

Data Science in Bioinformatics

Day 6

Let's dig a little more into the topics of the second day of student presentations. The topics of the talks included trimming and filtering of biological sequences and the alignment of sequences.

1) Trimming and Filtering

- a. Why is trimming and filtering important?
- b. Create a Conda environment with fastp.
- c. Use the samples downloaded from the last exercises and create a fastp report for filtering with the default parameters!
- d. What are the default filtering parameters for sequence quality and sequence length for fastp?
- e. What percentage of bases has a Phred score higher than 20 before and after filtering?

2) Alignment

- a. What is the difference between local and global alignment?
- b. What does pair-wise and multi-sequence mean?
- c. Which alignment tool out of the two presented (minimap2 and bwa) would better fit to the sequences previously downloaded and why?
- d. For an alignment you need a reference genome. What reference genome would you pick for the given sequences and where could you get it?
- e. Create an alignment with the sequencing tool you chose from above and the sequences created while filtering.
- f. Is there a way to see the parameters of the alignment without need of visualization?
- g. How many reads have been aligned?