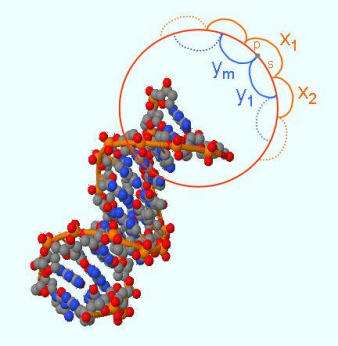


UFR Mathématique Informatique Strasbourg

Professeur Christian Michel

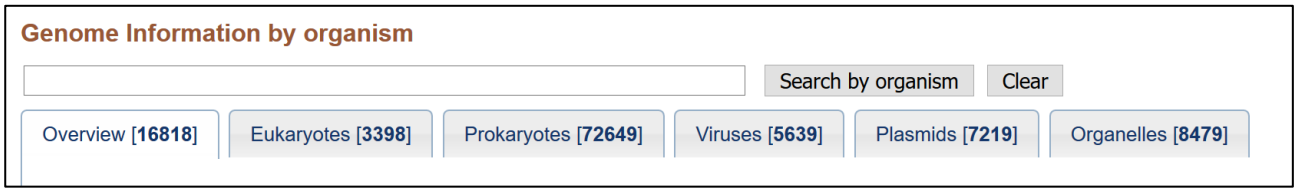
Projet Bioinformatique Master SDSC 2020 - 2021

BIG DATA EN BIOINFORMATIQUE

Parsing des régions fonctionnelles dans les génomes

1. Recherche des régions fonctionnelles (gènes, introns, rRNA, tRNA, etc) dans les génomes des eucaryota, bacteria, archaea, virus, plasmides et organelles (mitochondria, chloroplasts) dans la base de données de gènes GenBank

http://www.ncbi.nlm.nih.gov/genome/browse/
ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/



1.1. Stockage des régions dans des fichiers texte (extension txt)
Impératif: Respectez l'arborescence du site.

Results\Kingdom\Group\SubGroup\Organism\Organism CDS.txt

Results\Kingdom\Group\SubGroup\Organism\Organism intron.txt

Eukaryota

- Eukaryota
  - Animals
    - Amphibians
    - Birds
    - Fishes
    - Flatworms
    - Insects
    - Mammals
    - Other_Animals
    - Reptiles
    - Roundworms
  - Fungi
  - Other
  - Plants
  - Protists

Bacteria

- Bacteria
  - Acidobacteria
  - Aquificae
  - Bacteria_incertae_sedis
  - Caldiserica_Cryosericota_group
  - Chrysiogenetes
  - Coprothermobacterota
  - Deferribacteres
  - Dictyoglomi
  - Elusimicrobia
  - FCB_group
  - Fusobacteria
  - Nitrospirae
  - Proteobacteria
  - PVC_group
  - Spirochaetes
  - Synergistetes
  - Terrabacteria_group
  - Thermodesulfobacteria
  - Thermotogae

Archaea

- Archaea
  - Euryarchaeota
    - Archaeoglobi
    - Diaforarchaea_group
    - Methanomada_group
    - Methanopyri
    - Stenosarchaea_group
    - Thermococci
  - TACK_group
    - Candidatus_Korarchaeota
    - Crenarchaeota
    - Thaumarchaeota

Si il existe plusieurs NC pour un organisme, alors
Results\Kingdom\Group\SubGroup\Organism\Organism CDS NC0001.txt
Results\Kingdom\Group\SubGroup\Organism\Organism CDS NC0002.txt

IMPORTANT:
- Ne considérer que les identifiants NC (donc rejeter les identifiants NW, AC, etc.).
- Prendre tous les génomes.
- Utiliser les API de GenBank.
- Faire un menu permettant le choix des domaines (kingdoms).

1.2. Régions fonctionnelles

Faire un menu permettant le choix des régions fonctionnelles:
- CDS
- centromere
- intron
- mobile_element
- ncRNA
- rRNA
- telomere
- tRNA
- 3'UTR
- 5'UTR
- Option avec un texte libre choisi par l'utilisateur.

1.3. Structuration du fichier

1ère ligne: Exemples
CDS 777162..777467
CDS complement(6502701..6503480)
CDS join(1..10,30..50) 1..10
CDS join(1..10,30..50) 30..50

2ème ligne: la séquence nucléotidique avec une format FASTA.

```
CDS 777162..777467
ATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGTAACAGGAAGAAGCTTGCTTCTTTGCTGACG
AGTGGCGGACGGGTGAGTAATGTCTGGGAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAAT
ACCGCATAACGTCGCAAGACCAAAGAGGGGGACCTTCGGGCCTCTTGCCATCGGATGTGCCCAGATGGGATTA
GCTAGTAGGTGGGGTAACGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAAGATGACCAGCCACACTGG
AACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGC
AGCCATGCCGCGTGTATGAAGAAGGCCTTCGGGTTGTAAAGTACTTTCAGCGGGGAGGAAGGGAGTAAAGTTA
ATACCTTTGCTCATTGACGTTACCCGCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGG
AGGGTGCAAGCGTTAATCGGA
CDS join(1..10,30..50) 1..10
ATTGAACGCTGGCGGCAGGCCTAACACATGCAAGTCGAACGGTAACAGGAAGAAGCTTGCTTCTTTGCTGACG
AGTGGCGGACGGGTGAGTAATGTCTGGGAAACTGCCTGATGGAGGGGGATAACTACTGGAAACGGTAGCTAAT
ACCGCATAACGTCGCAAGACCAAAGAGGGGGACCTTCGGGCCTCTTGCCATCGGATGTGCCCAGATGGGATTA
GCTAGTAGGTGGGGTAACGGCTCACCTAGGCGACGATCCCTAGCTGGTCTGAGAAGATGACCAGCCACACTGG
AACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGCGCAAGCCTGATGC
AGCCATGCCGCGTGTATGAAGAAGGCCTTCGGGTTGTAAAGTACTTTCAGCGGGGAGGAAGGGAGTAAAGTTA
ATACCTTTGCTCATTGACGTTACCCGCAGAAGAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGG
AGGGTGCAAGCGTTAATCGGA
…
```

2. Acquisition des régions

2.1. Sélection des gènes
CDS (Coding Sequence) et CDS complement

2.2. Traitement des gènes
Pour les eucaryotes
- opérateur JOIN (cf. cours)
- opérateur COMPLEMENT (cf. cours)
- opérateur COMPLEMENT(JOIN) (cf. cours)

```
CDS             777162..777467
                /locus_tag="AM1_0801"
                /codon_start=1
                /transl_table=11
                /product="hypothetical protein"
                /protein_id="YP_001515158.1"
                /db_xref="GI:158333986"
                /db_xref="GeneID:5679627"
                /translation="MSPQPFQPPDEFESLLSTQRQTNADLERDLAELSEDSRRAVAD
                QRTMQKFFGILLVVGLALGAVTAVGVVHFIQWLRSTTNSEPQPPNQSWVDTSKGFKI
```

```
     CDS             complement(6502701..6503480)
                     /locus_tag="AM1_6411"
                     /codon_start=1
                     /transl_table=11
                     /product="hypothetical protein"
                     /protein_id="YP_001520659.1"
                     /db_xref="GI:158339482"
                     /db_xref="GeneID:5685188"
                     /translation="MSQVPNLNTLFQSAQADGVLSNASMQALNVVDIGAQIQAGLGT
                     VDDVMASEVVLVTIMPDDSGSIRFAGNGAVVRAGHNMVLDTLAMSPQQDQILVHNRY
                     NGAVLYPYCPVDQALRMDQHNYDPNLGTPLYDQTLVLLATVLAKAQAFIDNGVPART
                     SLIITDGADAHSRRSVREVKGVVEDMLRTEDHIIAAMGINDGQTDFKRVFREMGVRD
                     WILTPGNSQNEIRKAFQLFSQSVLRASQSAHNFNSWGGFGP"
ORIGIN
        1 aataaatact tacaggtatt ccacctgaaa ctctttctat gaatgacttt caagtctata
       61 tcctatattt atcctcaata aaatatgcac aatagatctc tactgagaaa actttatatt
```

```
  6503641 cacacagttg atcctgaccc ttctgcctaa agatggattc caggccaagt tgagatcgcc
  6503701 tccgtagact gcagaatcca ccac
//
```

```
CDS             join(861322..861393,865535..865716,866419..866469,
                871152..871276,874420..874509,874655..874840,
                876524..876686,877516..877631,877790..877868,
                877939..878438,878633..878757,879078..879188,
                879288..879533)
```

```
CDS             complement(join(880074..880180,880437..880526,
                880898..881033,881553..881666,881782..881925,
                883511..883612,883870..883983,886507..886618,
                887380..887519,887792..887980,888555..888668,
                889162..889272,889384..889462,891303..891393,
                891475..891595,892274..892405,892479..892653,
                894309..894461,894595..894620))
```

2.3. Tests sur les régions

Tests sur les opérateurs:
- les bornes inf et sup existent (valeurs existant dans la séquence);
- les bornes inf et sup sont des nombres;
- la borne inf est inférieure à la borne sup;
- les bornes inf et sup sont séparées par "..".

IMPORTANT: Si une région ne vérifie pas un des tests précédents, il est éliminé du parsing.

3. Programmation en Java

Structurer votre programme de façon qu'aux exécutions suivantes le parsing reprend l'arborescence locale et ne porte que sur les données génomiques modifiées (nouvelles ou modifiés).

IMPORTANT: L'exécution se déroule de la façon suivante:
- si l'arborescence locale des fichiers n'existe pas: génération de l'arborescence locale
- si l'arborescence locale des fichiers existe: mise à jour de l'arborescence locale

La mise à jour concerne l'ajout d'un nouveau génome de GenBank ou la suppression d'un génome supprimé de GenBank.

IMPORTANT: Il faut impérativement éviter que votre programme se bloque sur des données non conformes, un génome incomplet sans données ou un téléchargement (gestion des transferts, par exemple en mettant un délai de temporisation ou non).

Les points importants du programme doivent être commentés.

4. Modalités du projet

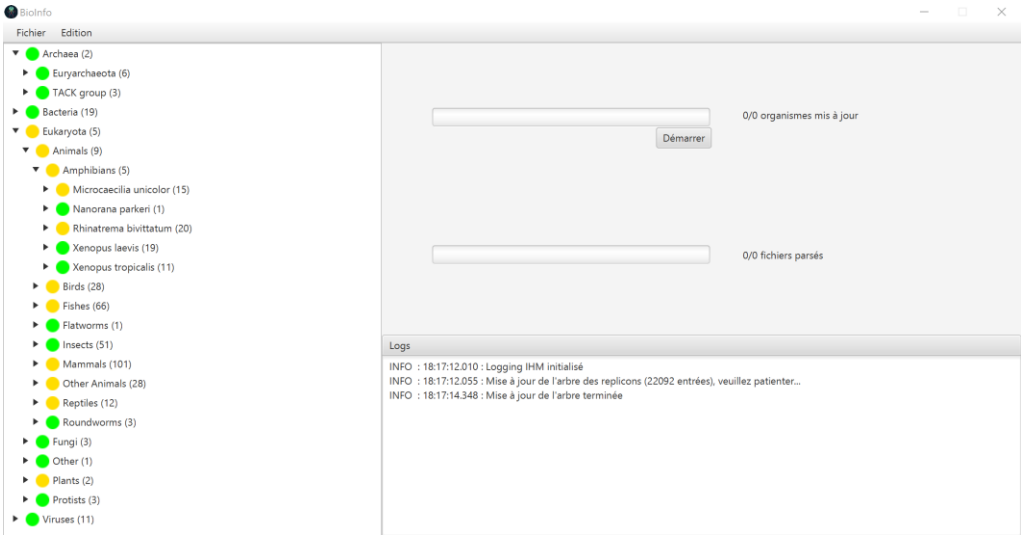
(i) Renvoyer par email à l'adresse c.michel@unistra.fr ou mettre sur un site de téléchargement:
- un dossier "Eclipse" pour les programmes sources pour Eclipse;
- un dossier "Jar" un fichier jar exécutable en double cliquant;
- un dossier "Results" (cf. l'arborescence Results\Kingdom\Group\SubGroup\Organism\).

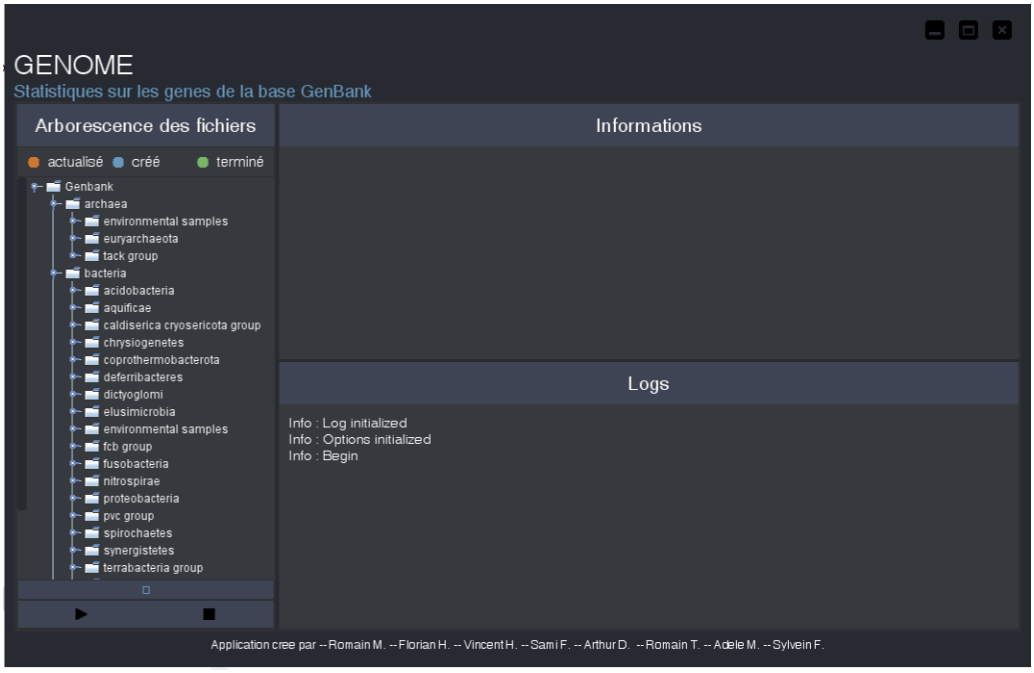
Ces trois dossiers sont compressés avec l'extension zip (impératif).

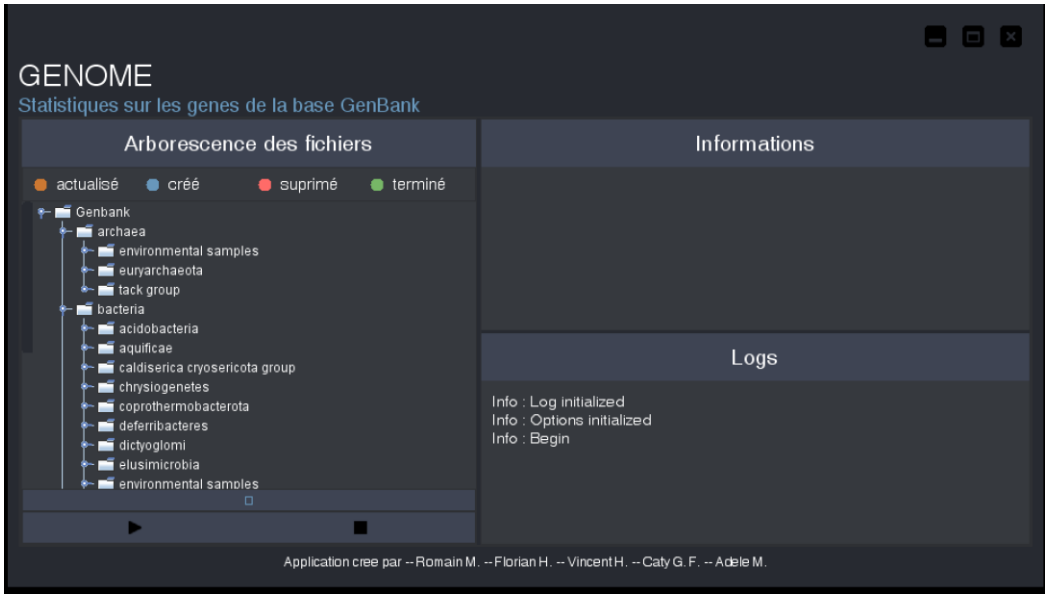
IMPORTANT ET IMPERATIF:
Le dossier "Results" doit également être un sous-répertoire du dossier "Eclipse"
Le dossier "Results" doit être également un sous-répertoire du dossier "Jar"

Donner dans un rapport:
- toutes les instructions et librairies permettant l'exécution du programme avec le logiciel Eclipse;
- toutes les informations sur l'arborescence locale et fichiers nécessaires;
- la classe contenant le main.

5. Exemples de logiciels







6. Info

https://www.ncbi.nlm.nih.gov/books/NBK21091/table/ch18.T.refseq_accession_numbers_and_mole/?report=objectonly

Table 1. RefSeq accession numbers and molecule types.

Accession prefix	Molecule type	Comment
AC_	Genomic	Complete genomic molecule, usually alternate assembly
NC_	Genomic	Complete genomic molecule, usually reference assembly
NG_	Genomic	Incomplete genomic region
NT_	Genomic	Contig or scaffold, clone-based or WGS[a]
NW_	Genomic	Contig or scaffold, primarily WGS[a]
NS_	Genomic	Environmental sequence
NZ_[b]	Genomic	Unfinished WGS
NM_	mRNA	
NR_	RNA	
XM_[c]	mRNA	Predicted model
XR_[c]	RNA	Predicted model
AP_	Protein	Annotated on AC_ alternate assembly
NP_	Protein	Associated with an NM_ or NC_ accession
YP_[c]	Protein	
XP_[c]	Protein	Predicted model, associated with an XM_ accession
ZP_[c]	Protein	Predicted model, annotated on NZ_ genomic records

[a] Whole Genome Shotgun sequence data.
[b] An ordered collection of WGS sequence for a genome.
[c] Computed.

Site à privilégier: ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/

Prendre NC

Site non recommandé: https://ftp.ncbi.nlm.nih.gov/genomes/genbank/

Prendre fichier GCF

What is the difference between a GenBank (GCA) and RefSeq (GCF) genome assembly?

A GenBank (GCA) genome assembly contains assembled genome sequences submitted by investigators or sequencing centers to GenBank or another member of the International Nucleotide Sequence Database Collaboration (INSDC). The GenBank (GCA) assembly is an archival record that is owned by the submitter. In rare cases where NCBI makes updates to the GenBank (GCA) assembly, for example, to remove contaminated sequences, the original submitter will be notified. GenBank (GCA) assemblies may include user-submitted or NCBI-generated annotation. User-submitted annotation can include annotation generated using NCBI's Prokaryotic Genome Annotation Pipeline (PGAP). PGAP annotation can be requested by the submitter during submission of the genome to GenBank or can be generated using the PGAP standalone software package.

A RefSeq (GCF) genome assembly represents an NCBI-derived copy of a submitted GenBank (GCA) assembly. RefSeq (GCF) assembly records are maintained by NCBI. In some cases the RefSeq (GCF) assembly may not be completely identical to the GenBank (GCA) assembly because NCBI staff may (1) remove short sequences or reported contaminants from the assembly or (2) add non-nuclear genome sequences (for example, mitochondrial or chloroplast genomes) to the assembly. All RefSeq (GCF) genome assemblies include annotation. In the majority of cases, this annotation is generated by the NCBI prokaryotic or eukaryotic genome annotation pipelines. In some cases, annotation is provided by the assembly submitter.