TIME_ESTIMATE: GUIDE

VLADIMIR SHCHUR

1. Installing

The `./time_estimate` is available as a part of the `argentum` project at `https://github.com/nvalimak/argentum`.

2. Brief overview

A proper command line is not implemented yet. The following arguments should be given even if they are not used.

```
./time_estimate [rangeMode] [initMethod] [max_iter] [exCycles] [counter] [output_mode]
```

`./time_estimate` reads ARG in enumerate format (see below) from the standard input. The enumerate format is one of the output formats for the `./main`.

There are three main running phases of `time_estimate`. The first phase initialising ARG times. The second phase improves time estimation through iterative process. The third phase outputs a summary.

Example:

```
gunzip -c data.txt.gz | ./time_estimate pop_map.txt 1 1 50 0 1000 0
```

3. Input format (tab delimited)

First line is of the form

```
ARGraph 6000 1317668 1311669
```

where the first number is the number of leaves, the second number is the maximal node id and the last number is the total number of nodes in the graph. In this example we have an ARG with 6000 leaves and 1311669 nodes (including leaves).

Each node is described in the following form

```
parent 16602 2 1
child 5528 4062 5235 337569 399302 1
child 1347 4062 5235 337569 399302 0
mutation 1347 4256 348494
```

Here the first line states that the node with id 16602 is described and this node has two child nodes and one mutation entry.

A line describing a child (in fact an oriented edge from a parent to a child)

```
child 5528 4062 5235 337569 399302 1
```

means: child node id is 5528, the edge spans positions $[4062, 5235]$ in SNP coordinates and $[337569, 399302]$ in base-pairs. 1 means that a recombination was inferred on the edge, 0 - otherwise.

A line describing mutation

```
mutation 1347 4256 348494
```

means that a mutation occurred on the edge connecting the (parent) node to the child with id 1347. Mutation coordinates are 4256 (in SNPs) and 348494 (in base-pairs).

4. Global parameters

`[rangeMode]` sets which coordinates (SNPs or BPs) will be used for the computations.
`[counter]` sets the progress output step for iterative operations.

5. Time initialisation and refinement

`[initMethod]` the method to initialize node times.

- `1` assigning times based on expectations of lengths of child edges;
- `2` assigning times in order to keep the condition time(parent) > time(child) (set [exCycles] = 1 in this case).

`[exCycles]` controls if cycles should be deleted (`1`) or not (`0`) from the ARG topology.
`[max_iter]` number of iterations for time update.

6. Output modes

`[output_mode]`

- `0` no output (may be useful for debug)
- `1, 2` getting coalescent times between two populations. (2 is much faster than 1)
- `3` getting ARG slice and outputting information for FastModularity and graph clustering scripts.
- `4` painting haplotypes based on clustering from FastModularity.
- `5` getting nodes from the slice in a certain time period and getting leaf distribution under them.
- `6` local tree likelihood.

7. Output modes 1 and 2

Mode `1` is rather slow as it converts ARG back to local tree. Mode `2` is much faster and uses ARG shared structure. In both modes, the following arguments should be added to the command line:

```
[pops_map.txt] [pop1] [pop2]
```

`[population_map.txt]` is the tab delimited file which maps a haplotype to a population in the format

```
51 1
934 1
1274 2
```

which means that haplotypes 51 and 934 are from the first population and haplotype 1274 is from the second population. The file does not necessary contain information about all nodes.

`[pop1]` `[pop2]` - IDs of populations to be compared (should be consistent with `[population_map.txt]` file).

8. Output modes 3, 4 and 5

The following parameters should be added to the command

```
[slice_left] [slice_right] [min_time] [max_time]
```

These parameters define a slice (a part) of the ARG: genomic region (`[slice_left]` and `[slice_right]`) and time period ((`[min_time]` and (`[max_time]`)

In mode `5` `[population_map.txt]` is not supported yet. It is assumed that the number of haplotypes is equal in each population.

VLSHCHUR@GMAIL.COM