Sequence comparison

a project for simple comparison of two biological sequences

Parts of the problem

- downloading and parsing existing sequences from databases
- actual comparison of the two sequences using multiple algorithms and parameters
- generating the PDF report with the results of the comparison

Downloading the sequences

- multiple database options are available as a source of the sequences
  - uniprot (Universal Protein Resources)
  - ncbi (National Center for Biotechnology information)
  - ebi (European bioinformatics institute)
  - ddb (DNA data bank of Japan)

- available formats - fasta

Downloading the sequences

- after the sequences are downloaded, they are normalized, so they can be compared better
- this is only a last resort option, and I don't think anyone would really use it

Comparison of the sequences

- simple histogram of characters
- global and local alignment, using match/mismatch option, gap opening/extension option, and also existing biological matrices
- available matrices: PAM40, PAM80, PAM120, PAM250, BLOSUM62

- the options can be chained - same sequences but multiple options used, for quicker analysis

Generating the PDF report

- used ReportLab library
- it is quite hard to get into (userguide of “only” 125 pages)
- but the lib is very efficient, with a lot of options (charts, tables, graphics, …)

Conclusion

- working download of sequences from online sources
- implemented comparing algorithms with different options usable from script arguments
- learned how to use a PDF generating library

All code is available from
https://github.com/gyfis/sequence-comparison

Thank you for your attention!