氏　　名	Phanucheep CHOTNITHI
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 2156 号
学位授与の日付	2020 年 3 月 24 日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Efficient Similarity Measures in NGS Genome Data Comparison for Phylogeny Reconstruction
論 文 審 査 委 員	主　査　教授　高須 淳宏
	教授　大山 敬三
	教授　合田 憲人
	准教授　相原 健郎
	教授　阿久津 達也 京都大学化学研究所

(Form 3)

Summary of Doctoral Thesis

Name in full Phanucheep CHOTNITHI

Title Efficient Similarity Measures in NGS Genome Data Comparison for Phylogeny Reconstruction

In the study of molecular evolution, the researchers in the field need to construct the phylogenetic tree to be analyzed in many applications. Phylogeny reconstruction is the process of studying the evolutionary relationships of a group of species. Such relationships can be represented by a branching diagram or a tree called Phylogenetic tree. Phylogenetic trees have been used in many applications in bioinformatics such as molecular evolution analysis, forensics, and medicine development. The very first step of phylogeny analysis is to construct an accurate phylogenetic tree for the genome sequences of interest. To construct the phylogenetic tree, pairwise distances between the genome sequences of species need to be computed. After that, the phylogenetic tree can be generated using a tree construction method on the distance matrix.

The phylogeny reconstruction requires the process of comparing and measuring the distance of genome sequences. With the growth of genomic data, next-generation sequencing (NGS) has become the mainstream format for genome sequence data due to its high-throughput sequencing. Next-generation sequencing is a method used to transform data from genome samples to a digitized data sequence which achieves a rapid throughput compared with traditional sequencing processes. Instead of one long sequence of genome data, NGS produces a large number of sequence fragments called *reads* per genome sample. NGS can be applied to various biological problems, including *denovo* whole-genome sequencing and RNA-seq. The NGS creates new challenges in many applications for genome sequence analysis. In the sequence comparison application, the traditional multiple-sequence alignment approach is not a solution for NGS data anymore because of short-read assembly and computational resource problems. Therefore, alignment-free methods are more suitable for NGS data comparisons. Most of the alignment-free methods are k-mer based algorithm. However, the characteristics of NGS data might make those k-mer based methods non-optimal since the k parameter is a crucial factor in distance measurement and phylogenetic tree results. This thesis proposes a novel approaches for parameter-free comparison of NGS short reads, which aims to eliminate the dependency of k parameters.

First, an alignment-free sequence comparison based on the number of the neighbor search result on the NGS read called d^{NS} is proposed to reduce the effect of

the overlap problem of the NGS data. Because most of the alignment-free methods have relied on a k-mer based distance measure, however, with the characteristic of NGS data, this might not be the optimal solution. NGS data contain the tremendous amount of overlap among each NGS read fragment with random. This affects the distance between NGS data of each input species. Instead of calculating the distance between NGS sets based-on k-mer, d^{NS} defines distance based on the number of neighbor search results. The proposed method is evaluated by comparing with two existing methods and show that the method can distinguish the difference between diverse species better than compared methods. According to the experiment, d^{NS} is effective for the input set, which each sequence is diverse from the other with high coverage. d^{NS} can construct the phylogenetic tree of a diverse species dataset with high accuracy, which indicates that d^{NS} can achieve sequence comparisons using NGS data. Also, d^{NS} is more robust concerning various values for k than the other k-mer based alignment-free methods, which indicates that d^{NS} is robust against the effects of NGS short-read overlap on the k-mer frequency distribution. However, d^{NS} performs decently for the closely related species dataset. Because this neighbor search-based alignment-free approach to sequence comparison is novel, there is plenty of scope for further development and possible improvements.

Finally, the other novel approach called d^{RA} is proposed, which is an effective parameter-free comparison of NGS short reads. Since the problems of k parameter have been discussed. d^{RA} focuses on k parameter-free approach. d^{RA} also provides an improvement from d^{NS} to calculate the distance of NGS sets of close species. To measure the distance between the NGS sets, the method contains two steps. The first step is the searching of the corresponding alignment pair of the NGS short read in the other NGS sets. Next, the information of the alignment pairs from the previous step is used to calculate the distance between NGS sets without k parameters. d^{RA} is also evaluated by conducting the experiments that compared the proposed method with existing methods. The results show that the novel read alignment approach can provide an accurate distance measurement on three simulated NGS datasets to construct the phylogenetic tree compared with other alignment-free methods. The phylogenetic trees constructed from the method are more similar to the benchmark tree provided by other researchers while requiring no parameter adjustment. The experiment on the multiple simulated NGS sets from the same dataset is also conducted to evaluate the effect on different reads randomness and coverage. d^{RA} manages to calculate the accurate distance between closely related species, which is an improvement from d^{NS}.

Form8・Separate Sheet
(様式8・別紙)

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏　名　Phanucheep CHOTNITHI

Title
論文題目　Efficient Similarity Measures in NGS Genome Data Comparison for Phylogeny Reconstruction

出願者は、生物の進化的関係の分析に用いる遺伝子データの類似度の計算法の研究を行い、その成果を博士論文としてまとめた。この研究では、次世代シーケンサ(NGS)によって得られる遺伝子配列データの類似度を計算する方法を提案し、従来手法と比較し系統樹の生成に効果があることを実験的に示している。

博士論文は６章より構成され英語で書かれている。第１章では、本論文のテーマである遺伝子配列の比較手法の意義について説明したのち、NGS によって得られる遺伝子配列の比較における技術的課題を論じ、本研究の貢献を述べている。第２章では、本研究の背景となる NGS の概要と生物の進化研究における役割を述べたのち、第３章では、本研究の関連研究である NGS データの類似度の計算法をサーベイしている。ここでは、関連研究を多重整列(multiple sequence alignment)を用いた方法と遺伝子部分配列を用いる方法(alignment-free comparison method)に分類し、それぞれの特性について論じている。続く２つの章で本論文の主たる貢献を述べている。第４章では遺伝子部分配列を用いた新たな類似度計算法を提案している。従来は遺伝子配列の特徴を固定長の部分配列(k-mer)の出現頻度ベクトルを用いて表すことが多いが、長い k-mer を用いた場合、高次元でスパースな特徴空間となる問題がある。本研究では、以下の二段階の類似度計算法を提案している。第一段階では、NGS によって得られる遺伝子部分配列(Read)間の類似度を k-mer の出現頻度ベクトルを用いて求める。第二段階では、Read の集合として表される遺伝子配列の特徴をデータ中からランダムに選んだ Read（参照 Read）を特徴素とし、その近傍の Read の数を特徴量とする特徴ベクトルで表す。この特徴ベクトルを用いた遺伝子配列間の類似度を提案している。参照 Read 数を調整することで特徴空間の次元を削減する方法となっている。続いて、系統樹の評価によく用いられる２つの評価データを用いて評価実験をおこなっている。ここでは、提案手法および比較手法の類似度を用いてクラスタ解析(Neighbor-Joining 法)を行い、得られた系統樹と評価データの基準となる系統樹との距離に基づいて各類似度の性能を評価している。次元の削減率と性能の比較を行い、read を特徴として用いることで従来手法よりも性能向上が図れること、また、特徴空間の次元を 1%程度に圧縮できることを示している。第５章では、第４章で提案された類似度計算法において必要になる k-mer の長さや参照 Read 数といったハイパーパラメタを必要としない計算法を提案している。まず、第一段階では、Read 間の類似度に編集距離を用いることで k-mer の長さに関するパラメタを不要にしている。第二段階では、最近傍 Read 対を特徴素し、その編集距離を特徴量とする特徴ベクトルを用いる。ここで、最近傍対の距離分布として Gauss 混合分布を仮定し、距離の短いグループを特徴素として用いることで次元数に

関するパラメタを不要にしている。評価実験では３つの評価データを用い、第４章と同じ方法で性能評価を行なっている。従来手法では生成される系統樹はk-merの長さに大きく影響を受けるが、提案手法は、従来手法で最も性能の良いパラメタを用いて得られた系統樹と同等以上の結果が得られたことが報告されている。第６章では、以上の結果をまとめるとともに今後の課題を示している。

公開論文発表会において、出願者はおよそ 45 分で博士論文の内容を説明し、その後、15 分程度の質疑が行われた。続いて、口述試験が行われ、審査委員からは関連研究との相違点、提案手法の特徴、実験手順の詳細、今後改善すべき点等について質問とコメントが寄せられ、出願者は適切に回答した。

口述試験後に審査委員で議論を行った。博士論文審査の結果、出願者は情報学分野の十分な知識と研究能力を持つと認められ、また研究内容は学位論文として十分なレベルの新規性、有効性があると認められた。本論文の内容に関し、情報処理学会論文誌に１編、査読付き国際会議に３編の論文が採択されている。以上より、審査委員会全員一致で、博士論文として十分な水準の研究であると認め、学位の授与に値すると判断した。