bio-twobit: Fast 2bit file reader in Ruby

kojix2

14 Jan 2022

Summary

bio-twobit is a fast 2bit file reader for ruby.

Code : https://github.com/ruby-on-bioc/bio-twobit

Statement of need

It is common to want to know the sequence of a specific position in the genome. 2bit file format is an efficient genome file format provided by the UCSC Genome Browser (Navarro Gonzalez et al. 2021). By using this file format, you can quickly access specific regions of the reference genome.

2-bit files can be read and written in R and Python. However, there are limited ways to access 2bit files from the Ruby language.

bio-ucsc-api (Mishima et al. 2012), one of the gems of BioRuby (Goto et al. 2010), has methods to read 2-bit files. But they are implemented in Ruby, so they are not fast enough.

Here we present bio-twobit. This is a binding to lib2bit (Ryan 2017), a library implemented in the C, and has a Python interface API called py2bit (Ryan 2018). bio-twobit speeds up access to the reference genome using the Ruby language, making it easier to use in web applications.

Benchmark

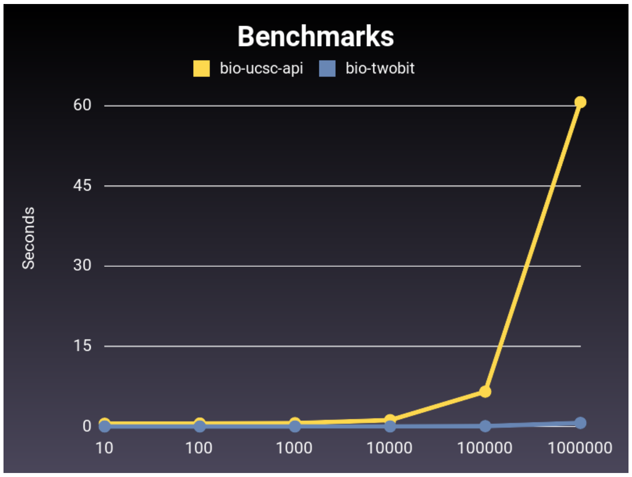

Figure 1: benchmark

The x-axis is the number of times "chr17:7579614-7579700" is called. Code used for benchmarking

Examples

Setup:

```ruby
require 'bio/twobit'
```

Load a 2bit file:

```ruby
hg38 = Bio::TwoBit.open("BSgenome.Hsapiens.UCSC.hg38/inst/extdata/single_sequences.2bit")
```

Print file information:

```ruby
hg38.info
# {"file_size"=>818064875,
# "nChroms"=>640,
# "sequence_length"=>3272116950,
# "hard_masked_length"=>161368694,
# "soft_masked_length"=>0}
```

Fetch a sequence:

```ruby
hg38.sequence("chr1", 50000, 50050)
# "AAACAGGTTAATCGCCACGACATAGTAGTATTTAGAGTTACTAGTAAGCC"
```

Fetch per-base statistics:

```ruby
hg38.bases("chr1", 10000, 10100)
# {"A"=>0.34, "C"=>0.49, "T"=>0.17, "G"=>0.0}

hg38.bases("chr1", 10000, 10100, fraction: false)
# {"A"=>34, "C"=>49, "T"=>17, "G"=>0}

hg38.bases("chr1")
# {"A"=>0.26940569141052323,
# "C"=>0.19302592242428676,
# "T"=>0.2701041550155312,
# "G"=>0.19325280952182064}
```

Fetch masked blocks

```ruby
hg38.hard_masked_blocks("chr1", 0, 1000000)
# [[0, 10000], [207666, 257666], [297968, 347968], [535988, 585988]]
```

Reference

Goto, Naohisa, Pjotr Prins, Mitsuteru Nakao, Raoul Bonnal, Jan Aerts, and Toshiaki Katayama. 2010. "BioRuby: Bioinformatics Software for the Ruby Programming Language." *Bioinformatics* 26 (20): 2617–19. https://doi.org/10.1093/bioinformatics/btq475.

Mishima, Hiroyuki, Jan Aerts, Toshiaki Katayama, Raoul J. P. Bonnal, and Koh-ichiro Yoshiura. 2012. "The Ruby UCSC API: Accessing the UCSC Genome Database Using Ruby." *BMC Bioinformatics* 13 (1): 240. https://doi.org/10.1186/1471-2105-13-240.

Navarro Gonzalez, Jairo, Ann S Zweig, Matthew L Speir, Daniel Schmelter, Kate R Rosenbloom, Brian J Raney, Conner C Powell, et al. 2021. "The UCSC Genome Browser Database: 2021 Update." *Nucleic Acids Research* 49 (D1): D1046–57. https://doi.org/10.1093/nar/gkaa1070.

Ryan, Devon. 2017. "Lib2bit."

———. 2018. "Py2bit." The deepTools ecosystem.