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)	審査の要旨
Type(English)	Exam Summary

（博士課程）

論文審査の要旨及び審査員

報告番号	甲第　　号		学位申請者氏名	鈴木　脩司	
論文審査 審査員		氏　名	職　名		氏　名	職　名
	主査	秋山 泰	教授	審査員	関嶋 政和	准教授
	審査員	佐藤 泰介	教授			
		宮崎 純	教授			
		村田 剛志	准教授			

論文審査の要旨（2000字程度）

本論文は「Faster Protein Sequence Homology Searches for Large-scale Metagenomic Data」（邦題：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）と題し，英文8章から成る．

第1章「Introduction」では，研究の背景を説明するとともに，本論文の全体構成を示している．アミノ酸配列相同性検索は文字列検索の技術を応用した生命情報分野の解析手法であり，様々な生命データ解析の基礎となる重要な手法となっている．現在，DNA配列の読み取り技術が急速に向上したことにより大量な配列情報に対する解析の需要が高まっているが，特にメタゲノム解析では検索精度の高いアミノ酸配列相同性検索を実現するために膨大な計算時間が必要である．このため，アミノ酸配列相同性検索の高速化は喫緊の課題となっている．

第2章「Sequence Homology Search」では，本論文で扱うアミノ酸配列相同性検索に関する基礎事項や関連研究を説明している．

第3章「A Protein Sequence Homology Search Algorithm Using a Query Suffix Array and a Database Suffix Array」では，高速なアミノ酸配列相同性検索を行うために suffix array を用いた可変長文字列比較のアルゴリズムを提案している．このアルゴリズムでは，文字列間の類似度指標を基準としてクエリの部分文字列毎に検索するべき対象文字列の長さを変更し，従来手法である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に対してデータ構造として 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というアイデアを導入している．このアルゴリズムをGHOSTXソフトウェアとして実装し，従来手法であるBLASTXと比較したところ典型的な口腔内や土壌のメタゲノム解析のタスクにおいて最大約165倍の速度向上が得られることを実際のデータを利用し実験的に示している．

第4章「A Protein Sequence Homology Searches with Clustering Subsequences Technique」では，アミノ酸配列相同性検索をさらに高速化するために，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，配列間距離に関する三角不等式を用いて，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ている．このアルゴリズムを GHOSTZ ソフトウェアとして実装し，BLASTX よりも最大約 285 倍の速度向上が得られることを実際のメタゲノムのデータを利用し実験的に示している．

第5章「A GPU-Accelerated Protein Sequence Homology Search」では，graphic processing unit(GPU)に適したアミノ酸配列相同性検索のアルゴリズムを提案している．このアルゴリズムをGHOSTMソフトウェアとして実装し，BLASTXの1 CPU thread利用時と比較して1 GPUを利用した場合に最大約130倍，4 GPUを利用した場合に最大約407倍の速度向上が得られることを実際のメタゲノムのデータを利用し実験的に示している．

第6章「A Protein Sequence Homology Searches with Clustering Subsequences Technique on GPUs」では，第5章で提案したGHOSTMをさらに高速化するために，第4章で提案したGHOSTZと組み合わせたアルゴリズムを提案している．このアルゴリズムでは 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っている．このアルゴリズムをGHOSTZ-GPUソフトフェアとして実装し，12 CPU thread と 3 GPU を利用した場合，GHOSTZの12 CPU thread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のデータを利用し実験的に示している．

第7章「A Large-scale Protein Sequence Homology Search on Massively Parallel Computing System」では，アミノ酸配列相同性検索を並列計算機で効率的に実行するために，Message Passing Interface (MPI) を利用した並列化を行い，効率的に実行するための実装上の工夫と，性能評価について述べている．

第8章「Conclusion and Future Work」では，本論文で得られた結果を総括するとともに，残された課題やさらなる研究の発展性について論じている．

補遺では，本論文を通じて検索精度を評価する際に利用した厳密解との比較による正解率とは別の指標として，類似度に基づく指標を導入し，この指標を用いた場合の検索精度の評価について補足的に述べている．

以上を要するに本論文は，生命情報解析における基礎的技術であるアミノ酸配列相同性検索の問題に対して，大幅な性能改善を得られる新規手法を提案するものであり，工学上，及び，工業上貢献するところが大きい．

よって本論文が博士（工学）の学位論文として十分価値あるものと認める．