Introduction

Guided by a phylogenetic tree, the simulation program SIMALI starts with an ancestral sequence (root sequence), and simulates evolutionary processes to produce current sequences. Besides mutations such as substitution, deletion, and insertion, we simulate inversions. Transposition is possible to occur, since overlapped inversions result in transpositions. Duplication is not simulated at this stage.

The simulation program records the true relationship among the generated sequences (i.e. orthologous segments), which can be regarded as the true alignment. Based on that, the extent of agreement for alignments produced by aligners could be determined.

Installation

1. Get the gzipped tar file of source code;
2. Decompress the files in a clean directory;
3. Type `make` to build the program.

Usage

- `simali simali.para`
  `simali.para` contains parameters that will be used in the program.
- `simali simali.para N=10`
  Optional argument N is the number of dataset you need.
  Default value is 1.
- `simali --help`
  Display help message and quit the program.

Input

The guiding tree in the program includes nine mammalian species:

```
(((human chimp) baboon)(rat mouse))((cow pig)(cat dog)))
```

After decompressing the tar file, there will be a parameter file `simali.para` in the directory together with source code. `simali.para` contains parameters that will be used in the program.

- `TheAlphabet`
  "ACGT" for DNA sequences, i.e. TheAlphabet = ACGT
- `SequenceLen`
  length of root sequence, e.g. SequenceLen = 20000
- `TheDNAmodel = HKY`
  name of the model. The program only handles HKY substitution model in this version!
- `TransitionBias`
  Ts/Tv ratio, e.g. TransitionBias = 2.0
- `TheFreq`
  frequency of each character in alphabet, e.g. TheFreq = [0.3,0.2,0.2,0.3]
- `TheTree`
  the phylogenetic tree used in the program. The tree should be in NEWICK format, with evolution distance (edge length).
- `TheInsFunc, TheDelFunc, TheInvFunc`
  length distribution functions of insertion, deletion, and inversion.
- `TheInsertThreshold, TheDeleteThreshold, TheInvertThreshold`
  probabilities of the occurence of insertion, deletion, and inversion.

- `InsertionMultiplier, DeletionMultiplier, InversionMultiplier`
  multipliers of insertion, deletion, and inversion associated to each node in the tree.

- `MinInvLength`
  minimum length of inversion.

Information followed the above parameters is the interspersed repeats families.

Output

The output of the program includes simulated sequences in FASTA format and a true alignment in Jim Kent's MAF format. Each genereated sub-directory `datasetXXX` contains one dataset.