Thesis for the Degree of Ph.D. in Science

Algorithms for comparing and visualizing genome-scale datasets

January 2013

Graduate School of Science and Technology
Keio University

Kristoffer Popendorf

主　論　文　要　旨

報告番号	㊙甲 乙 第　　号	氏 名	ポペンドフ　クリストファー

主論文題目:

Algorithms for comparing and visualizing genome-scale datasets

(ゲノム規模データセットの比較解析および視覚化するためのアルゴリズム)

（内容の要旨）

近年の超並列シークエンシング技術の発展によって、ゲノムに関するデータを新しく生成するコストと時間が急速に下がり、ゲノムを解読するプロジェクトの数が急激に増えている。現在2,547種の真核生物と12,460種の原核生物のゲノム解読プロジェクトが進行中である。また、ヒトやチンパンジー，マウスなどの57種の脊椎動物のゲノムがドラフト配列として完成されたので、ヒトに関連するゲノム解析を強力に推し進めることができる時代になった。しかし処理するべきデータの量と配列の数が指数的に増えているため、従来の解析手法をそのまま適用すると非現実的なコストがかかってしまう。

本論文では、ゲノム規模データセットを解析するための2つのアルゴリズムを提案する。第1のアルゴリズムは、複数のゲノムから相同領域を並列計算機で検索するためのアルゴリズムの構築である。第2のアルゴリズムは、近年開発された次世代シークエンサーから生成される大量のリードデータを解析した結果を視覚的に分かりやすい形で高速に表示するアルゴリズムである。

第1章では、ゲノム解読と比較ゲノム解析、次世代シークエンスデータとその解析および問題点について述べた。

第2章では、相同領域を検索するために開発したMurasakiと呼ばれるアルゴリズムについて述べた。複数の大規模な全ゲノムでも効率的に比較できるようにするために、計算機クラスターを用いた並列計算によって高速にゲノム比較が計算できる手法を開発した。並列計算の効率を保証するために、ハッシュ関数の計算および計算ジョブの振り分け方に関する新しいアルゴリズムを実装した。脊椎動物のゲノムデータを用いた検証実験により、相同領域の検出精度、並列計算の効率のいずれにおいても既存手法よりも優れていることを確認した。

第3章では、大量のシークエンスデータを解析した結果を高速に表示するアルゴリズムについて述べた。次世代シークエンサーから生成された何千万本の短いリード配列は、リードマッピングのプログラムを用いて参照ゲノムに配置される。配置されたリードマッピングの形状から生物学的な解析を行うためには、全ゲノム領域中の形状を高速に視覚化するためのソフトウエアが必要である。哺乳類規模の大きなゲノムに大量のリードデータをマッピングした場合、ゲノムレベルで俯瞰する表示から一塩基レベルでの詳細な表示までをスムーズに可視化するプログラムは存在しない。そこで、Samscopeと呼ばれるコンピュータグラフィックスの手法を取り入れた新しい表示プログラムを開発した。チキンゲノムのセントロメア解析において、Samscopeはタンパク質結合部位をゲノム配列から視覚的に発見することに威力を発揮することが示された。

第4章では、本研究を総括するとともに、提案した2つのアルゴリズムについて他のゲノム解析問題への応用可能性を議論した。

SUMMARY OF Ph.D. DISSERTATION

School	Student Identification Number	SURNAME, First name
School of Fundamental Science and Technology	80745137	POPENDORF, Kristoffer

Title

Algorithms for comparing and visualizing genome-scale datasets

Abstract

Recent years have seen a massive explosion in the number, complexity, and raw volume of new sequencing data thanks to advances in modern sequencing technology. In particular the advent of massively parallel sequencing, or colloquially "Next Generation Sequencing" or "NGS," has opened up a new world of sequencing applications that were once impractical at best. Whole bacteria can now sequenced in a matter of days, new mammalian genomes can be sequenced for a fraction of what they once cost, and well known species like *homo sapiens* can be re-sequenced to discover novel genetic variants for less than a $1000. In the first chapter of this dissertation, we review the current state of genomics sequencing technology, its applications, and current challenges.

One of the products of this sequencing explosion has been a wealth of newly sequenced genomes, including 57 vertebrates. With such rich data concerning some of our closest evolutionary relatives, comparative genomics studies promise to provide great insight into our physiology and development though analysis of similarities of whole genomes across multiple species. However, existing comparative genomics tools are capable of dealing with a few chromosomes at one time, and require excessive computational resources to keep pace with the vast number of genomes rapidly becoming available. To address this problem, we introduce a new approach to parallel sequence similarity search which offers efficient use of cluster computing resources to provide the scalability necessary to analyze current and future genome projects. We've named this algorithm Murasaki, and its details are described in Chapter 2.

Another application of NGS technology has been in areas of transcriptome, regulation, and variant analysis. Two relatively new applications unique to NGS use the massive number of reads available from NGS to assay RNA products by sequencing the RNA itself (RNA-Seq), or capture and sequence the DNA bound to specific transcription factors or DNA-binding proteins (ChIP-Seq). The data from these experiments can be hard to understand because of the scale of the data involved is overwhelming and requires some practical reduction to find features of interest before conducting a more detailed investigation. Existing techniques for visualizing NGS data has been limited to examining small regions and/or offered limited support RNA-Seq/ChIP-Seq features. To address this problem we propose a new algorithm and data format implemented in our program, Samscope, described in chapter 3.

In chapter 4 we summarize the impact of these new approaches, and examine their potential future areas of development.