ruby-libssw

kojix2

14 July 2022

Summary

Ruby-libssw is the Ruby binding of libssw, a library that uses the Smith-Waterman algorithm to find the best pairwise alignment of two sequences. ruby-libssw was created using fiddle, the Ruby standard library. Ruby-libssw can be used to create local alignments of nucleotide and amino acid sequences in the Ruby language.

Code : https://github.com/kojix2/ruby-libssw

Statement of need

Sequence alignment effectively investigates the relationships between DNA, RNA, and amino acid sequences. libssw ("SSW Library: An SIMD Smith-Waterman C/C++ Library for Use in Genomic Applications | PLOS ONE" n.d.) performs local alignment of base sequences (DNA, RNA) and amino acid sequences (proteins). libssw uses Smith-Waterman's algorithm. The Smith-Waterman algorithm is accurate but takes a long time to calculate. libssw uses SIMD (Single-Instruction Multiple-Data) to perform parallel operations at the processor to increase speed. libssw is often run within an application for genomic analysis. libssw has wrappers for the C ++, Python, Java, and R languages. But until now, there was no wrapper for the Ruby language. So I created Ruby-libssw.

Benchmark

TODO: https://gist.github.com/ktym/7a4799fb055436dc2139308dcab802f0

Examples

```ruby
require 'libssw'

ref_str  = "AAAAAAAAACGTTAAAAAAAAAA"
ref_int  = SSW::DNA.to_int_array(ref_str)
# [0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 3, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

read_str1 = "ACGTT"
read_str2 = SSW::DNA.revcomp(read_str1)
# "AACGT"
read_int1 = SSW::DNA.to_int_array(read_str1)
# [0, 1, 2, 3, 3]
read_int2 = SSW::DNA.to_int_array(read_str2)
# [0, 0, 1, 2, 3]

mat = SSW.create_scoring_matrix(SSW::DNA::Elements, 2, -2)
# mat = [2, -2, -2, -2,  0,
```

```
#       -2,  2, -2, -2,  0,
#       -2, -2,  2, -2,  0,
#       -2, -2, -2,  2,  0,
#        0,  0,  0,  0,  0]

profile1 = SSW.init(read_int1, mat)
align1   = SSW.align(profile1, ref_int, 3, 1, 1, 0, 0)
pp align1.to_h
# {
#  :score1       => 10,
#  :score2       => 0,
#  :ref_begin1   => 8,
#  :ref_end1     => 12,
#  :read_begin1  => 0,
#  :read_end1    => 4,
#  :ref_end2     => 0,
#  :cigar        => [80],
#  :cigar_len    => 1,
#  :cigar_string => "5M"
# }

profile2 = SSW.init(read_int2, mat)
align2   = SSW.align(profile2, ref_int, 3, 1, 1, 0, 0)
pp align2.to_h
# {
#  :score1       => 10,
#  :score2       => 0,
#  :ref_begin1   => 7,
#  :ref_end1     => 11,
#  :read_begin1  => 0,
#  :read_end1    => 4,
#  :ref_end2     => 0,
#  :cigar        => [80],
#  :cigar_len    => 1,
#  :cigar_string => "5M"
# }

puts SSW.build_path(read_str1, ref_str, align1)
# 5M
# ACGTT
# |||||
# ACGTT
```

Reference

"SSW Library: An SIMD Smith-Waterman C/C++ Library for Use in Genomic Applications | PLOS ONE." n.d. Accessed May 24, 2022. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0082138.