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　５ 年　 ６ 月 ２１ 日現在

機関番号：６３８０１
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ０６６１２

研究課題名（和文）Improving efficiency of sequence databases by applying the NAF format

研究課題名（英文）Improving efficiency of sequence databases by applying the NAF format

研究代表者
クリュコフ　キリル（Kryukov, Kirill）

国立遺伝学研究所・Biological Networks Laboratory・特命准教授

研究者番号：２０８０６２０２

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：Short version, see English version for details. Achievements of this project: (1) Advancing the Nucleotide Archival Format. (2) Sequence Compression Benchmark. (3) Using NAF for the GenomeSync database. (4) Supporting NAF in bioinformatic tools. (5) Publishing 9 papers related to this project.

研究成果の学術的意義や社会的意義
Genome data is increasingly used across many fields of science. NAF greatly increases efficiency of working with such data compared to previous formats. This project applied, improved and advanced NAF towards becoming the fundamental infrastructure tool for the next generation of genome databases.

研究成果の概要（英文）：The achievements of this project: (1) Continued development, maintenance, and popularization of the Nucleotide Archival Format (NAF). Additions: Improved compression strength, improved customization of decompressed format, support for storing multiple files, added Bioconda installation option. (2) Evaluation of performance of various compressors in the Sequence Compression Benchmark - the most comprehensive benchmark of available compressors for biological sequence data. This benchmark clearly shows that NAF is a superior format for storing and working with sequence data. The benchmark paper has 25 Google Scholar citations. (3) Distributing NAF-compressed genome sequences via the GenomeSync database - one of the largest genome databases. Now GenomeSync offers convenient access to over 640,000 genomes, thanks to the efficiency of the NAF format. (4) Supported NAF in bioinformatic tools such as Genome Search Toolkit and Primer Tester. (5) 9 papers were published related to this project.

研究分野：Bioinformatics

キーワード：Data compression　NAF　GenomeSync

Report on the result of the Kakenhi Project titled "Improving efficiency of sequence databases by applying the NAF format"

Achievements of this project

1. **NAF format and tools.** This project enabled the continued development, maintenance, and popularization of the Nucleotide Archival Format (NAF) tools. Several new features have been added. (1) Improved compression strength, with the addition of the "--long" command line option to the *ennaf* compressor. (2) Added more flexibility to the decompressed output format with the new command line option "--sequences" for the *unnaf* decompressor. (3) Added Bioconda installation option, to make it easier to install NAF tools on user's computer. NAF compressor is being increasingly used in the field for compressing large sequence data, and the original NAF paper in Bioinformatics now has 33 Google Scholar citations.

2. **Sequence Compression Benchmark.** We evaluated the performance of various compressors and constructed Sequence Compression Benchmark - the most comprehensive benchmark of available compressors for biological sequence data. It currently compares 555 settings of 52 compressors (including 31 specialized sequence compressors), and uses 28 test datasets. The test data include entire genomes, from small to huge sizes, as well as single gene datasets, bacterial and mitochondrial datasets, protein databases, virus databases, and multiple

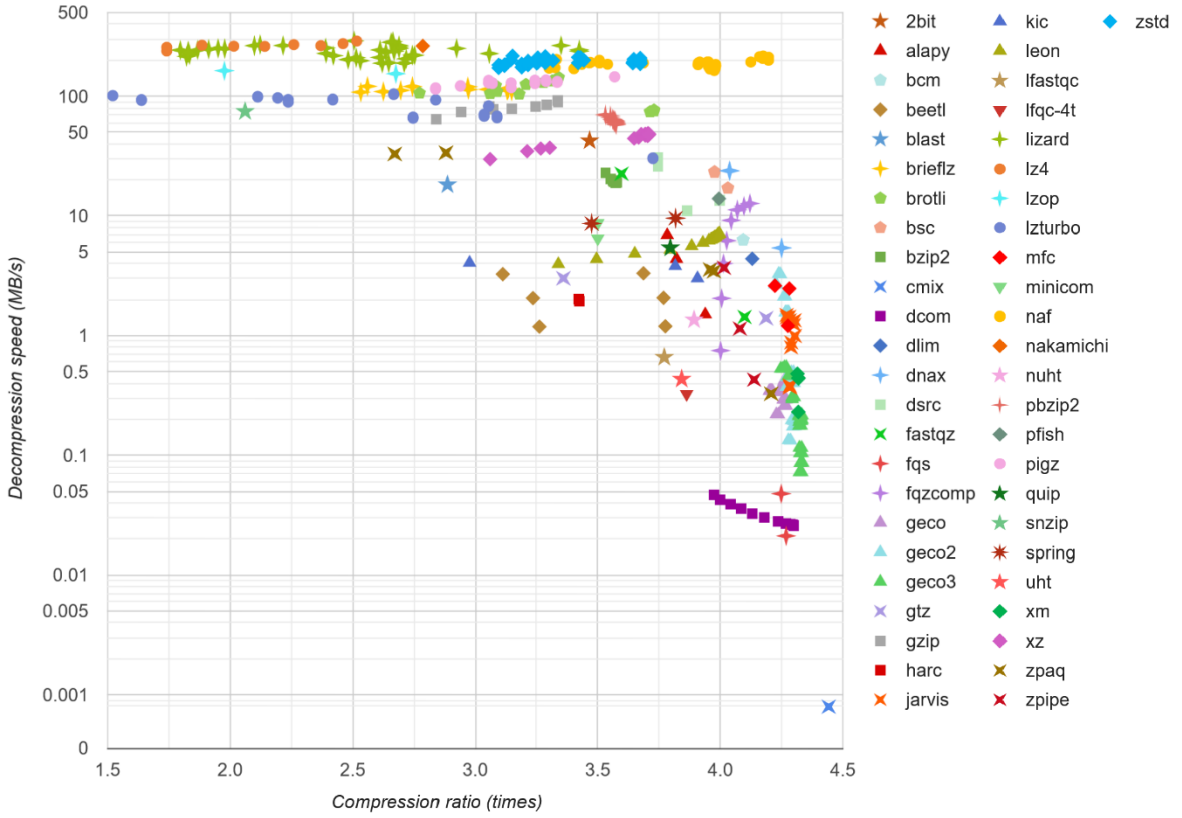

Fig. 1. A chart dynamically generated by the Sequence Compression Benchmark website. This chart shows compression strength (ratio) plotted against decompression speed, for various compressors. As a test data, a set of smaller genomes is used, to be able to include even compressors that can't work with large data.

sequence alignments. With this diverse data, the benchmark clearly shows that NAF is a superior format for storing and working with sequence data, especially when considering a combination of metrics such as Compression Strength and Decompression Speed - the two most important parameters of a sequence compressor. Importantly, the benchmark website (http://kirr.dyndns.org/sequence-compression-benchmark/) provides a dynamic interface for constructing custom comparisons based on particular user-selected combination of compressors, test data, and performance measures (Fig. 1). The benchmark paper in GigaScience already has 25 Google Scholar citations.

3. **GenomeSync.** Our GenomeSync database (https://genomesync.org/) distributes the available public genome data, compressed into the NAF format. Now GenomeSync offers convenient access to over 640,000 genomes, or 10.4 TB of data. Thanks to the efficiency of the NAF format, in compressed form this data occupies just 2.2 TB. GenomeSync more than doubled in the size of stored data over the course of this project (Fig. 2). GenomeSync provides an important example of usefulness of NAF compression for storing large data.

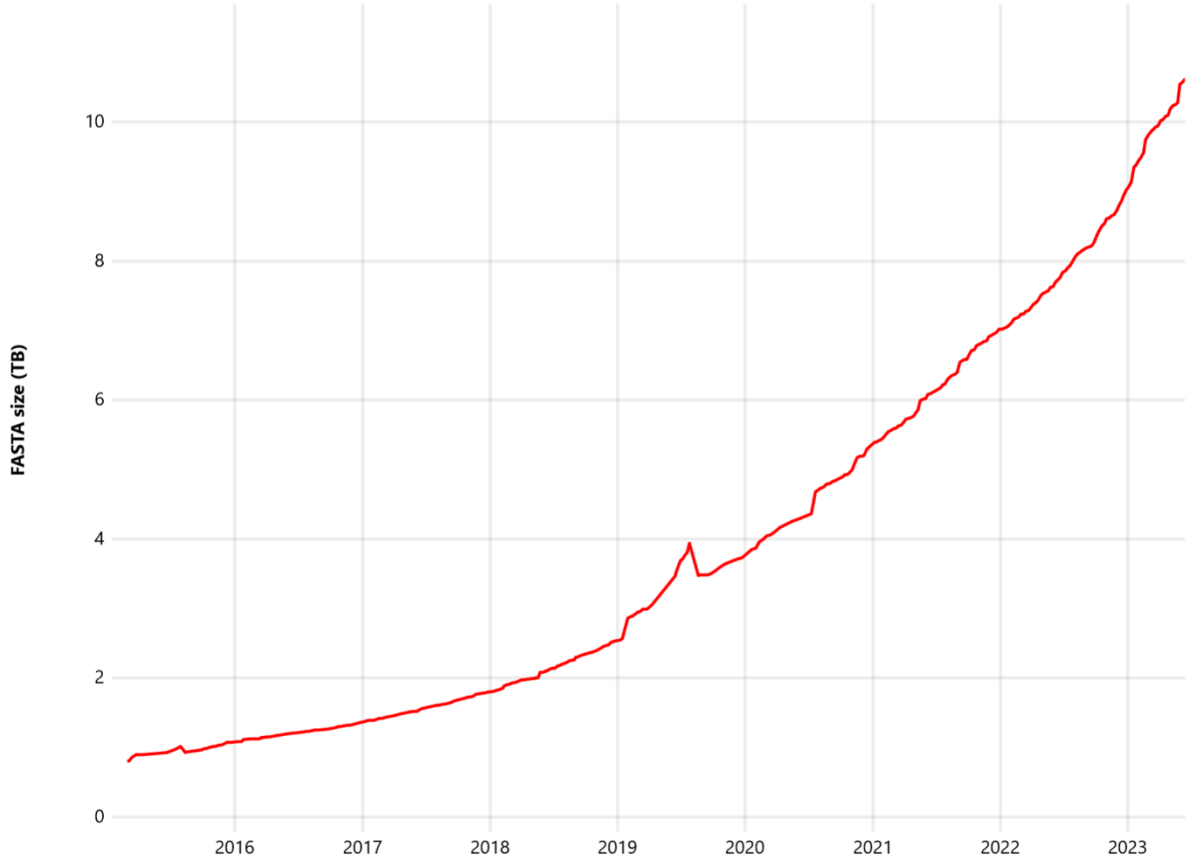

Fig. 2. Changes in the size of data stored in GenomeSync over time. This chart shows the uncompressed size of data, the compressed size is much smaller thanks to the NAF compression.

4. **Other tools using NAF.** NAF is now supported in two bioinformatic tools: Genome Search Toolkit and Primer Tester. Genome Search Toolkit is a software package for analyzing metagenomic datasets based on a comparison with genome database. Primer Tester is a tool for scanning genome data for matches to a particular set of primer sequences.

5. **SARS-CoV-2 genome compression.** We evaluated NAF compression for improving the efficiency of distributing SARS-CoV-2 genome data. Millions of SARS-CoV-2 genomes have been sequenced and deposited to databases. This massive data is being downloaded and used by research groups around the world for rapid response to the changing situation with the emergence and spread of new variants of the virus. We have successfully demonstrated the massive superiority of the NAF format for storing and distributing SARS-CoV-2 genomes. Compared to the solutions currently used by sequence databases, distributing SARS-CoV-2 genomes in NAF format would provides an increase in efficiency ranging from 3.7 to 52.2 times. We recently published these findings in Patterns.

6. **Publications.** In total 9 papers related to this project have been published, including journals: GigaScience, Patterns, BMC Microbiology, Scientific Reports, Infectious Diseases, BMC Medical Genomics.

Summary

Genome data is increasingly used across many fields of science. The large genome databases accumulate enormous amounts of sequence data, but it's often challenging to obtain and efficiently utilize this massive data. NAF greatly increases efficiency of working with sequence data, compared to previous formats. In this project we applied, improved and advanced NAF, in both core functions and in its applications, towards becoming the fundamental infrastructure tool for the next generation of genome databases. Thanks to this project, now NAF is a mature technology, ready to be used for accelerating genome data analysis in scientific, medical, and industrial projects.

５．主な発表論文等

〔雑誌論文〕　計9件（うち査読付論文　6件／うち国際共著　7件／うちオープンアクセス　5件）

１．著者名 Kirill Kryukov, Lihua Jin, So Nakagawa	４．巻 3
２．論文標題 Efficient compression of SARS-CoV-2 genome data using Nucleotide Archival Format (NAF)	５．発行年 2022年
３．雑誌名 Patterns	６．最初と最後の頁 100562
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.patter.2022.100562	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Kirill Kryukov, Tadashi Imanishi, So Nakagawa	４．巻 2632
２．論文標題 Nanopore sequencing data analysis of 16S rRNA genes using GenomeSync-GSTK system	５．発行年 2023年
３．雑誌名 Methods in Molecular Biology	６．最初と最後の頁 215-226
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-1-0716-2996-3_15	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Kirill Kryukov, So Nakagawa, Yoshiyuki Matsuo, Kiichi Hirota, Tadashi Imanishi	４．巻 Dec 2021
２．論文標題 Metagenomic analysis of bacterial 16S rRNA sequences	５．発行年 2021年
３．雑誌名 Experimental Medicine	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Shinnosuke Komiya, Yoshiyuki Matsuo, So Nakagawa, Yoshiharu Morimoto, Kirill Kryukov, Hidetaka Okada, Kiichi Hirota	４．巻 15
２．論文標題 MinION, a portable long-read sequencer, enables rapid vaginal microbiota analysis in a clinical setting	５．発行年 2022年
３．雑誌名 BMC Medical Genomics	６．最初と最後の頁 68
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s12920-022-01218-8	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 Kirill Kryukov, Mahoko Takahashi Ueda, So Nakagawa, Tadashi Imanishi	4．巻 9
2．論文標題 Sequence Compression Benchmark (SCB) database - a comprehensive evaluation of reference-free compressors for FASTA-formatted sequences	5．発行年 2020年
3．雑誌名 GigaScience	6．最初と最後の頁 giaa072
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/gigascience/giaa072	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 Yoshiyuki Matsuo, Shinnosuke Komiya, Yoshiaki Yasumizu, Yuki Yasuoka, Katsura Mizushima, Tomohisa Takagi, Kirill Kryukov, Tadashi Imanishi, Aisaku Fukuda, Yoshiharu Morimoto, Yuji Naito, Hidetaka Okada, Hidemasa Bono, So Nakagawa, Kiichi Hirota	4．巻 21
2．論文標題 Full-length 16S rRNA gene amplicon analysis of human gut microbiota using MinION nanopore sequencing confers species-level resolution	5．発行年 2021年
3．雑誌名 BMC Microbiology	6．最初と最後の頁 35
掲載論文のDOI（デジタルオブジェクト識別子） 10.1186/s12866-021-02094-5	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 Ayumu Ohno, Kazuo Umezawa, Satomi Asai, Kirill Kryukov, So Nakagawa, Hayato Miyachi, Tadashi Imanishi	4．巻 11
2．論文標題 Rapid profiling of drug-resistant bacteria using DNA-binding dyes and a nanopore-based DNA sequencer	5．発行年 2021年
3．雑誌名 Scientific Reports	6．最初と最後の頁 3436
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41598-021-82903-z	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 Yoshiki Shiraishi, Kirill Kryukov, Katsuyoshi Tomomatsu, Fumio Sakamaki, Shigeaki Inoue, So Nakagawa, Tadashi Imanishi, Koichiro Asano	4．巻 53
2．論文標題 Diagnosis of pleural empyema/parapneumonic effusion by next-generation sequencing	5．発行年 2021年
3．雑誌名 Infectious Diseases	6．最初と最後の頁 450-459
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/23744235.2021.1892178	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 大野　歩，　中川　草，　Kirill Kryukov，　今西　規	4．巻 49
2．論文標題 ナノポアDNAシークエンサーを用いた 迅速な細菌同定法	5．発行年 2020年
3．雑誌名 臨床化学	6．最初と最後の頁 265-270
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計3件（うち招待講演　3件／うち国際学会　0件）

1．発表者名 Kirill Kryukov
2．発表標題 GenomeSync: streamlining access to current genome data
3．学会等名 Genome Concept Centennial Conference（招待講演）
4．発表年 2021年

1．発表者名 Kirill Kryukov
2．発表標題 Sequence Data Compression. History, methods, best practices, perspectives.
3．学会等名 EvoGen Reading Club（招待講演）
4．発表年 2021年

1．発表者名 Kirill Kryukov
2．発表標題 GenomeSync and GSTK: Toolkit for precision analysis of medical metagenome sequence data
3．学会等名 Genome Analysis for Precision Medicine of Infectious Diseases, 2021（招待講演）
4．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

NAF on GitHub
https://github.com/KirillKryukov/naf
GenomeSync
https://genomesync.org/
高速かつ高効率にシーケンスデータを圧縮 / 解凍する NAF
http://kazumaxneo.hatenablog.com/entry/2019/03/07/073000

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関