Introduction to Bioinformatics

Course NR. 22111

GenBank

Stefan Olevinskiy
s246026

Polina Krasikova
s245850

Data in GenBank

Question 1:

a) How many genes are contained in this entry?

Two genes – Alpha-A and Alpha-D

b) From which organism does the DNA originate?

From *Columbia livia* (Rock pigeon).

c) What kind of information is contained within the HEADER and within the FEATURE block?

HEADER: Contains a unique accession ID, description, organism (taxonomy), reference publications, etc. (Information that applies to **all genes** in the entry.)

FEATURE block: Contains a description and the coordinates of individual features, such as genes or promoters. CDS (Coding sequence): The protein-coding region of a gene, consisting of exons and introns within the coding region.

Question 1.2:

a) What happened to the alpha-globin genes? Can they still be found?

The alpha-globin genes (alpha-D and alpha-A) are still present in the sequence. The FASTA format show the raw DNA sequence, so the genes themselves are there, but no annotations (like coding regions, exons, or introns) are visible.

b) Which part of the GenBank entry has been converted?

In Fasta format from the GenBank entry has been converted the ORIGIN block, which contains the actual DNA sequence. The HEADER information is reduced to a single line starting with > (accession ID and description), and the FEATURE block annotations are not included in the FASTA file.

QUESTION 1.3:

Does the downloaded file have UNIX or Windows line-endings?

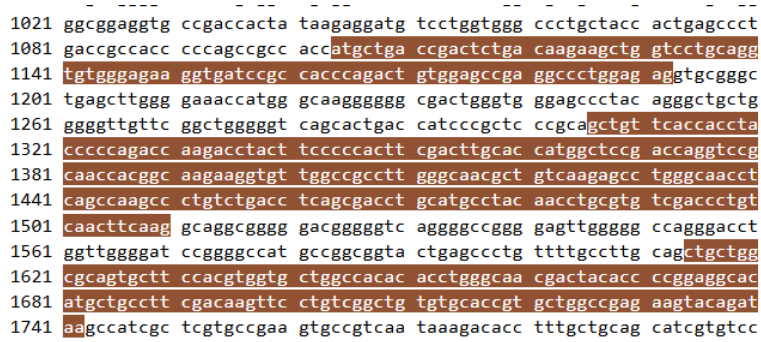
The file shows EOL: LF in Geany, it means the file uses Linux/UNIX-style line endings.

QUESTION 1.4:

a) What do these numbers mean?

The numbers indicate the positions of the coding sequences (CDS) within the DNA sequence.

For example, join(1104..1192,1306..1510,1614..1742) means the CDS is split across multiple segments (exons) in the DNA.



The .. shows the start and end positions of each segment. The join() indicates that the exons are joined together in the final mRNA after splicing.

b) How many coding exons does each gene contain?

The first CDS (1104..1192, 1306..1510, 1614..1742) has 3 coding exons.

The second CDS (4915..5009, 5474..5602, 5165..5369) also has 3 coding exons.

QUESTION 1.5:

The numbers in the sequence title show the start and end positions of each coding exon in the DNA; the join() indicates that these exons are spliced together to form the complete CDS.

Searching in GenBank

Question 2.1.1:

a) How many search results were returned?

Searching only the nucleotide sequences, 252405 entries were returned.

b) Are they all from Human? If no, give a counterexample. (Would you have expected them to be all human?)

No, of course, they are not all human – the first result is from Octodon Degus – a cute rodent.

c) Are they all insulin? If no, give a counterexample.

No, of course not. They are all nucleotide sequences containing the word insulin somewhere in the entry. On the first page, for example, I already see Insulin-like Growth Factor-2.

"By default the search term is matched against ALL POSSIBLE fields in theGenBank entries - including almost all text in the HEADER and FEATURE table. It's even possible to pick up entries where the match is to one of the authors names and not a gene name! ".

QUESTION 2.1.2:

a) What have your search for "insulin" been expanded into?

"Insulin[All Fields]"

QUESTION 2.1.3:

a) How many search results were returned?

19745 results were returned.

b) Can you find the human insulin entry? (If yes, write down its title and Accession)

Yes, it is titles Homo sapiens insulin (INS) gene, complete cds. For the genomic dna, its accession is AH002844. For just the coding transcript, it is NM_000207.

c) How was your search interpreted by the system (the SEARCH DETAILS box)?

""Homo sapiens"[Organism] OR human[All Fields]) AND insulin[All Fields]"

QUESTION 2.2:

a) How many hits do we have now?

We have 5609 hits this time.

b) Are they all from Human? If no, give a counterexample.

Yes, they are.

c) Do they all appear to be insulin genes? If no, give a counterexample.

Once again, they are not. The first page contains IGF-2, insulin-like growth factor binding proteins, etc.

QUESTION 2.3:

a) How many hits are found when "Keyword" is set to insulin?

Just 9.

b) How many hits are found when "Protein Name" is set to insulin?

15.

c) Find the correct Human Insulin gene entry (the correct hit). Click on it and write down its Accession codes (there are more than one!), Locus name and Definition (title).

Once again, it is AH002844, but also J00265 J00268 – these are secondary accessions and they're from older partial submissions that were later merged into AH002844.

QUESTION 2.4:

a) Which search term did you end up using?

“INS[Gene Name] AND complete cds[Title]) NOT mRNA[Title]”

b) How many search results do you get now?

26.

"Free exercise"

1. **Find the Rat and Mouse Insulin gene**

   (((((("Mus"[Organism] OR "Mus musculus"[Organism]) OR ("Rattus"[Organism] OR "Rattus norvegicus"[Organism])) AND insulin[Keyword]) NOT mRNA[Title]) NOT factor[Title]) NOT receptor[Title]

2. **Find the alpha-globin gene from *Capra hircus* - (Remember: Alpha-globin is part of hemoglobin).**

   "Capra hircus"[Organism] AND Alpha-globin[Title] NOT mRNA[Title]

3. **Find the human insulin receptor gene. Avoid partial genes / single exons in the results.**

   "Homo sapiens"[Organism] AND complete[Title]) AND insulin receptor[Keyword]