氏名（本籍）	髙橋　誉文（広島県） たかはし　よしふみ
学位の種類	博士（情報工学）
学位記番号	甲第116号
学位授与年月日	平成27年9月25日
学位授与の要件	広島市立大学大学院学則第36条第2項及び広島市立大学学位規程第3条第2項の規定による
学位論文題目	Studies on high-performance data processing in bio-databases バイオデータベースにおける高性能なデータ処理に関する研究
論文審査委員	主　査　教　授　北　上　始 副　査　教　授　高　橋　健　一 副　査　教　授　鷹　野　優 副　査　准教授　田　村　慶　一

論文内容の要旨

蛋白質は酵素，抗体やホルモンなど，生命活動の生体機能にかかわる重要な物質である．また，塩基配列(DNAやRNA)を初めとして，アミノ酸配列や蛋白質立体構造などを含むバイオデータベースの構築・利用を前提としたコンピュータ利用は，生命科学・医療および創薬の開発などの発展に大きく貢献するものと考えられてきている．現在，国際的に組織されている主要なデータバンクには，塩基配列データバンクおよび蛋白質立体構造データバンクがあり，日米欧による国際協力体制で運営されている．これらのデータバンクでは，塩基配列を格納する塩基配列データベースおよび蛋白質立体構造を3次元座標配列として格納する蛋白質立体構造データベースをそれぞれ構築している．さらに，これらのデータバンクでは，バイオインフォマティシャンを含む研究者コミュニティの育成を初めとして，データベースの利用を可能にするツールの構築やWeb検索サービスなどを実施してきている．

近年，塩基配列データベースを初めとして，アミノ酸配列データベースや蛋白質立体構造データベースに登録されるデータ件数が急増してきているため，このようなバイオデータベースに対する高性能なデータ処理として類似検索の高速化や高精度化の研究が盛んに行われている．

バイオデータベースの高性能なデータ処理の研究には，主として，塩基配列やアミノ酸配列に関する類似配列検索および立体構造に関する類似構造検索の二つに分類される．さらにそれぞれの研究は，索引構造を用いて高速に類似検索を行うアプローチと，整列化を用いて高精度に類似検索を行うアプローチの二つの研究に分類される．このような分類を踏まえ，本研究では，バイオデータベースに対する高性能なデータ処理の実現をめざし，以下の三つの問題点に取り組む．

（１） 類似配列検索の高速化

塩基配列やアミノ酸塩基配列は文字列データである．類似配列検索の高速化をめざし，索引構造として文字列の接尾辞木が研究されている．この索引構造の研究では，小規模データを対象としているため，大規模なデータベースを想定した研究が十分に行われていないという問題点がある．

（２） 類似構造検索の高速化

蛋白質立体構造は座標配列データである．類似構造検索の高速化をめざし，座標配列の接尾辞木が研究されている．この索引構造の研究では，小規模データを対象としているため，大規模なデータベースを想定した研究が十分に行われていないという問題点がある．

（３） 類似構造検索の高精度化

蛋白質立体構造データベースから類似構造検索をするには，二つの構造間の構造整列化を繰り返し実施しながら，問合せ構造に類似すると蛋白質構造をデータベースから見つけ出すことが重要となる．構造整列化を実施する方法には，RMSD や動的計画法，CMO 問題を解く方法が研究されている．しかしながら，EO を用いた発見的解法には，解の大域的な探索が難しい CMO 問題に対して局所解へ陥りやすいという問題点がある．

上記（１）の問題点については，マルチコア CPU 上においてメモリ上ではなくディスク上に構築されたディスクベースの接尾辞木を用いた類似配列検索のための高性能な並列化モデルを提案する．そのために，配列データベース全体をパーティションと呼ばれる二つ以上のサブデータベースに分割するデータ分割型並列化を採用する．パーティションそれぞれでディスクベースの接尾辞木を構築し，一つのディスクベースの接尾辞木に対する類似配列検索を一つのタスクとして定義する．さらに，CPU のコア間における競合を防ぐためにマルチプルバッファリング管理システムを導入する．評価実験を行い，従来の手法より最大で約 35%の高速化を達成し，効率の良い類似配列検索が可能となった．

上記（２）の問題点については，幾何学的接尾辞木をメモリ上ではなくディスク上に構築する方法と，幾何学的接尾辞木の構築と検索を並列化する方法を提案する．そのために，前者についてはディスク上に幾何学的接尾辞木を構築するためのバッファ管理システムを構築する．後者については，幾何学的接尾辞木の構築方法を逐次構築法からトップダウン構築法に変更し，座標配列の全件をまとめて構築する．さらに，幾何学的接尾辞木に対するデータ分割法やタスク分割法による問題分割法のほかに，マスタワーカ法や分散型ワーカ法による負荷分散法を組み合わせた並列化を行う．評価実験を行った結果，逐次構築法よりも高速な幾何学的接尾辞木の構築と検索を行うことができた．

上記（３）の問題点については，改良版 EO を用いた CMO 問題の発見的解法を提案する．そのために，世代交代に改良版 EO，初期個体の作成に動的計画法，改良版 EO における状態遷移に最良移動戦略の三つを用いている．提案手法の評価実験を行った結果，EO による発見的解法よりも評価の高い最良解が得られた．

論文審査の結果の要旨

平成27年8月10日（月）13：00から14：30まで博士学位論文発表会（公聴会）を開催した。申請者が論文内容について説明を行い、その後、論文内容および専門知識に関する質疑応答を行った。

ビッグデータ時代において、DNA塩基配列データベースを初めとして、アミノ酸配列データベースやタンパク質立体構造データベースに登録されるデータ件数が急増してきているため、このようなバイオデータベースに対する高性能なデータ処理として類似検索の高速化や高精度化の研究が盛んに行われている。本論文では、バイオデータベースに対する高性能なデータ処理の実現をめざし、以下の二つの方法を提案している。

第１の方法は、文字列データや座標配列データから成るデータベース対して高速な類似検索法を実現する方法である。この方法では、DNA塩基配列やアミノ酸配列を文字列データとみなし、タンパク質立体構造を座標配列データとみなした上で、これらのデータがそれぞれ含まれるデータベースにおいて高速な類似検索を可能にする索引構造として、接尾辞木に着目している。従来の接尾辞木の研究では、大規模なデータベースを想定した研究が十分に行われていないという問題点がある。この問題を解決するために、大規模データがディスク上に格納されていることを前提に、隠れ配列法を初めとして、接尾辞木の構築や検索に適したバッファ管理法や並列処理法を導入することにより、接尾辞木に基づく高速処理手法を提案し、評価実験によりその有効性を確認している。

第２の方法は、高精度な類似構造検索を達成する上で重要な仕組みである構造整列化をCMO問題とみなして、解決する方法である。CMO問題は解の大域的な探索が難しいが、進化的アルゴリズムの１つであるEOにより発見的に解く方法が研究されている。しかしながら、この方法は、局所解に陥りやすいという問題点がある。この問題を解決するために、世代交代に改良版EO、初期個体の生成に動的計画法、改良版EOにおける状態遷移に最良移動戦略の三つを用いることにより、従来の手法よりも高精度な解を得る方法を提案し、評価実験によりその有効性を確認している。

本研究の新しい研究成果は、2編（フルペーパ）の情報処理学会論文誌および3編の査読付き国際会議論文等で公表済みであり、本論文を総合的に評価した結果、博士学位論文審査は合格と判定した。