様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月　 ３ 日現在

機関番号：３２６４４
研究種目：研究活動スタート支援
研究期間：2017～2018
課題番号：１７Ｈ０７１２３

研究課題名（和文）Medical Sequence Classifier - a system for quick and accurate sequence-based diagnostics

研究課題名（英文）Medical Sequence Classifier - a system for quick and accurate sequence-based diagnostics

研究代表者

クリュコフ　キリル（KRYUKOV, Kirill）

東海大学・医学部・奨励研究員

研究者番号：２０８０６２０２

交付決定額（研究期間全体）：（直接経費）　　2,100,000円

研究成果の概要（和文）：感染症病原体の迅速かつ正確な診断は、治療結果を成功させるために重要です。 このプロジェクトで私は非常に効率的なシーケンス分類システムを開発しました。 このシステムは、それらのサンプルに含まれる病原体を検出するために医学的メタゲノムシーケンスサンプルに適用することができる。 また、分類子が参照として使用するGenomeSyncデータベースも開発しました。 私の医療メタゲノム研究システムは、研究、現地調査、医療サンプルの検査に有用性を示しています。 結果は査読付き文献に記載されている。 私の分類器システムは、次世代シーケンシングを用いた感染症の調査に広く採用されると期待しています。

研究成果の学術的意義や社会的意義
Our society is increasingly threatened by pandemics and spread of drug-resistant pathogens. My sequence classifier will help in achieving rapid response to these threats, by providing fast and accurate sequence based diagnostics.

研究成果の概要（英文）：Rapid and accurate diagnostics of infectious disease pathogen is important for successful treatment outcome. In this project I developed a highly efficient sequence classifier system. This system can be applied to medical metagenomic sequence samples in order to detect pathogens contained in those samples. I also developed the GenomeSync database that is used by the classifier as a reference. My system for investigating medical metagenomes has shown utility in research studies, field trips and tests on medical samples. The results have been described in peer-reviewed literature. I expect that my classifier system will be widely adopted for investigating infectious diseases using next generation sequencing.

研究分野：Bioinformatics

キーワード：Metagenome　NGS sequencing　classification　diagnostics　DNA　sequence

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

Rapid and accurate diagnostics of infectious disease pathogen is important for successful treatment outcome. Recently sequencing-based approaches show promise as efficient diagnostic tools. However their performance is still lacking in both speed and accuracy. The currently available DNA classifiers are severely limited, providing either modest accuracy at the cost of long computation time, or high speed with low accuracy. Also, existing classifiers use small databases that don't contain recently sequenced genomes. Therefore there is an urgent need for developing a better techniques for sequence-based diagnostics of infections.

２．研究の目的

(1) The purpose of this project was developing a computational system for classifying metagenomic reads obtained from medical samples, with the purpose of detecting and describing the pathogenic microbes contained in those samples. Such system would allow quick diagnostics of dangerous infections, and help choosing the treatment.
(2) The system has to support variety of sequencing platforms, including Illumina HiSeq and MiSeq, IonPGM, and Oxford Nanopore's MinION. It has to be able to detect bacterial and fungal pathogens, and work with both amplicon-based and metagenomic approaches.
(3) The classifier system has to provide graphical report showing the composition of a sample in an easy to interpret way.

３．研究の方法

(1) Development of GenomeSync - a reference genome database
- ① GenomeSync is one of the largest databases of genome sequences. It is used for classifying metagenomic reads with high accuracy. GenomeSync is available at http://genomesync.org/.
- ② I developed a new highly effecient data compression method NAF. I now applied this method to compress the GenomeSync database. This increases the speed of transfering and accessing the genomes.

(2) Development of Genome Search Toolkit (GSTK)
- ① GSTK allows performing searches against the GenomeSync database using variety of homology search tools, including BLAST, BWA, LAST, Centrifuge, Minimap2.
- ② GSTK then performs taxonomic classification of the reads.
- ③ GSTK produces classification reports for each sample.
- ④ GSTK can work on computers ranging from laptops for field work to supercomputers for accurate analysis of large datasets.

４．研究成果

During this project I developed a computational system for analyzing medical metagenomic DNA datasets. With improved functionality, speed and accuracy this system has now been applied in several projects. In these projects system is used for analyzing actual medical samples in order to achieve rapid diagnostics of diseases.

The early version of our system was tested on a mock bacterial community and on pleural effusion samples from a patient with empyema. The improved system was applied to detecting bacterial pathogens in blood samples from sepsis patients. These studies are collaborations within Tokai University. Recently our system is also used at Kyoto University Hospital and Kansai Medical University for rapid identification of bacterial pathogens. A portable application of our system was successfully tested in the field application in Zambia. Currently the system is undergoing further tests and applications, and more papers are in preparation.

GenomeSync

I continued to develop and improve the genome database that is used as a reference in my classifier system. This database, called GenomeSync, currently contains 2,962 Gbp of sequence in 234,358 genomes, making it one of the largest reference genome resources. This database is specifically designed to facilitate metagenome analyses.

Genome Search Toolkit

The main component of our system is Genome Search Toolkit (GSTK) - a set of tools for performing analysis of metagenome datasets. This toolkit allows precise and rapid classification of metagenomic reads by comparing them with the reference genomes and taxonomy database. The comparison can be done using a wide selection of supported homology search tools, including BLAST, BWA, LAST, Centrifuge and Minimap2.

Genome Search Toolkit performs taxonomic classification of metagenomic sequences. Each sequence read is placed in a taxon, and then the number of reads for each taxon is counted. GSTK then visualizes this data in circular Krona charts.

Related projects

We applied my metagenome analysis pipeline to understanding viral sequences in eukaryotic genomes. We constructed pEVE - a database for predicted endogenous viral elements in eukaryotic genomes.

Another important outcome of this project was the development of a new compression method for DNA data. This method is now used in our GenomeSync database. The method was recently published online in Bioinformatics.

Conclusion

My system for investigating medical metagenomes has shown utility in research studies, field trips and tests on medical samples. The results have been described in peer-reviewed literature. We are preparing additional publications for this method. I expect that my classifier system will be widely adopted for investigating infectious diseases using next generation sequencing.

５．主な発表論文等

〔雑誌論文〕（計　5 件）

① Kryukov K, Ueda MT, Nakagawa S, Imanishi T, "Nucleotide Archival Format (NAF) enables efficient lossless reference-free compression of DNA sequences" Bioinformatics btz144, 2019, 査読有

② Tanaka H, Matsuo Y, Nakagawa S, Nishi K, Okamoto A, Kai S, Iwai T, Tabata Y, Tajima T, Satoh M, Kryukov K, Imanishi T, Hirota K, "Application of the MinION™ portable DNA sequencer in clinical microbiology evaluation: a case report." JA Clinical Reports, 5, 24, 2019, 査読有

③ Kai S, Matsuo Y, Nakagawa S, Kryukov K, Matsukawa S, Tanaka H, Iwai T, Imanishi T, Hirota K, "Rapid bacterial identification by direct PCR amplification of 16S rRNA genes using the MinIONTM nanopore sequencer", FEBS Open Bio., 9(3), 548-557, 2019, 査読有

④ Kryukov K, Ueda MT, Imanishi T, Nakagawa S, "Systematic survey of non-retroviral virus-like elements in eukaryotic genomes" Virus Research, 262, 30-36, 2019, 査読有

⑤ Watanabe N, Kryukov K, Nakagawa S, Takeuchi JS, Takeshita M, Kirimura Y, Mitsuhashi S, Ishihara T, Aoki H, Inokuchi S, Imanishi T, Inoue S, "Detection of pathogenic bacteria in the blood from sepsis patients using 16S rRNA gene amplicon sequencing analysis", PLoS ONE 13(8), e0202049, 2018, 査読有

〔学会発表〕（計　5 件）

① Kryukov K, Ueda MT, Nakagawa S, Imanishi T "NAF - data compression for next generation of genome databases" The 13-th Annual Meeting of the Japanese Society of Genome Microbiology, March 6-8, 2019, Hachioji, Japan.

② Kryukov K, Imanishi T "GenomeSync – an automatically synchronizing local database of genome sequences" The 41-st Annual Meeting of the Molecular Biology Society of Japan, November 28-30, 2018, Tokyo, Japan.

③ Kryukov K, Ueda MT, Imanishi T, Nakagawa S "Non-retroviral virus-like elements in eukaryotic genomes" Society of Molecular Biology and Evolution 2018 (SMBE 2018), July 8-12, 2018, Yokohama,

Japan.

④ Kryukov K "Phylogenetic trees explain mysteries of complete genomes" Neighbor-Joining Symposium, June 2, 2018, Mishima, Japan.

⑤ Kryukov K, Imanishi T "Classifying metagenomic reads using GenomeSync and Genome Search Toolkit" The 91st Annual Meeting of Japanese Society for Bacteriology, March 27-29, 2018, Fukuoka, Japan.

〔図書〕（計　　件）

〔産業財産権〕
○出願状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等

GenomeSync database: http://genomesync.org/

pEVE database: http://peve.med.u-tokai.ac.jp

NAF Compression: http://kirill-kryukov.com/study/naf/

NAF Compressor and Decompressor on GitHub: https://github.com/KirillKryukov/naf

６．研究組織

なし