とよばれるリード格納索引構造を用いて、その上で、superbubble とよばれるゲノム上の重要なトポロジカルなパタンを発見・解析する高速なアルゴリズムの構築に成功した。

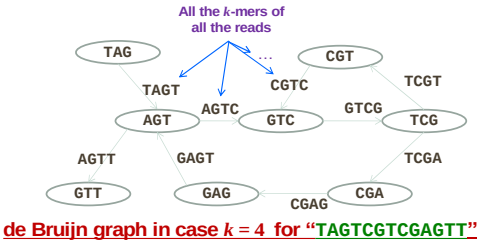


図４　de Bruijn graph

(7) 次世代シークエンサーデータからの構造変異解析の高精度化

さらに、次世代シークエンサーデータに関しては、参照ゲノムへのリードが複数個所へマッピングされることを活用して、その構造多型をより高精度に予測する技術の開発に成功した。

５．主な発表論文等

〔雑誌論文〕（計　９件）

① Taku Onodera, Tetsuo Shibuya, Fast Classification of Protein Structures by an Alignment-free Kernel, LNCS 9954, pp. 68-79, 2016, DOI: 10.1007/978-3-319-46049-9_7

② Yang Li, Tetsuo Shibuya, Malphite: A Convolutional Neural Network and Ensemble Learning Based Protein Secondary Structure Predictor, 2015, pp. 1260-1266, IEEE Press, ISBN 978-1-4673-6799-8/15. doi:10.1109/BIBM.2015.7359861

③ Yoichi Sasaki, Tetsuo Shibuya, Kimihito Ito, and Hiroki Arimura, Efficient Approximate 3-Dimensional Point Set Matching Using Root-Mean-Square Deviation Score, 2015, LNCS 9371, pp. 191-203 (DOI: 10.1007/978-3-319-25087-8_18)

④ Mohammad A. Eita, Tetsuo Shibuya, and Amin A. Shoukry, Locating Controlling Regions of Neural Networks Using Constrained Evolutionary Computation, 2015 IEEE Congress on Evolutionary Computation (CEC2015), pp. 1581-1588, IEEE Press, ISBN 978-1-4799-7492-4/1, 2015.

⑤ Wing-Kin Sung, Kunihiko Sadakane, Tetsuo Shibuya, Abha Belorkar and Iana Pyrogova, An O(m log m)-time algorithm for detecting superbubbles, IEEE/ACM Transactions on Computational Biology and Bioinformatcs, 12(4), pp. 770-777, July/August 2015. doi:10.1109/TCBB.2014.2385696

⑥ Taku Onodera, Tetsuo Shibuya, The Gapped Spectrum Kernel for Support Vector Machines, LNCS 7988, pp. 1-15, 2013.

〔学会発表〕（計８件）

① Tetsuo Shibuya, Algorithmic Challenges for Bio Big Data, UTokyo-IITM Workshop, March 15th, 2017, Chennai, India.

② Tetsuo Shibuya, Algorithm Design Paradigm Shift Needed for Bio Big Data, Genomic Medicine 2015, Ho Chi Minh, Vietnam, July 21, 2015.

③ Tetsuo Shibuya, Algorithmic Challenges to Bio Big Data, The 11th International Workshop on Advanced Genomics, Tokyo, Japan, May 22, 2015.

④ Tetsuo Shibuya, Designing Faster Algorithms For Big Data, UK-Japan Workshop on Big Data, Feburuary 6, 2014.

⑤ Tetsuo Shibuya, Fast Indexing Methods for Protein 3-D Structure Searching, NII Shonan Seminar 29: Compact Data Structures for Big Data, 2013.

〔図書〕（計 1 件）

① 渋谷哲朗, アルゴリズム, 東京大学工学教程・情報工学, 東京大学工学教程編纂委員会（編）, 丸善出版, 2016 年 11 月 20 日発行.

〔産業財産権〕

○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
特になし

６．研究組織
(1)研究代表者
渋谷 哲朗（SHIBUYA, Tetsuo）
東京大学・医科学研究所・准教授
研究者番号：60396893

(2)研究分担者
なし

(3)連携研究者

なし

(4)研究協力者
なし