様式 C-19

科学研究費補助金研究成果報告書

平成　２３　年　６　月　１　日現在

機関番号：12601
研究種目：若手研究（Ｂ）
研究期間：２００８　～　２０１０
課題番号：２０７００２６４
研究課題名（和文）　タンパク質３次元構造データベースに対する高速かつ柔軟な検索手法とその応用
研究課題名（英文）　Fast and Flexible Searching Methods for Protein 3-D Structure Databases and Their 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

研究成果の概要（和文）：　本研究では、タンパク質立体構造データベースに対する類似構造検索アルゴリズムの研究を行い、従来からの立体構造類似検索の手法の多くを抜本的に改善する、実用的にも理論的にも従来手法と比べきわめて高速な超高速線形時間基本探索アルゴリズムを開発することに成功するなど、今後の構造生物学の発展に大きく寄与する様々なアルゴリズムを開発し、その成果は国内外から高い評価を受けた。

研究成果の概要（英文）：　We explored research on similarity search algorithms for protein 3-D structure databases, and succeeded in developing various algorithms which must benefit the future structural biology a lot and are highly praised in the international research community. Especially we developed a very fast fundamental linear-time protein 3-D structure database searching algorithm, which is practically and theoretically much faster than any of previous algorithms.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	1,600,000	480,000	2,080,000
2009 年度	900,000	270,000	1,170,000
2010 年度	800,000	240,000	1,040,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：　バイオインフォマティクス
科研費の分科・細目：　情報学・生体生命情報学
キーワード：アルゴリズム・検索技術・データベース・生物情報・タンパク質・３次元立体構造

１．研究開始当初の背景

近年、NMR (nuclear magnetic resonance、核磁気共鳴)技術等の進展などに伴い、タンパク質やペプチドなど様々な生体化合物について分子構造だけでなく立体構造がわかるようになり、さらにそれに加えて計算機によって大量に予測された立体構造が得られるようになってきたことから、これらの生体化合物の大規模な立体構造データベースが作成されるようになってきた。これらの生体化合物の様々な機能の多くはその立体構造に

依存しているといわれ、類似した立体構造を持つ化合物は類似した機能を持つことも多いことが知られている。また、近年では、機能は未知だが構造のわかっているようなタンパク質も数多く知られるようになった。従って、このような立体構造のデータベースの中から、ある与えられた化合物の構造と類似するものを検索することができれば、その化合物の機能等を推測するのに極めて有用であることが推測できる。アミノ酸配列や核酸配列などの一次構造に関してはそのような類似した配列をデータベースの中から効率的に検索をする接尾辞木(suffix tree)、FASTA, BLAST, PatternHunter といった様々なデータ構造やアルゴリズムが、実用的に分子生物学の研究で使われており、今日ではなくてはならないものになっている。しかしながら、立体構造データベース検索に関しては、なかなか決定的なアルゴリズムは存在していなかった。

２．研究の目的

これまで行われている立体構造検索の手法には、RMSD 等の構造間の厳密な類似度に基づいて類似する構造を探すアプローチと、形そのものの類似度よりも、さらに大まかな立体構造のαヘリックスやβシートといった大局的な構造を数値化し、それの類似度の高いものを探そう、といったアプローチの 2 つがあるが、本研究提案当初、特に基本的な前者においてすら、決定的なアルゴリズムは存在しない状況であった。

これに対し、本研究提案直前の 2006 年に、厳密解法の 1 手法として、幾何的接尾辞木(geometric suffix tree)というデータ構造が、本計画提案者である渋谷によって提案された。これは、通常配列を対象としている接尾辞木というデータ構造をタンパク質などの鎖状生体分子の 3 次元立体構造に応用したものであり、しかも巨大な立体構造データベースの中から、厳密に RMSD が一定の閾値を超えないすべての部分構造の検索を高速に行うことのできることで脚光を浴びていた全く新しい手法であった。本研究の目的は、この手法を中心に、検索速度の向上およびより柔軟な検索手法の開発をめざすことであった。

３．研究の方法

(1) 新しいアルゴリズム設計理論の研究
本研究では、まず、タンパク質立体構造検索アルゴリズムを新たに構築するにあたって、その基礎となるアルゴリズム設計理論の構築を行った。

(2) 超高速基本検索アルゴリズムの設計
前述設計理論に基づき、タンパク質立体構造データベースに対する高速検索アルゴリズムの設計を行った。

(3) アルゴリズムの検証
前述のアルゴリズムの実際の性能を、実データベース上で実装し、実際に計算機実験を行うことで確認を行った。

(4) アルゴリズムの拡張
さらに、上記アルゴリズムを、より柔軟な検索を実現するために、様々な改良を行った。

４．研究成果

本研究では、タンパク質立体構造データベース検索のため、全く新しいアルゴリズム設計パラダイム SMAD を創造し、さらにそれに基づいて、全く新しい超高速基本的検索アルゴリズムの開発に成功した。このアルゴリズムは、従来手法と比べ、理論的に精度の犠牲が全くないにも関わらず、理論的にも実際にもきわめて高速なアルゴリズムであり、その成果は、計算生物学における最高峰の査読付国際会議である RECOMB（13th Annual International Conference on Research in Computational Molecular Biology）において最優秀論文（Best Paper Award）に選ばれるなど国際的にも確固とした高い評価を受けた。また、このアルゴリズムを設計するために構築したアルゴリズム設計パラダイム SMAD に関しても、我が国の情報科学に関する研究について顕著な功績のあった者に授与される第十回船井学術賞を船井情報科学振興財団より受賞するなど、きわめて高い内外の評価を受けた。

(1) アルゴリズム設計パラダイム SMAD
本研究では、まず、タンパク質立体構造データベースに対するアルゴリズムの設計および性能評価のための新しい理論基盤を構築することに成功した。

通常、アルゴリズムの理論的性能は、最悪性能および平均性能で測られることが多い。その理論的性能をあげることにより、実際の性能もあげることにつながるため、それらの性能を解析するための理論基盤の整備は必須である。

本研究で扱うタンパク質立体構造検索を RMSD とよばれる尺度を用いて行うにあたっては、データベースのサイズが N, 検索したい構造のサイズを m とした時、これまで知られている最良の最悪計算量は $O(N\log m)$ であり、しかも、理論的にもそれが最良で、改善の見込みはなかった。一方、平均計算量に関しては、そもそも、タンパク質立体構造データベ

ースに対するアルゴリズムの平均計算量を考えるための理論的基盤がこれまで存在せず、そのような基盤を構築しない限り、解析すらできない状況であった。そのため、タンパク質立体構造検索が構造生物学あるいはバイオインフォマティクスの分野にとってきわめて重要な問題であるにもかかわらず、長い間、本研究で新たに開発された新しいアルゴリズムが登場するまで、非効率的なアルゴリズムが基本アルゴリズムとして改善されることなく広くそのまま使われてきた。

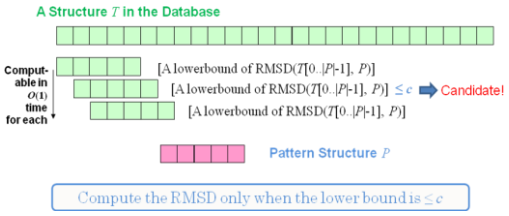
本研究では、これに対し、タンパク質立体構造データベースに対するアルゴリズムの平均性能を理論検証・アルゴリズム設計をするための全く新しい理論解析の枠組み SMAD (Statistical Model-based Algorithm Design)を構築した。(雑誌論文③) これは、データベース中のタンパク質分子が統計的挙動を分子物理学理論に基づいた理想的鎖分子であると考え、解析を行うものである。この枠組みの構築によってはじめて、タンパク質立体構造データベース上で、平均理論性能を議論することが可能となり、それを高速化するためのアルゴリズムを設計することが可能となった。

なお、このアルゴリズム設計思想 SMAD は、きわめて汎用性が高く、より一般的に他の複雑なデータベースでも同様なアルゴリズム設計が可能であるなど、情報科学の新しい見地を開くものでもあり、その功績に対しては、我が国の情報科学に関する研究について顕著な功績のあった者に授与される第十回船井学術賞を船井情報科学振興財団が授与されている。

(2) 超高速線形時間探索アルゴリズム

前述のとおり、タンパク質立体構造データベース上で、RMSD に基づいた類似検索の最悪性能は $O(N \log m)$ がほぼ最良であり、それより高速な検索は不可能であるとこれまで考えられてきた。しかしながら、平均性能に関しては、これまで考えられておらず、SMAD の設計思想を用いれば、より高速なアルゴリズムを設計できる可能性があった。

本研究では、これに対し、前述のアルゴリズム設計パラダイム SMAD に基づいて、文字列処理における Karp-Rabin アルゴリズムの 3



3 次元立体構造検索アルゴリズム

次元立体構造データベース版ともいえる、平均性能 $O(N)$ を実現する超高速検索アルゴリズムを設計することに成功した。(雑誌論文③)

さらに、実際にも、タンパク質立体構造データベースＰＤＢ上での実験結果からは、このアルゴリズムの実際の性能が、これまでの最良の従来手法と比べ、状況により異なるが、数倍～数十倍高速であることが示された。

検索長	40	80	120	160	200
本手法	58.9	25.5	17.3	14.2	12.9
一般手法	447.0	442.0	415.2	378.9	342.5
FFT	531.9	463.1	399.8	330.6	293.0

PDB 上での平均検索時間の本手法・旧手法（一般手法およびＦＦＴ）の比較（秒）

このように、この研究成果は、構造生物学における基本的なアルゴリズムに対する革命であり、計算生物学における最高峰の査読付国際会議である RECOMB（13th Annual International Conference on Research in Computational Molecular Biology）において最優秀論文（Best Paper Award）に選ばれるなど国際的にも極めて高い評価を受けた。(学会発表②)

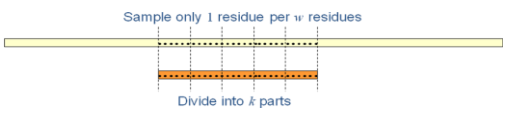
(3) 高速検索のための索引構造

さらに本研究では、前項のアルゴリズムをさらに高速化するための前処理アルゴリズムも開発することに成功した。(雑誌論文③)
前述の線形時間検索アルゴリズムは、アルゴリズムの高速化のために、RMSD の理論的下限値を計算するが、この前処理は、十分小さな下限値を持つデータベース中の部分構造を二分探索によって検索することを可能とする画期的なデータ構造であり、文字列に対する有名なデータ構造である接尾辞配列のタンパク質立体構造データベース版ともいえるものである。

(4) 線形時間からのさらなる理論的高速化

文字列処理においては、Boyer-Moore アルゴリズム等、線形時間よりも高速な平均計算量で検索できるアルゴリズムが知られている。

本研究でも、前述のアルゴリズムに、さらにサンプリング技術、高次元近点検索技術を組み合わせることによって、理論的に同様の



さらなる高速化技術

線形時間よりも高速なアルゴリズムを構築することに成功した。(雑誌論文⑤)

(5) 構造の曖昧比較問題の難しさの解明

ここまで扱ったアルゴリズムは、塩基間にギャップを全く考えない状況におけるタンパク質立体構造データベースの類似検索を扱ったものである。しかしながら、実際の構造生物学においては、より難しい、塩基間のギャップ(すなわち塩基の挿入・削除)を考慮して構造を比較する必要がある場合も多い。しかしながら、そのような問題は定式化にもよるが非常に難しい問題である。特に、Contact Map Problem として知られる構造比較問題は NP 困難問題としても知られる問題で、そのことから、そのようなギャップを考慮した構造比較問題は極めて難しいと思われてきた。

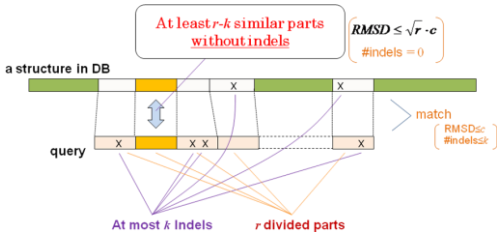
その一方で、Contact Map Problem は、構造比較問題をグラフ上の問題に落としたもので、実際に RMSD の値を最適化するような定式化ではなく、実際のところ、RMSD をギャップを考慮して最適化する問題、すなわち、タンパク質立体構造間曖昧比較問題がどれほど難しい問題かは、長い間オープンプロブレムであり、不明であった。

本研究では、この構造間曖昧比較問題を高次元へ拡張した問題が、ＮＰ困難な問題であることを世界で初めて示した。(雑誌論文①)これによって、この問題の難しさの一端を解明することに成功した。

(6) 線形時間構造曖昧検索アルゴリズム

前述のとおり、タンパク質立体構造間曖昧比較問題は、NP 困難かどうかは不明ではあるものの、極めて難しい問題である。さらに、そのような曖昧比較によって得られる RMSD の値に基づいた構造データベース検索は、曖昧比較をデータベース中のすべての部分構造に対して行う必要があり、極めて困難な問題といえる。

しかしながら、本研究では、本研究で開発したアルゴリズム設計パラダイム SMAD と、組み合わせパタンマッチング技術としてよく用いられるクエリー分割による高速化技術を組み合わせることで、理論的に線形時間でそのような曖昧検索を可能とするアルゴリズムの開発に成功した。(雑誌論文①)



3 次元立体構造曖昧検索アルゴリズム

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　５件)

① Tetsuo Shibuya, Jesper Jansson, and Kunihiko Sadakane, Linear-Time Protein 3-D Structure Searching with Insertions and Deletions, 査読有, BMC Algorithms for Molecular Biology, Vol. 5:7, 2010.

②Tetsuo Shibuya, Geometric Suffix Tree: Indexing Protein 3-D Structures, 査読有, Journal of the ACM, Vol. 57, No. 3, Article No.15, pp. 1-17, 2010.

③Tetsuo Shibuya, Searching Protein 3-D Structures in Linear Time, 査読有, Journal of Computational Biology, Vol. 17, No. 3, pp. 203-219, 2010.

④ Tetsuo Shibuya, Fast Hinge Detection Algorithms for Flexible Protein Structures, 査読有, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 7, No. 2, 2010, pp. 333-341.

⑤Tetsuo Shibuya, Searching Protein 3-D Structures in Faster Than Linear Time, 査読有, Journal of Computational Biology, Vol. 17, No. 4, pp. 593-602, 2010.

〔学会発表〕(計　３件)

①　渋谷哲朗、線形時間タンパク質３次元構造探索アルゴリズム、情報処理学会　アルゴリズム研究会、査読無、IPSJ SIG Notes SIGAL 123-4, pp. 25-32, 2009.

②Tetsuo Shibuya, Searching Protein 3-D Structures in Linear Time, 査読有, Proc. 13th Annual International Conference on Research in Computational Molecular Biology (RECOMB 2009), LNCS 5817 (LNBI 5541), pp. 1-15, 2009.　最優秀論文賞

③ Tetsuo Shibuya, Jesper Jansson and Kunihiko Sadakane, Linear-Time Protein 3-D Structure Searching with Insertions and Deletions, 査読有, 9th Workshop on Algorithms in Bioinformatics (WABI 2009), LNCS, vol. 5724, pp. 310-320, 2009.

〔図書〕(計　０　件)

〔産業財産権〕

○出願状況（計　０　件）
名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　０　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
設計パラダイム SMAD の紹介を以下のページ
で簡単に紹介している。
http://shibuyalab.hgc.jp/SMAD.html

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
該当なし

（　　）

研究者番号：

(3)連携研究者
該当なし

（　　）

研究者番号：