COMP364: Manipulating Rfam data with Biopython

Jérôme Waldispühl, McGill University

Rfam

The Rfam database is a collection of RNA families, each represented by multiple sequence alignments, consensus secondary structures and covariance models (CMs).

http://rfam.sanger.ac.uk/

Download the seed alignment in Stocholm format of the 5S ribosomal rRNA (Rfam ID: RF00001).

Using Biopython, read and parse this alignment. Print all sequences in your terminal with the format:

1. Id
2. Sequence (without gaps!)
3. Length of the sequence

At the end of your program, return the average length.

SeqUtils

Visit the documentation of the SeqUtils module of Biopython at:

http://biopython.org/DIST/docs/api/Bio.SeqUtils-module.html

- Using SeqUtils & matplotlib, plot the molecular weight vs. the sequence's length of each sequence in your family (WARNING: ignore sequences with ambiguous nucleotides).
- Extract one sequence of your alignment and plot its GC skew with a window size of 20.