論文 / 著書情報
Article / Book Information

題目(和文)	環境中の微生物叢解析のための超高速な配列解析ツールの開発
Title(English)	
著者(和文)	矢野雅大
Author(English)	Masahiro Yano
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10498号, 授与年月日:2017年3月26日, 学位の種別:課程博士, 審査員:山田 拓司,黒川 顕,伊藤 武彦,山口 雄輝,中島 信孝
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第10498号, Conferred date:2017/3/26, 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	生命情報 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	矢野 雅大	指導教員（主）： Academic Advisor(main)	山田 拓司
		指導教員（副）： Academic Advisor(sub)	黒川 顕

要旨（英文 800 語程度）

Thesis Summary（approx.800 English Words ）

[Background & Purpose]

Metagenomics is widely applied for studying microbial community dynamics in a variety of habitats, such as natural environments or human bodies. In metagenomics, there are two major analytical methods to investigate microbial communities: shotgun metagenomic DNA sequencing, and amplicon sequencing using the 16S rRNA gene. In both forms of analyses, homology searching is the core technology, and it is feasible to detect homologous sequences or regions between the metagenomic and database sequences and annotate the functional and taxonomical information. In recent years, with the rapid expansion of data produced by next-generation sequencing, homology searching for both shotgun and amplicon metagenomic analysis has become difficult in terms of computational cost.

Owing to the trade-off between calculation speed and sensitivity, most of the homology searching tools, such as BLAST, have high accuracy but are slow. Genome resequencing tools, such as Bowtie2, are ultrafast but fail to detect weak homology. One of the efficient ways to lower the computational cost of homology searching is to reduce the query sequence and accelerate the calculation. For amplicon metagenomic analysis, many sequence clustering algorithms have been developed to reduce the amount of query sequences. In general, sequence clustering tools with greedy algorithms, such as UCLUST, accomplish the clustering quickly, but the accuracy is compromised. Conversely, hierarchical clustering algorithms, such as ESPRIT-TREE, have high accuracy but are slow.

Therefore, to overcome these problems and make metagenomics analysis easier, we developed two types of tools for homology searching and sequence clustering. The first is a novel, ultrafast, and sensitive massive homology search tool named CLAST that uses Graphics Processing Units (GPUs). The second is a novel, ultrafast, and highly accurate sequence clustering tool named Saturn.

[Methods & Results]

CLAST detects short perfect matches (seeds) between query sequences and database sequences, and then performs sequence alignment in the area where multiple seeds are found. CLAST has three significant features. First, CLAST tries to

detect homologous regions between multiple database sequences and multiple query sequences in parallel. Second, CLAST does not create a sequence index separately. Third, CLAST can perform global alignment (global mode) and local alignment (local mode). The first feature allows for CLAST to exploit GPUs effectively, the second allows for CLAST to use massive and frequently updated sequence database easily, and the third makes CLAST suitable to use for motif searching (local mode) and taxonomic assignment (global mode).

In the performance test that simulated metagenomic analysis, CLAST was 80.8 times as fast as BLAST, 9.6 times as fast as BLAT, and approximately as fast as Bowtie 2. In the test, the CGA-ratio (the ratio of each tool successfully predicting the genus of query sequences from the database) of BLAST and CLAST (local mode) was 54?63% and that of CLAST (global mode) was 89?92%. In addition to these results, the sensitivity of CLAST (local mode) was comparable to that of BLAST. It is noteworthy that the amount of memory consumption of CLAST was less than 4 GB in any case.

Saturn performs greedy clustering with two significant features. First, Saturn counts the numbers of duplicate sequences and then clusters the sequences in descending order of those numbers. Second, Saturn calculates a virtual centroid sequence for each cluster and measures the cluster radius, using the virtual centroid sequence. The first feature enables Saturn to estimate true cluster centers without massive calculations, and that make the results of Saturn more appropriate.

In the performance test, Saturn was at most 5.15 times as fast as DNACLUST and 1.39 times as fast as UCLUST. In this test, Normalized Mutual Information (NMI) between the results of ESPRIT-TREE and Saturn was larger than the NMI between the results of ESPRIT-TREE and other greedy clustering tools in any case. This indicates that the clustering result of Saturn resembles that of ESPRIT-TREE more than those of any other greedy clustering tools we tested.

[Discussion & Outlook]

In the performance test of CLAST, we demonstrated that CLAST is far faster than most of the existing homology search tools, has enough sensitivity for metagenomic analysis, and the amount of memory consumption of CLAST is less than 4 GB. This performance indicates that CLAST enables analysis of massive metagenomic sequences even with a desktop PC with one GPU. In the performance test of Saturn, we demonstrated that Saturn is one of the fastest sequence clustering tools, and its clustering accuracy is comparable to those of hierarchical clustering tools. This performance indicates that Saturn enables clustering of millions of 16S rRNA gene sequences in hundreds of seconds with high accuracy.

In this study, we developed CLAST and Saturn, which accomplished extremely fast and highly sensitive calculations for metagenomic analysis. Our study overcame the problem of the computational cost for analyzing metagenomic sequences, and will contribute to the goal of further unveiling microbial communities and their dynamics.