Lecture 5
Practical part

Warmup task

Make groups of 1-3 people such that each group is able to:

- Log in to kirss.at.mt.ut.ee
- Create a directory for your group
- Copy file lecture5.fasta to your directory
- View the contents of this file
- Run command to see information about the sequence
  `infoseq -sequence lecture5.fa`
- `Find the tasks in tasks.txt`

Tasks

TASK 1 - Analyzing CpG islands with EMBOSS

- How high is GC-content? (infoseq, geecee, dreg, wordcount)
- How many CpG-s are there? (dreg, wordcount)
- How many would you expect from GC-content, if there is no dependence between two consecutive nucleotides?
- Check your expectation by shuffling the sequence and counting CpG (shuffleseq, dreg, wordcount)

TASK 2 - Finding genes with EMBOSS

- Find all open reading frames and the corresponding translations into protein sequences (getorf)
- Find all open reading frames with >300 nucleotides
- Narrow down the list of potentially functional open reading frames (tcode, marscan)

Tasks

TASK 3 - Clustering vector data with R

- Copy file capitals.txt to your directory
- Run R, input the data there using capitals=read.table("capitals.txt")
- Run K-means with different values of K (kmeans)
- Calculate all pairwise distances, try different measures (dist)
- Cluster with hierarchical clustering, try different methods (hclust)
- Plot the result to PNG (png, plclust)

TASK 4

- Find all 7-mers that occur at least 10 times (wordcount)
- Calculate pairwise Hamming or Levenshtein distances
- Cluster with hierarchical clustering

Help

- geecee -help
- http://emboss.sourceforge.net/apps/release/6.2/emboss/apps/geecee.html
  (google emboss geecee)
- geecee (specify filename after)
- geecee -sequence filename
- geecee -sequence filename -outfile otherfile
- http://emboss.sourceforge.net/apps/release/6.2/emboss/apps/
- EMBOSS-6.3.1/doc/programs/text/