Prof. Dr. Daniel Huson
Zentrum für Bioinformatik
Fachbereich Informatik
Mathematisch-Naturwissenschaftliche Fakultät

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Grundlagen der Bioinformatik   SoSe 2018

Assignment 08   Submit electronically in Ilias by 25.6.2018, 10h

1   Implementation of a simple Wright-Fisher simulator (6 points)

Using the file `PopGenSimulator.java` found on the course website, implement a population genetics simulator based on the Wright-Fischer model. The simulator should represent the extant (present-day) population of $2N$ individuals as strings `A01`, `A02`, `A03,...`. In addition, the program is given a parameter k that determines the size of a *sample* within the population.

The simulator operates backward in time, for each current individual choosing its parent based on the coalescent model (as discussed in the lecture).

Note that a parent node can have more than one child. A parent is given the smallest label of any of its labeled children. If it doesn't have any labeled child, then give it label `---`.

After each iteration of your simulation, determine how many different ancestors of the *sample* (first k individuals of present-day population) currently exist. At the beginning of the simulation, this number will be k. During the simulation, it will decrease each time two ancestors of the sample choose the same parent. Once this number hits 1, you have found the MRCA of the initial k individuals.

Each generation is output as a *generation-number* 0, -1, -2 etc followed by the list of *individuals* and then *number of lineages* remaining for the sample of k.

Example output for $k = 4$ and $2N = 8$:

```
 0 A01 A02 A03 A04 A05 A06 A07 A08  4
-1 A04 --- A01 A02 A03 A05 --- A06  4
-2 --- --- A01 --- --- --- A05 A03  2
-3 --- --- --- A01 A03 --- A05 ---  2
-4 A05 --- --- --- --- --- --- A01  1
```

2   Time to MRCA of different sample sizes (4 points)

Using your program, plot the time for the set of first k individuals to find their MRCA as a function of population size. Using an interesting range of values for k (four different values) and $2N$ (four different values).