Notes for Genome Track 2, Step 2

Manuel Gil

Menu

A. Current Situation
B. Next Step

A. Current Situation

As pointed out in our last meeting, we plan to work according to the following steps.

1. Simulated data with PIP, getting to know the cluster
2. Infer MSA, parameters, and derived statistics
3. Summarise, interpret, present
4. Thesis specific

At this point, you should have performed Step 1, i.e. you should have simulated data according to the following specification:

- Topologies:
  - 8 leaves
  - Two forms: perfectly balanced, and ladder topology
- Branch lengths, each branch:
  - 0.001 substitutions/site (Phuong)
  - 0.01 substitutions/site (Jithin)
  - 0.1 substitutions/site (ELdhose)
- Substitution model:
  - K80, a nucleotide model with one parameter only, the transition/transversion rate ratio κ
  - https://en.wikipedia.org/wiki/Models_of_DNA_evolution#K80_model_(Kimura_1980)
  - $\kappa = 2.0$
- Indel model:
  - PIP
  - Intensity $\zeta = \lambda \cdot \mu = \{10,100,200\}$
  - Asymptotic expected sequence length: $\eta = \lambda / \mu = 1000$ nucleotides
- Replicates: 100

Further, you should be able to use the cluster:
https://collab.zhaw.ch/Kooperation/3221/SitePages/Home.aspx

B. Next Step

This week, we are moving to Step 2, which has two parts. First, you will infer MSAs and model parameters from the simulated data using our progressive aligner. In the second part, you will analyse the results.

Settings for Inference

All inference should be performed with the correct tree as guide-tree, and with the correct values for κ, λ and μ (or ζ and η, respectively).

Analyses of Results

Given the inferred MSAs, we would like you to analyse the following quantities:

- A1: Distance to true MSA (Max will point you to a program for this).
- A2: Difference between the length of the true MSAs and the length of the inferred MSAs
- A3: Difference between the number of gaps in the the true MSAs and the inferred MSAs
- A4: Difference between the indel lengths in the the true MSAs and the inferred MSAs

To extract the quantities for A2—A4 you will have to write a bit code. The three of you are encouraged to collaborate to produce it.

Produce one plot for each quantity A1—A4, with ζ as x-axis and the mean (and standard deviation as error-bars) over the 100 replicates of the quantity in question.

Tentative deadline: next week. Please contact me, if you do not manage to generate the plots in a total of about 13h.