Laborator 4 - Bioinformatică

Formatul GenBank

1. conţine informaţii despre genomul secvenţiat;

2. conţine informaţii despre autori;

3. secvenţa de ADN este ultima intrare în fişier şi începe după cuvântul ORIGIN;

4. secvenţa de ADN se termină la apariţia şirului “//”;

5. secvenţa de ADN (o catenă) este scrisă pe linii; fiecare linie a secvenţei de ADN conţine 60 de nucleotide grupate câte 10, iar grupurile de nucleotide sunt separate printr-un spaţiu (ultima linie din secvenţă poate avea mai puţin de 60 de nucleotide);

6. genele codificate în secvenţa de ADN sau în complementara acesteia sunt semnalate de cuvântul *gene* (când este precedat de spaţii) şi/sau CDS;
7. imediat după semnalarea genei, apare locaţia acesteia; locaţia unei gene în secvenţa de ADN se poate găsi în următoarele situaţii:
   - *6981..9121* → începe la poziţia 6981 din secvenţa de ADN din fişier şi se termină la poziţia 9121
   - *complement(2630..3110)* → începe la poziţia 2630 şi se termină la poziţia 3110 din secvenţa de ADN complementară celei din fişier (catena complementară din dublul helix)
   - *join(1223..1567, 2110..2304)* → gena este alcatuită din două segmente din secvenţa de ADN găsită în fişier – segmentul care începe la poziţia 1223 şi se termină la poziţia 1567 şi segmentul care începe la poziţia 2110 şi se termină la poziţia 2304 (gena poate fi alcatuită şi din mai mult de două segmente, acestea fiind separate prin ",")
   - *complement (join(1223..1567, 2110..2304))* → gena este alcătuită din mai multe secvenţe de pe catena de ADN complementară celei găsite în fişier
   - *7193..>8061* → semnul ">" arată că nu se ştie exact unde se termină gena; pentru a extrage locaţia unei regiuni codificatoare trebuie tratate toate aceste situaţii

8. Cuvintele cheie încep de la caracterul al 6-lea (coloana 6 în fişier) şi au o lungime maximă de 15 caractere. Locaţia începe de la caracterul al 22-lea (coloana 22 în fişier). (ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt)

Aplicaţii. Formatul GenBank

1. Să se acceseze baza de date *Genome* de la NCBI (National Center for Biotechnology Information): http://www.ncbi.nlm.nih.gov/sites/genome si sa se salveze secventa cu numarul de acces NC_014378 in format GenBank. Pentru afisarea intregii secvente in fisierul GenBank, inainte de salvarea fisierului, sa se selecteze optiunea „Show sequence" de la „Customize view".

2. Parsarea GenBank

   Să se implementeze funcţia *extractDNA(fin, fout*), care citeşte secvenţa de ADN dintr-un fişier în format GenBank ('*sequence.gb*') şi o scrie apoi într-un fişier în format FASTA ('*sequence.adn*').

3. Pe baza scheletului de cod din fişierul *schelet_genbank.py*, să se extragă din secvenţa de ADN toate genele indicate în fişierul GenBank şi să se scrie câte una pe linie în fişierul *'sequence.genes'*. Trebuie tratate toate cazurile posibile ale poziţiei unei gene: complement, join, etc. Sa se utilizeze fisierul FASTA creat la 2.

   Trebuie completate următoarele funcţii:

   - *join(str, fADN)* – care tratează cazul în care o genă este alcatuită din mai multe subsecvenţe din ADN şi întoarce gena în urma concatenarii secvenţelor care o compun.
   - *extractGeneFromFile(a,b,fADN)* – care extrage din fişierul cu numele *fADN* (format FASTA) secvenţa de ADN cuprinsă între poziţia *a* şi poziţia *b* şi întoarce gena ca un string.

Observaţii:

Să se utilizeze funcţiile *split()* şi *strip()*:

- *sir.split(',')* – care întoarce o listă cu subşirurile din şirul *sir* separate prin separatorul *','*
- *sir.strip()* – care elimină toate spaţiile, taburile, caracterele *'\n'* de la capetele şirului *sir*
- *sir.lstrip()* – care elimină toate spaţiile de la stânga şirului
- *sir.rstrip()* – care elimină toate spaţiile de la dreapta şirului
- *sir.strip(chars)* – elimină de la capetele şirului *sir* toate caracterele conţinute în şirul *chars*
- *sir.lstrip(chars)* – elimină din stânga şirului *sir* toate caracterele conţinute în şirul *chars*
- *sir.rstrip(chars)* – elimină din dreapta şirului *sir* toate caracterele conţinute în şirul *chars*
- *sir.replace('c1','c2')* – înlocuieşte caracterul *c1* cu *c2* din şirul *sir*
- *sir.upper()* - transformă literele din şirul *sir* în litere mari
- *sir.find('subs')* – întoarce indicele de început al subşirului *subs* în şirul *sir*