

1	2	3	$\sum (7)$

**Blatt 11**

(Abgabe am 1. Februar 2016)

**1. Theoretical Assignment - Coverage statistics****1a. fraction sequenced**

The used equation is from the script (equation 1).

$$f = 1 - e^{-c} \quad (1)$$

As a result of using the given coverages  $c = 1, 2, 4, 8, 14$ , table 1a is deduced. The last column shows the bases which are not covered when the sequence length is  $L = 248956422 \text{ bp}$  like in the human chromosome 1.

coverage	fraction covered f	fraction not covered	bases not covered [bp]
1	0.63212	0.36788	91586089
2	0.86466	0.13534	33693762
4	0.98168	0.01832	4650882
8	0.99966	0.00034	84645
14	0.99995	0.00005	12448

The *Treponema pallidum* chromosome has a length of 1,139,633 bp. If now less than one base can be missed, the not covered fraction is  $\leq \frac{1}{1,139,633} = 7.16 \cdot 10^{-6}$ . This results in a positive fraction of at least 0.999993. The coverage is determined with the transformed equation 2 and has a value of 11.87.

$$c = -\ln(1 - f) \quad (2)$$

**1b. times sequenced**

If there is a coverage of  $c = 10$ , a base has been exactly been sequenced three times with a probability  $p(3)=0.015$  according to the poisson distribution (equation 3).

$$p(x) = \frac{c^x}{x!} e^{-c} \quad (3)$$

$$p(x) = \frac{10^3}{3!} e^{-10} = 0.015 \quad (4)$$

The next question was at most three times covered. This can be calculated with equation 6.  $F(3)$  has a value of  $F(3) = 0.018$ .

$$F(n) = \sum_{x=0}^3 p(x) \quad (5)$$

$$F(3) = \frac{10^0}{0!} * e^{-10} + \frac{10^1}{1!} * e^{-10} + \frac{10^2}{2!} * e^{-10} + \frac{10^3}{3!} * e^{-10} = 0.018 \quad (6)$$

**1c.**

expected number of gaps mean contig size

## 2. Theoretical Assignment - *Application of the arrival statistic for unitigs*

The probability whether a unitig is unique can be done with a formula derived from the Lander-Waterman model for shotgun sequencing. This formula is

$$\frac{e^{-c} c^k}{k!}, \text{ with } c = \frac{\rho R}{L}$$

for non-oversampled unitigs and

$$\frac{e^{-2c} (2c)^k}{k!}$$

for unitigs that are the result of collapsing two repeats. In these formulas  $R$  is the number of reads,  $L$  is the length of target sequence,  $\rho$  is the length of the unitig and  $k$  is the number of reads contained by the unitig.

In this task we consider a source sequence of length  $L = 2Mb$  and  $R = 22000$  reads. The length of the unitig is  $\rho = 3000bp$  and it consists of  $k = 100$  reads. With the formulas mentioned above, the probability of this specific unitig to be unique is

$$\frac{e^{-\frac{3000 \cdot 22000}{3000000}} \cdot \left(\frac{3000 \cdot 22000}{3000000}\right)^{100}}{100!} \approx 5.221 \cdot 10^{-34}$$

for the case that our unitig is not oversampled and

$$\frac{e^{-2 \frac{3000 \cdot 22000}{3000000}} \cdot \left(2 \frac{3000 \cdot 22000}{3000000}\right)^{100}}{100!} \approx 1.846 \cdot 10^{-13}$$

for the case that our unitig is the result of collapsing to repeats.

### 3. Theoretical Assignment - *On distances*

#### 3a.

Consider a tree  $T$  constructed with  $D$  and the four nodes  $i, j, k, l \in T$ . Since  $D$  is ultrametric, the following four inequalities have to hold:

$$d(i, j) \leq \max \{d(i, k), d(j, k)\}$$

$$d(i, j) \leq \max \{d(i, l), d(j, l)\}$$

$$d(k, l) \leq \max \{d(i, k), d(i, l)\}$$

$$d(k, l) \leq \max \{d(j, k), d(j, l)\}$$

W.l.o.g we can assume that  $d(i, k) = d(j, k) = d(i, l) = d(j, l)$ . With that assumption we can rewrite the inequalities as

$$d(i, j) + d(i, j) \leq d(i, k) + d(j, l)$$

$$d(k, l) + d(k, l) \leq d(i, l) + d(j, k)$$

And also the following inequality is valid:

$$d(i, j) + d(k, l) \leq \max \{d(i, k) + d(j, l), d(i, l) + d(j, k)\}$$

As one can see, that is the Four-Point-Condition (4PC) and a metric fulfill 4PC if and only if it is additive. So  $D$  is a tree metric  $\square$

To show the back direction one consider the four elements  $A, B, C, D \in X$  of a taxa  $X$  and the following distance matrix:

$$D = \begin{bmatrix} & B & C & D \\ A & 7 & 6 & 5 \\ B & & 3 & 6 \\ C & & & 5 \end{bmatrix}$$

One can see in the script of this lecture that  $D$  is a tree metric. But the Three-Point-Condition (3PC) is not fulfilled as can be easily shown. To fulfill 3PC the following inequality has to be valid:

$$d(A, B) \leq \max \{d(A, C), d(B, C)\}$$

but

$$7 \not\leq \max \{6, 3\}$$

So  $D$  does not fulfill 3PC and thus  $D$  is not ultra metric.  $\square$

#### 3b.

There are different methods to estimate the (evolutionary) distance between two sequences. The easiest one is to simply calculate the Hamming distance. This is called the observed or  $p$ -distance.  $p$ -distance is just suitable for closely related species, since it is not able to detect superimposed mutations, i.e. mutations that happened at a position that was already mutated, previously. A correction for observed distance was formulated by Jukes and Cantor. The Jukes-Cantor transformation corrects the  $p$ -distance to take, among others, superimposed events into account. This

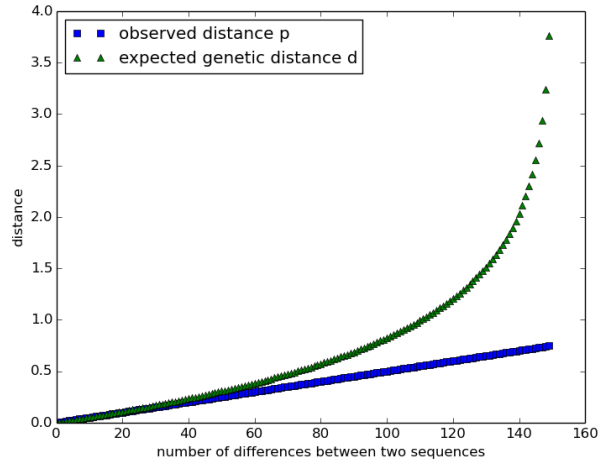


Figure 1: Behavior of observed distance  $p$  and expected genetic distance  $d$  of two non-saturated sequences. Both sequences are of length 200.

correct distance is called the expected genetic distance  $d$ . An important note is that JC transformation does not work for sequences that are saturated w.r.t. each other. So Figure 1 shows the relationship between  $d$  and  $p$  for non-saturated sequences of length 200. As one can see the observed distance  $p$  is constantly underestimating the genetic distance for more distant sequences. Especially, if both sequences are close to be saturated, the difference between  $d$  and  $p$  shows that  $p$  cannot give a reliable result. So for sequences/species that are evolutionary close to each other both measurements  $d$  and  $p$  give almost the same result. But the further away the last common ancestor of both sequences the larger is the difference between observed distance and expected genetic distance.