Integrative Transcriptomics
Prof. K. Nieselt,
Institute for Bioinformatics and Medical Informatics Tübingen
Prof. S. Nahnsen,
Institute for Bioinformatics and Medical Informatics Tübingen

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Lecture: Grundlagen der Bioinformatik — SoSe 2022

Assignment 6 (20 points)

Hand out: Thursday, June 2
Hand in due (**two weeks!**) : Thursday, June 16, 18:00
Direct inquiries via the ILIAS forum or to your respective tutor at:
Mathias Witte Paz: iizwi01@uni-tuebingen.de
theresa-anisja.harbig@uni-tuebingen.de
meret.haeusler@student.uni-tuebingen.de
jules.kreuer@student.uni-tuebingen.de
simon.heumos@qbic.uni-tuebingen.de

Prepare for the exam

During the following two weeks we also leave you time to prepare for the midterm exam. For this we have prepared some preparatory exam questions, you find these on ILIAS (in the folder 'Additional Material'). Please note that the answers to the questions shall not be handed in, we will also not hand out model answers.

Theoretical Assignments

1. **Most recent common ancestor (MRCA)** (1P)

   On p.120 we computed for an example, under the assumptions of the Wright Fisher model, the size of the gene pool $t = 15$ generations ago of a population of size $2N = 10.000$. Derive a general formula to compute t for the MRCA. Compute t for $2N = 10.000$.

2. **The coalescence rate** (4P)

   (a) What is the coalescence rate for a sample of 6 (and population size 2N)? What is the expected time you have to wait to go from 6 to 5 lineages? And from 5 to 4, 4 to 3, 3 to 2 and 2 to 1? Draw the coalescent tree (with correct branch lengths) for a sample of 6, using the expected waiting times (in N generation units). (Hint: p. 125).

   (b) Similarly to the formula we derived for the expected total height of a coalescent tree $E(T_{\text{MRCA}})$ (s. p. 118/9 in script), derive a closed formula for the expected total length, $E(T_{\text{total}})$ of all the branches in the genealogy.

3. **Conclusions from Population genetics of Humans** (5P)

   Describe and discuss why and how population genetics of humans has led scientists to the conviction that the word 'race' in the context of humans should be banned. You may want to also read the Jena declaration in this context (Altogether approx. 250-300 words).

Practical Assignments

To solve the following task, please download the file `materialA06.zip` from Ilias.

4. Implementation of a simple Wright-Fisher genealogy simulator (10P)

Using the file `PopGenSimulator.py` found in the material folder, implement a very simple Wright-Fischer genealogy simulator. This simulator represents the initial gene population of $size = 2N$ individuals as different letters a b c ....

The simulator operates backward in time, for each current individual choosing its parent as discussed in the lecture. The waiting times between coalescence events are exponentially distributed. To simulate the waiting times, you may want to make use of the following formula:

$$T_k = -\binom{k}{2}^{-1} \ln(U_k), k = n, \ldots, 2$$

where U_k are independent uniform random variables on [0, 1].

A parent is given the smallest label of any of its children. If it doesn't have any children, then give it the label '-'.

Each generation is printed as generation number 0, -1, -2 etc., followed by the string of individuals. (Careful: in your program when using the formula for the continuous time, random waiting times might become shorter than a generation, you may need to pay attention in your implementation).

Terminate when the MCRA of all initial individuals has been found (what is the condition for that?)

Example output:

```
 0 abcdefgh
-1 ab-d-fg-
-2 a-f---gd
-3 g-f--a--
-4 --f---a-
-5 ---a----
```

(a) Run your implementation three times for each of the sizes $2N = 4, 6, 10$.

(b) Using your program, plot the time for all initial individuals to find their MRCA as a function of population size, for the sizes $4, 6, 10$ (median of 3 runs per size) and also plot the value given by the theory.
Hand in the code and the results of all runs (i.e., 9 runs).