論文 / 著書情報
Article / Book Information

題目(和文)	大量メタゲノム情報に対するアミノ酸配列相同性検索の高速化
Title(English)	Faster Protein Sequence Homology Searches for Large-scale Metagenomic Data
著者(和文)	鈴木脩司
Author(English)	Shuji Suzuki
出典(和文)	学位:博士（工学）, 学位授与機関:東京工業大学, 報告番号:甲第9884号, 授与年月日:2015年3月26日, 学位の種別:課程博士, 審査員:秋山　泰,佐藤　泰介,宮崎　純,村田　剛志,関嶋　政和
Citation(English)	Degree:, Conferring organization: Tokyo Institute of Technology, Report number:甲第9884号, Conferred date:2015/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	論文要旨
Type(English)	Summary

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	計算工学　専攻	申請学位（専攻分野）： Academic Degree Requested	博士（工学） Doctor of
学生氏名： Student's Name	鈴木　脩司	指導教員（主）： Academic Advisor(main)	秋山　泰
		指導教員（副）： Academic Advisor(sub)	

要旨（和文 2000 字程度）
Thesis Summary（approx.2000 Japanese Characters ）

本論文は大量なデータに対するメタゲノム解析を実行可能とすることを目的として，(1)高速なアミノ酸配列相同性検索アルゴリズムの提案，(2)graphic processing unit(GPU)による高速化，(3)大規模計算機を利用した並列実行について報告する．

類似したアミノ酸配列を持つタンパク質同士には進化的な関係があり，互いによく似た機能を持つことが知られている．このため，類似した配列の情報を用いることでタンパク質の機能を予測することができ，クエリとなるタンパク質のアミノ酸配列と類似する配列を巨大なデータベースの中から見つけ出すアミノ酸配列相同性検索は，生命情報解析の基礎となる手法となっている．

近年，DNA 配列を読み取る機器である DNA シーケンサの改良が進んだことにより，短時間に大量の短い DNA 配列断片を得ることができるようになった．このため，大量の配列情報を利用した解析を行いたいという要求が高まっている．しかし，微生物の集団から培養を経ずに直接 DNA 配列を読み取って解析するメタゲノム解析の場合，最新の DNA シーケンサが出力するデータがあまりにも大量であるため，従来用いられてきた検索精度の高いアミノ酸配列相同性検索である BLASTX では，多くの処理時間が必要となる．このため，アミノ酸配列相同性検索の高速化が喫緊の課題となっている．

(1)の高速なアミノ酸配列相同性検索アルゴリズムの提案では，suffix array による可変長文字列比較を用いたアルゴリズムとデータベースの部分文字列クラスタ情報を用いたアルゴリズムを提案した．Suffix array による可変長文字列比較を用いたアルゴリズムでは，文字列間の類似度指標を基準にしてクエリの部分文字列毎に検索すべき対象文字列の長さを変更し，BLASTX よりも平均的に長い部分文字列を高速に検索する．さらに，クエリとデータベースの両方でデータ構造として suffix array を用いることで，複数回出現する部分文字列に関してはまとめて検索を行う．このアルゴリズムを GHOSTX として実装し，典型的な口腔内や土壌のメタゲノムのデータを用いて BLASTX と比較したところ最大約 165 倍の速度向上が得られることを示した．また，データベースが年々巨大化しているのに伴い冗長な部分文字列が増加している．この冗長な部分文字列に対して効率的なアミノ酸配列相同性検索を行うために，データベースの部分文字列を予めクラスタリングしておき，このクラスタ情報と，文字列間距離に関する三角不等式を利用して詳細なスコア計算を行う回数を削減して高速化するアルゴリズムを提案した．このアルゴリズムを GHOSTZ として実装し，実際のメタゲノムデータを用いて GHOSTX と比較したところ最大約 2 倍の速度向上が得られることを示した．

高速なアミノ酸配列相同性検索アルゴリズムの提案に加え，(2)の GPU による高速化では，GHOSTZ のアルゴリズムを基にして GPU を用いて検索を行う GHOSTZ-GPU を開発した．GHOSTZ-GPU は GPU のメモリアクセスの最適化や CPU と GPU の非同期処理の利用により，さらなる高速化を行い，12 CPU threads と 3 GPUs を利用した場合，GHOSTZ の 12 CPU threads 利用時よりも最大約 7 倍の速度向上が得られることを実際のメタゲノムデータを用いて示した．

また，(3)の大規模計算機を利用した並列実行では，Message Passing Interface(MPI)を利用し，提案した高速なアミノ酸配列相同性検索を複数ノード上で実行可能とした．実際のメタゲノムデータを用いて TSUBAME2.5 の 128 ノードを利用して実験したところ，BLASTX を MPI によって並列実行する mpiBLAST と比べ，GHOSTX を MPI によって並列実行する GHOST-MP は約 89 倍の速度向上が得られることを示した．これにより，TSUBAME2.5 や「京」などのスーパーコンピュータを利用し，大規模なデータを利用した高速なアミノ酸配列相同性検索が実行可能となった．

本研究により最新の DNA シーケンサが出力する全データを利用した大規模なメタゲノム解析が可能となり，環境中の微生物の関係などを詳細に解析するための新たな解析支援ツールとして広く利用されることが期待される．

備考：論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note：Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).

（博士課程）
Doctoral Program

論 文 要 旨

THESIS SUMMARY

専攻： Department of	計算工学 専攻	申請学位（専攻分野）： Academic Degree Requested	博士 （工学） Doctor of
学生氏名： Student's Name	鈴木 脩司	指導教員（主）： Academic Advisor(main)	秋山 泰
		指導教員（副）： Academic Advisor(sub)	

要旨（英文 300 語程度）
Thesis Summary（approx.300 English Words ）

Sequence homology search is an approach for establishing structural and functional similarity with existing genes or proteins using a variety of databases containing a large number of DNA and protein sequences and the associated biological information. Sequence homology search is used in metagenomics. However, because of improvements in DNA sequencing technology, the volume of sequence data and the number of queries used in this analysis have been increasing rapidly in recent years, and the speed of sequence homology search has become insufficient.

In this dissertation, we propose fast protein sequence homology search algorithms that can be applied to metagenomics using the latest DNA sequencing output. We used three approaches: development of novel protein sequence homology search algorithms, acceleration of protein sequence homology search with graphics processing unit (GPU), and parallelization of protein sequence homology search using modern supercomputing environments.

We propose a novel protein sequence homology search algorithm that finds similarities between a query and database sequences based on the suffix arrays of these sequences. We used a subsequence search method relying on a similarity-based optimal length. This algorithm designated as GHOSTX provides approximately 165 times faster protein sequence homology search than BLASTX in the analysis of metagenomic data. In addition, we propose a novel protein sequence homology search method based on database subsequence clustering, designated as GHOSTZ. This method clusters similar subsequences retrieved from a database to reduce alignment candidates based on triangle inequality, and its performance in the analysis of metagenomic data is approximately two times faster than that of GHOSTX.

In addition, we applied the GPUs and massively parallel computing systems, TSUBAME and the K computer, for protein sequence homology search and show that these approaches provide a significant acceleration of protein sequence homology search.

DNA sequencing technology is constantly improving, resulting in generation of vast amounts of sequence data. This explosion of sequence volume makes computational analysis with contemporary tools more difficult. Here, we offer the algorithms, which may provide a potential solution to this problem.

備考 ： 論文要旨は、和文 2000 字と英文 300 語を 1 部ずつ提出するか、もしくは英文 800 語を 1 部提出してください。
Note : Thesis Summary should be submitted in either a copy of 2000 Japanese Characters and 300 Words (English) or 1copy of 800 Words (English).