Grundlagen der Bioinformatik

SoSe 2022

Tutor: Theresa/ Mathias

1	2	3	4	$\sum$

Marina Dittschar & Clarissa Auckenthaler

Blatt 6

(Abgabe am 16.06.2022)

Theoretical Assignments

Task 1: Most recent common ancestor (MRCA) (1)

The number of ancestral genes at the time point that the MRCA is found is one.
Therefore, we can insert 1 for n in the formula and reformulate the equation for t:

$$n = 2N \cdot (1 - e^{-1})^t \tag{1}$$

$$1 = 2N \cdot (1 - e^{-1})^t \tag{2}$$

$$\frac{1}{2N} = (1 - e^{-1})^t \tag{3}$$

$$t = log_{1-e^{-1}}(\frac{1}{2N}) \tag{4}$$

Now, we can calculate the number of generations at the point of thee most recent ancestor by inserting 2N = 10000:

$$\begin{aligned} t &= log_{1-e^{-1}}(\frac{1}{2N}) \\ t &= log_{0.63}\frac{1}{2N} \\ t &= log_{0.63}\frac{1}{10000} \\ t &= 19.934 \end{aligned} \tag{5}$$

Task 2: The coalescence rate (4)

a) What is the coalescence rate for a sample of 6 (and population size 2N)?

The probability of the coalescence event in one generation is:

$$p = \binom{k}{2}\frac{1}{2N} \tag{6}$$

because our population size is 2N the rate parameter is: $\lambda = p = \binom{k}{2}$
For the sample of 6 follows $\lambda = \binom{6}{2} = 15$

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?

$$E[H_n] = 2(1 - \frac{1}{n}) \tag{7}$$

6 to 5:

$$\begin{aligned} E[H_{56}] &= E[H_6] - E[H_5] \\ &= 2(1 - \frac{1}{6}) - 2(1 - \frac{1}{5}) \\ &= \frac{1}{15} \\ \longrightarrow j &= 2N * \frac{1}{15} \end{aligned}$$

5 to 4:

$$\begin{aligned} E[H_{45}] &= E[H_5] - E[H_4] \\ &= 2(1 - \frac{1}{5}) - 2(1 - \frac{1}{4}) \\ &= \frac{1}{10} \\ \longrightarrow j &= 2N * \frac{1}{10} \end{aligned}$$

4 to 3:

$$\begin{aligned} E[H_{34}] &= E[H_4] - E[H_3] \\ &= 2(1 - \frac{1}{4}) - 2(1 - \frac{1}{3}) \\ &= \frac{1}{6} \\ \longrightarrow j &= 2N * \frac{1}{6} \end{aligned}$$

3 to 2:

$$\begin{aligned} E[H_{23}] &= E[H_3] - E[H_2] \\ &= 2(1 - \frac{1}{3}) - 2(1 - \frac{1}{2}) \\ &= \frac{1}{3} \\ \longrightarrow j &= 2N * \frac{1}{3} \end{aligned}$$

2 to 1:

$$\begin{aligned} E[H_{12}] &= E[H_2] - E[H_1] \\ &= 2(1 - \frac{1}{2}) - 2(1 - \frac{1}{1}) \\ &= 1 \\ \longrightarrow j &= 2N * 1 \end{aligned}$$

Height of the tree is:

$$\begin{aligned} T_{MRCA} &= T_5 + T_4 + T_3 + T_2 + T_1 \\ T_{MRCA} &= \frac{1}{15} + \frac{1}{10} + \frac{1}{6} + \frac{1}{3} + 1 = 1,6 \longrightarrow j = 2N * 1,6 = 3,2N \end{aligned} \tag{8}$$

Draw the coalescent tree (with correct branch lengths) for a sample of 6, using the expected waiting times (in N generation units).

The Expected coalescent tree with the above calculated values is shown in figure 1.

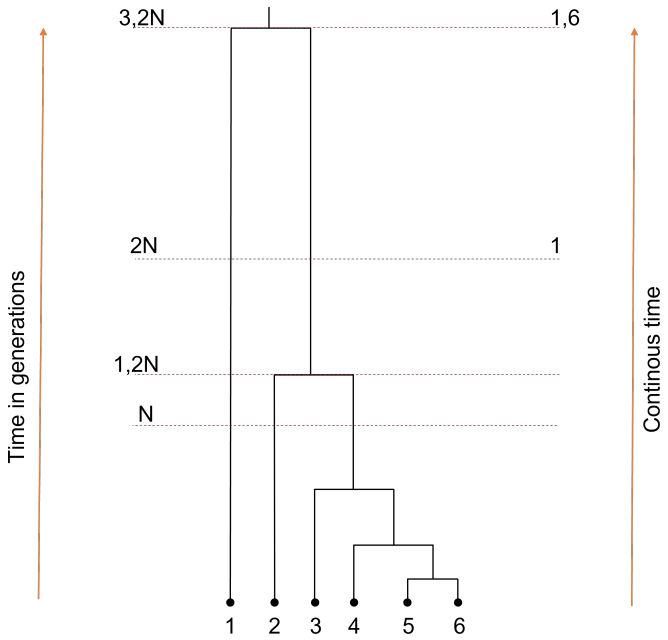

Figure 1: Our Coalescent tree with correct branch lengths

b) Similarly to the formula we derived for the expected total height of a coalescent tree $E(T_{MRCA})$ (s. p. 118/9 in script), derive a closed formula for the expected total length, $E(T_{total})$ of all the branches in the genealogy. When we derive a general formula for the expected total branch length of a tree, we need to remember that with each coalescence event the total number of genes decreases by one and that therefore, in the timespan T_n for a given n there are n branches present. We can add up all these, which results in the Formula:

$$E[T_{tot}] = \sum_{i=2}^{n} iE[T_i] \tag{9}$$

$$= \sum_{i=2}^{n} i \cdot \frac{1}{\binom{i}{2}} \tag{10}$$

$$= \sum_{i=2}^{n} i \cdot \frac{2}{i(i-1)} \tag{11}$$

$$= \sum_{i=2}^{n} \frac{2}{i-1} \tag{12}$$

$$= \sum_{i=1}^{n-1} \frac{2}{i} \tag{13}$$

$$= 2(\frac{1}{2} + \frac{1}{3} + ... + \frac{1}{n-1}) \tag{14}$$

Going from this function, we can simplify as [3]:

$$\sum_{i=1}^{n-1} \frac{2}{i} \approx 2(log(n) + \gamma)) \tag{15}$$

Where γ is the Euler constant. Please also note that this is an approximate solution.

Task 3: Conclusions from Population genetics of Humans (5)

A human race theory has been around for decades [4], and many scientists have studied it to figure out whether the theory is reflected in human genetics or whether it is based on random traits. Since the 1000 Genome Project was conducted, there is much more information about the genetic composition of the human genome [2]. It has been scientifically proven that there are no identifiable genetic commonalities within a given population. Most of the variation is within a population, not outside of it [1][4]. For example, in the genealogical tree of the human species, East-Africans are more closely related to Europeans than Africans from other regions and therefore, all people are of African descent from a genealogical perspective [1]. In addition to this, skin colour is known to have changed in response to exposure to sunlight or other environmental factors rather than as a response to hereditary processes [1]. Therefore, the common racist classification of people by the colour of their skin is non-justifiable. This also clearly refuted Ernst Haeckel‘s classification of the human races, which from a genetic point of view is based only on arbitrary characteristics (e.g., skin colour, eye colour and shape, hair colour and structure)[1]. While genetic characteristics could be found in pets (e.g. dogs) within the race, this is also due to human interference, for example, dogs are bred within a race and the similarities were not developed naturally. Therefore, also in this context the term "breeding" instead of race is more adequate.
The refutation of the human races from the genetic point of view does not mean that there is no genetic variation between groups of humans, but it is not based on the above-mentioned characteristics and there is "not even a single base pair" [1] that would support this thesis.

Practical Assignments

Task 4: Implementation of a simple Wright-Fisher genealogy simulator (10)

For Task 4 we used Python 3.8.8, and the following libraries: **sys, random, itertools, math, numpy and matplotlib**.
Enter the following code in the command line to run our code:

```
python auckenthaler_dittschar_PopGenSimulator.py 4 6 10
```

Each integer behind the filename stands for one initial size of the population. If you want to change, add or remove arguments, you can do that freely. Note that the code saves an image *"auckenthaler_dittschar_median_vs_expected_trees.png"* into the folder of the file. You can also see one good run of our code in the figure 2, here the median of 3 runs of the randomly generated vs. the expected tree sizes is not to far apart.

Please note, that you have to close the figure-window, to run the code again.

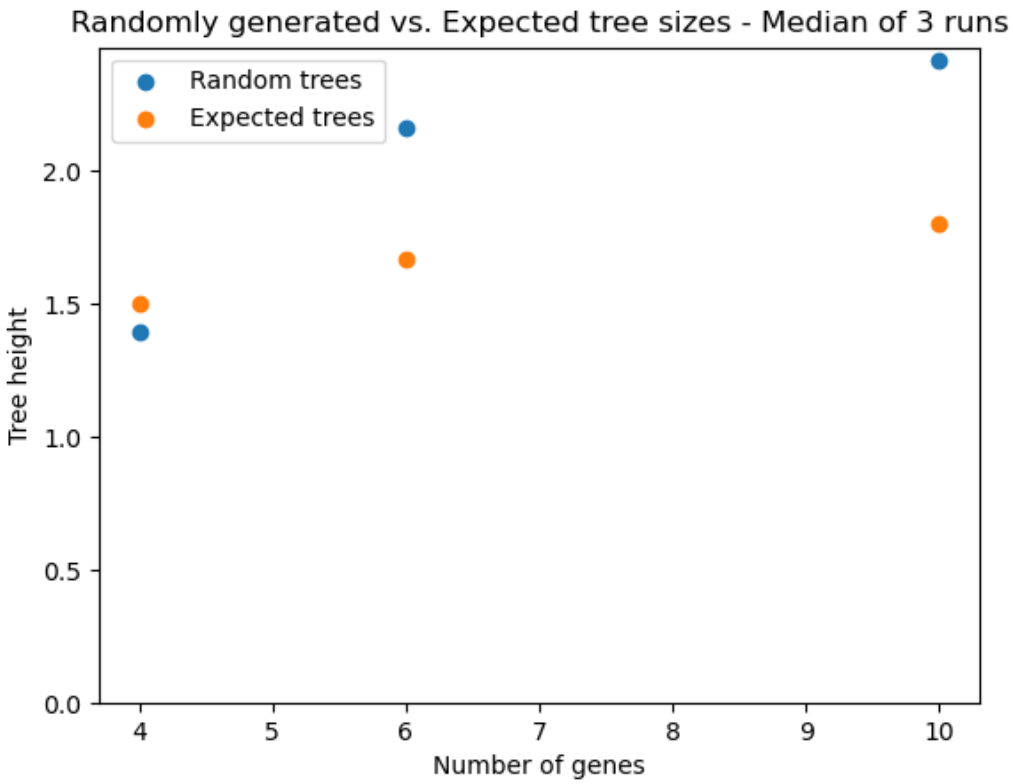

Figure 2: One good example run of our code, from the randomly generated vs. expected tree sizes - Median of 3 runs

References

[1] Martin S. Fischer and Uwe Hoßfeld and Johannes Krause and Stefan Richter. Jena declaration: The concept of race is the result of racism, not its prerequisite. Friedrich-Schiller-Universität Jena, 2019.

[2] Charmaine D. M. Royal and Georgia M. Dunston. Changing the paradigm from 'race' to human genome variation. Nature genetics, 36(11 Suppl):S5–7, 2004.

[3] Harish Nagarajan Vineet Bafna and Nitin Udpa. Algorithms for genetics: Basics

of wright fisher model and coalescent theory. https://cseweb.ucsd.edu/classes/wi16/cse280A-a/lectures/Coalescent_Notes_NH_Nu.pdf, 2022.

[4] Michael Yudell, Dorothy Roberts, Rob DeSalle, and Sarah Tishkoff. Science and society. taking race out of human genetics. Science (New York, N.Y.), 351(6273):564–565, 2016.