Optimizing k-mer length for bacterial strain identification

Introduction
Second generation sequencing technologies have made it possible to describe whole environmental or clinical samples containing a large number of species. Conventional methods such as PCR or 16S RNA sequencing are slow or might not give sufficient information.

Our approach for identification is to use strain specific k-mers (substrings of length k). Shorter length produces less k-mers and might not cover the whole genome but as k-mer length increases, giving us more k-mers, so does the database and searching time.

Goals
- Find out how word size affects specific k-mer distribution over the genome
- Propose the optimum length.

Workflow
1. **Genome sequence**
   - K-mer sequences
   - Gene locations
   - Map k-mers
   - Find in which region k-mer is located
   - Calculate weighted column heights
   - Draw histogram

   **K-mer distribution, genome divided into 1523 segments**
   - Gene density
   - Unconserved regions
   - Conserved region

   Comparing histograms with different word size it can be concluded that as word size increases strain-specific k-mer count overall becomes larger. However more of the genome will not be covered. K-mers are quite uniformly distributed over the genome with some more denser variable regions and empty highly conserved regions.

2. **Genome sequence**
   - K-mer sequences
   - Count times every nucleotide is covered
   - Draw barplots

   **Number of nucleotides**
   - Single nucleotide coverage
   - 284520 nucleotides covered 22 times (22 k-mers)

   **Typical regions**
   - Word size 16
   - Word size 18
   - Word size 20
   - Word size 22

   **Comparison**

Results
- Word size 16 is too short – all specific regions are not yet found
- With word size 18 most specific regions are found and further on these regions only get connected – no new independent regions that are not found
- Longer word size k-mers cover single nucleotides more than previously uncovered nucleotides

Conclusions
- With increasing word size database size increases faster than previously uncovered nucleotide coverage
- Word size 18 appears to be enough
- Due to the fact that bacteria genomes vary greatly word size 20 is recommended