Tutorial

Phylogeny

Kerstin Reuter

August 26, 2014

Exercise 1.1: MSA - Multiple Sequence Alignment

(a) Use a database of your choice to search for the MecA protein sequence of the reference genome (NC_017340 (04-02981)). Apply a BLAST search to obtain different MecA sequences of *S. aureus*. Save all sequences found by BLAST in FASTA format.

(b) Go to http://www.ebi.ac.uk/Tools/msa/mafft/ and upload your FASTA file containing the MecA proteins. Have a look at the multiple alignment. What does it mean if ”-” is present in a sequence?

(c) Have a look at the phylogenetic tree. How can you find the most similar sequences with the help of the phylogenetic tree.

(d) Assume you align several protein sequences. Is it always necessary that the amino acids are highly conserved (always the same letter in all columns) to conclude that the proteins are similar/homolgs?

(e) Assume you would like to have an idea of the localization of the active center of a protein but you only have the protein sequence. How can you use a multiple alignment to solve this problem? Why is it possible to use a multiple alignment for this problem?

Exercise 1.2: Phylogeny: General

(a) Why is it useful to make a phylogenetic analysis of isolated MRSA strains?

(b) Why is it suitable to use several reference genomes for the phylogenetic analysis?

Exercise 1.3: SNP Matrix File

In this exercise, you should inspect the SNP matrix file. The file can be downloaded from our webpage.

(a) What is contained in the SNP matrix file/which information is stored?

(b) How was this file created?

(c) Why is this kind of data suitable for phylogenetic analysis?

(d) Why do all sequences have the same length (different strains do not necessarily have the exact same genome size, do they?)?

Exercise 1.4: SeaView and FigTree

SeaView is installed and accessible via CLI (command line interface). That means, open the Terminal (system tools → XTerm) and type *"seaview"*. You can open FigTree with the *"figtree"* command.

(a) Open SeaView and load the SNP matrix file. How are the sequences represented? What is the meaning of the colors and why are they useful in this context?

(b) Based on the way the SNP matrix file was generated, what can you say about each column. Is a fully conserved column possible?

(c) Next, you should built a phylogenetic tree. Start with parsimony and default settings. You should get a new window containing the phylogenetic tree and some options on the top. Using the *File* option you can save the tree in different formats. Save the rooted tree you just generated.

(d) Assume, you want to use the tree, that you generated in (c), in a publication. Therefore, you need a nice layout which can be established using FigTree. A good layout can also facilitate the interpretation.
Open FigTree and open your rooted tree.
Have a look at different layouts (*Rectangular tree layout*, *Polar tree layout*, *Radial tree layout*) and familiarize yourself with the different layouts.
Which strain was used as root by default? Can you identify the main clusters in each layout? Use the *Highlight* optin to highlight the clusters. Save the rectangular and polar tree layout with your highlighted clusters as PDF.
Can you conclude how those strains might have evolved? What do isolates t003 and t504 have in common? How do you interpret the circumstance how the other 4 German strains are clustered?

(e) Next, you should investigate what happens if you change the rooting of the tree.
Construct your phylogenetic tree using parsimony and default settings (like before). But this time re-root the tree with reference genome NC_017340_ref. Save the re-rooted tree, open it with FigTree again, use the *Highlight* option to mark different clusters and save it in your favourite layout. Is there a difference with respect to the tree you analysed before?

(f) This time you should investigate what happens if you change the settings for the tree construction. Apply different algorithms (distance methods (e.g. the neighbor-joining algorithm (*NJ*) which was explained in the lecture), different evolutionary distances etc.) and save the trees in a layout which enables a good interpretation.

(g) You used various phylogenetic and distance methods. How did the tree structure change when applying different measures? Can you conclude something from this? Try to interpret your results.