Tutorial 1

Databases

Kerstin Reuter

August 25, 2014

As *S. aureus* reference genome, NC_017340 (04-02981) should be used in the following.

Exercise 1.1: NCBI - National Center for Biotechnology Information (morning)

In this exercise you should find out more about the *S. aureus* MecA gene.
Therefore, go to NCBI (http://www.ncbi.nlm.nih.gov/).

(a) Search for the gene sequence from the reference genome and save the sequence in FASTA format.

(b) Name the locus tag.

(c) At which position do you find this gene in the reference genome? Which genes are next to this one?

(d) Find the respective protein ID and the protein sequence.

Exercise 1.2: PATRIC - Pathosystems Resource Integration Center (morning)

Go to PATRIC (http://patricbrc.org/).

(a) Can you say something about the *S. aureus* lineage?

(b) Have a look at the different strains. How many *S. aureus* strains does the database contain (from Whole Genome Sequencing? In total?)? Have a look at all genomes (you will get a nice table view). How many strains do you find which are responsible for the toxic shock syndrome and isolated in USA?

(c) Next, you should have a look at *antibiotic resistance*. Which databases are queried by PATRIC and how many genes are contained in each of them? Have a look at the databases. In the following concentrate on the reference strain. How many resistance genes do you find for the reference strain, in ARBD and CARD respectively? Name some antibiotic resistance genes. What is the function of MecR1? Open the gene as well as the protein sequence of MecR1 in FASTA format.

(d) How many drug targets can you find for the reference strain? Name some genes.

(e) Now, have a look at *virulence factors*. How many virulence factors do you find for the reference genome? Name some virulence factors that prevent phagocytosis. Get some information about virulence factor *hlb*.

Exercise 1.3: BLAST - Basic Local Alignment Search Tool (afternoon)

When running a BLAST search algorithm parameters can be specified (e.g. the *expected threshold* or the *word size*). In this exercise you will BLAST the sequences of the resistance marker gene MecA against different *S. aureus* strains.

(a) How is the *expected threshold* defined? Why is an E-value of 10 not suitable? What are suitable E-values?

(b) How does the *word size* influence the runtime and accuracy of the search?

(c) Go to the BLAST homepage (http://blast.ncbi.nlm.nih.gov/Blast.cgi). You should execute a BLAST search of the MecA *S. aureus* gene, you saved the sequence for in the first exercise. Blast against *ALL microbial genomes.* Do you find any significant alignments? Have a look at the sequences and familiarize yourself with the representation of the alignments.
Which strains do you find? What about the E-values? In general, which genes do you find via a BLAST search w.r.t. homology?

(d) Assume you isolated unknown proteins in the wet lab from unknown *S. aureus* strains and want to know the functions. The protein sequences can be downloaded from our homepage (filename: unknown_s_aureus_protein1.fasta, unknown_s_aureus_protein2.fasta and unknown_s_aureus_protein3.fasta). How can you use BLAST to get an idea of the function? If you find promising candidates, how sure are you, that your protein could exhibit this function?
(*Hints: beside blastn (nucleotide blast) there is also protein blast (blastp). Remember that you know the taxa to search for.*)

Exercise 1.4: Bonus*

Try to find the same MecA protein (penicillin-binding protein) of Exercise 1.1 in the Swissprot/Uniprot database. Can you find a template with three-dimensional structure for the MecA protein that provides structural insight into the binding-mode for penicillin?