COMP364: Manipulating GenBank data with Biopython

Jérôme Waldispühl, McGill University
Go to Genbank and download the record of the "Saccharomyces cerevisiae TCP1-beta gene" at:


• Calculate the results of transcription and translation of this gene. Comment your results. Does it look like the real coding sequence?
More on representations

What is repr()? Use your genbank record “seq_record” and print:

• seq_record
• repr(seq_record)
• seq_record.seq
• repr(seq_record.seq)

The repr module provides a means for producing object representations with limits on the size of the resulting strings. This is used in the Python debugger and may be useful in other contexts as well.
More on GenBank Format

Print the content of the GenBank objet. What do you see?

Genbank format contains many data that can be retrieve using the appropriate fields. For instance, “organism”, “references” and “features”.

Lets have a look at the field “.features”. Print its content. It looks like a list of objects... Using a for loop, enumerate all the items inside the features and print them.

It contains information about the mRNA and coding sequence produced by this gene.
Coding region (CDS)

Filter only the features corresponding to the coding regions (i.e. such that .features.type is “CDS”)

Retrieve the start, end and strand direction of the coding gene:
- seq_record.features.location.start (start of transcription)
- seq_record.features.location.end (end of transcription)
- seq_record.features.location.strand (direction)

Retrieve the offset and final product (protein sequences) with:
- seq_record.features.qualifiers['codon_start']
- seq_record.features.qualifiers['translation']

Print these data. Use them to produce the protein sequence form the dna and compare your results to the expected ones.
(N.B: The last one is part of your homework!)
Some utilities on Protein Sequences

Use the ProtParam module of SeqUtils (http://biopython.org/DIST/docs/api/Bio.SeqUtils ProtParam-module.html) and show:

- the amino acids count,
- the amino acids percent,
- the flexibility,
- the secondary structure fraction.

When suitable, use Matplotlib.