1 Data representation

1.1 sequence file

A sequence file is a text file of l lines, each line have only one integer, represent the character(chromatin state) of this site. Usually the file name ends with ".seq".

1.2 compressed sequence file

A compressed sequence file is a compressed representation of a sequence file. It is a k line text file, in which each line has two integers: "(character) (repeated time)".

Usually the file name ends with ".sseq".

1.3 sequence database

A sequence database is composed of n compressed sequence files and one index file. The compressed sequence files should locate in the same folder.

The index file stores the paths to the compressed sequence files, each file name takes one line.

A sequence database could be a set of genome regions, an epigenome, a mixture of several epigenomes, or whatever you can come up with.

Usually an index file has no suffix.

2 C codes

Matching score and attention score are defined in "/Ccode/custom/CustomFunction.h(.c)". A string "para" is passed into these functions. "para" could be a value, a path to a file, or anything.

Here we provide 3 types of attention scores in our package.

2.1 folder structure

There are 6 useful folders here. They are "/Ccode/bin", "/Ccode/main", "/Ccode/include", "/Ccode/custom", "/Ccode/Alignment", "/Ccode/DataProcessing".

"/Ccode/bin" contains executable files and bash scripts. "/Ccode/main" contains interface codes. The other folders have some basic library functions.

2.2 compile

The script "compile.bash" compiles the C codes into executable files to folder "/Ccode/bin". Before running this script, all codes in the other folders should have a link in folder "/Ccode/main".

"FakeChromosomeGenerator.cpp" uses random generator specified by c++11, hence the compiler should support c++11. In linux, g++ version 4.8 or higher supports c++11, remember to substitute "g++-5" in "compile.bash" to your compiler.

2.3 bash scripts and executable files

2.3.1 TowRegions.out

Usage: ./TwoRegions.out seqfile1 seqfile2 para

This program implement smith-waterman algorithm to compare two epigenetic state sequences. "seqfile1" and "seqfile2" are two compressed sequence files.

The output includes the best local match regions and the matching score.

2.3.2 DatabaseSearch.out

Usage: ./DatabaseSearch.out Paths_Search Para_Search Para_align

Do search of a query sequence set from a database. Both query sequence set and the database are stored as "sequence database". There are examples of "Paths_Search", "Para_Search" under "/Ccode/bin". "Para_align" is the same as the "para" described at the beginning of this section

2.3.3 DatabaseSearch_baseline.out

Usage: ./DatabaseSearch_baseline.out Path_Search_baseline Para_Search_baseline

2.3.4 FakeChromosomeGenerator.out

Usage: ./FakeChromosomeGenerator.out chromosome fake_chromosome

This program generate fake chromosome by Markov rule. Fake chromosome has the same length to the original. The state transition probability matrix to generate fake chromosome is computed from the original chromosome.

"chromosome" is a compressed sequence file(of epigenetic states), "fake_chromosome" is the file to output.

2.3.5 cut_sseq.out

Usage: ./cut_sseq.out Para_Cut

Cut a long sequence into regions, each region stored in a file.

2.3.6 CutFolder_Init.sh

2.3.7 FakeGenomeGenerator.sh

Usage: ./FakeGenomeGenerator.sh index genomefolder newindex newgenomefolder

A genome has more than one chromosomes, this script iteratively run FakeChromosomeGenerator.out on the chromosomes. The genome is a sequence database structure.

2.3.8 GenomeSearch_Path.sh

Usage: ./GenomeSearch_Path.sh Para_GenomeSearch PathToThisScript

Automatically cut query genome into regions, and do search for each query region from another genome.