**https://github.com/ZeinabRabea/SuffixAligner**

**SuffixAligner**

SuffixAligner is a python-based aligner for long noisy reads generated from third-generation sequencing machines. SuffixAligner exploits the nature of biological alphabet that has a fixed-size and a predefined lexical ordering to construct a suffix array for indexing a reference genome. FM-index is used to efficiently search the indexed reference and locate the exact matched seeds among reads and the reference. The matched seeds are arranged into windows/clusters and the ones with the maximum number of seeds are reported as candidates for mapping positions, more details about SuffixAligner can be found in:

* (Indexing) <https://doi.org/10.1016/j.jksuci.2022.04.015>
* (Mapping)

#### System requirements

64-bit machine with python in general, numpy, and pandas libraries.

#### ~~Installation~~

1. ~~Downlowd files~~
2. ~~Run code f5~~

#### Quick usage guide

**Indexing:**

Run

Find\_SA\_for\_Genome(L=100,l=100,G\_file='genome\_file.fasta')

Or Run

Find\_SA\_for\_Genome()

\*[G\_file] Reference Genome File [default:example1.fasta]

\*[L] Divide genome to substrings of length L [default:100]

\*[l] Overlap length [default:100]

output

[G\_file].SA.txt the suffix array in string format

**Mapping:**

Run

Mapping(Type="r",G\_file="example1.fasta",

R\_file="Read1.fastq",

SA\_file="example1.fasta.SA.txt",

Start=0,End=10)

Or run

Mapping(Type="s",G\_file=" example1.fasta",

SA\_file="example1.fasta.SA.txt",

Sam\_file="example1.bwa.sam",

Start=0,End=10)

Or Run

Mapping()

\*[Type] "r" for read, "s" for sam file [default:r]

\*[G\_file] Reference Genome File [default:example1.fasta]

\*[R\_file] Read File [default:example1.fasta]

\*[SA\_file] Suffix array File [default:example1.fasta]

\*[Start] read number which Start from [default:0]

\*[End] number of read which end with [default:10]

\*[Sam\_file] Sam file which generated from other aligner

#### ~~Notes~~

* ~~If the gap size parameter is missing, LightAssembler invokes its parameters extrapolation module to compute the starting gap based on the sequencing coverage and the error rate of the dataset.~~
* ~~The maximum read length for this version is 1024 bp.~~
* ~~The maximum supported read files for this version is 100 files.~~

#### Read files

SuffixAligner align sequencing reads given in ***fastq*** format to reference genome given in ***fasta*** format. Also, SuffixAligner can read directly the sam files given in ***sam*** format which genertated from other aligner and search for solution for unmapped read.

#### Outputs

The output of SuffixAligner is suffix array in text format from the step of indexing in the file:

[G\_file].SA.txt

The output of the step of mapping is sam file in the file:

[R\_file].[Start].[End].sam when input is read file

Or

[Sam\_file].[Start].[End].sam when input is sam file

SuffixAligner also reports the following on the screen:

* Genome length
* Number of read
* Number of read in sam file
* Number of unmapped read in sam file